# Bounding Average Time Separations of Events in Stochastic Timed Petri Nets with Choice†

Aiguo Xie, Sangyun Kim and Peter A. Beerel
Department of Electrical Engineering – Systems
University of Southern California
Los Angeles, CA 90089-2562

**Abstract**

This paper presents a technique to estimate the average time separation of events (TSE) in stochastic timed Petri nets that can model time-independent choice and have arbitrary delay distributions associated with places. The approach analyzes finite net unfoldings to derive closed-form expressions for lower and upper bounds on the average TSE, which can be efficiently evaluated using standard statistical methods. The mean of the derived upper and lower bounds thus provides an estimate of the average TSE which has a well-defined error bound. Moreover, we can often make the error arbitrarily small by analyzing larger net unfoldings at the cost of additional run-time. Experiments on several asynchronous systems demonstrate the quality of our estimate and the efficiency of the technique. The experiments include the performance analysis of a full-scale Petri net model of Intel's asynchronous instruction length decoding and steering unit RAPPID containing over 900 transitions and 500 places.

## 1 Introduction

In some applications, well-designed asynchronous systems can achieve significantly higher average performance than their synchronous counterparts. Designing such systems, however, can be challenging because such systems must often be highly concurrent and optimized for the average case. Consequently, a variety of CAD tools have been developed to support synthesis, verification, and system optimization.

Many of these tools find the time separation of user-specified pairs of system events, or more simply, the time separation of events (TSE). Some synthesis and formal verification tools analyze system models with interval delays and derive extreme values (or bounds) of TSEs. These tools enable timing assumptions to be effectively and reliably used to optimize the system design (e.g., [MM93]). Performance analysis and performance-driven synthesis tools, on the other hand, often analyze models with either fixed delays (which represent average component delays) or stochastic delays (which represent the distribution of component delays). These tools calculate, estimate, or bound the *average* TSEs and can be applied at the architectural/system level, estimating average latency, throughput, or response time, or at the gate/transistor level, identifying the typically late arriving inputs.

†This work is funded in part by NSF Grant CCR-9812164.

For the purpose of performance analysis, stochastic delays are more expressive than fixed or interval delays. Unfortunately, developing a technique that can analyze sufficiently general system models with stochastic delays has up to now been an elusive goal. Existing efficient techniques have been limited in either the type or size of system model they can analyze. For instance, efficient techniques to analyze both the average and extreme TSEs in models with only fixed delays and without choice (e.g., Marked graphs) are well-known [Bur91, Kar78]. More complex techniques exist for analyzing the extreme and average TSEs for models with interval delays but no choice [EB97, HB95a, MM93]. For models with interval delays and a restricted form of choice called *free-choice* (whose decisions can be decoupled from the system timing behavior), efficient analysis techniques have been developed using net unfolding to analyze extreme TSEs [HB94] but not average TSEs. For models with more general forms of choice, all known techniques must explore the complete timed state space of the system (e.g., [AD94, BM98]). Consequently, these techniques suffer from the state explosion problem and can thus be applied only to relatively small systems. For models with stochastic delays the options are much more limited. Even for marked graphs with general delay distribution, a Markovian analysis of the complete timed state space [HV87, KGBA94, XB97] is often required. Despite some recent approaches to mitigate the state explosion problem [Buc94, HMPS96, XB], the techniques can handle only small systems. For large, more complex systems, random simulation is the only known method for obtaining an estimate of average TSEs. A typical random simulation method generates a long sample path (i.e., a sequence of timed markings) of the model from which one computes a sequence of TSEs. The estimate of the average TSE is obtained by simply averaging the TSEs in the sequence.

The principal difficulty with random simulation is in analyzing the accuracy of the resulting estimate without which one cannot decide when to stop the simulation. There are many statistical methods developed for such analysis including Monte-Carlo approach (e.g., [MFJ90]), regenerative method (e.g., [GI93]) and standardized time series method (e.g., [GI90]) which, under certain circumstance, can give a *confidence interval* of the estimate. That is, the methods computes a probability $\alpha$ with which the estimate is no far from the average TSE by a given error $\epsilon$. The fundamental mathematical tool behind all these methods is the well-known Central Limit Theorem [GS92] which applies only to a sequence of *independent* and *identically distributed* (i.i.d) random variables. The problem that since most systems we are interested in are inherently cyclic, their timed Petri net models are necessarily cyclic as well. As a consequence, timed markings (and thus the TSEs) along a sample path generated by an execution of such a model are dependent on each other. This violates the assumption of the Central Limit Theorem, which makes the application of existing statistical methods difficult.

Early work on characterizing estimates from simulation was limited to *regenerative* timed Petri nets. In such a net, a sample path can be partitioned into sub-sequences of timed markings at the so called *regeneration time instances* such that timed markings from different sub-sequences are independent. Consequently, TSEs derived from different sub-sequences are also independent, which may facilitate the application of the standard regenerative methods. However, the requirement that the net to be regenerative can be quite restrictive. To handle more general nets, one recent work [Haa97] proposed using standardized time series method to yield a confidence interval by partitioning a long sample path into a fixed number of sub-sequences of roughly equal length (in time). Unfortunately, the method assumes that the means of TSEs derived from different sub-sequences are independent, which does not generally hold and thus limits the applicability of the method.

We recently proposed a new approach to estimating the average TSEs for models with stochastic delays [XB98a] using Monte-Carlo approach. To overcome the difficulty due to the dependence of TSEs within a random simulation, we did not apply statistical methods to estimates the average TSE (denoted by $\overline{\gamma}$) itself, but use them to estimate two other quantities, $U$ and $L$, which we prove serves lower and upper bounds on the average TSE. More precisely, $U(L)$ is the limiting running average of a sequence of *i.i.d* random variables where each of the random variables is derived from a *finite* random simulation (or *unfolding*) of the net. By taking the average of such a sequence of a finite length, we get an unbiased estimate $\hat{U}(\hat{L})$ of $U(L)$. In addition, using simple statistical methods such as Monte-Carlo approach, we get the confidence interval of $\hat{U}(\hat{L})$. If it is further desired, we can take the mean of $\hat{U}$ and $\hat{L}$ to be an estimate of $\overline{\gamma}$, and the distance between $\hat{U}$ and $\hat{L}$, $|\hat{U} - \hat{L}|$ is its error bound[1]. Figure 1 illustrates the relationship among $\hat{L}$, $\hat{U}$ and the estimate of the average TSEs.

For a user specified confidence level $\alpha$, the confidence interval of $\hat{U}$ and $\hat{L}$ can often be made arbitrarily small by increasing the length of the random sequence being averaged[2]. Moreover, the distance between $\hat{U}$ and $\hat{L}$ can be rapidly reduced to almost zero by increasing the length of the finite random simulations of the net (and thus additional run-time) that generate the above random sequences. Consequently, this approach yields very accurate estimates of average TSEs which have a well-defined error bound.

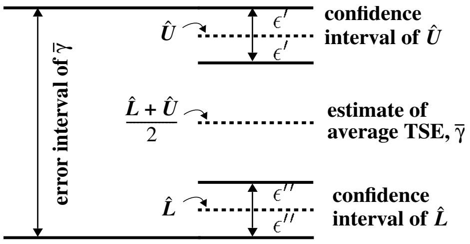



Figure 1: Relationship among the estimates.

Our preliminary efforts were limited to stochastic timed marked graphs (i.e., having no choice) but demonstrated very promising results [XB98a]. The approach yields much sharper bounds than those obtained by the state-of-art bounding technique for average cycle time of events [CCS91] which are often too loose to be useful in practice. In all the circuits examined, the time complexity appeared to grow only sub-quadratically in the system size and the resulting error bounds were very narrow. For example, the method yielded upper and lower bounds of almost-zero difference on the throughput of a 32-stage micropipeline in a few seconds on a Ultra Sparc 10. In comparison, the previous state-of-the-art technique using Markovian analysis [XB] cannot handle micropipelines of more than 8 stages within several hours.

This paper extends our approach to stochastic timed Petri nets with choice. Its contributions include:

- Definition of the average TSE on stochastic Petri nets with unique- and free-choice based on an infinitely long random simulation and characterization of a large class of transition pairs for which the average TSE is a finite constant.

[1] Strictly speaking, the error bound is $|\hat{U} - \hat{L}| + \epsilon' + \epsilon''$.

[2] This interval is in reverse proportional of the square root of the sequence length.

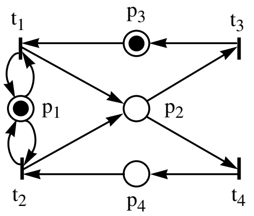
- Development of closed-form expressions for bounds on the average TSE based on dynamically generated finite (with probability 1) net unfoldings.
- Explanation of how our upper and lower bounds can often be made arbitrarily close to each other using larger net unfoldings at the expense of additional run-time.
- Demonstration of how these closed-form expressions can be efficiently computed with arbitrary high accuracy and confidence using simple statistical methods.
- Demonstration of the efficiency of the approach in a variety of examples, including a performance analysis of a full-scale stochastic timed Petri model of Intel's asynchronous instruction length decoding and steering unit, RAPPID, (with over 900 transitions) in less than 45 minutes of CPU time.

The organization of this paper is as follows. Section 2 reviews stochastic timed Petri nets. Section 3 provides an overview of our approach which is detailed in Sections 4, 5, and 6. Section 7 describes our case studies and Section 8 gives some conclusions and a brief description of some future work. All the proofs are included in the appendix.

# 2 Stochastic timed Petri nets

We start by reviewing general concepts and properties of untimed Petri nets and then introduce stochastic delays and choice probabilities specific to stochastic timed Petri nets. Further details on Petri nets may be found in [Mur89]. This section also defines the average TSEs in stochastic timed Petri nets and identifies a class of transition pairs for which the average TSEs are guaranteed to converge to a finite number.

## 2.1 Petri net models




(a) The Petri net,

$$
\begin{array}{ll}
X(p_1) = 5, & \\
X(p_2) \sim \texttt{exp}(2.5), & \\
X(p_3) \sim \texttt{uniform}(2,4), & \mu(p_2,t_3) = 0.4, \\
X(p_4) \sim \texttt{uniform}(5,10). & \mu(p_2,t_4) = 0.6 = 1 - \mu(p_2,t_3).
\end{array}
$$

(b) its stochastic assumption.

Figure 2: A stochastic Petri net example.

As is usual, we denote a Petri net by a triple $N = (P, T, F)$ where $P$ is the set of places, $T$ the set of transitions, and $F \subseteq (P \times T) \cup (T \times P)$ the flow relation. The preset of $x \in P \cup T$ denotes the set $\bullet x = \{y \in P \cup T \mid (y, x) \in F\}$, and its poset $x\bullet = \{y \in P \cup T \mid (x, y) \in F\}$. A *marked graph* is a Petri net in which every place has at most one input and output transition, i.e., $|p\bullet| \leq 1 \geq |p\bullet|, \forall p \in P$. A place $p$ is a choice place if $|p\bullet| > 1$. $N$ is *free-choice* (FC) net if $\forall p_1, p_2 \in P$, $p_1\bullet \cap p_2\bullet \neq \emptyset \Rightarrow |p_1\bullet| = |p_2\bullet| = 1$. Equivalently, in a FC net, if two transitions share an input place, they may not have any other input places. A Petri net is *extended free-choice* (EFC) if $\forall p_1, p_2 \in P$, $p_1\bullet \cap p_2\bullet \neq \emptyset \Rightarrow p_1\bullet = p_2\bullet$. An EFC net can be translated into a FC net [Bes87b]. A choice place $p$ is *unique-choice* if no two output transitions of $p$ are ever enabled simultaneously. Figure 2 shows an example of such a net, where $p_1$ is unique-choice and $p_2$ is free-choice. Note that tokens in $p_1$ never needs to choose which of its two output transitions to fire since $p_3$ and $p_4$ can never be simultaneously marked.

A marking is a mapping $M : P \rightarrow \{0, 1, 2, \cdots\}$. The number of tokens in place $p$ under marking $M$ is denoted by $M(p)$. A transition $t$ is *enabled* at marking $M$ if all its input places have at least one token, i.e., $M(p) \geq 1, \forall p \in \bullet t$. When $t$ is enabled, it may fire. The firing of $t$ removes one token from each place in its preset and deposits one token to each place in its poset, leading to a new marking $M'$, denoted by $M[t\rangle M'$. A sequence of transitions $\sigma = t_0 t_1 \cdots t_{m-1}$ is a firing sequence from a marking $M_0$ iff $M_k[t_k\rangle M_{k+1}$ for $k = 1, \cdots, m-1$. In that case, one also writes $M_0[\sigma\rangle M_m$ and says $\sigma$ has a length of $m$ which we denote by $|\sigma|$.

A *marked* Petri net $\Sigma$ is a tuple $(N, M_0)$, where $N$ is a Petri net and $M_0$ is the initial marking of $N$. Sometimes, we use $N_\Sigma$ to stress $N$ is the underlying net of $\Sigma$. The set of reachable markings of $\Sigma$ is denoted by $R(M_0)$. $\Sigma$ is *live* iff every transition of $N$ will eventually be enabled from every $M \in R(M_0)$. It is *safe* if no place is ever marked with more than one token, i.e., $M(p) \leq 1$ for $\forall M \in R(M_0), \forall p \in P$. A live and safe (LS) marked Petri net has no source or sink places and no source or sink transitions [Bes87a]. Thus, a LS Petri net can be partitioned into a set of strongly connected components each evolving independently of others. Below, we assume the Petri net is strongly connected. In particular, we restrict ourselves to LS Petri nets with free-choice and unique-choice places.

Finally, we write $\vec{\sigma}$ to be the firing counter vector of $\sigma$ which indicates the total number of times each transition is fired in a firing sequence $\sigma$. $\vec{\sigma}$ is called a *T-invariant* if there is a reachable marking $M$ such that $M[\sigma\rangle M$.

## 2.2 Stochastic assumptions

In this paper, we associate delays with places. That is, a token flowing into a place $p$ must experience a random delay associated with $p$ before it is *available* to be consumed by an output transition of $p$. The delays experienced by different tokens are independent. The firing of a transition is assumed to be instantaneous. Let $X(p)$ be the random variable denoting the delay associated with place $p$. Let $F_{X(p)} : R \rightarrow [0, 1]$ denote its distribution function, i.e., $F_{X(p)}(x) = \text{Prob}(X(p) \leq x)$. Unlike some other research in stochastic timed Petri nets, we do not put any restriction on the distribution functions except that they all have finite means and variances, i.e., $\mathbf{E}X(p) < \infty, \mathbf{Var}X(p) < \infty, \forall p \in P$.

For each free-choice place $p \in P$, we assume there is a probability mass function (p.m.f.) $\mu$ that resolves the choice. That is, if $t \in p\bullet$, $\mu(p, t)$ is the probability that $t$ consumes the token in $p$ each time $p$ is marked. In addition, we have $\sum_{t \in p\bullet} \mu(p, t) = 1$.

*Example* Figure 2.(b) lists a possible stochastic assumption on delay distributions and choice p.m.f. on the Petri net in Figure 2.(a). For example, the delay in place $p_1$ is 5, the delay in $p_2$ is exponentially distributed with a parameter 2.5, the delay in $p_3$ is uniformly distributed between 2 and 4, and so forth. In addition, with probability 0.4, $t_3$ fires each time $p_2$ receives a token, and with the remaining probability, $t_4$ fires. The other choice place $p_1$ is not assigned a p.m.f. since it is unique-choice. ■

## 2.3 Timed executions and average TSEs

To describe a possible run of a stochastic Petri net $\Sigma$, we use the notion of *timed executions*. Roughly speaking, a timed execution is an unfolding (similar to the ones in [Maz89, McM95, HB94]) of the net with all the choice resolved and all the places assigned a delay value. For example, Figure 3 shows a timed execution of the Petri net in Figure 2. In this execution, $t_3$ fires after $p_2$ is marked for the 1st, 3rd and 6th times whereas $t_4$ fires after $p_2$ is marked for the 2nd, 4th and 5th times. The numbers along the (instanced) places denote the delay values assigned to them. The instanced places and their assigned delay values represent the constraints among the occurrences of transitions. Let $t^{(k)}$ and $p^{(k)}$ denote the $k$-th occurrence of a transition $t$ and $k$-th instance of a place $p$, respectively. We call an occurrence of a transition an *event*. In Figure 3, we have dropped the instance indices of places for brevity.

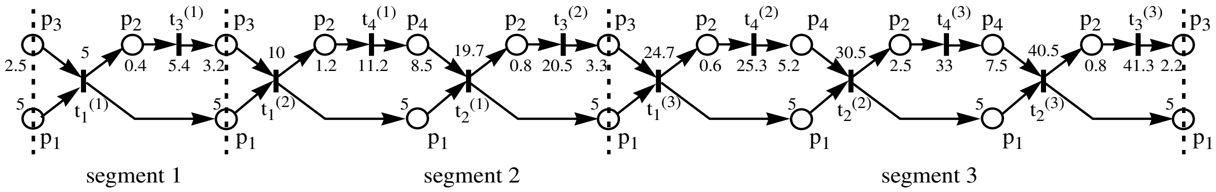



Figure 3: A timed execution of the stochastic Petri net in Figure 2.

Formally, a timed execution $\pi$ of $\Sigma$ is a triple $(N_\pi, d_\pi, \ell)$ where $N_\pi = (P_\pi, T_\pi, F_\pi)$ is an acyclic event graph (or marked graph), $d$ a function $P_\pi \rightarrow \mathbf{R}$ denotes the (constant) delay value of each place in $P_\pi$ and a labeling function $\ell : P_\pi \cup T_\pi \rightarrow P \cup T$ maps each instanced places and transitions to their corresponding ones in $\Sigma$.

To generate a timed execution, we first instance all initially marked places by assuming tokens flow into them at time 0, and set the token available times by randomly assigning delays to the places according to their delay distributions. Then, we iteratively advance the clock to fire the earliest (time) enabled transitions. Let function $\tau$ denote the occurrence time of instanced transitions (or events). For the timed execution shown in Figure 3, the corresponding functions $d$ and $\tau$ are illustrated along the instanced places and transitions.

Given a timed execution, the occurrence time of event $t^{(k)}$ is determined as follows:

$$\tau(t^{(k)}) = \max_{\substack{(s^{(j)}, p) \in F_\pi, \\ p \in \bullet t^{(k)}}} \tau(s^{(j)}) + d(p) \tag{1}$$

where the term $\tau(s^{(j)}) + d(p)$ reduces to $d(p)$ if $\bullet p = \emptyset$ in $\pi$, i.e., if $p$ is a source place of $\pi$.

Note that for Petri nets with only free-choice and unique-choice, the set of underlying structures resulting from all possible timed executions coincides with that from all possible untimed processes [HB95b, SY96]. This fact will be used later to decouple the decision from timing.

For a given timed execution $\pi$, the time separation of event pair $(s^{(k)}, t^{(k+\varepsilon)})$, denoted by, $\gamma^{(k)}(s,t,\varepsilon)$, is the time difference between their occurrences. That is,

$$\gamma^{(k)}(s,t,\varepsilon) = \tau(t^{(k+\varepsilon)}) - \tau(s^{(k)}) \tag{2}$$

where $\tau$ is the function associated with $\pi$.

Let $\Pi_\Sigma$ be the set of all possible timed executions of $\Sigma$. If we consider $\pi$ as randomly taken from $\Pi_\Sigma$, then $\gamma^{(k)}(s,t,\varepsilon)$ is a random variable. Consequently, the sequence $\{\gamma^{(k)}(s,t,\varepsilon) : k = 1, 2, \cdots\}$ is a random process. It is the average of this TSE sequence that we are concerned in this paper, namely, $\lim_{n\rightarrow\infty} \frac{1}{n} \sum_{k=1}^{n} \gamma^{(k)}(s,t,\varepsilon)$ which we define as the average TSE of the separation triple $(s,t,\varepsilon)$.

## 2.4 Proof of existence

Before we describe how to bound the average TSE, it seems prudent to question whether the above limit in the definition of average TSEs exists. For example, one might imagine that this limit grows unboundedly as time proceeds, or a situation where it oscillates and never converges to a constant.

For marked graphs, we showed in [XB98a] that the limit exists almost surely for any transition pair $(s, t)$ and any finite $\varepsilon$. That is, their corresponding average TSE is a finite constant for every possible timed execution with probability 1 (w.p.1). However, this is not always true if the Petri net has choices. Consider, for example, transitions $t_3$ and $t_4$ of the Petri net in Figure 2. The difference between the numbers of occurrences of $t_3$ and $t_4$ will be infinite w.p.1 as time progresses. Consequently, their average TSE diverges w.p.1 provided that the net has at least one place with a positive mean delay. For the above reason, we restrict ourself to transition pairs that guarantee finite average TSEs.

To formally characterize the class of transitions pairs we consider, the notion of *steady markings* is useful. A steady marking is one that can be reached from all reachable markings. Since the reachability graph of a LS Petri net with only free-choice and unique-choice is irreducible [BV84], its unique terminal strongly connected component contains all the steady markings of the net. In this paper, we require the targeted transition pair $(s, t)$ to satisfy the following condition.

**Condition 1** *If $M$ is a steady marking and $\sigma$ is a firing sequence such that $M[\sigma\rangle M$, then $\vec{\sigma}(s) = \vec{\sigma}(t)$.*

The condition requires that the net fires both transitions the same amount of times in order to traverse any cycle of steady markings. In practice, we expect this condition to be guaranteed by the user. However, it can also be checked using structural analysis for free-choice nets and a reachability analysis [XB98b] for nets with both choice-free and unique-choice.

The following theorem verifies that the running average of TSEs of a transition pair satisfying Condition 1 converges for the Petri nets considered in this paper. The theorem generalizes the *weak ergodicity* property of the cycle time sequence of the consecutive instances of a same transition [CCS91, BM95] to that of the TSE sequence of a pair of different transitions.

**Theorem 1** *Let $\Sigma$ be a LS Petri net that has only free-choices and unique-choices and satisfies stochastic assumptions given in Section 2.2. For any transition pair $(s, t)$ for which Condition 1 holds, its corresponding average TSE with a fixed occurrent offset $\varepsilon$ is a finite constant $\overline{\gamma}(s, t, \varepsilon)$ almost surely and in mean. That is,*

$$\mathrm{Prob}\Big(\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \gamma^{(k)}(s, t, \varepsilon) = \overline{\gamma}(s, t, \varepsilon)\Big) = 1, \tag{3}$$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbf{E}\gamma^{(k)}(s, t, \varepsilon) = \overline{\gamma}(s, t, \varepsilon). \tag{4}$$

At first glance, Condition 1 may seem rather restrictive. However, if there is a cycle of steady markings along which $s$ an $t$ fire different amount of times, then the difference of occurrence times of $s$ and $t$ is likely to grow unboundedly as time proceeds. According to the definition of average TSE, their average TSE $\overline{\gamma}(s, t, \varepsilon)$ is then likely to be infinite (positive or negative) for every fixed $\varepsilon$. As an example, let us consider transitions $t_3$ and $t_4$ of the Petri net in Figure 2(a). They do not satisfy Condition 1 since from the initial marking $M_0$ as shown, the firing sequence $t_1 t_3$ leads the net back to $M_0$ which can be checked as a steady marking. As a result, one can show that their average TSE does not exist since for every random firing sequence $\sigma$, $\lim_{|\sigma|\to\infty} |\vec{\sigma}(t_3) - \vec{\sigma}(t_4)| \to \infty$ almost surely, unless the choice p.m.f of place $p_2$ evaluates strictly the same at $t_3$ and $t_4$, i.e., $\mu(p_2, t_3) = \mu(p_2, t_4) = \frac{1}{2}$.

In some cases, it may be of interest to compute average time distance between some special occurrences of transition pair $(s, t)$ which does not satisfy Condition 1. We state the following two cases where the *average time distance* still exists and can be computed by the approach described in this paper.

**Case 1** *The interested average time distance is between an occurrence of $s$ to the $\varepsilon$-th occurrence of $t$ following the occurrence of $s$.*

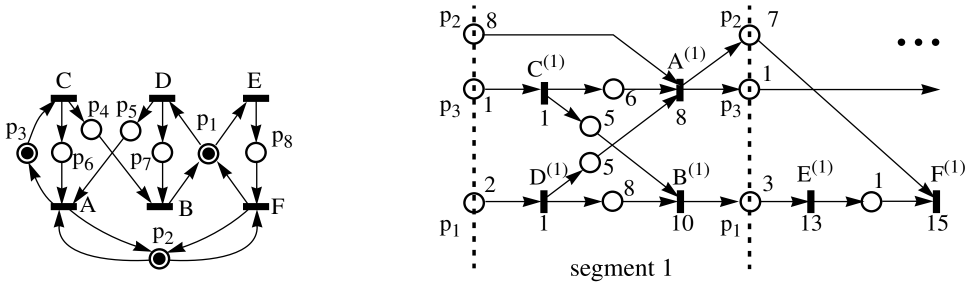
Such an average time distance may make sense if $s$ occurs less frequently than $t$, although it does not verify the definition of average TSE set earlier. If one can skip the irrelevant occurrence of $t$, the condition 1 is met. For instance, the pair $(t_3, t_4)$ in the above example can be considered within this case.

**Case 2** *The pair $(s, t)$ is such that for every T-invariant $\vec{v}$, if $\vec{v}(s) > 0$, $\vec{v}(t) > 0$, and vice versa. Moreover, $\vec{v}(s)/\vec{v}(t) = c$ which is a constant.*

In this case, one may split the occurrences of each of the two transitions as if they are the occurrences of some new transitions, and consider the average TSE of a pair of new transitions corresponding to $s$ and $t$. The new transition pair satisfies Condition 1.

Below, we assume the interested transition pair satisfies Condition 1 which serves as the basis of the other two cases.

## 3 Overview of our approach




(a) A Petri net, (b) a timed execution of the net.

Figure 4: Deducing bounds on TSEs in a given timed execution using only limited history.

To introduce our approach, we first explain how we bound a TSE instance in a given timed execution using limited history. For this purpose, a timed execution is partitioned into segments. Roughly speaking, a segment is a portion of a timed execution that starts and ends at a same marking. We argue that bounds on a TSE instance can be derived using the knowledge of the segment which contains the TSE instance.

Let us consider analyzing the time separation from $A^{(1)}$ to $B^{(1)}$ in segment 1 of the timed execution shown in Figure 4(b) whose underlying net is shown in Figure 4(a). According to the timing in the figure, $A^{(1)}$ occurs at time 8 and $B^{(1)}$ occurs at time 10. Thus, their time separation is 2. In our approach, we will ignore the exact token available times in the source places $p_1$, $p_2$ and $p_3$, and deduce TSE bounds using only the timing within segment 1. In particular, by assuming the token available times in the source places to be anywhere within $(-\infty, \infty)$, a simple longest path analysis yields an upper bound of 3 on the TSE instance. The duality of the problem gives a lower bound.

To obtain bounds on the average TSEs, our approach analyzes a finite number of timed execution segments using the path analysis alluded to above. More precisely, by randomly generating timed execution segments and analyzing each of them separately, and averaging the results, we indeed get valid bounds on the average TSE with arbitrarily high precision. In order to determine the number of segments needed to guarantee the required precision, we use a simple statistical method i.e., Monte-Carlo approach. Finally, we observe that the bounds can often be dramatically improved by using more history, i.e., using more segments prior to the one considered.

The proof of the correctness of this approach requires the development of two key facts. First, the structure of the $k$-th segment in a random timed execution is independent of index $k$ (the location of the segment in the timed execution). From this, we can reason about segments that are infinitely far in the future. Secondly, multiple TSEs corresponding to a given transition pair may start in one segment and they are not independent. To overcome this dependence, we introduce a notion of a grouped TSE which is the sum of TSEs that start in a same segment, and describe an approach to compute bounds on the average grouped TSE. We then show that when divided by the average number of TSE pairs in a group, these bounds become the bounds on the average TSE.

Section 4 formally defines segments and grouped TSEs. Section 5 describes the path analysis method used to obtain the actual bounds in more detail.

# 4 Partitioning infinite timed executions

## 4.1 Definition of a segment

Let $\pi = (N_\pi, d, \ell)$ be a timed execution of a stochastic Petri net $\Sigma = (N, M_0)$. It is well-known that every (untimed) reachable marking of $\Sigma$ induces a cut $\xi$ (a set of instanced places) that partitions the event graph $N_\pi$ into two parts [Bes90]. The portion of the event graph in between two different cuts due to the same reachable marking $M$ is called a *segment*. More precisely, if $\sigma$ is a firing sequence that starts from a cut $\xi$ and ends at another cut $\xi'$ such that $\ell(\xi) = \ell(\xi) = M$, then the portion of $N_\pi$ between $\xi$ and $\xi'$ is a segment and denote it by $S(\xi, \sigma)$. We say $S(\xi, \sigma)$ is *minimal* if it does not contain any other segment starting from $\xi$. For the timed execution shown in Figure 3, there are three minimal segments corresponding to the reachable marking where places $p_1$ and $p_3$ are marked. This example also shows that for Petri nets with choices, there can be many (minimal) segments whereas for those without choices (i.e, marked graphs), the minimal segment is unique. In the sequel, when we write segments, we refer to minimal ones.

One simple property of a random time execution of a Petri net considered in this paper is that the structures of its segments are independent of each other. This is simply because the structure of a segment is determined by the choices made on the places inside the segment and choices made in different segments are independent. As a result, the sequence of segments generated by a random time execution has a property that the structure of a segment is not determined by the location of the segment in the sequence. Formally, let $\mathcal{S}(M)$ denote the set (typically infinite) of all possible segments starting from a cut due to marking $M$, i.e., $\mathcal{S}(M) = \{S(\xi,\sigma) \subseteq N_\pi \mid \ell(\xi) = M, \forall\sigma \in \pi, \forall\pi \in \Pi_\Sigma\}$. Let $\ell' : S(M) \rightarrow \{1, 2, \cdots\}$ gives each distinct segment structure a number (label). Function $\ell'$ maps the sequence of segments of a random execution into a random process of segment labels. These random labels are independent and identically distributed. This simple fact will allow us to reason about an infinite execution by considering all possible finite executions of as little as one segment in length.

### 4.2 Grouped TSEs and their weak ergodicity

Although the structures of segments are independent, as mentioned earlier, multiple TSE instances of a transition pair $(s,t)$ may start in one segment and they can be dependent on each other. For example, for the time execution shown in Figure 3, there are two TSE instances of pair $(t_4, t_2)$ starting from segment 3 and their TSE values not independent. This motivates us to consider all the TSE instances starting from a same segment together as a *group*.

To formally define a TSE group, let us consider a TSE sequence $\{\gamma^{(j)}(s,t,\varepsilon) : j \geq 1\}$ generated by a random timed execution $\pi$ of $\Sigma = (N, M_0)$. The segment sequence $\{S^{(l)} \in \mathcal{S}(M_0) : l \geq 1\}$ of $\pi$ partitions the TSE sequence into sub-sequences. Two TSEs $\gamma^{(i)}(s,t,\varepsilon)$ and $\gamma^{(j)}(s,t,\varepsilon)$ are contained in one sub-sequence if their corresponding source events $s^{(i)}$ and $s^{(j)}$ are in a same segment.

For the timed execution shown in Figure 3, if the time separation triple $(t_4, t_2, 0)$ is considered, then its first TSE group is $\{(t_4{}^{(1)}, t_2{}^{(1)})\}$ (due to segment 2). Its second TSE group is $\{(t_4{}^{(2)}, t_2{}^{(2)}), (t_4{}^{(3)}, t_2{}^{(3)})\}$ since $t_4{}^{(2)}$ and $t_4{}^{(3)}$ are both in the segment 3.

We denote the index of the last TSE in group $k$ by $\eta^{(k)}(s)$. That is,

$$\begin{aligned}\eta^{(k)}(s) = \max\{&i > \eta^{(k-1)}(s) : \\ &\exists l, s^{(\eta^{(k-1)}(s)+1)} \in S^{(l)}, s^{(i+1)} \notin S^{(l)}\},\end{aligned}$$

where $\eta^{(0)}(s) \stackrel{\triangle}{=} 0$. For convenience, we define the number of TSEs in group $k$ the *length* of the group, denoted by $\alpha^{(k)}(s)$. That is.

$$\alpha^{(k)}(s) = \eta^{(k+1)}(s) - \eta^{(k)}(s).$$

Finally, we define the sum of all TSEs in group $k$ as $k$-th *grouped TSE* denoted by $\beta^{(k)}(s,t,\varepsilon)$. That is

$$\beta^{(k)}(s,t,\varepsilon) = \sum_{l=\eta^{(k-1)}(s)+1}^{\eta^{(k)}(s)} \gamma^{(l)}(s,t,\varepsilon).$$

In the above example, we have $\alpha^{(1)}(t_4) = 1$, $\alpha^{(2)}(t_4) = 2$, and $\beta^{(1)}(t_4, t_2, 0) = \tau(t_2{}^{(1)}) - \tau(t_4{}^{(1)}) = 19.7 - 11.2 = 8.5$, $\beta^{(2)}(t_4, t_2, 0) = (\tau(t_2{}^{(2)}) - \tau(t_4{}^{(2)})) + (\tau(t_2{}^{(3)}) - \tau(t_4{}^{(3)})) = (30.5 - 25.3) + (40.5 - 33) = 12.7$.

We will show that the sequence of grouped TSEs satisfies a similar weak ergodic property as the sequence of TSEs, i.e., its average converges to a finite number, which will enable us to estimate the average TSE.

The following theorem verifies that both the group length sequence and the grouped TSE sequence exhibit similar ergodicity property to that of the TSE sequence stated in Theorem 1.

**Theorem 2** *Let $\Sigma$ be a stochastic LS Petri net that has only free-choices and unique-choices, and $(s,t)$ be a transition pair for which Condition 1 holds. Then, for any fixed $\varepsilon$, the average of the grouped TSE sequence (group length sequence) $\{\beta^{(k)}(s,t,\varepsilon) : k \geq 1\}$ $(\{\alpha^{(k)}(s) : k \geq 1\})$ converges to a finite constant $\overline{\beta}(s,t,\varepsilon)$ $(\overline{\alpha}(s))$. That is,*

$$\frac{1}{n}\sum_{k=1}^{n}\beta^{(k)}(s,t,\varepsilon) \rightarrow \overline{\beta}(s,t,\varepsilon), \tag{5}$$

$$\frac{1}{n}\sum_{k=1}^{n}\alpha^{(k)}(s) \rightarrow \overline{\alpha}(s). \tag{6}$$

*The convergence in (5) and (6) takes almost surely and in mean. Moreover, $\overline{\beta}(s,t,\varepsilon) = \overline{\alpha}(s)\overline{\gamma}(s,t,\varepsilon)$ where $\overline{\gamma}(s,t,\varepsilon)$ is the average TSE of the triple $(s,t,\varepsilon)$.*

This theorem also states the relationship between the average TSE $\overline{\gamma}(s,t,\varepsilon)$ and the average grouped TSE $\overline{\beta}(s,t,\varepsilon)$. Note that since the labels of segment structures are i.i.d for a random time execution, the average group length $\overline{\alpha}(s)$ is merely the average length of group $k$ ($k$ fixed) over all possible executions, i.e., $\overline{\alpha}(s) = \mathbf{E}\alpha^{(k)}(s)$. In Section 6, we use simple statistical method to evaluated $\mathbf{E}\alpha^{(k)}(s)$ as well as the bounds on $\overline{\beta}(s,t,\varepsilon)$, and consequently we get bounds on $\overline{\gamma}(s,t,\varepsilon)$.

# 5 Deriving the bounds

As mentioned earlier, we obtain bounds on the grouped TSE via path analysis of random segments. The intuition behind the bounds is that we ignore the time when tokens in the source places of the segments become available. To do this, we identify a set of *reference events* which can be considered as synchronization points for the targeted event pair. To obtain an upper bound on the time separation, we determine the time separation under the assumption that each synchronization point is the critical one and take the largest separation obtained.

## 5.1 The duality of bounds

In many timing analysis problems, finding a lower bound on a delay variable can often be transformed to finding an upper bound on a related delay variable. This is also true in our case.

In particular, for a given $\varepsilon$, and any $j \in \{1,2,\cdots\}$ s.t. $j > -\varepsilon$, we have $-\gamma^{(j)}(s,t,\varepsilon) = -t^{(j+\varepsilon)} + s^{(j)} \triangleq \gamma^{(j+\varepsilon)}(t,s,-\varepsilon)$ from the definition of $\gamma$. Thus, if one finds an upper bound $U$ on $\overline{\gamma}(t,s,-\varepsilon)$, then $-U$ is a lower bound on $\overline{\gamma}(s,t,\varepsilon)$. Because of this duality, we shall be concerned only with the upper bounds from now on.

## 5.2 The upper bound: $\varepsilon = 0$ case

The desired goal is to find an upper bound on $\overline{\beta}(s,t,\varepsilon)$. To ease the exposition of the technique, we take the occurrence offset-index $\varepsilon$ to be 0. Extension of the result to the case where $\varepsilon \neq 0$ is presented in the next subsection.

Our approach first derives an upper bound $U_\beta{}^{(k)}(s,t,\varepsilon)$ on $\beta^{(k)}(s,t,\varepsilon)$ for a fixed $k \geq 1$, which implies that $\mathbf{E}U_\beta{}^{(k)}(s,t,\varepsilon)$ is an upper bound on $\mathbf{E}\beta^{(k)}(s,t,\varepsilon)$. We then show $\mathbf{E}U_\beta{}^{(k)}(s,t,\varepsilon)$ is *independent* of $k$, which implies it is also an upper bound on $\overline{\beta}(s,t,\varepsilon)$ following the weak ergodicity property of the grouped TSE sequence stated in Theorem 2.

We need the following notations to formulate the upper bound $\mathbf{E}U_{\beta^{(k)}}(s,t,\varepsilon)$. Let $\pi = (N_\pi, d, \ell)$ be a random timed execution of Petri net $\Sigma$. For every path $\rho \in N_\pi$, we denote by a random variable $\delta(\rho)$ the sum of random delays assigned to all the places along path $\rho$, namely, $\delta(\rho) = \sum_{p\in\rho} d(p)$. Since $\pi$ is considered as a random timed execution, the delay $d(p)$ assigned to every place $p \in P_\pi$ is also a random variable whose distribution function is the same as that of $X(\ell(p))$.

Let $D$ be a particular delay assignment $D \in \otimes_{p\in P_\pi} X(\ell(p))$, $d_D(p)$ denote the delay assigned to place $p \in P_\pi$ under $D$, and $\delta_D(\rho)$ denote the value of $\delta(\rho)$ under $D$. For any $x, y \in P_\pi \cup T_\pi$, we denote by $\mathcal{P}(x,y)$ the set of all paths leading from $x$ to $y$, i.e., $\mathcal{P}(x,y) = \{\rho \in N_\pi | x \rightsquigarrow y\}$. From the timing relation (1) (cf., Section 2.3), we note that if there is a path from event $x$ to $y$, $y$ must occur after $x$ by at least the maximum sum of delays on all the paths from $x$ to $y$. That is, whenever $\mathcal{P}(x,y) \neq \emptyset$, we have

$$\tau(x) + \max_{\rho\in\mathcal{P}(x,y)} \delta(\rho) \leq \tau(y). \tag{7}$$

Under a given delay assignment $D$, we say that $x$ is *critical* for $y$ if (7) holds in equality. In that case, there is a path $\rho^* \in \mathcal{P}(x,y)$ such that $\delta_D(\rho^*) = \tau_D(y) - \tau_D(x)$, and we call $\rho^*$ a critical path from $x$ to $y$.

The criticality of timing along paths can be described using a concept of *reference sets*. Roughly speaking, a reference set for event $e$ in a timed execution $\pi$ is a subset of events in $\pi$ such that every path from a source place of $N_\pi$ to $e$ contains at least one event in $R$, and every event in $R$ has a path to $e$.

For example, in Figure 3, it can be checked that $\{{t_1}^{(1)}\}$ is a *minimal* reference set for all the events in the execution. On the other hand, $\{{t_3}^{(1)}\}$ is not a reference set for all other events except for ${t_3}^{(1)}$ itself, because there is at least one path from ${t_1}^{(1)}$ to all other events which does not contain ${t_3}^{(1)}$.

The significance of a reference set is that we can determine the occurrence time of an event $e$ by knowing only the occurrence times of the events contained in a reference set of $e$ plus the delay values of places following the events in the reference set. That is, the following timing relation holds for $e$, where $R$ is a reference set of $e$.

$$\tau(e) = \max_{e'\in R}[\tau(e') + \max_{\rho\in\mathcal{P}(e',e)} \delta(\rho)]. \tag{8}$$

The term $\max_{\rho\in\mathcal{P}(e',e)} \delta(\rho)$ in (8) measures the maximum (random) delay on any path from event $e'$ to $e$ which will be repeatedly referred to later. For convenience, we write:

$$\delta^*(e',e) = \max_{\rho\in\mathcal{P}(e',e)} \delta(\rho), \tag{9}$$

where $\delta^*(e',e) \stackrel{\triangle}{=} -\infty$ if there is no path from $e'$ to $e$, i.e., $\mathcal{P}(e',e) = \emptyset$. For completeness, we define an event $e$ itself to be a path of delay 0, and consequently $\delta^*(e,e) = 0$.

Suppose the TSEs of $k$-th group start in the $j$-th segment of $\pi$, i.e., $S^{(l)}$. Since the set of source places of a segment is a cut of $N_\pi$, the poset of the source places of $S^{(l)}$ must contain a reference set for every event $e$ in the segment (in fact, for every event in $S^{(l')} if l' \geq l$). For convenience, if event $e \in S^{(l')} (l' \geq l)$, let us denote by $R(e,l)$ its reference set contained in the poset of the source places of segment $S^{(l)}$.

To give an upper bound on the $k$-th grouped TSE $\beta^{(k)}$, we first give an upper bound on every TSE in the group. Recall that we used $\alpha^{(k)}$ to denote the number of TSE pairs in group

$k$ and the last TSE instance of the group is indexed by $\eta^{(k)}$ in the global TSE sequence. For every TSE $\gamma^{(m)}(s,t,\varepsilon)$ in the group (thus $\eta^{(k-1)} < m \leq \eta^{(k)}$), Equation (10) computes an upper bound $U_\gamma{}^{(m)}(s,t,0)$ on it, which we state as a lemma.

**Lemma 1** *Suppose the m-th TSE of the triple $(s,t,0)$, i.e., $\gamma^{(m)}(s,t,0)$, occurs in segment $S^{(l)}$. Then, it is upper bounded by $U_\gamma{}^{(m)}(s,t,0)$ defined below where $R(t^{(m)},l)$ is the reference set for event $t^{(m)}$ contained in the poset of the source places of $S^{(l)}$.*

$$U_\gamma{}^{(m)}(s,t,0) = \max_{e \in R(t^{(m)},l)} [\delta^*(e,t^{(m)}) - \delta^*(e,s^{(m)})] \tag{10}$$

The critical fact concerning the above upper bound $U_\gamma{}^{(m)}(s,t,0)$ on $\gamma^{(m)}(s,t,0)$ is that it is independent of the occurrence time of the events in the reference set of $s^{(m)}$. In other words, it relies only on the structure of segment $S^{(l)}$ and the delays of the places within the segment. In particular, it does not depend on the history of the timed execution prior to this segment.

Following Lemma 1, an upper bound on the $k$-th grouped TSE $\beta^{(k)}$ is merely the sum of upper bounds on the TSEs with the group:

$$\beta^{(k)}(s,t,0) \leq U_\beta{}^{(k)}(s,t,0) = \sum_{m=\eta^{(k-1)}+1}^{\eta^{(k)}} U_\gamma{}^{(m)}(s,t,0) \tag{11}$$

By taking the expectation of $U_\beta{}^{(k)}(s,t,0)$, it is now straightforward to show that the average of $\beta^{(k)}(s,t,0)$ over all possible timed executions is upper bounded by $\mathbf{E}U_\beta{}^{(k)}(s,t,0)$. Finally, coupling this upper bound with the i.i.d property of the segment sequence described in Section 4.1, we are now ready to state our main result of this subsection which gives the upper bound on the average grouped TSE $\overline{\beta}(s,t,0)$.

**Theorem 3** *Let $(s,t)$ a transition pair of a stochastic Petri net $\Sigma$ for which Condition 1 holds. Then, the average grouped TSE $\overline{\beta}(s,t,\varepsilon=0)$ is upper bounded by $\mathbf{E}U_\beta{}^{(1)}(s,t,0)$. That is,*

$$\overline{\beta}(s,t,0) \leq \mathbf{E}U_\beta{}^{(1)}(s,t,0) = \mathbf{E} \sum_{0<m\leq\eta^{(1)}} U_\gamma{}^{(m)}(s,t,0), \tag{12}$$

*where*[3] *$U_\gamma{}^{(m)}(s,t,0)$ is given by (10).*

The above theorem states that we can give an upper bound on the average grouped TSE $\overline{\beta}(s,t,0)$ by analyzing only one TSE group of a fixed index (the first group), and over all possible timed executions.

Before closing this section, we discuss several possible algorithms to compute the upper bound on a grouped TSE, i.e., $U_\beta{}^{(k)}(s,t,0)$ in (11). Since a segment structure is acyclic, one possible approach is to use the TSE approximate algorithm described in [CD97] by assigning a delay interval $(-\infty,\infty)$ to every source place of the segment being considered. The algorithm has a complexity no lower than $O(|T(S)|^2)$ where $|T(S)|$ is the number of events in segment $S$.

However, since all the internal places of a segment in a given timed execution has a fixed delay (only the delay values of source places whose delay values are unknown), the problem can be solved more efficiently. In particular, using a simple longest path analysis from a fixed event $e$, the

[3] Note also that the expectation operator in the rightmost-side of (12) does not pass the summation sign since the number of terms under the summation sign, i.e., $\alpha^{(k)} = \eta^{(k)} - \eta^{(k-1)}$, is a random variable.

term $\delta^*(e, t^{(m)})$ in (10) can be determined in $O(|T(S)|+|P(S)|)$ time where $|P(S)|$ is the number of places in segment $S$. Thus, $U_\gamma{}^{(m)}(s,t,0)$ in (10) can be determined in $O((|T(S)|+|P(S)|)*|R|)$ time where $|R|$ is the size of the referent set of $e$. As a result, $U_\beta{}^{(k)}$ in (11) can be determined in $O((|T(S)|+|P(S)|)*|R|)$ time, which is typically much faster than using the method in [CD97]. Alternatively, the longest path analysis can be conducted from each of the interested events back towards their reference events. This way, $U_\beta{}^{(k)}$ is determined in $O((|T(S)|+|P(S)|)*\eta^{(k)})$ time where $\eta^{(k)}$ is the length of $k$-th TSE group.

### 5.3 The upper bound: $\varepsilon \neq 0$ case

We now extend our upper bound on $\overline{\beta}(s,t,\varepsilon=0)$ derived in the previous subsection to the case where $\varepsilon \neq 0$. First, we argue that an upper bound on every TSE in group $k$ can be obtained which is similar to the one in the case of $\varepsilon = 0$.

Note that the only difference between the two cases is that when $\varepsilon \neq 0$, the source and destination events of particular TSE pair $(s^{(n)}, t^{(n)})$ may fall into different segments in the segment sequence. Generally, let us consider the triple $(s,t,\varepsilon)$ and its $k$-th group in a random timed execution $\pi$. Suppose again that this group starts in segment $S^{(l)}$ of the event graph $N_\pi$. Note that all the $\alpha^{(k)}$ source events (i.e., $s^{(m)}, \eta^{(k-1)} < m \leq \eta^{(k)}$) are within the segment $S^{(l)}$. Since transitions $s$ and $t$ satisfy Condition 1, we count that there are $\min\{\alpha^{(k)}, |\varepsilon|\}$ destination events falling outside segment $S^{(k)}$.

Let us first consider the case where $\varepsilon > 0$. In that case, the destination events are either within segment $S^{(l)}$ or in segments following $S^{(l)}$. Therefore, the poset of the source places of segment $S^{(l)}$ constrains a reference set for every destination events (namely, $t^{(m+\varepsilon)}, \eta^{(k-1)} < m \leq \eta^{(k)}$) of the TSE pairs in the $k$-th group. Following the notation in the previous subsection, we denote such a reference set for destination event $t^{(m+\varepsilon)}$ by $R(t^{(m+\varepsilon)}, l)$. Consequently, similar to the result of Lemma 1 in the $\varepsilon = 0$ case, we again have an upper bound $U_\gamma{}^{(m)}(s,t,\varepsilon)$ on the $m$-th TSE $\gamma^{(m)}(s,t,\varepsilon)$ but using the reference set $R(s^{(m+\varepsilon)}, l)$ (instead of $R(s^{(m)}, l)$) where

$$U_\gamma{}^{(m)}(s,t,\varepsilon) = \max_{e \in R(t^{(m+\varepsilon)}, l)} [\delta^*(e, t^{(m+\varepsilon)}) - \delta^*(e, s^{(m)})] \tag{13}$$

for every $m \in (\eta^{(k-1)}, \eta^{(k)}]$. Following a similar argument as in the $\varepsilon = 0$ case (Theorem 3), we deduce an upper bound on $\overline{\beta}(s,t,\varepsilon)$. More precisely, we have

$$\overline{\beta}(s,t,\varepsilon) \leq \mathbf{E}U_\beta{}^{(0)}(s,t,\varepsilon) = \mathbf{E} \sum_{0 \leq m \leq \eta^{(0)}} U_\gamma{}^{(m)}(s,t,\varepsilon). \tag{14}$$

The case here $\varepsilon < 0$ is no more different than the $\varepsilon > 0$ case except that some of the destinations events now appear in the segments prior to segment $S^{(l)}$. We count that the first destination event of the TSE pairs in $k$-th group to be $t^{(\eta^{(k-1)}+\varepsilon+1)}$ since the corresponding first source event is $s^{(\eta^{(k-1)}+1)}$. Let $S^{(l')}(l' < l)$ be the segment containing event $t^{(\eta^{(k-1)}+\varepsilon+1)}$. Then, the poset of the source place of segment $S^{(l')}$ contains a reference set for every source and destination events of the TSE pair in $k$-th group. Following the similar argument to the one made in the $\varepsilon > 0$ case, we upper bound $\overline{\beta}(s,t,\varepsilon < 0)$ by $\mathbf{E}U_\beta{}^{(l)}(s,t,\varepsilon)$. To be precise, we have,

$$\mathbf{E}U_\beta{}^{(l)}(s,t,\varepsilon) = \mathbf{E} \sum_{\eta^{(l-1)} < m \leq \eta^{(l)}} U_\gamma{}^{(m)}(s,t,\varepsilon), \tag{15}$$

where $l$ is a random variable determined as $l = \min\{j+1 | \eta^{(j)} \geq -\varepsilon\}$. That is, $l$ indicates the first segment in the segment sequence of $\pi$ such that all its previous segments contain at least $|\varepsilon|$

occurrences of transition $t$. In addition, each of the terms ${U_\gamma}^{(m)}(s,t,\varepsilon)$ in (15) is now determined by (16) below,

$$ {U_\gamma}^{(m)}(s,t,\varepsilon) = \max_{e \in R(t^{(m+\varepsilon)}, l')} [\delta^*(e, t^{(m+\varepsilon)}) - \delta^*(e, s^{(m)})] \quad (16) $$

with $l'$ being the index of the segment that contains the 0-th group of the TSE sequence.

# 6 Evaluating the bounds

In the previous section, we derived bounds on the average grouped TSE $\overline{\beta}(s,t,\varepsilon)$ in a closed-form expression as given in (12). The expression takes an expectation (i.e.,$\mathbf{E}{U_\beta}^{(1)}(s,t,\varepsilon)$) of a potentially complicated random variable. In some special cases, analytical methods are available to obtain the exact value of the expectation [XB98a]. In general cases, the formula can be estimated using statistical methods to a sufficiently high accuracy in the probabilistic sense. In our current tool, we use the Monte-Carlo approach (e.g.,[MFJ90]).

## 6.1 Monte-Carlo sampling

For convenience, let us write $W$ to denote the above random variable for which the expectation is taken, and let $F_W$ denote its distribution function.

The intuition behind Monte-Carlo simulation approach is the Central Limit Theorem [GS92] which deals with the partial sum of *i.i.d* random variables. In our case, if $W_1, W_2, \cdots, W_n$ are independent random variables all having the same distribution function $F_W$, then for a large $n$, the random variable $\frac{1}{n}S_n$ approaches to $\mathbf{E}W$ where $S_n = W_1 + W_2 + \cdots + W_n$. More precisely, $\frac{1}{n}S_n$ approaches a *normal* distribution with mean $\mathbf{E}W$ and variance $\frac{1}{n}\sigma_W^2$ where $\sigma_W$ is the variance of $W$. Since $\frac{1}{n}\sigma_W^2$ decreases to 0 as $n$ increases, any *realization* of the random variable $\frac{1}{n}S_n$ is a good (unbiased) estimate of $\mathbf{E}W$ when $n$ is large. In fact, the following result is well-known (e.g.,[MFJ90]). For any given *relative error interval* $I_e$ and a confidence level $L_c$, we have $P\{|(\frac{1}{n}S_n - \mathbf{E}W)/\mathbf{E}W| < I_e\} > L_c$ as long as

$$ n > (\frac{z_{x/2}\sigma_W}{I_e\mathbf{E}W})^2, \quad (17) $$

where $x = 1 - L_c$ and $z_{x/2}$ is defined such that the tail probability to its right under the standard normal distribution is $x/2$. The procedure of this realization of $\frac{1}{n}S_n$ is commonly known as *Monte-Carlo* sampling. It involves a random realization of $n$ i.i.d random variables (i.e., $W_1$ through $W_n$) also called as a *random sample* of $W$ with a size of $n$. (17) is referred to as the stopping criterion.

In practice, the true variance of $W$, i.e., $\sigma_W^2$ is not known and is commonly replaced by the variance of the sample, i.e., $S^2 = \frac{1}{n-1}\sum(W_i - \frac{1}{n}S_n)^2$. Similarly, $\mathbf{E}W$ in (17) is replaced by the mean of the sample, i.e., $\frac{1}{n}S_n$ itself. As a consequence, $z_{x/2}$ in (17) with $x = 1 - L_c$ is replaced by $t_{x/2}$ which is defined such that the tail probability to its right under the student $t$-distribution [MFJ90] is $x/2$.

## 6.2 Improving the bounds

In many cases, the upper and lower bounds on the average grouped TSE $\overline{\beta}(s,t,\varepsilon)$ derived in the previous subsections are very close to each other, which combined yield a very good estimate of $\overline{\beta}(s,t,\varepsilon)$, and consequently a very good estimate of $\overline{\gamma}(s,t,\varepsilon)$. As one of the distinct features of

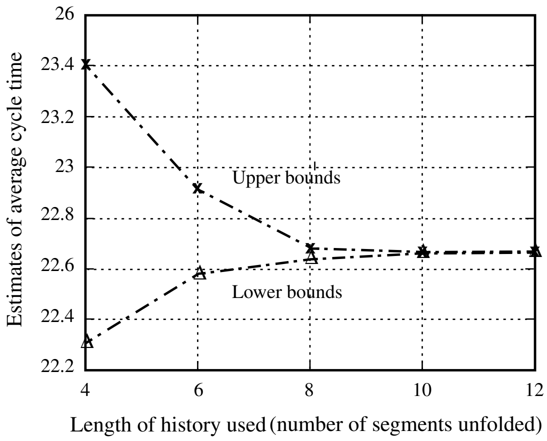



Figure 5: Improving the bounds on the average cycle time of an 8-stage micropipeline model (with choice) by pushing the reference set further apart from the TSE group that formulates the bounds.

our technique, the above bounds can be made even closer to each other by slightly increasing the amount of (consecutive) segments used in the derivation of the bound. That is, we can often improve the bounds by slightly pushing the reference set further apart from the TSEs pairs being considered. As an example, Figure 6.2 illustrates this feature on a micropipeline model with choice which is described in the next section.

# 7 Experimental results

The proposed technique was implemented in C. We experimented with two scalable examples, the first a modified micropipeline FIFO where each stage has choice and the second a Petri net model of Intel's asynchronous instruction length decoding and steering unit design called RAPPID.

## 7.1 Micropipelines with choice

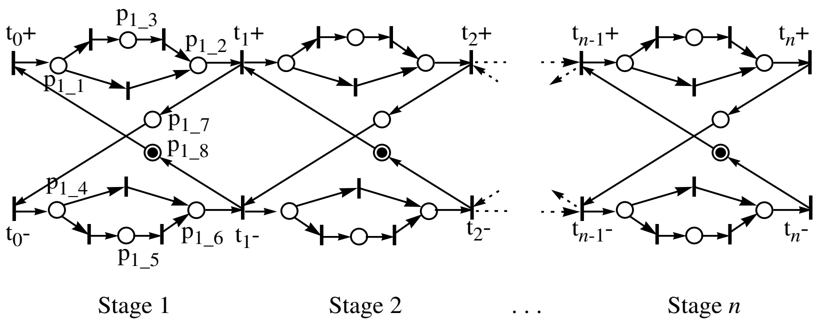



Figure 6: A micropipeline where each stage exhibits choice.

The micropipeline with choice example we tested is depicted in Figure 7.1. For stage $i$, places $p_{i\_1}$ and $p_{i\_4}$ are assigned delays which are uniformly distributed between 0 and 10, places $p_{i\_2}$

and $p_{i\_6}$ are assigned delays which are uniformly distributed between 1 and 2. Places labeled by $p_{i\_3}$ and $p_{i\_5}$ have a fixed of 2. The other two places $p_{i\_7}$ and $p_{i\_8}$ have a fixed delay of 0.5. Each transition out of all choice places receives a probability of 0.5.

We used our tool to study the average cycle time (and thus the average throughput) of these micropipelines as a function of micropipeline depth. The results are described in Table 1.

| $n^\dagger$ | $\|T\|^\dagger$ | $\|P\|^\dagger$ | Bounds $\hat{L}$ | Bounds $\hat{U}$ | CPU secs | $m^\dagger$ |
|---|---|---|---|---|---|---|
| 2 | 18 | 16 | 19.48 | 19.48 | 4.0 | 2026 |
| 4 | 34 | 32 | 21.64 | 21.67 | 38.0 | 9026 |
| 8 | 66 | 64 | 22.65 | 22.66 | 8.1 | 961 |
| 16 | 130 | 128 | 23.16 | 23.23 | 34.3 | 860 |
| 32 | 258 | 256 | 23.90 | 23.96 | 148.8 | 401 |

Table 1: Estimates of average cycle time, i.e., $\overline{\gamma}(t_i+, t_i+, 1)$ with $i$ fixed. The relative error interval is set to 1% and the confidence level is set to 99%. † $n$: number of stages, $|T|$: number of transitions, $|P|$: number of places, $m$: sample size.

As expected, the results indicate the larger the number of stages, the larger the average cycle time and thus the lower average throughput. What is perhaps more impressive, however, is the low complexity the tool demonstrates and the precision of the results. Bounds of width less than 1% are obtained within a few minutes even for a 32-stage micropipeline. Recall, in contrast, that the previous state-of-the-art Markovian analysis technique ran out of memory for much simpler, choice-free micropipelines of more than 8 stages [XB].

We also used our tool to estimate various average response times and latencies between different stages. For example, the average response time from a request of the environment to the corresponding acknowledgment back from the first stage can be computed as $\overline{\gamma}(t_0+, t_1+, 0)$. Similarly, the average latency through the whole micropipeline can be computed as $\overline{\gamma}(t_0+, t_n+, 0)$. The accuracy of the resulting estimates of these quantities and the required CPU times (which are not shown) are similar to those in the average cycle time estimation listed above.

Under certain timing assumptions (e.g., the delay through the environment is no smaller than that of any stage), the above micropipeline models can also be analyzed to obtain an upper bound on their *amortized response times* using the result presented in [EB97]. That is, given the extreme values of the delays through the environment and stages, the theory in [EB97] yields an intuitive analytical upper bound on the average response time irrespective of the delay distributions. In comparison, our approach requires the specification of delay distributions and a numerical evaluation, but it can often yield a much sharper upper bound as well as a lower bound on the average response time, and the mean of our bounds provides a very accurate estimate (cf. also, [XB98a]).

### 7.2 A Petri net model of RAPPID

The RAPPID design is a very aggressive asynchronous design aimed at decoding the length of x86 instructions and steering them to the subsequent instruction decoder with a high average throughput. The instructions are fed into the design through a number of parallel byte units

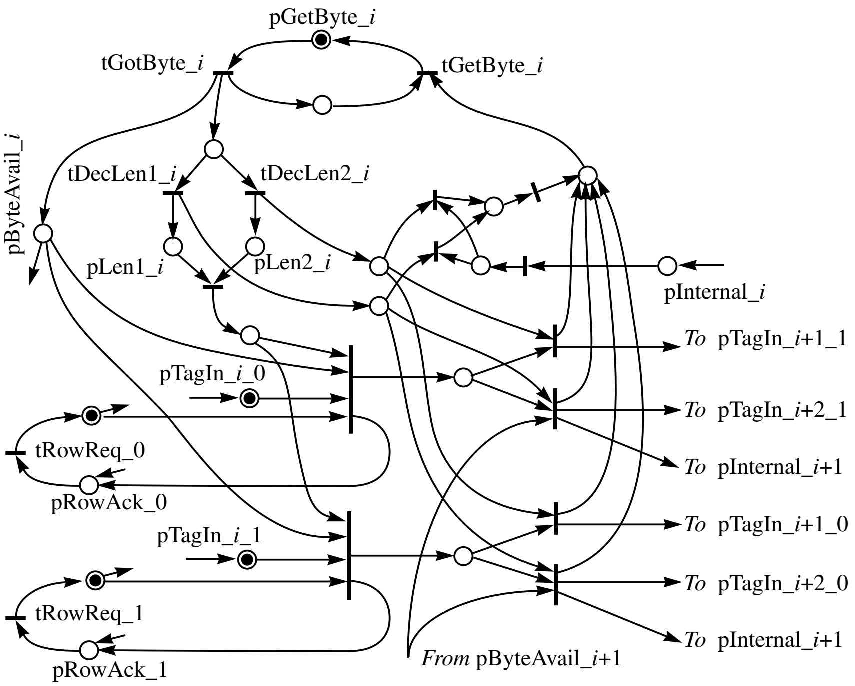



Figure 7: A Petri net model of 1 column of the RAPPID instruction length decoding and steering unit. This column has 2 Tag Unit rows and can handle instructions of length up to 2 bytes.

associated with columns. Each column speculatively decodes the byte under the assumption that an instruction starts at its associated byte. Each column has a number of tag units responsible for receiving a tag indicating this column is the start of the next instruction and routing the token to a tag unit associated with the column which contains the start of the subsequent instruction. The tag units are arranged in rows each associated with an output buffer for the instruction. To alleviate a bottleneck in the output buffers, tag units in one row send its token to tag units in the next row. Both columns and rows are designed in a ring fashion so that they seamlessly handle an infinite number of instructions. The real design has 16 columns, 4 rows, and is optimized to decode instructions of a length up to 7 bytes, and presented in another paper at this conference [RSA+].

Here, we examine a somewhat simplified live safe Petri net model of the design where the delays of places are only loosely related to the real delays in the design. In addition, our model ignores the handling of very rare instructions (e.g., length changing prefixes) and branches. The central driving units for the model is the actual length decoding logic in the columns. These units are optimized for the common instruction lengths and consequently have longer delay for less common instructions, as previously described in [CBG+98]. In our Petri net model, depicted in Figure 7, this logic is modeled with a free-choice place with probability mass function that matches the relative frequency of instruction lengths given in [CBG+98]. Even for a given instruction length the decoding time can vary, thereby motivating the use of stochastic delay models for the decoding of each length.

Once decoded, the column broadcasts a signal to multiple tag units associated with its column indicating its instruction is ready to be transmitted to the output buffer associated with the tag unit. In our Petri net model, the tag unit in the top row of the first column is initially marked with a token to indicate that the first instruction will start at the first column and should be routed to the first output buffer. Once all bytes of the first instruction are available, this tag unit sends a token to a downstream tag unit in the second row of the column where the next instruction starts. For example, if the first column decodes the length of the first instruction to be 2 bytes, the tag unit sends a signal to the tag unit in the second row of the third column. In our Petri net model, we use unique-choice places to route tokens to downstream tag units as well as the consumption of tokens representing all available bytes of an instruction.

| # Columns | # Rows | Length | # Trans. | # Places | Bounds Lower | Bounds Upper | Run-time CPU secs | Sample size |
|---|---|---|---|---|---|---|---|---|
| 2 | 1 | 2 | 29 | 34 | 1.292 | 1.330 | 28.4 | 5,712 |
| 2 | 2 | 2 | 36 | 40 | 2.571 | 2.660 | 33.1 | 2,809 |
| 2 | 3 | 2 | 43 | 46 | 3.899 | 3.940 | 40.8 | 2,304 |
| 2 | 4 | 2 | 50 | 52 | 5.230 | 5.233 | 42.1 | 1,681 |
| 4 | 1 | 2 | 57 | 66 | 0.701 | 0.722 | 44.5 | 2,704 |
| 4 | 2 | 2 | 70 | 76 | 1.349 | 1.371 | 139.7 | 4,761 |
| 4 | 3 | 2 | 83 | 86 | 2.055 | 2.062 | 53.8 | 1,680 |
| 4 | 4 | 2 | 96 | 96 | 2.722 | 2.730 | 142.2 | 2,401 |
| 8 | 1 | 7 | 273 | 210 | 0.961 | 0.983 | 96.9 | 1,765 |
| 8 | 2 | 7 | 338 | 228 | 1.908 | 1.913 | 282.0 | 2,104 |
| 8 | 3 | 7 | 403 | 246 | 2.904 | 2.928 | 476.1 | 2,209 |
| 8 | 4 | 7 | 468 | 264 | 3.803 | 3.819 | 1,052 | 2,209 |
| 16 | 1 | 7 | 545 | 418 | 0.570 | 0.583 | 336.5 | 2,401 |
| 16 | 2 | 7 | 674 | 452 | 1.076 | 1.112 | 1,087 | 2,304 |
| 16 | 3 | 7 | 803 | 486 | 1.574 | 1.604 | 1,412 | 1,849 |
| 16 | 4 | 7 | 932 | 520 | 2.072 | 2.110 | 2,644 | 1,369 |

Table 2: Evaluated bounds on the average decoding cycle time of the RAPPID model. The relative error interval is set to 5% and the confidence level is set to 99% during Monte-Carlo sampling.

Using our tool, we experimented with a number of different RAPPID-based architectures by varying the number of columns, rows, and the lengths handled. For each architecture, the tool was asked to report the average time separation between consecutive occurrences of transition `tRowReq0`, i.e., $\overline{\gamma}(\texttt{tRowReq0}, \texttt{tRowReq0}, 1)$. This average TSE divided by the number of rows in the particular design provides an estimate of its average decoding cycle time and thus its throughput.

We specified uniform delays of between 0.1 and 0.6 time units for pGotBuff, between 0.5 and 1.5 time units for pLen1, pLen2, and pLen3, between 1.0 and 2.5 time units for pLen4 and pLen5, and between 2 and 4 time units for pLen6 and pLen7. The pRowReq was given a fixed delay of 0.5 time units and all other places where given fixed delays of 0.1 time units. Experimental results are tabulated in Table 2. The length of the history used in the experiments (i.e., the number of segments prior to the one considered) was set between 4 to 16 segments depending on the sizes of the individual RAPPID models. Since we did not have sufficient time to match these delay models with the real design, we would rather not stress the significance of the performance trends.

Instead, we prefer to stress the low run-time of the tool and the accuracy of the estimates. In all the examples tested, the width of the derived error bounds are less than 5% and are typically less than 2%. The largest model contains over 900 transitions and 500 places and yet can be analyzed with 5% error tolerance and 99% confidence in less than 45 minutes. These experiments demonstrate that the tool may be used to trade off the silicon area of the design with the system throughput.

## 8 Conclusions

This paper describes a new technique to obtain precise estimates of average TSEs in stochastic timed Petri nets. The technique dynamically generates finite net executions to derive bounds on average TSEs which are evaluated by Monte-Carlo approach. The tool implemented using this technique can analyze dramatically larger and more complex models than the previous state-of-the-art technique. As an example, it successfully analyzed a full-scale Petri net model of Intel's asynchronous instruction length decoding and steering unit called RAPPID which contains over 900 transitions and 500 places.

Nevertheless, the technique has significantly more room for improvement. More sophisticated statistical methods such as *stratified and importance sampling* [Ros90] can be used to replace Monte-Carlo approach to reduce the required size of the random sample. In addition, an extension of the result to Petri nets that can model arbitration is being considered.

We believe this tool is not only useful for architectural exploration and analysis but also may be useful if integrated with performance-driven controller synthesis. Towards the latter, the technique may be extended to analyze the criticality of each components and thereby identify performance bottlenecks of the architecture. To do so, we may need analyze TSEs due to multiple transition pairs, and keep track of more statistics of these TSEs rather than simply their running averages.

## References


[AD94] R. Alur and D. L. Dill. A theory of timed automata. *Theoretical Computer Science*, 126:183–235, 1994.

[Bes87a] E. Best. Some classes of live and safe Petri nets. In *Concurrency and nets,* Special volume in the series, "Advances in Petri nets", pages 71–94. Springer-Verlage, 1987.

[Bes87b] E. Best. Structural theory of Petri nets: The free choice hiatus. In *Lecture Notes in Computer Science, LNCS 254*, pages 168–206. Springer-Verlage, 1987.

[Bes90] E. Best. Partial order behavior and structure of Petri nets. *Formal Aspects of Computing*, 2:123–138, 1990.

[BM95] F. Baccelli and J. Mairesse. Ergodic theorems for stochastic operators and discrete event networks. Technical Report No. 2641, INRIA, 1995.

[BM98] W. Belluomini and C. J. Myers. Verification of timed systems using POSETS. In *Proc. International Workshop on Computer Aided Verification*, pages 403–415, 1998.

[Buc94] P. Buchholz. Exact and ordinary lumpability in finite Markov chains. *Journal of Applied Probability*, 31:59–75, 1994.

[Bur91] S. M. Burns. *Performance Analysis and Optimization of Asynchronous Circuits.* PhD thesis, California Institute of Technology, 1991.

[BV84] E. Best and K. Voss. Free choice systems have home states. *Acta Informatica*, 21:89–100, 1984.

[CBG+98] W.-C. Chou, P. A. Beerel, R. Ginsoar, R. Kol, C. J. Myers, S. Rotem, K. Stevens, and K. Y. Yun. Average-case optimized technology mapping of one-hot domino circuits. In *Proc. International Symposium on Advanced Research in Asynchronous Circuits and Systems (ASYNC)*, pages 80–91. IEEE Computer Society Press, March 1998.

[CCS91] J. Campos, G. Chiola, and M. Silva. Properties and performance bounds for closed free choice synchronized monoclass queueing networks. *IEEE Transactions on Automatic Control*, 36(12):1368–1382, Dec. 1991.

[CD97] S. Chakraborty and D. L. Dill. Approximate algorithms for time separation of events. In *Proc. International Conf. Computer-Aided Design (ICCAD)*, pages 190–194, November 1997.

[EB97] J. Ebergen and R. Berks. Response time properties of some asynchronous circuits. In *Proc. International Symposium on Advanced Research in Asynchronous Circuits and Systems (ASYNC)*, pages 76–86. IEEE Computer Society Press, April 1997.

[GI90] P. W. Glynn and D. L. Iglehart. Simulation output analysis using standardized time series. *Math. Oper. Res.*, 3:1–16, 1990.

[GI93] P. W. Glynn and D. L. Iglehart. *Regenerative Stochastic Simulation.* Academic Press, New York, 1993.

[GS92] G. R. Grimmett and D. R. Stirzaker. *Probability and Random Processes (2nd Edition).* Oxford Science Publications, 1992.

[Haa97] P. J. Haas. Estimation methods for stochastic Petri nets based on standardized time series. In *International Workshops on Petri Nets and Performance Modelds*, pages 194–204, June 1997.

[HB94] H. Hulgaard and S. M. Burns. Bounded delay timing analysis of a class of CSP programs with choice. In *Proc. International Symposium on Advanced Research in Asynchronous Circuits and Systems (ASYNC)*, pages 2–11. IEEE Computer Society Press, November 1994.

[HB95a] H. Hulgaard and S. M. Burns. An algorithm for exact bounds on time separation of events in concurrent systems. *IEEE Transactions on Computers*, 44(11):1306–1317, November 1995.

[HB95b] H. Hulgaard and S. M. Burns. Efficient timing analysis of a class of Petri nets. In *Proc. International Workshop on Computer Aided Verification*, pages 423–436, 1995.

[HMPS96] G. D. Hachtel, E. Macii, A. Pardo, and F. Somenzi. Markovian analysis of large finite state machines. *IEEE Transactions on Computer-Aided Design*, 15(12):1479–1493, December 1996.

[HV87] M. A. Holliday and M. Y. Vernon. A generalized timed Petri net model for performance analysis. *IEEE Transactions on Software Engineering*, 13(12):1297–1310, December 1987.

[Kar78] R. M. Karp. A characterization of the minimum cycle mean in a diagraph. *Discrete mathematics*, 23:309–311, 1978.

[KGBA94] P. Kudva, G. Gopalakrishnan, E. Brunvand, and V. Akella. Performance analysis and optimization of asynchronous circuits. In *Proc. International Conf. Computer Design (ICCD)*, pages 221–225, October 1994.

[Kin73] J. Kingman. Subadditive ergodic theory. *Annals of Probability*, 1:883–909, 1973.

[Maz89] A. Mazurkiewicz. Basic notions of trace theory. In *Linear Time, Branching Time and Partial Order in Logics and Models for Concurrency*, pages 285–363. Springer-Verlage, 1989.

[McM95] K. McMillan. A technique for state space search based on unfolding. *Formal Methods in System Design*, 6(1):45–66, January 1995.

[MFJ90] I. R. Miller, J. E. Freund, and R. Johnson. *Probability and Statistics for Engineers*. Prentice Hall, 1990.

[MM93] C. J. Myers and T. H.-Y. Meng. Synthesis of timed asynchronous circuits. *IEEE Transactions on VLSI Systems*, 1(2):106–119, June 1993.

[Mur89] T. Murata. Petri nets: Properties, analysis and applications. *Proceedings of the IEEE*, 77:541–580, April 1989.

[Ros90] S. M. Ross. *A Course in Simulation*. MacMillian Coll Div., 1990.

[RSA+] S. Rotem, K. Stevens, B. Agapiev, C. Dike, M. Roncken, R. Ginosor, R. Kol, P. A. Beerel, K. Y. Yun, and C. J. Myers. Rappid: An asynchronous instruction length decoding and steering unit. to appear in Proc. International Symposium on Advanced Research in Asynchronous Circuits and Systems (ASYNC), April 1999.

[SY96] A. Semenov and A. Yakovlev. Verification of asynchronous circuits using time Petri-net unfolding. In *Proc. ACM/IEEE Design Automation Conference*, pages 59–63, 1996.

[XB] A. Xie and P. A. Beerel. Accelerating Markovian analysis of asynchronous systems using state compression. to appear in *IEEE Transactions on Computer-Aided Design*.

[XB97] A. Xie and P. A. Beerel. Symbolic techniques for performance analysis of asynchronous systems based on average time separation of events. In *Proc. International Symposium on Advanced Research in Asynchronous Circuits and Systems (ASYNC)*, pages 64–75. IEEE Computer Society Press, April 1997.

[XB98a] A. Xie and P. A. Beerel. Bounding average time seperations of events in stochastic marked graphs, October 1998. submitted to *ACM/IEEE DAC'99*.

[XB98b] A. Xie and P. A. Beerel. Efficient state classification of finite state Markov chains. *IEEE Transactions on Computer-Aided Design*, 17(12):1334–1338, December 1998.

# Appendix: Proofs

Our proof of the weak ergodicity of TSE sequence stated in Theorem 1 uses (1) the weak ergodicity of cycle time sequence of transitions [BM95] and (2) Kingman's *subadditive ergodic theorem* [Kin73].

In [BM95], Baccelli et al proved the weak ergodicity of average cycle time sequence in a wide class of discrete event systems (DES) which enter the so-called *monotone-separable* framework. Roughly speaking, the monotonicity here means delaying one event will not cause any forthcoming events events to occur earlier. The separability here is trivial, which requires that the system returns to a background (or empty) state whenever the system becomes idle. Other requirements for a system to fall into the framework include *causality* in a usual sense that the initiation of an activity should not occur earlier than its completion. Stochastic timed Petri nets with free-choice and unique choice enter this framework. Those with more general choice such as arbitration choice (timing-dependent choice) may not satisfy the above monotonicity requirement, thus may fall outside the framework. Therefore, for the stochastic Petri nets considered in this paper, the average cycle time sequence of any transition converges to a finite number almost surely and in mean. Note that the average cycle times of different transitions are not necessarily the same.

For completeness, we include the Kingman's ergodic theorem here.

**Theorem 4** (Kingman's Ergodic Theorem) *Let $\{Y_{l,n}, l < n \in \mathbf{Z}\}$ be a doubly-indexed stationary random process such that*

$$\begin{aligned} \textit{boundedness:}\quad \mathbf{E}Y_{0,n} &\geq -Cn, n > 0, \textit{ for some finite constant } C > 0, \\ \textit{subadditivity:}\quad Y_{l,n} &\leq Y_{l,m} + Y_{m,n}, \forall l < m < n. \end{aligned}$$

*There exists a constant $\phi$ such that the sequence $\{\frac{Y_{0,n}}{n}, n > 0\}$ converges in mean and almost surely, i.e., $\lim_{n\to\infty} \frac{\mathbf{E}Y_{0,n}}{n} = \phi$, and $\lim_{n\to\infty} \frac{Y_{0,n}}{n} = \phi$, w.p.1.*

**Theorem 1** *Let $\Sigma$ be a LS Petri net that has only free-choices and unique-choices and satisfies stochastic assumptions given in Section 2.2. For any transition pair $(s,t)$ for which Condition 1 holds, its corresponding average TSE with a fixed occurrent offset $\varepsilon$ is a finite constant $\overline{\gamma}(s,t,\varepsilon)$ almost surely and in mean. That is,*

$$\text{Prob}(\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \gamma^{(k)}(s,t,\varepsilon) = \overline{\gamma}(s,t,\varepsilon)) = 1, \tag{18}$$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbf{E}\gamma^{(k)}(s,t,\varepsilon) = \overline{\gamma}(s,t,\varepsilon). \tag{19}$$

*Proof* (*Sketch*) We first prove for $\varepsilon = 0$ case, check the theorem for the case where $\varepsilon \neq 0$ is then trivial.

Let $s, t$ be two transitions of $\Sigma$ that verify Condition 1. From the *coupling argument* [GS92] plus the stationarity of a random timed execution of $\Sigma$, we deduces that there exists a positive integer $K < \infty$ such that their TSE sequence $\gamma^{(k)}(s,t,0) = \{\tau(t^{(k)}) - \tau(s^{(k)})\}$ becomes stationary almost surely. In addition, for all $0 < l < m$, we let $Z_{l,m} = \sum_{k=l+1}^{m} \gamma^{(k)}(s,t,0)$. Clearly, $Z$ is additive since $Z_{l,n} = Z_{l,m} + Z_{m,n}$ for all $0 < l < m < n$.

Next, we show that for large $n$, $\mathbf{E}Z_{L,n}$ is bounded from below by $-c\dot{n}$ for some finite constant $c$ and some finite integer $L$. From Condition 1, we know that $t$ never can have fired infinitely many times than $s$. That is, there is a positive integer $h$ such that for any firing sequence $\sigma$ of $\Sigma$, $|\#_\sigma(s) - \#_\sigma(t)| \leq H$. Consequently, $\tau(t^{(k)}) \geq \tau(s^{(k+h)})$ for all $k > H$. It can then be checked that for all $n > L = \max\{K, H\}$

$$Z_{L,n} = \sum_{k=L+1}^{n} \tau(t^{(k)} - \tau(s^{(k)}) \geq - \sum_{k=n-H+1}^{n} \tau(s^{(k)}). \tag{20}$$

Taking expectation on both sides of (20), the desired bounded of $\mathbf{E}Z_{L,n}$ follows the weak ergodicity of cycle time of a transition as described earlier.

As such, $Z_{L,n}$ verifies the conditions of Kingman's ergodic theorem, and we deduce that the sequence $\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} \gamma^{(k)}(s,t,0) = \lim_{n\to\infty} \frac{1}{n} Z_{0,n}$ equals to a constant almost surely and in mean. For the case where $\varepsilon \neq 0$, the theorem can be checked by similar arguments. ■

**Theorem 2** *Let $\Sigma$ be a stochastic LS Petri net that has only free-choices and unique-choices, and $(s,t)$ be a transition pair for which Condition 1 holds. Then, for any fixed $\varepsilon$, the average of the grouped TSE sequence (group length sequence) $\{\beta^{(k)}(s,t,\varepsilon) : k \geq 1\}$ ($\{\alpha^{(k)}(s) : k \geq 1\}$) converges to a finite constant $\overline{\beta}(s,t,\varepsilon)$ ($\overline{\alpha}(s)$). That is,*

$$\frac{1}{n}\sum_{k=1}^{n} \beta^{(k)}(s,t,\varepsilon) \;\to\; \overline{\beta}(s,t,\varepsilon), \tag{21}$$

$$\frac{1}{n}\sum_{k=1}^{n} \alpha^{(k)}(s) \;\to\; \overline{\alpha}(s). \tag{22}$$

*The convergence in (5) and (6) takes almost surely and in mean. Moreover, $\overline{\beta}(s,t,\varepsilon) = \overline{\alpha}(s)\overline{\gamma}(s,t,\varepsilon)$ where $\overline{\gamma}(s,t,\varepsilon)$ is the average TSE of the triple $(s,t,\varepsilon)$.*

*Proof* The proof of this theorem is straightforward with the result of Theorem 1 where we showed the similar convergence modes for the (running) average of the TSE sequences.

Let us first show that the average grouped TSE sequence $\{\frac{1}{n}\sum_{k=1}^{n} \beta^{(k)}(s,t,\varepsilon) : n \geq 0\}$ converges almost surely for very triple $(s,t,\varepsilon)$ such that $(s,t)$ verifies Condition 1 and $\varepsilon$ being finite. Towards this end, let us consider a random timed execution $\pi$ of $\Sigma$. We introduce a random variable $h(n)$ which denotes the total number of TSE pairs that start in any of the first $n$ TSE groups of $\pi$, namely, $h(n) = \sum_{k=1}^{n} \alpha^{(k)}(s)$. Apparently, $h(n) \to \infty$ if $n \to \infty$ since $\alpha^{(k)}(s) \geq 1$ for every $k$. Now,

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} \beta^{(k)}(s,t,\varepsilon) &= \lim_{n\to\infty} \frac{h(n)}{n} \cdot \frac{1}{h(n)} \sum_{k=1}^{n} \beta^{(k)}(s,t,\varepsilon) && (23)\\
&= \lim_{n\to\infty} \frac{h(n)}{n} \cdot \frac{1}{h(n)} \sum_{l=1}^{h(n)} \gamma^{(l)}(s,t,\varepsilon) && (24)\\
&= \lim_{n\to\infty} \frac{h(n)}{n} \cdot \overline{\gamma}(s,t,\varepsilon), \text{ almost surely.} && (25)
\end{aligned}$$

Equation (25) follows (24) since $\lim_{h(n)\to\infty} \frac{1}{h(n)} \sum_{l=1}^{h(n)} \gamma^{(l)}(s,t,\varepsilon)$ *exists* and equals to the constant $\overline{\gamma}(s,t,\varepsilon)$ almost surely according to Theorem 1.

Notice that $\alpha^{(k)}(s)$ and $\alpha^{(j)}(s)$ are independent and identically distributed (i.i.d) random variables since the labeling of segments $S^{(k)}$ and $S^{(j)}$ are i.i.d variables (as described in Section

4.1). According to the Central Limit Theorem [GS92], we have

$$\lim_{n\to\infty} \frac{h(n)}{n} = \lim_{n\to\infty} \frac{1}{n} \sum_{n\to\infty} \alpha^{(n)} = \mathbf{E}\alpha^{(0)}(s), \text{ almost surely and in mean,} \tag{26}$$

which proves the part of the theorem concerning the group length sequence.

With (26), the almost surely convergence of the average grouped TSE sequence follows (25). In addition, we see $\overline{\beta}(s,t,\varepsilon) = \overline{\alpha}(s)\overline{\gamma}(s,t,\varepsilon)$ where $\overline{\alpha}(s) = \mathbf{E}\alpha^{(0)}(s)$.

Using similar arguments, we check the grouped TSE sequence also converges in mean by the fact that both the average group length sequence and the average TSE sequence converge in mean. ∎

We know prove Lemma 1 and Theorem 3 concerning our upper bound on the average grouped TSE.

**Lemma 1** *Suppose m-th TSE of the triple $(s,t,0)$, i.e., $\gamma^{(m)}(s,t,0)$, occurs in segment $S^{(l)}$. Then, it is upper bounded by ${U_\gamma}^{(m)}(s,t,0)$ defined below where $R(t^{(m)},l)$ is the reference set for event $t^{(m)}$ contained in the poset of the source places of $S^{(l)}$.*

$${U_\gamma}^{(m)}(s,t,0) = \max_{e\in R(t^{(m)},l)} [\delta^*(e,t^{(m)}) - \delta^*(e,s^{(m)})] \tag{27}$$

*Proof* The lemma is proved by a common practice of critical path analysis. Let $R(t^{(m)},l)$ be the reference set for event $t^{(m)}$ contained in the poset of the source places of $S^{(l)}$. From Condition 1, we know that $s^{(m)}$ is also in segment $l$, and denote its reference set contained the poset of $S^{(l)}$ by $R(s^{(m)},l)$.

From timing relation (8) and (9), we have

$$\begin{aligned} \tau(t^{(m)}) &= \max_{e\in R(t^{(m)},l)} [\tau(e) + \delta^*(e,t^{(m)})], \text{ and} && (28)\\ \tau(s^{(m)}) &= \max_{e\in R(s^{(m)},l)} [\tau(e) + \delta^*(e,t^{(m)})]. && (29) \end{aligned}$$

Thus,

$$\begin{aligned} \tau(t^{(m)}) - \tau(s^{(m)}) &= \max_{e\in R(t^{(m)},l)} [\tau(e) + \delta^*(e,t^{(m)})] - \max_{e\in R(s^{(m)},l)} [\tau(e) + \delta^*(e,t^{(m)})] && (30)\\ &\le \max_{e\in R(t^{(m)},l)} [\tau(e) + \delta^*(e,t^{(m)}) - (\tau(e) + \delta^*(e,s^{(m)}))] && (31)\\ &= \max_{e\in R(t^{(m)},l)} [\delta^*(e,t^{(m)}) - \delta^*(e,s^{(m)})] && (32) \end{aligned}$$

where (31) follows (30) properly because $\delta^*(e,e')$ is defined to be $-\infty$ if there is no path leading from $e$ to $e'$ (cf. (9)). ∎

**Theorem 3** *Let $s$ and $t$ be any two transitions of a stochastic Petri net $\Sigma$ for which Condition 1 holds. Then, their average grouped TSE $\overline{\beta}(s,t,\varepsilon=0)$ is upper bounded by $\mathbf{E}{U_\beta}^{(0)}(s,t,0)$. That is,*

$$\overline{\beta}(s,t,0) \le \mathbf{E}{U_\beta}^{(0)}(s,t,0) = \mathbf{E} \sum_{0\le m\le \eta^{(0)}} {U_\gamma}^{(m)}(s,t,0), \tag{33}$$

*where ${U_\gamma}^{(m)}(s,t,0)$ is given by (10).*

*Proof* Following Lemma 1, an upper bound on the $k$-th grouped TSE $\beta^{(k)}$ is merely the sum of upper bounds on the TSEs with the group, i.e., $\beta^{(k)}(s,t,0) \leq U_\beta{}^{(k)}(s,t,0) = \sum_{m=\eta^{(k-1)}+1}^{\eta^{(k)}} U_\gamma{}^{(m)}(s,t,0)$. By taking the expectation of $U_\beta{}^{(k)}(s,t,0)$, it is now straightforward to show that the average of $\beta^{(k)}(s,t,0)$ over all possible timed executions is upper bounded by $\mathbf{E}U_\beta{}^{(k)}(s,t,0)$. Finally, coupling this upper bound with the strong stationarity of the segment sequence described in Section 4.1, we achieve an upper bound on the average grouped TSE $\overline{\beta}(s,t,0)$ as given in the theorem. Specifically, we merely need the convergence *in mean* of the average grouped TSE sequence shown in Theorem 2 in the last statement. ■

**Remark 1** It might be useful to note that in a random timed execution, the grouped TSEs $\beta^{(k)}(s,t,\varepsilon)$ and $\beta^{(j)}(s,t,\varepsilon)$ are generally not independent for $j \neq k$. However, their upper bounds $U_\beta{}^{(k)} = \sum_{m=\eta^{(k-1)}+1}^{eta^{(k)}} U_\gamma{}^{(m)}(s,t,0)$ and $U_\beta{}^{(j)} = \sum_{m=\eta^{(j-1)}+1}^{eta^{(j)}} U_\gamma{}^{(m)}(s,t,0)$ (derived from Lemma 1) are independent (and identically distributed). It is this key fact that makes it possible for us to give bounds on the average (grouped) TSE sequence in a random infinitely long timed execution.