Bounding Average Time Separations of Events in Stochastic Marked Graphs

Aiguo Xie and Peter A. Beerel

November 6, 1998

Abstract

Stochastic timed marked graphs are graphical models of concurrent systems such as asynchronous circuits, embedded systems, queuing networks, manufacturing systems, and many automatic control systems. Unlike earlier works in which delays must be fixed or exponential, we allow the models to include arbitrary delay distributions as long as they have finite means. For such models, one important problem is to determine the average Time Separations of Events (TSE's). For example, an efficient means of finding TSE's in such models of asynchronous circuits facilitates both performance analysis as well as performance-driven synthesis. Towards this end, we present a novel technique to obtain upper and lower bounds on the average TSE for arbitrary pairs of system events. The bounds are formulated using a finite segment of the infinite unfolding of the marked graph and can be efficiently evaluated either using statistical sampling or, in some special cases, analytical methods. The resulting bounds are typically much sharper than using any other known method. The efficiency of technique and the quality of the bounds are demonstrated on several asynchronous pipelines.

1 Introduction

The purpose of this paper is to present a novel technique to bound a quantity called the *average time separation of events* (TSE), a generic problem in analyzing timed concurrent and distributed systems. In this paper, we focus on bounding average TSE's of such systems that can be modeled using stochastic timed *marked graphs* (e.g., [1]), also referred to as timed *event-graphs* (e.g., [2, 3]) or *decision-free Petri nets* (e.g., [4]). Although marked graphs are a restrict class of Petri nets [5], they constitute an adequate model for many real-world systems, including many asynchronous circuits and embedded systems. In different application domains, the TSE's are typically characterized either for their *extreme values* or for their *average values*. Extreme values of TSE's can be very useful in verifying correct operations (e.g., [6, 7]). On the other hand, the average TSE's are particularly important in analyzing system performance such as average throughput, latency, and resource utilization (e.g, [8, 4, 9, 10]) as well as possibly yield estimation. For instance, the average throughput of an asynchronous pipeline is the inverse of the average time separation of consecutive output requests.

Over the last two decades, the problem of computing the average TSE's in basic classes of Petri nets has been studied extensively (e.g., [11, 4, 3, 2, 12, 13, 14, 15, 9, 16, 10]) although the majority of the work was on average system throughput. The pioneer work by Karp [11] solved this problem for mean cycle time for marked graphs with deterministic delays. In the stochastic cases where delays are not fixed, the problem has been addressed using Markovian analysis (e.g., [8, 9, 10]) and event-driven simulation. Despite some recent results, the Markovian analysis techniques are still often limited to small systems because of the *state explosion problem* ([17, 18, 19]). Because of the inherent cyclic nature (hence the memory) of general concurrent systems, simple event-driven simulation approaches suffer from lack of confidence in analyzing the steady-state system behavior (e.g., [20]). Consequently, the state-of-art techniques to compute the exact value of average TSE's is not adequate for large systems.

Because of these difficulties with finding the exact value of the average cycle time, many researchers have resorted to efficiently finding lower and upper bounds of the average cycle time. For special marked graphs (i.e., with tree-like structures), Ebergen and Berks give nearly exact bounds on *amortized* response time [16]. Campos et al [13] give average throughput bounds using linear programming and later extended their results to free-choice Petri nets [14, 21]. A distinct feature of their bounds is that they are independent of higher moments (except the first moment) of the delays. In practice, these bounds can be efficiently computed at structure level instead of state-level although exponential run time may also be hit in the worst-case. Unfortunately, in many real applications, their bounds are too loose to be useful.

The technique proposed in this paper also gives lower and upper bounds on average TSE's, but the results tend to be much sharper, and typically are very close to the true value of the average TSE. It is well-known that the all possible executions of a Petri net can be captured by an infinite unfolding of the net [22]. The challenge we are faced with is to characterize the average delay determined by this infinite unfolding in finite time. Fortunately, for marked graphs, the infinite unfolded graph has a repeated structure. Consequently, we show that we can analyze a finite unfolding of the graph to obtain lower and upper bounds on the TSE of the targeted events. In particular, the paper presents a method to identify the degree of unfolding sufficient to find the bounds and the proof that such a finite unfolding always exists. More specifically, to obtain the bounds from this finite unfolding we identify a set of reference events from which the delays to the targeted pair of events are analyzed. By ignoring the time separation between reference events we obtain bounds that are independent of the long-term history of the net. Consequently, we can prove that our derived bounds are valid for the infinite unfolding. The principle penalty for ignoring the time separation between reference events is that sometimes it leads to wide bounds. However, this penalty can be reduced (often arbitrarily) by analyzing a larger unfolding. In all the experiments we have done, the technique yields sharp lower and upper bounds, and exhibits a time complexity that is polynomial in system size.

We also note that the idea of bounding stochastic measures of sequential systems using statistical methods is not new. In particular, Kozhaya and Najm's used statistical methods to bound the power estimation of synchronous sequential circuits (modeled as finite state machines) [23]. Their method is similar to ours in that the infinite execution of the sequential circuit is analyzed using a finite execution. In their case they ignore the initial state of the sequential circuit whereas in our case we ignore the time separation between reference events. Another difference between our two works is that our analysis yields closed form equations which sometimes can be computed using analytical techniques.

The organization of the remaining sections is as follows. Section 2 gives a quick review of timed marked graph as a basic class of Petri nets, our stochastic delay models, and shows the existence of average TSE's. A detailed description of our technique is given in Section 3. Section 4 briefly discusses several approaches to evaluate the derived formulae of the bounds. The next section is devoted to an easy extension of the technique described in Section 3 to further sharpen the bounds. Experimental results are described in Section 6. The paper is concluded in Section 7 where we discuss possible extensions to more general Petri nets. For the review purpose and the sake of theoretical completeness, an appendix is included for the proofs of all the lemmas and theorems new to this work.

2 Stochastic timed marked graphs

We start by a quick recall of the basic structure, firing rules, and properties of Petri net models, in particular, marked graphs. (For further details on Petri nets, see, e.g., [5].) Next, we discuss our stochastic delay models and show the existence of average TSE's in stochastic timed marked graphs.

2.1 Marked graphs

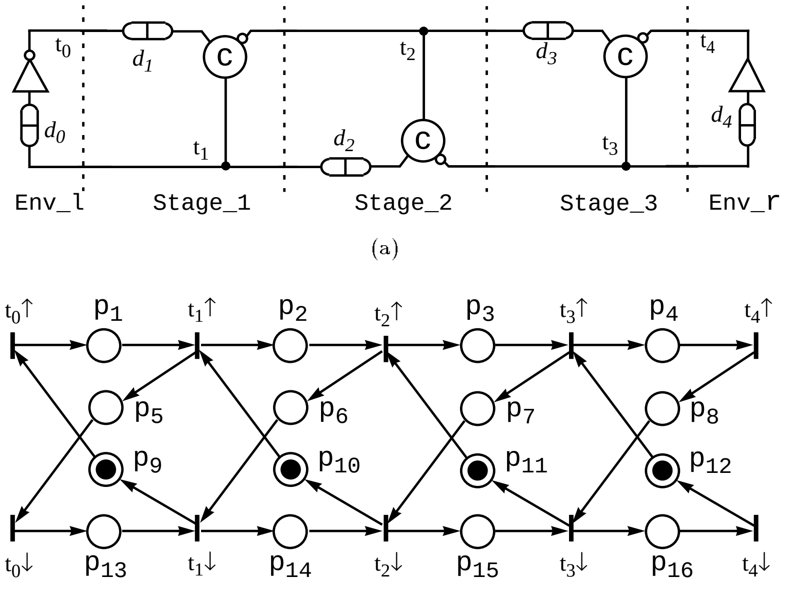
As is usual, we denote a Petri net by a triple (P, T, F) where P is the set of places, T the set of transitions, and $F : (P \times T \cup T \times P) \rightarrow \{0, 1\}$ the flow function (or the incident matrix). The preset of $x \in P \cup T$, denoted by $\bullet x$, is the set $\{y \in P \cup T \mid (y, x) \in F\}$, and its poset $x\bullet = \{y \in P \cup T \mid (x, y) \in F\}$. A Petri net is a *marked graph* if $|\bullet p| = |p\bullet| = 1, \forall p \in P$, i.e., every *place* has a unique input and output transition.

A *marking* is a mapping $M : P \rightarrow \mathbf{N} = \{0, 1, 2, \cdots\}$. The number of tokens in place p under marking M is denoted by $M(p)$. If $M(p) > 0$, one says p is in the *support set* of M. A transition $t \in T$ is enabled at marking M iff all its input places have at least one taken, 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 $\bullet t$ and deposits one token at each place in $t\bullet$, leading to a new marking M', denoted by $M[t\rangle M'$.

A *marked* net is a tuple $\Sigma = (N, M_0)$ where $N = (P, T, F)$ is a Petri net and M_0 is a marking of N, also called the initial marking of Σ. Sometimes, we call N the underlying net of Σ, denoted by N_Σ. $R(M_0)$ denotes all reachable marking of N_Σ from M_0. Σ is *live* iff every transition can be enabled from every $M \in R(M_0)$. It is *safe* if no place ever contains more than one token, i.e., $M(p) \leq 1$ for every $M \in R(M_0)$. It is a well-known result in the literature [24] that a live and safe (LS) marked Petri net has no source or sink places and no source or sink transitions. This implies that a LS Petri net can be partitioned into a set of strongly connected components each evolving independently of others. In the sequel, we assume the entire net is strongly connected. In particular, we restrict ourselves to live and safe marked graphs (LSMG's).

A sequence $(v_1, v_2, \cdots, v_{m+1})$ is a (directed) *path* if $v_i \in P \cup T$ and $(v_i, v_{i+1}) \in F$ for every $1 \leq i \leq m$. A circuit is a path with its head and tail being identical. Below, when we write *circuits*, we refer to elementary ones, i.e., those not containing any other circuits. It is again well-known in the literature [25] that a marked graph (G, M_0) is live and safe *iff* every circuits has at least one token and every place belongs to a circuit that has only one token. We close this subsection by a LSMG example.

Example Figure 1(b) shows a marked graph that models the behavior of the control circuit of an asynchronous micropipeline [26] (Figure 1(a)). The pipeline has three stages between two environments



(a)

(b)

Figure 1: An asynchronous micropipeline: (a). the control circuit, and (b). its marked graph model.

(a sender, i.e., Env_l, and a receiver, i.e., Env_r). Initially, the outputs of C-elements are all low. When Env_l has data ready, it sends a *request* to stage_1 by flipping the logic value of signal t_0. When stage_1 finishes processing the data, it latches the result. Meanwhile, it sends a request to stage_2 and simultaneously, sends an acknowledge back to Env_l by flipping the logic value of t_1. Following that, stage_2 starts processing and when it is done, it latches the result, and meanwhile it sends a request to the stage_3 and an *acknowledge* back to stage_1. When stage_2 is busy processing data, stage_1 may possibly be processing new data if it receives another request from Env_l, demonstrating potential concurrency of the system. However, stage_1 cannot latch its new result until it receives an acknowledge from stage_2 to avoid ruining the data being processed at stage_2. This latter constraint shows the potential *data blocking*. The above behavior is captured by the marked graph in Figure 1(b). With the initial marking as shown, we may check that the marked graph is live and safe. ■

2.2 Stochastic modeling and timing constraints

In this paper, we associate delays with places[1]. 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 the output transition of p. Let $X(p)$ be the random variable denoting the delay associated with place p. The delays experience by different tokens in the same place are independent. We also assume that $X(p)$ and $X(q)$ are independent if p and q are different places, i.e., $p \neq q$. For each (random) delay variable $X(p)$, we let $F_{X(p)} : R \rightarrow [0, 1]$ denote its distribution function. That is, $F_{X(p)}(x) = Prob(X(p) \leq x)$. We shall not put any restriction

[1] Sometimes, delays are associated with transitions or both in the literature. For marked graphs, one can always model these cases in our place-delay semantics by adding extra places and transitions.

on the distribution functions except that they all have finite first moments, namely, $\mathbf{E}X(p) < \infty, \forall p \in P$. When a transition is enabled and all the tokens in its preset places are available, it must fire. The actual firing of a transition is assumed to be instantaneous.

As an example, in the micropipeline example (Figure 1), let us assume the C-elements (inside stages), the inverter and the buffer (inside the environments) all have unit delay. The (random) delay variables ($d_0 \sim d_3$) represent the data processing delays in corresponding stages and environments. It is not difficult to determine the delay on each places in marked graph model (Figure 1(b)). For instance, $X(p_1) = d_1 + 1$, $X(p_2) = d_2 + 1$, etc.

In a marked graph, since each place has a unique input and output transition, the basic timing constrains among the firing of transitions can be easily expressed through the delays on places as follows. Suppose $\tau(t^{(k)})$ denotes the time when transition $t \in T$ fires for its k-th time. In particular, we define $\tau(t^{(0)}) = 0$ if $t^\bullet \cap I \neq \emptyset$. Then, for every $k > 0$, we have,

$$\tau(t^{(k)}) = \max_{s \in T, s\bullet \cap \bullet t \neq \emptyset} \tau(s^{(k-\epsilon)}) + X(s\bullet \cap \bullet t). \qquad (1)$$

where ϵ is the initial occurrence index offset. That is, $\epsilon = 1$ if $s\bullet \cap \bullet t \in M_0$, and 0 otherwise.

2.3 Average cycle time and time separation of events

The k-th occurrence time of a transition $t \in T$, namely, $\tau(t^k)$ is a random variable. The k-th time separation between transitions s and t with an occurrence offset ε is also a random variable which we denote by $\overline{\gamma}^{(k)}(s, t, \varepsilon)$. That is,

$$\gamma^{(k)}(s, t, \varepsilon) = \tau(t^{(k+\varepsilon)}) - \tau(s^{(k)}), k \in \mathbf{N}, \varepsilon \geq -k.$$

The above time separation sequence, i.e., $\{\gamma^{(k)}(s, t, \varepsilon) : k \geq 0\}$ forms a stochastic process. There are cases, for instance when all delays are deterministic, in which this sequence can be periodic in k. However, what we are interested in this paper is the *average* of this sequence over all k, namely, the average time separation between s and t as defined below.

Definition 1 *Let (G, M_0) be a stochastic timed MG. The average time separation between transitions s and t with occurrence index-offset ε is $\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \gamma^{(k)}(s, t, \varepsilon)$.*

Note that the average *cycle time* of a transition u is a special case of the average TSE defined above with $s = t = u$ and $\varepsilon = 1$. It is well-known that all the transitions in a live stochastic marked graph has the (finite) same cycle time [12, 27]. We shall show that the limit in the above definition in fact converges for every transition pair (s, t) and every finite ε. Theorem 1 states this result which generalizes the *weak ergodicity property* of the cycle time (of one transition) [12, 27] to that of the time separations of two arbitrary transitions (see the Appendix for a proof of the theorem).

Theorem 1 *Let s, t be two arbitrary transitions of a LS stochastic timed MG (G, M_0) as defined in Section 2.2. Their average time separation (with occurrence offset ε) converges to a constant $\overline{\gamma}(s, t, \varepsilon)$ almost surely and in mean, i.e.,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \gamma^{(k)}(s, t, \varepsilon) = \frac{1}{n} \sum_{k=1}^{n} \mathbf{E}\gamma^{(k)}(s, t, \varepsilon) = \overline{\gamma}(s, t, \varepsilon) \quad \textit{a.s., and in mean.}$$

Consequently, under fairly weak assumptions on delay models, we see that the average of the time separation sequence of arbitrary transitions converges almost surely to some constant. Bounding this constant for arbitrary event pairs is the focus of this paper. For convenience, we refer to (s, t, ε) as an *event separation triple*, meaning we are interested in the average TSE $\overline{\gamma}(s, t, \varepsilon)$.

3 Bounding the average TSE's

One obvious challenge in bounding the average TSE's is that they are defined over infinite execution of the timed marked graph. In this section, we show that the average TSE's can be bounded by identifying and analyzing a special finite segment (meaning that we do not have to explicitly analyze the infinite execution). We show such a finite segment can always be found for any event separation triple (s, t, ε). All the lemmas and theorems appearing in this section can be found in the Appendix.

3.1 The duality of the bounds

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

We note that for any $j, \varepsilon \in \mathbf{N}$ s.t. $j > -\varepsilon$, from the definition of γ, we have $-\gamma^{(j)}(s, t, \varepsilon) = -t^{(j+\varepsilon)} + s^{(j)} \triangleq \gamma^{(j+\varepsilon)}(t, s, -\varepsilon)$. Therefore, if one finds an upper bound, say, 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 finding the upper bounds from now on.

3.2 Unfolding

The *untimed* behavior of a marked graph (G, M_0) (either in terms of transition firing sequence or marking sequence) can be directly reasoned through net unfolding [22]. In this subsection, we discuss some simple properties of the unfolded graph of (G, M_0) which serve as key bases of later sections.

The unfolding of (G, M_0) results an acyclic marked graph $H = (P_H, T_H, F_H)$. All source places of H are marked initially. The sets P_H and T_H represent the instances of places in P and transitions in T, respectively. The flow relation F_H captures the causality of the firing of transitions as the net evolves. As an example, Figure 3(a) shows a finite unfolding of the marked graph in Figure 1 where every transition is instanced exactly once. It may be useful to note that from the unfolded graph, one may enumerate all possible transition sequences leading from one reachable marking to another.

Figure 2 gives a simple procedure that unfolds (G, M_0) once. It instances every transition of G exactly once, and returns an unfolded graph $G^{(0)} = (T^{(0)}, P^{(0)}, F^{(0)})$. The (inverse) labeling function $\ell : T^{(0)} \cup P^{(0)} \rightarrow T \cup P$ maps the instanced transitions and places to their corresponding ones in G.

Lemma 1 *Procedure 1 terminates for every LSMG (G, M_0). Moreover, when it returns, $M = M_0$.*

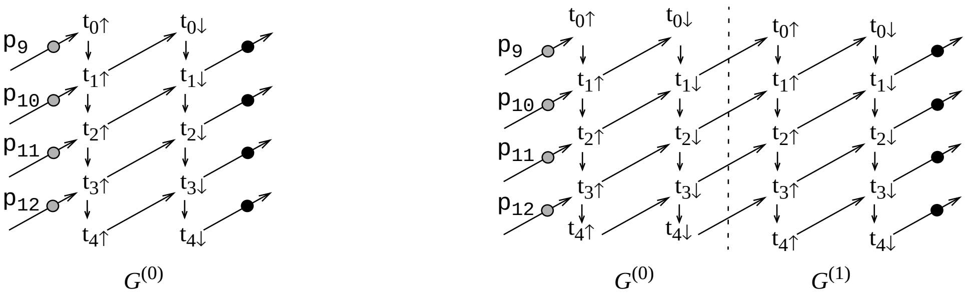
Lemma 1 verifies that when all the transitions are instanced exactly once, the sink places of the unfolded graph coincide with the support set of the initial marking (recall that a place p is in the support set of M if $M(p) > 0$). This is intuitive because for a marked graph if a firing sequence which fires all transitions

```
Procedure 1  Unfold($(G, M_0)$)                /* $G = (P, T, F)$ */
    $P^{(0)} \leftarrow \emptyset, T^{(0)} \leftarrow \emptyset, F^{(0)} \leftarrow \emptyset, M \leftarrow M_0$;      /* initialize */
    foreach ($p \in P$ s.t. $M_0(p) > 0$)
        $P^{(0)} \leftarrow P^{(0)} \cup \{p^{(0)}\}$, $\ell(p^{(0)}) \leftarrow p$;      /* make a copy of $p$ and label it */
    while $T \neq \emptyset$ {
        pick a $t \in T$ s.t. $t$ is enabled at $M$;
        $T^{(0)} \leftarrow T^{(0)} \cup \{t^{(0)}\}$, $\ell(t^{(0)}) \leftarrow t$;      /* make a copy of $t$, and label it */
        foreach ($p \in \bullet t$)
            $F^{(0)} \leftarrow (p^{(0)}, t^{(0)})$, $M(p) \leftarrow 0$;      /* add an edge from $p^{(0)}$ to $t^{(0)}$, erase token in $p$ */
        foreach ($p \in t \bullet$) {
            $P^{(0)} \leftarrow P^{(0)} \cup \{p^{(0)}\}$, $\ell(p^{(0)}) \leftarrow p$;      /* make a copy of $p$, and label it */
            $F^{(0)} \leftarrow (t^{(0)}, p^{(0)})$, $M(p) \leftarrow 1$; }      /* add an edge from $t^{(0)}$ to $p^{(0)}$, put a token in $p$ */
        $T \leftarrow T \setminus \{t\}$; }      /* $t$ has been instanced */
    return $G^{(0)} = (T^{(0)}, P^{(0)}, F^{(0)})$ and $\ell$;
```

Figure 2: The unfolding procedure.

an equal number of times, it leads back to the same marking [28]. We call the above unfolding the 0th unfolding of G.

Example The 0th unfolding of the marked graph in Figure 1(b) is shown in Figure 3(a). It starts with the initial marking (shaded) and ends with the same marking (dark) after instancing all the transitions once. In the figure, we have omitted the instanced places, assuming there is a place instanced on each edge. In addition, we have dropped the occurrence indices of transitions in order to simplify the exposition. (They should be clear from the indices of the unfolding subgraph, e.g., $G^{(0)}$.) ■



(a) Unfolding once,

(b) unfolding twice.

Figure 3: Unfolding the marked graph in Figure 1(b).

Starting from the marking obtained from the 0th unfolding, we may repeat the unfolding process and obtain the 1st unfolded graph $G^{(1)}$ (e.g., Figure 3(b)) and so forth. For all $j, k \in \mathbf{Z}_+$ such that $j \leq k$, let us define an unfolding segment $H(j,k)$ to be the unfolded subgraph generated from j-th to k-th unfolding, i.e.,

$$H(j,k) = \cup_{i=j}^{k} G^{(i)} \tag{2}$$

For convenience, we define a *shift* operator Δ on a *subgraph* of an unfolded segment. It changes the indices of all the elements of the subgraph by a constant. For example, $\Delta_l H(j,k)$ results in a same graph as $H(j,k)$ except the indices of all the nodes and edges are added by an amount l. In particular, the fact that $G^{(j)}$ (i.e., $H(j,j)$) and $G^{(j+1)}$ (i.e., $H(j+1,j+1)$) are isomorphic (due to Lemma 1) is equivalent to saying $H(j,j) = \Delta_{-1} H(j+1,j+1)$. More generally, we have the following simple lemma which states that the unfolded graph of a LSMG has repeated unfolding segments.

Lemma 2 *For all $j,k,l \in \mathbf{N}$, the unfolding segments $H(j+l,k+l)$ and $H(j,k)$ are identical under the shift operator, i.e., $H(j+l,k+l) = \Delta_l H(j,k)$.*

3.3 Reference sets

In this subsection, we derive an upper bound on average TSE $\overline{\gamma}^{(k)}(s,t,\varepsilon)$ for any fixed $k > 0$. In the next subsection, we will show that the derived upper bound is indeed an upper bound on $\overline{\gamma}(s,t,\varepsilon)$ by showing it is an upper bound on $\overline{\gamma}^{(k)}(s,t,\varepsilon)$ for all sufficiently large k. To derive this upper bound, we need the concept of *reference sets* of the separation triple (s,t,ε) which will be formally introduced below. Roughly speaking, a reference set cuts every path from the source places of the unfolded graph $H(0,\infty)$ to $t^{(k+\varepsilon)}$, and every event in the reference set has a path leading to $s^{(k)}$. Our key idea is that under any delay assignment, we can derive an upper bound on $\overline{\gamma}^{(k)}(s,t,\varepsilon)$ by considering only the delays on paths from reference events to the targeted events $s^{(k)}$ and $s^{(k+\varepsilon)}$, but not any further history nor the occurrence time of the reference events.

For all $i,j \in \mathbf{N}$ s.t. $i \leq j$, let $\mathcal{D}(i,j)$ be the space generated by all possible delay assignments of places inside the unfolding segment $H(i,j)$. That is, $\mathcal{D}(i,j) = \otimes_{p \in H(i,j)} X(\ell(p))$. (Recall that function $\ell(.)$ maps an instanced place to its corresponding place in the origin marked graph.) Then, $\mathcal{D} = \lim_{j\to\infty} \mathcal{D}(0,j)$ is the space generated by delay assignments of places of all possible timed executions[2]. In order to stress a particular delay assignment D under which random quantities are referred, we subscript them with D. For instance, $X_D(\ell(p))$ means the delay value of place $p \in H(0,\infty)$ under delay assignment D, and ${\gamma_D}^{(j)}(s,t,\varepsilon)$ denotes the value of random variable $\gamma^{(j)}(s,t,\varepsilon)$ under D.

Let $\rho \in H(0,\infty)$ be a path in the unfolded graph. We denote by a random variable $\delta(\rho)$ the sum of random delay variables associated with all the places along path ρ, namely, $\delta(\rho) = \sum_{p \in \rho} X(\ell(p))$. Let u,v be any two nodes in $H(0,\infty)$. Denote by $\mathcal{P}(u,v)$ the set of all paths from u to v, namely, $\mathcal{P}(u,v) = \{\rho \in H(0,\infty) \mid x \overset{\ell}{\leadsto} y\}$.

From timing constraint 1, we verify that if there is a path from events u to v, v must occur after u by at least the maximum sum of delays on all the paths from u to v. Formally, $\forall u,v \in H(0,\infty)$, s.t., $\mathcal{P}(u,v) \neq \emptyset$, we have,

$$\tau(u) + \max_{\rho \in \mathcal{P}(u,v)} \delta(\rho) \leq \tau(v). \tag{3}$$

Under a given delay assignment D, we say that u is *critical* for v if (3) holds in equality. In that case, there is a path $\rho^* \in \mathcal{P}(u,v)$ so that $\delta_D(\rho^*) = \tau_D(v) - \tau_D(u)$ and we call ρ^* a critical path from u to v. It is understood that any subpath of a critical path is also critical under the same delay assignment. Let S

[2] Roughly speaking, a timed execution of a timed marked graph is a tuple (N,f) where N is an unfolding of the graph and f indicates the occurrence time of each event in N as well as the delay assigned to the places in N.

be the source nodes of $H(0,\infty)$, then at least one node in S is critical for every transition $t \in H(0,\infty)$. The criticality of timing along the paths can be described using the concept of *cut sets* defined below.

Definition 2 (*Cut set*) *A cut set for a transition $t \in H(0,\infty)$ is a set of transitions $C(t)$ in $H(0,\infty)$ such that any path leading from a source place of $H(0,\infty)$ touches at least one transition in $C(t)$. Moreover, every transition in $C(t)$ has a path leading to t. Formally, $\forall p \in M_0, \forall \rho \in \mathcal{P}(p^{(0)},t)$, $\rho \cap C(t) \neq \emptyset$ and $\forall u \in C(t)$, $u \rightsquigarrow t$. The cut set is minimal if it does not contain any other cut set of t.*

Example Considering the unfolded graph in Figure 3(b), $\{{t_{1\downarrow}}^{(0)}, {t_{2\downarrow}}^{(0)}\}$ is a cut set for ${t_{1\uparrow}}^{(1)}$, but not for ${t_{1\downarrow}}^{(1)}$ because it does not cut the path from p_{12} to ${t_{1\downarrow}}^{(1)}$ through transition ${t_{3\uparrow}}^{(0)}, {t_{4\uparrow}}^{(0)}, {t_{3\downarrow}}^{(0)}$, and ${t_{2\uparrow}}^{(1)}$, nor is it a cut set for transition ${t_{0\uparrow}}^{(0)}$ because ${t_{2\downarrow}}^{(0)} \not\rightsquigarrow {t_{0\uparrow}}^{(1)}$. Note that the set is minimal for ${t_{1\uparrow}}^{(1)}$. ■

Lemma 3 *Let $C(t)$ be a cut set for transition $t \in H(0,\infty)$. Then, for any delay assignment $D \in \mathcal{D}$,*

$$\tau_D(t) = \max_{u \in C(t)} [\tau_D(u) + \max_{\rho \in \mathcal{P}(u,t)} \delta_D(\rho)] \tag{4}$$

The above lemma shows that the occurrence time of a transition in the unfolded graph can be completely determined by the occurrence time of transitions in any of its cut sets plus the delays assigned on the places on paths leading from cut set transitions to it. In other words, any other knowledge regarding the timing of unfolded graph further in the history is then irrelevant.

We now define the *reference sets* of a pair event separation pair which is the key concept in the remaining sections.

Definition 3 *Let $(s^{(j)}, t^{(j+\varepsilon)})$ be any valid transition pair. A subset of transitions $\mathcal{R}$ in $H(0,\infty)$ is a reference set for the pair if*

1. *R is a cut set of $t^{(j+\varepsilon)}$,*
2. *$\forall u \in \mathcal{R}, u \rightsquigarrow s^{(j)}$.*

A minimal reference set is one that does not contain any other reference set.

Intuitively, $\mathcal{R}$ is a set of transitions that cut every path from a source place of the unfolded graph to $t^{(j+\varepsilon)}$. Moreover, every transition in $\mathcal{R}$ has paths leads to $s^{(j)}$. Note that if R is a minimal reference set, it is necessary to be a minimal cut set of $t^{(j+\varepsilon)}$.

Example Consider the unfolded graph in Figure 3(b). We can check $\{{t_{1\downarrow}}^{(0)}, {t_{2\downarrow}}^{(1)}, {t_{3\downarrow}}^{(0)}\}$ to be a (minimal) reference set for pair $({t_{1\uparrow}}^{(1)}, {t_{1\downarrow}}^{(1)})$. It is not a reference set for pair $({t_{0\uparrow}}^{(1)}, {t_{0\downarrow}}^{(1)})$ because it is not a cut set for either ${t_{0\uparrow}}^{(1)}$ or ${t_{0\downarrow}}^{(0)}$. In fact, for this limited unfolding, there is no subset of transitions in $G^{(0)}$ that is a reference set for $({t_{0\uparrow}}^{(1)}, {t_{0\downarrow}}^{(1)})$. ■

For any two transitions $u, v \in H(0,\infty)$, we define a random variable $\delta^*(u,v)$ which measures the maximum delay on any path from u to v. That is,

$$\delta^*(u,v) = \max_{\rho \in \mathcal{P}(u,v)} \delta(\rho). \tag{5}$$

Example Let us consider transition ${t_{0\uparrow}}^{(0)}$ and ${t_{0\uparrow}}^{(1)}$. There are two paths connecting the two events, namely,

$$\begin{aligned}\mathcal{P}({t_{0\uparrow}}^{(0)}, {t_{0\uparrow}}^{(1)}) &= \{({t_{0\uparrow}}^{(0)}, {p_1}^{(0)}, {t_{1\uparrow}}^{(0)}, {p_5}^{(0)}, {t_{0\downarrow}}^{(0)}, {p_{13}}^{(0)}, {t_{1\downarrow}}^{(0)}, {p_9}^{(0)}, {t_{0\uparrow}}^{(1)}),\\ &\quad\; ({t_{0\uparrow}}^{(0)}, {p_1}^{(0)}, {t_{1\uparrow}}^{(0)}, {p_2}^{(0)}, {t_{2\downarrow}}^{(0)}, {p_6}^{(0)}, {t_{1\downarrow}}^{(0)}, {p_9}^{(0)}, {t_{0\uparrow}}^{(1)})\}\end{aligned}$$

Thus, $\delta^*({t_{0\uparrow}}^{(0)}, {t_{0\uparrow}}^{(1)}) = \max\{X(p_1) + X(p_5) + X(p_{13}) + X(p_9), X(p_1) + X(p_2) + X(p_6) + X(p_9)\}$. We will repeatedly use such random variables defined by δ^*. ■

The following lemma gives an upper bound of the average TSE of (s, t, ε) at its k-th occurrence.

Lemma 4 *Let $\mathcal{R}$ be a reference set of transition pair $(s^{(j)}, t^{(j+\varepsilon)})$. Then,*

$$\mathbf{E}\gamma^{(j)}(s, t, \varepsilon) \leq \mathbf{E}\max_{u \in \mathcal{R}} [\delta^*(u, t^{(j+\varepsilon)}) - \delta^*(u, s^{(j)})] \tag{6}$$

What is important to note is that the bound given by Lemma 4 is independent of the occurrence time of the events in the reference set. It only relies on the structure of the paths leading from events in the reference set to the targeted event pair and the delay distributions on the places of these paths. This implies that if there is a reference set for every sufficiently large j such that the structure between the reference set and the corresponding targeted events is identical (under the shift operator Δ) for every j, we may effectively upper bound the average TSE of (s, t, ε) by analyzing a finite segment of the infinite unfolding of the marked graph.

The proof of the lemma heavily uses the fact that under any delay assignment, there is at least one transition in the reference set that is *critical* for the event $t^{(j+\varepsilon)}$ and the corresponding timing equality given by Lemma 3, i.e., $\tau(t^{(j+\varepsilon)}) = \tau(u) + \delta^*(u, t^{(j+\varepsilon)})$ if u is critical for $t^{(j+\varepsilon)}$. However, since $\tau(s^{(j)})$ is at least $\tau(u) + \delta^*(u, s^{(j)})$ due to the fact that u has a path to $s^{(j)}$, we know $\gamma^{(j)}(s, t, \varepsilon)$ is at most $\delta^*(u, t^{(j+\varepsilon)}) - \delta^*(u, s^{(j)})$, which eliminates the occurrence time of u. Thus, under any circumstance, $\gamma^{(j)}(s, t, \varepsilon)$ is bounded from above by $\max_{u \in \mathcal{R}} [\delta^*(u, t^{(j+\varepsilon)}) - \delta^*(u, s^{(j)})]$. The desired result follows immediately if we take the expectation of the above bound. For convenience, we write

$$U^{(j)}(\mathcal{R}, s, t, \varepsilon) = \mathbf{E}\max_{u \in \mathcal{R}} [\delta^*(u, t^{(j+\varepsilon)}) - \delta^*(u, s^{(j)})] \tag{7}$$

in order to emphasize the fact that the resulting bound is derived from the reference set $\mathcal{R}$.

3.4 The bounds

In the previous subsection, we derived an upper bound for the expected j-th time separation of the targeted event pair, i.e., $\overline{\gamma}^{(j)}(s, t, \varepsilon)$, based on one of its reference sets (if there is one). In this subsection, we derive an upper bound on the average time separation of the event pair over all j, i.e., $\overline{\gamma}(s, t, \varepsilon)$. Towards this end, we show the existence of a reference set $R(s^{(j)}, t^{(j+\varepsilon)})$ for pair $(s^{(j)}, t^{(j+\varepsilon)})$ for every j no less than some constant $\pi(s, t, \varepsilon)$. This ensures we can always find an upper bound for $\overline{\gamma}^{(j)}(s, t, \varepsilon)$ for sufficiently large j using the result in the previous subsection. Finally, we show that such upper bounds for all sufficiently large j are the same by showing the shiftability of the corresponding reference sets under the shift operator Δ. Consequently, we can obtain an upper bound by analyzing a finite segment for only a fixed j. The remainder of this section formalizes this result.

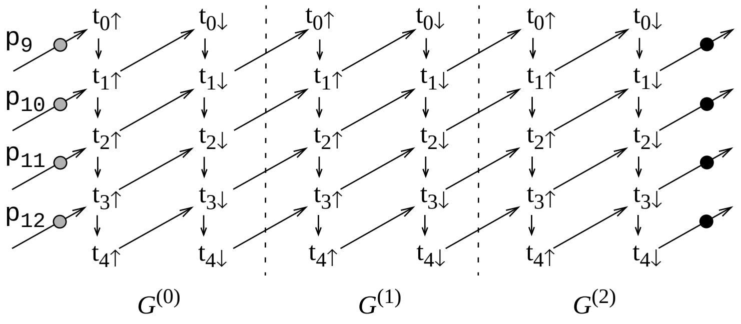


Figure 4: The unfolding segment $H(0,3)$ of the marked graph in Figure 1(b).

3.4.1 Existence of R sets and their shiftability

That constant $\pi(s,t,\varepsilon)$ we mentioned above is related to the token distribution under the initial marking M_0. We need to define two quantities related to the token distribution before we state the result of this subsection.

Let u, v be two arbitrary transitions of a LSMG G which has an initial marking M_0. We define $\phi(u,v)$ to be the minimum number of tokens under the initial marking in any path ρ of G leading u to v. More precisely,

$$\phi(u,v) = \min_{\rho \in \mathcal{P}} \mu_{M_0}(\rho)$$

where $\mu_M(\rho)$ denotes the number of tokens on the path ρ under the marking M. For instance, $\phi(t_{3\uparrow}, t_{1\uparrow}) = 1$ and $\phi(t_{3\downarrow}, t_{1\uparrow}) = 2$ in our example marked graph (Figure 1(b)).

In other words, $\phi(u,v)$ measures the minimum number of tokens one must encounter in order to traverse from transition u to v under M_0. Among all transitions of G, there must be a transition u^* that maximizes this measure. That is,

$$\phi(u^*, v) = \max_{u \in T} \phi(u,v).$$

We use $\Phi(v)$ to denote this number. That is, $\Phi(v) = \phi(u^*, v)$. The importance of $\Phi(v)$ lies in the following fact. It can be shown that for any transition u of G, there must be path from $u^{(j)}$ to $v^{(j+\Phi(v))}$ for every $j \geq 0$. For instance, in Figure 1(b), we may check that $\Phi(t_{1\uparrow}) = 2$. As a result, any transition in the unfolded graph $G^{(0)}$ must have a path leading to $\Phi({t_{1\uparrow}}^{(2)})$.

With the foregoing definitions, we state the following theorem which shows that for sufficiently large k, there always exists a reference set for event separation triple (s,t,ε).

Theorem 2 *Let s,t be arbitrary transitions of G. For any $j \geq \pi(s,t,\varepsilon) = \max(\Phi(s), \Phi(t) - \varepsilon, 1 - \varepsilon)$, there is a subset of transitions A in $G^{(0)}$ which is a reference set for pair $(s^{(j)}, t^{(j+\varepsilon)})$. Moreover, $\Delta_k A$ is a reference set for the pair $(s^{(j+k)}, t^{(j+\varepsilon+k)})$ for every $k > 0$.*

Intuitively, since $j + \varepsilon \geq \Phi(t)$ and $j + \varepsilon \geq 1$ as assumed by the theorem, there must be a subset of transitions, say A, of $G^{(0)}$ which is a cut set for $t^{(j+\varepsilon)}$. Moreover, since $j \geq \Phi(s)$, every transition in set A has a path leading to $s^{(j)}$. This verifies A is a reference set for the pair $(s^{(j)}, t^{(j+\varepsilon)})$. More importantly, it is shown that this reference set is structurally shiftable. We emphasize that the quantity $\pi(s,t,\varepsilon)$ defined in this theorem identifies a sufficient amount of net unfolding that guarantees the existence of a shiftable reference set for the separation triple. Thus, in some cases, a reference set can be found with less unfolding.

Example In the marked graph of our micropipeline example (Figure 1(b)), let us consider transition pair $({t_{1\uparrow}}^{(j)}, {t_{1\downarrow}}^{(j)})$ where we have set $\varepsilon = 0$. We verify that $\Phi(t_{1\uparrow}) = \Phi(t_{1\downarrow}) = 2$. Therefore, for every $j \geq \max\{\Phi(t_{1\uparrow}) = 2, \Phi(t_{1\downarrow}) - \varepsilon = 2 - 0, 1 - 0\} = 2$, the pair has a reference set in $G^{(0)}$. For instance, if we unfold the marked graph one more time from Figure 3 (b) as shown in Figure 4. We see $\{{t_{1\downarrow}}^{(0)}, {t_{2\downarrow}}^{(0)}, {t_{3\downarrow}}^{(0)}\}$ is a reference set for $({t_{1\uparrow}}^{(2)}, {t_{1\downarrow}}^{(2)})$. Besides, $\{{t_{1\downarrow}}^{(m)}, {t_{2\downarrow}}^{(m)}, {t_{3\downarrow}}^{(m)}\}$ must be a reference set for $({t_{1\uparrow}}^{(2+m)}, {t_{1\downarrow}}^{(2+m)})$ at every $m > 0$. ■

3.4.2 Bounds on $\overline{\gamma}$

The following lemma shows the invariance property of the probability distribution of the upper bound of the TSE given in (4) under the operator Δ.

Lemma 5 *Let $R \subseteq G^{(0)}$ be a reference set for pair $(s^{(j)}, t^{(j+\varepsilon)})$ where $j \geq \pi(s,t,\varepsilon)$. For any $m > 0$, we have*

$$U^{(j+m)}(\Delta_m R, s, t, \varepsilon) = U^{(j)}(R, s, t, \varepsilon). \tag{8}$$

This lemma ensures that the upper bound on the average TSE $\gamma^{(k)}(s,t,\varepsilon)$ given in (7) is valid for every sufficiently large k. In other words, this bound is shiftable. The result follows from the shiftability of the reference set when it is sufficiently far from the targeted events plus the fact that the unfolded graph has repeated segments (both structurally and i.i.d delay distributions on repeated places).

We are now ready to state our main theorem.

Theorem 3 *Let $R \subseteq G^{(0)}$ be a reference set for the pair $(s^{(\pi(s,t,\varepsilon))}, t^{(\pi(s,t,\varepsilon)+\varepsilon)})$. Then,*

$$\overline{\gamma}(s,t,\varepsilon) \leq \mathbf{E} \max_{u \in R} [\delta^*(u, t^{(\pi(s,t,\varepsilon)+\varepsilon)}) - \delta^*(u, s^{(\pi(s,t,\varepsilon))})]. \tag{9}$$

The result follows Lemma 5 and weak ergodicity theorem in Section 2.3 which states that the average TSE equals the average of the expected TSE at all the occurrences of the separation triple.

The bounds given in (9) involves taking the expectation of maximum of a set of random variables. Note that the expectation sign does not pass through the maximum sign in general. That is maximization is not linear under expectation, which would otherwise make the evaluation of our bounds much easier. In particular, there is no easy analytical solution to the expectation of such complicated random variables. Nevertheless, this difficulty can be overcome by the techniques discussed in the next section.

4 Evaluating the bounds

We can image several different approaches to evaluate the upper bound derived in the previous section (given by (9)). For some special graph structures and delay distributions, we may come up with

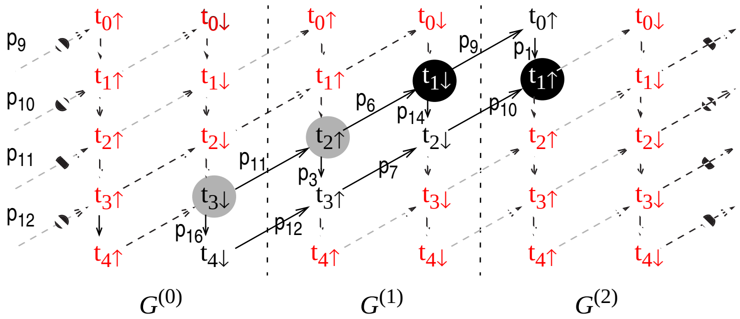


Figure 5: Bounding average TSE $\overline{\gamma}(t_{1\uparrow}, t1_{\downarrow}, 0)$ in a micropipeline.

exact (analytical) result for the bounds. More generally, our bounds can be efficiently evaluated using standard statistical techniques such as Monte-Carlo sampling to get arbitrarily high confidence level and small error interval. Below, we briefly discuss the two approaches.

4.1 Analytical solution

First, let us demonstrate a possible analytical solution to the bounds using our micropipeline example.

Consider bounding $\overline{\gamma}(t_{1\downarrow}, t_{1\uparrow}, 1)^3$. For convenience, we redraw the unfolded graph of the corresponding marked graph in Figure 4.1. Similar to the examples in previous sections, we see that $\{{t_{1\downarrow}}^{(0)}, {t_{2\downarrow}}^{(0)}, {t_{3\downarrow}}^{(0)}\}$ is a reference set for $({t_{1\downarrow}}^{(1)}, {t_{1\uparrow}}^{(2)})$. There are in fact many reference sets for the event pair. For instance, $R = \{{t_{3\downarrow}}^{(0)}, {t_{2\uparrow}}^{(1)}, {t_{1\downarrow}}^{(1)}\}$ is also a reference set for $({t_{1\downarrow}}^{(1)}, {t_{1\uparrow}}^{(2)})$ (cf., the solid part of Figure 4.1) . In addition, we check that $\Delta_k R = \{{t_{3\downarrow}}^{(k)}, {t_{2\uparrow}}^{(k+1)}, {t_{1\downarrow}}^{(k+1)}\}$ is a reference set for $({t_{1\downarrow}}^{(k+1)}, {t_{1\uparrow}}^{(k+2)})$ at every $k \geq 0$, which means the reference set R is shiftable. Therefore, an upper bound on $\overline{\gamma}(t_{1\downarrow}, t_{1\uparrow}, 1)$ is $\mathbf{E}\max_{v\in R}[\delta^*(v, {t_{1\uparrow}}^{(2)}) - \delta^*(v, {t_{1\downarrow}}^{(1)})]$.

Suppose each control circuit element (i.e., the C-element, inverter and the buffer) takes the same constant delay c. In the figure, we have $X(p_{11}) = X(p_{12}) = X(p_6) = X(p_7) = X(p_{10}) = c$. Besides, $X(p_{16}) = c + d_4, X(p_3) = c + d_3, X(p_{14}) = c + d_2, X(p_1) = c + d_1$ and $X(p_9) = c + d_0$, where d_0 through d_4 are the (random) data processing delays of the environments and the stages as shown in Figure 1. By inspection on the relevant paths in the figure, we see that $\max_{v\in R}[\delta^*(v, {t_{1\uparrow}}^{(2)}) - \delta^*(v, {t_{1\downarrow}}^{(1)})] = c + \max\{d_0 + d_1, d_2, d_3, d_4\}$ and hence an upper bound on $\overline{\gamma}(t_{1\downarrow}, t_{1\uparrow}, 1)$ is $c + \mathbf{E}\max\{d_0 + d_1, d_2, d_3, d_4\}$.

Let us assume the data process delays at each stage and the left-side environment are independent and uniformally distribution in an interval (d, D) where $(d \leq D)$. That is, $d_i \sim \text{uniform}(d, D)$. Suppose further the right-side environment always has data ready (i.e., $d_0 = 0$) so that the pipeline is fed as fast as possible. With these assumptions, we know from the *order statistics* (e.g., [29]) that $\frac{1}{D-d}[\max\{d_0 + d_1, d_2, d_3, d_4\} - d]$ has a *beta*(4,1) distribution. According to the special property of the *beta* family [29], we obtain the following result: $\mathbf{E}\frac{1}{D-d}[\max\{d_0 + d_1, d_2, d_3, d_4\} - d] = \frac{4}{5}$. Consequently, $\overline{\gamma}(t_{1\downarrow}, t_{1\uparrow}, 1)$ is upper bounded $c + \frac{1}{5}d + \frac{4}{5}D$.

This result can be easily generalized to an n-stage micropipeline $(n \geq 0)$ with similar delay assump-

[3] It can be shown that the average throughput of the micropipeline is simply the reciprocal of $\overline{\gamma}(t_{1\downarrow}, t_{1\uparrow}, 1)$.

tions where we argue that $\overline{\gamma}(t_{1\downarrow}, t_{1\uparrow}, 1) \leq c + \frac{1}{n+2}d + \frac{n+1}{n+2}D$. Note that as n increases, the upper bound gets close to $c + D$ from below. From a theoretical point of view, it might also be interesting to notice that as $n \rightarrow \infty$, the above pipeline keeps a *positive* average throughput which is at least $\frac{1}{c+D}$.

With a similar reasoning, we arrive at a lower bound on $\overline{\gamma}(t_{1\downarrow}, t_{1\uparrow}, 1)$ equal to $c + \max\{\mathbf{E}(d_0 + d_1), \mathbf{E}d_2, \cdots, \mathbf{E}d_{n+1}\}$ for an n-stage micropipeline where d_1 through d_n are the (random) data processing delays at the stages, d_0 and d_{n+1}, the data processing delay at the right- and left-side environments. If we make similar delay assumptions as before (i.e., assuming identical and uniformally distributed data processing delays), we have $\overline{\gamma}(t_{1\downarrow}, t_{1\uparrow}, 1) \geq c + \frac{1}{2}(d + D)$.

4.2 Statistical simulation

The above analytical approach to evaluating the derived bounds can usually be applied only to restricted types of delay distributions. Moreover, for large systems, the procedure to apply such an analytical approach tends to be tedious and hard to automate. On the other hand, standard statistical techniques can be used to estimate our bounds with a sufficiently high quality in the probabilistic sense. In our experiment, we used the well-known Monte-Carlo simulation approach (e.g., [30]) which we briefly outline below. More advanced statistical techniques such as *stratified and importance sampling* [31] may be applied as well.

Recall that at the beginning of Section 3.3, we discussed the space $D(i, j)$ generated by all possible assignments of delays on places in the unfolding segment $H(i, j)$, $0 \leq i \leq j$. Our upper bound (9) $U(R, s, t, \varepsilon)$ on $\overline{\gamma}(s, t, \varepsilon)$ is the expectation of a random variable which in turn is a function of the random delay assignment drawn from $D(0, k)$ where $k = \max\{\pi(s, t, \varepsilon), \pi(s, t, \varepsilon) + \varepsilon\}$ (cf. Theorem 3). That is, $U(R, s, t, \varepsilon) = \mathbf{E}W$ where random variable $W = \max_{u \in R}[\delta^*(u, t^{(\pi(s,t,\varepsilon)+\varepsilon)}) - \delta^*(u, s^{(\pi(s,t,\varepsilon))})]$. Let F_W denote the distribution function of W.

The intuition behind Monte-Carlo simulation approach is the Central Limit Theorem [32] which deals with the partial sum of *i.i.d* random variables. In our case, if $W_1, W_2, \cdots$ are independent random variables all having distribution function F_W, then for 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 σ_W is the variance of W. Since $\frac{1}{n}\sigma_W^2$ decrease to zero as n increases, any *realization* of the random variable $\frac{1}{n}S_n$ is a good (unbiased) estimate of $\mathbf{E}W$. In fact, the following result is well-known (e.g.,[30]). For any given relative error interval β and a confidence level $1 - \alpha$, we have $P\{|\frac{1}{n}S_n - \mathbf{E}W/\mathbf{E}W| < \beta\} > \alpha$ as long as

$$n > (\frac{z_{\alpha/2}\sigma}{\beta \mathbf{E}W})^2, \tag{10}$$

where $z_{\alpha/2}$ is defined such that the tail probability to its right under normal distribution is $\alpha/2$. The process of this realization of $\frac{1}{n}S_n$ is commonly know as Monte-Carlo simulation approach. The process involves a random realization of n i.i.d random variables (W_1 through W_n) also called as a *random sample* of W with size n. Equation (10) is referred to as stopping criterion.

In practice, the true variance of W, i.e., σ_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 (10) is replaced by the mean of the

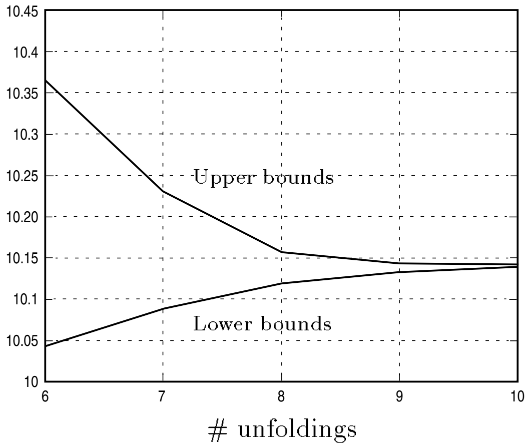


Figure 6: The convergence of lower and upper bounds on $\overline{\gamma}(t_{1\downarrow}, t_{1\uparrow}, 1)$ in an 8-stage micropipeline as the unfolding amount increase (Monte-Carlo sampling with a relative error interval of 1% and a confidence level of 99%).

sample, i.e., $\frac{1}{n}S_n$ itself. As a consequence, $z_\alpha/2$ in (10) has to be replaced by $t_\alpha/2$ which is defined such that the tail probability to its right under the student t-distribution [30] is $\alpha/2$.

The relative error interval and the confidence level can be chosen to trade the quality of the estimate with the run time. In practice, a relative error interval of 1% and confidence level of 99% are sufficient.

5 Improving the bounds

With the unfolding amount suggested by Theorem 3, the resulting bounds are typically much sharper than using other known method (e.g., [13, 16]). Equipped with statistical evaluation technique, we can make the bounds even sharper by simply increasing the amount of unfolding. In fact, the lower and upper bounds can converge very quickly as the unfolding amount slightly exceeds the number suggested by Theorem 3. Figure 5 shows the convergence behavior of the lower and upper bounds in one of our micropipeline experiments (to be discussed in the next section) for which the suggested unfolding amount by Theorem 3 is 6.

Note also that the time complexity typically increases linearly in the amount of unfolding. This is mainly because the required sample size is roughly independent of the problem size, as will be shown in the next section. This latter fact is also observed in many other applications using statistical evaluation, e.g., [33].

6 Experiments

We implemented our bounding technique in C on a Sun UltraSparc10 with 640Mb of main memory. The technique was tested on asynchronous micropipelines [26] and self-timed ring circuits [34]. For a comparison of the resulting bounds, we also implemented the state-of-the-art technique proposed by

# stages	Uniform distributions				Beta distributions			
	Campos et al's		Ours		Campos et al's		Ours	
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
2	14	32	17.28	17.29	8	20	9.798	9.800
4	14	60	19.29	19.30	8	36	11.11	11.14
8	14	116	20.19	20.24	8	68	12.01	12.01
16	14	228	20.69	20.78	8	132	12.44	12.48
32	14	452	21.05	21.09	8	260	12.68	12.72

Table 1: Computed bounds on the average cycle time of transitions in micropipelines. Note that all transitions in a timed marked graph have same average cycle time $\overline{\gamma}$. The relative error interval is set to 0.5% and the confidence level is set to 99.5% during Monte Carlo sampling.

Campos et al [13, 14] using Lindo [35] for the required linear programming. This section discusses the experimental results. Since Campos et al's technique cannot be easily bound quantities other than average cycle time, only average cycle time bounds are reported.

Our first set of experiments is on bounding the average cycle time of a series of asynchronous micropipelines whose circuit structure and timed marked graph model have appeared in the earlier sections (cf., Figure 1). The experiments are made by varying the number of stages in the pipeline as well as the distributions of component delays. Apparently, one can thus design many such experiments. In particular, we vary the number of stages from 2 to 32 and choose `uniform` and `beta` families [29] for the distribution of data processing delay of pipeline stages. Among the many possible reference sets, we chose the one that corresponds to a column of transitions in the first unfolding (i.e., $G^{(0)}$) excluding the top and bottom transitions.

The left-half of Table 1 lists the bounds on the average cycle time of the micropipelines where the data processing delay in each stage is uniform distributed between 1 and 11 (units). The other half of the Table lists the results of the same micropipelines but the stages experiencing a `beta`-shape processing delay also ranging from 1 and 11. The scaled `beta` distribution has a parameter pair (1,4) [29]. The two environments are set to their highest speed, i.e., they do not experience data processing delays. All the control components (C-elements, buffer and inverters) are assumed to have a unit delay. In the Monte Carlo sampling, we set a relative error interval of .5% and confidence level of 99.5%.

In all 10 experiments, our technique achieves which is much shaper than Campos et al's. We note that the technique suggested by Ebergen and Berks [16] (which is specialized in tree-like pipeline structures) yields an upper bound of 24 for all the ten experiments.[4] While significantly better than Campos et al's upper bound, they are not as sharp as ours.[5]

Table 2 lists the number of unfoldings performed, the sample size in Monte Carlo sampling and the run time for each experiment. The sample size remains roughly the same as the number of stages increases. The overall run time grows almost quadratically in the number of stages. We also note that

[4] Strictly speaking they upper bound the *amortized worst case* response time.

[5] Note that Ebergen and Berks' technique does not compute lower bounds on the armortized worst case.

		Uniform distributions		Beta distributions	
# stages	# unfoldings	Sample size	Run time (secs)	Sample size	Run time (secs)
2	4	19,359	0.58	30,277	1.05
4	6	9,217	1.01	21,610	2.63
8	10	5,626	2.96	17,425	10.35
16	18	4,490	15.50	15,877	61.70
32	36	4,260	182.1	14,885	499.9

Table 2: Run time statistics of our bounding technique in the micropipeline experiments.

due to the strong cyclic structure of the micropipelines, state-of-art Markovian approach (e.g., [19]) suffers from the state explosion problem and cannot handle the micropipeline models of more than 8 stages within a reasonable amount of time.

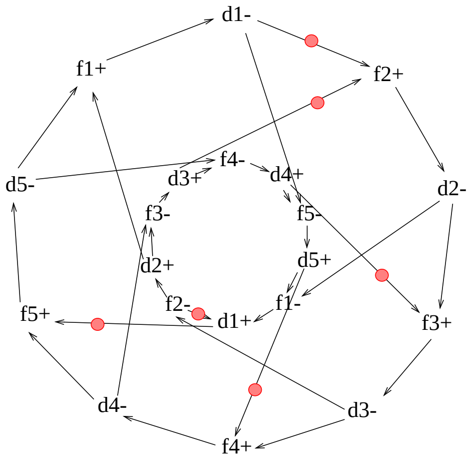


Figure 7: The marked graph that models a 5-stage self-timed ring circuit.

Our second set of experiments are a series of self-timed ring structures due to Williams [34]. Figure 7 shows the timed marked graph model of a 5-stage ring with the **PS0** configuration where each stage contains a precharging functional block. (Places are omitted in the figure for a brevity.) With the initial marking shown in Figure 7, there is one data-token flowing around the ring. Due to space limitations, we refer the readers to [34] for a detailed description of the model.

Similar to the micropipeline experiments above, we vary the number of stages from 3 to 32. (At least 3 stages are required for a working ring structure [34].) For the stage *evaluation* delays (including the *completion* detection), we consider `uniform` and `beta` distributions similar to the data processing delays in the micropipeline experiments, but ranging from 5 to 10. The *precharge* delay of each stage is assumed to be the same. We choose the poset of the initial marking to be the reference set. (In fact, the poset of the initial marking is always guaranteed to be a reference set for a sufficiently unfolded net.) Table 3 lists the bounds on the average cycle time of the ring (i.e., the average time for a data-token

# stages	Uniform distributions				Beta distributions			
	Campos et al's		Ours		Campos et al's		Ours	
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
2	26.5	28.5	26.50	26.50	22	24	22.00	22.11
4	30	38	30.00	30.04	24	32	23.96	24.00
8	60	76	59.98	60.00	48	64	47.95	48.00
16	120	152	119.93	120.08	96	128	95.89	96.14
32	240	304	239.91	240.07	192	256	191.88	192.10

Table 3: Computed bounds on the average cycle time of transitions in self-timed rings. The relative error interval is set to 0.5% and the confidence level is set to 99.5% during Monte Carlo sampling.

to circle through the ring once). Again, the bounds obtained by our technique are much sharper than those using Campos et al's.

7 Conclusions

In summary, this paper presented a technique that bounds the average time separation of events in stochastic timed marked graphs that can be evaluated using in some special cases analytical methods and more generally using Monte Carlo simulation. The time complexity appears to be low thereby allowing us to handle much larger systems than previously possible via Markovian approaches and the bounds obtained are often much sharper than achievable using previously known techniques.

The next step in this research is to extend this approach to system models that can handle choice, e.g., free-choice Petri nets. The challenge here is that the unfolding of such nets must involve resolving choices (possibly according to given probability distributions) and may not have an easily identifiable repeated structure.

In addition, we hope to explore the integration of these techniques within synthesis routines, thereby facilitating performance-driven architectural-level and gate-level synthesis.

References

[1] J. L. Peterson. *Petri Nets Thoery and the Modeling of Systems.* Prentice-Hall, 1981.

[2] G. Cohen, D. Dubois, J.-P. Quadrat, and M. Viot. A linear-system-theoretic view of discrete-event processes and its use for performance evaluation in manufacturing. *IEEE Transactions on Automatic Control*, 30(3):210–220, March 1985.

[3] H. P. Hillion and A. H. Levis. Timed event-graphs and performance evaluation of systems. In *8th European Workshop on Applied Theory of Petri Nets*, June 1987.

[4] C. V. Ramamoorthy and G. S. Ho. Performance evaluation of asynchronous concurrent systems using Petri nets. *IEEE Transactions on Software Engineering*, 6(5):440–449, September 1980.

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

[6] 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.

[7] K. McMillan and D. L. Dill. Algorithms for interface timing verification. In *International Conference on Computer Design, ICCD-1992*. IEEE Computer Society Press, 1992.

[8] M. Holiday and M. Vernon. A generalized timed Petri net model for performance analysis. *IEEE Transactions on Software Engineering*, 13:1297–1310, December 1987.

[9] 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.

[10] 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, 1997.

[11] R. M. Karp. A characterization of the minimum cycle mean in a diagraph. *Discrete mathematics*, 23, 1978.

[12] F. Baccelli, G. Cohen, G. J. Olsder, and J.-P. Quadrat. *Synchronization and Linearity: An algebra for discrete event systems*. John Wiley & Sons, 1992.

[13] J. Campos, G. Chiola, J. M. Colom, and M. Silva. Properties and performance bounds for timed marked graphs. *IEEE Transactions on Circuits and Systems–I: Fundamental theory and applications*, 39(5):386–401, May 1992.

[14] 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.

[15] A. Dasdan and R. K. Gupta. Faster maximum and minimum mean cycle algorithms for system performance analysis. Technical report, UC Irvine, February 1997.

[16] 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, 1997.

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

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

[19] A. Xie and P. A. Beerel. Accelerating Markovian analysis of asynchronous systems using string-based state compression. In *Proc. International Symposium on Advanced Research in Asynchronous Circuits and Systems (ASYNC)*, pages 247–260. IEEE Computer Society Press, 1998.

[20] T.-L. Chou and K. Roy. Accurate power estimation of cmos sequential circuits. *IEEE Transactions on VLSI Systems*, 4(3), Sept. 1996.

[21] M. Silva and J. Campos. Structural performance analysis of stochastic Petri nets. In *IEEE International Computer Performance and Dependability Symposium*, pages 61–70, 1995.

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

[23] J. N. Kozhaya and F. N. Najm. Accurate power estimation for large sequential circuits. In *Proc. International Conf. Computer-Aided Design (ICCAD)*, pages 488–493, 1997.

[24] 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.

[25] F. Commoner, A. W. Holt, S. Even, and A. Pnueli. Marked directed graphs. *Journal of Computer and System Sciences*, 5:511–523, 1971.

[26] I. E. Sutherland. Micropipelines. *Communications of the ACM*, 32(6):720–738, 1989.

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

[28] T. Murata. Circuit theorectic analysis and synthesis of marked graphs. *IEEE Transactions on Circuits and Systems*, 24:400–405, July 1977.

[29] G. Casella and R. L. Berger. *Statistical Inference.* Duxbury Press, Wadsworth Publishing Company, California, 1990.

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

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

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

[33] R. Burch, F. N. Najm, P. Yang, and T. Trick. A monte carlo approach for power estimation. *IEEE Transactions on VLSI Systems*, 1(1):63–71, March 1993.

[34] Ted E. Williams. *Self-Timed Rings and their Application to Division.* PhD thesis, Stanford University, June 1991.

[35] L. Schrage. *Lindo: User's Manual.* The Scientific Press, 1991.

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

Appendix: Proofs

We include the appendix the proofs of all Lemmas and Theorems in the paper. Due to space constraints, however, if this paper is accepted, the appendix will be omitted in the final version and placed in a technical report made available via the web.

We will use the well-known results regarding the liveness and safeness of Petri nets (in particular, marked graphs) stated in the following three lemmas.

Lemma 6 [24] *For a live and safe (marked) Petri net, there is no source or sink places and no source or sink transitions. That is,* $\forall x \in P \cup T$, $\bullet x \neq \emptyset$, $x \bullet \neq \emptyset$.

Lemma 7 [25] *Let* (G, M_0) *be a marked graph. It is live iff every circuit is initially marked, i.e.,* $\mu_{M_0}(\xi) \geq 1$, $\forall \xi \in C(G)$. *It is live and safe iff every place belongs to a circuit* $\xi \in C(G)$ *for which* $\mu_{M_0}(\xi) = 1$.

Lemma 8 [12] *For a connected marked graph, a firing sequence results in the initial marking if and only if it fires all transitions a same number of times.*

Our proof of the weak ergodicity of time separation sequence of arbitrary two events, Theorem 1, uses the well-known weak ergodicity theorem by on the event cycle time, that we formally state below.

Theorem 4 *[12, 27] (Weak ergodicity) For a LS stochastic timed MG* (G, M_0) *as defined above, there exists a constant* $0 \leq c < \infty$ *such that (11) holds for every* $t \in T$.

$$\lim_{k\to\infty} \frac{1}{k}\tau(t^{(k)}) = \lim_{k\to\infty} \frac{1}{k}\mathbf{E}\tau(t^{(k)}) = c \quad \textit{a.s. and in mean,} \tag{11}$$

Theorem 1 *Let* s, t *be two arbitrary transitions of a LS stochastic timed MG* (G, M_0) *as defined in Section 2.2. Their average time separation (with occurrence offset* ε*) converges to a constant* $\overline{\gamma}(s,t,\varepsilon)$ *almost surely and in mean, i.e.,*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} \gamma^{(k)}(s,t,\varepsilon) = \overline{\gamma}(s,t,\varepsilon) \quad \textit{a.s., and in mean.}$$

Moreover, $(-h+\varepsilon)\cdot c \leq \overline{\gamma}(s,t,\varepsilon) \leq (h+\varepsilon)\cdot c$ *where* h *is the minimum number of places in all the circuits containing* s, t*, and* c *is the average cycle time of* G.

Proof (Sketch) We will use a special case of Kingman's ergodic theorem [36] plus the result from Theorem 4. The theorem is first proved for $\varepsilon = 0$. Extension to any fixed ε is argued similarly.

Let s, t be as given in the assumption. From the *coupling argument* [12], there is a positive integer $K < \infty$ almost surely such that for any $t, s \in T$, the sequence $\gamma^{(k)}(s,t,0) = \{\tau(t^{(k)}) - \tau(s^{(k)}) : k > K\}$ is stationary.

Next, for all $0 < l < m$, let us define $Z_{l,m} = \sum_{k=l+1}^{m} \gamma^{(k)}(s,t,0)$. Clearly, Z is additive, i.e., for all $0 < l < m < n$, $Z_{l,n} = Z_{l,m} + Z_{m,n}$. We shall further show that for large n, $\mathbf{E}Z_{K,n}$ is bounded from below by $-(c \cdot h)\cdot n$ where c is the average cycle time of the graph.

For any $k > h$, we have

$$\tau(t^{(k)}) \geq \tau(s^{(k-h)}) \tag{12}$$

because t cannot have fired h times more than s at any instant under the safeness assumption. It can then be checked that,

$$Z_{K,n} = \sum_{k=K+1}^{n} \tau(t^{(k)}) - \tau(s^{(k)}) \geq - \sum_{k=n-h+1}^{n} \tau(s^{(k)}). \tag{13}$$

By taking expectation on both sides of (13), the boundedness of $\mathbf{E}Z_{K,n}$ follows Theorem 4 when n is large.

As such, $Z_{K,n}$ verifies the conditions of Kingman's ergodic theorem, from which we deduce that $\lim_n \frac{1}{n} \sum_{k=1}^{n} \gamma^{(k)}(s,t,0) = \lim_n \frac{1}{n} Z_{0,n}$ equals to a constant almost surely. Moreover, this constant is lower bounded by $-c \cdot h$. It is further upper bounded by $c \cdot h$ because (12) is also valid if s and t are interchanged.

For the case where $\varepsilon \neq 0$, the theorem can be checked with similar arguments. In particular, the sequence $\{\gamma^{(n)}(s,t,\varepsilon)\}$ is almost surely stationary for large n because $\{\gamma^{(n)}(r,r,1)\}$ is almost surely stationary for all $r \in T$ and large n. The only difference is that the average time separation is now bounded by $c \cdot (h + \varepsilon)$ from above and $c \cdot (-h + \varepsilon)$ from below. ■

We now prove Lemma 1 through Lemma 4.

Lemma 1 *Procedure 1 terminates for a every LSMG (G, M_0). Moreover, when it returns, $M = M_0$.*

Proof Suppose there is a transition $t \in T$ that cannot be instanced into $G^{(0)}$. Since G is choice free, this mean there is a place in $\bullet t$ that is never instanced into $G^{(0)}$, which implies transition $\bullet p$ is never instanced. In addition, $^\bullet p$ must not be t because otherwise p and t form a circuit, forcing $p \in I$ by Lemma 7. We may repeat the same argument on transition $\bullet p$ and deduce that no transition of T can ever be instanced (enabled), contradicting the assumption that G is live. Finally, since the procedure instances every transition exactly once, the final marking M returns to the initial marking M_0 by Lemma 8. ■

Lemma 2 *For all $j, k, l \in \mathbf{N}$, $H(j + l, k + l) = \Delta_l H(j, k)$.*

Proof Following Lemma 1, for any $i, j \in \mathbf{Z}_+$, $H(i,i) = G^{(i)}$ is checked to be identical to $G^{(j)} = H(j,j)$ under Δ with a proper shift amount. ■

Lemma 3 *Let $C(t)$ be a cut set for transition $t \in H(0, \infty)$. Then, for any delay assignment $D \in \mathcal{D}$,*

$$\tau_D(t) = \max_{u \in C(t)} [\tau_D(u) + \max_{\rho \in \mathcal{P}(u,t)} \delta_D(\rho)] \tag{14}$$

Proof For any $u \in C(x)$, we have $\tau_D(t) \geq \tau_D(u) + \max_{\rho \in \mathcal{P}(u,t)} \delta_D(\rho)$. Therefore,

$$\tau_D(t) \geq \max_{u \in C(t)} \tau_D(u) + \max_{\rho \in \mathcal{P}(u,t)} \delta_D(\rho). \tag{15}$$

Further, under given D, there is a critical path, say $\rho^* \in \mathcal{P}(v,t)$, from some source node $v \in S$ to t. Let $w \in \rho^* \cap C(t) \neq \emptyset$. By sub-criticality argument, $\tau_D(w) + \delta_D(\rho_1^*) = \tau_D(t)$ where $w \overset{\rho_1^*}{\rightsquigarrow} t$ and $\rho_1^* \subseteq \rho^*$. Thus, (15) holds only with equality. ■

Lemma 4 *Let $\mathcal{R}$ be a reference set of transition pair $(s^{(j)}, t^{(j+\varepsilon)})$. Then,*

$$\mathbf{E}\gamma^{(j)}(s,t,\varepsilon) \le \mathbf{E}\max_{u\in\mathcal{R}}[\delta^*(u,t^{(j+\varepsilon)}) - \delta^*(u,s^{(j)})] \tag{16}$$

Proof First, let $\mathcal{R}'$ denote a cut set of $s^{(j)}$ s.t. $\mathcal{R} \subseteq \mathcal{R}'$. From Lemma 3, we have

$$\tau(t^{(j+\varepsilon)}) = \max_{u\in\mathcal{R}}[\tau(u) + \delta^*(u,t^{(j+\varepsilon)})], \text{ and} \tag{17}$$

$$\tau(s^{(j)}) = \max_{u\in\mathcal{R}'}[\tau(u) + \delta^*(u,s^{(j)})] \ge \max_{u\in\mathcal{R}}[\tau(u) + \delta^*(u,s^{(j)})] \tag{18}$$

where we have used the notion of δ^*. Deducting (18) from (17), we get

$$\gamma^{(j)}(s,t,\varepsilon) \le \max_{u\in\mathcal{R}}[\tau(u) + \delta^*(u,t^{(j+\varepsilon)})] - \max_{u\in\mathcal{R}}[\tau(u) + \delta^*(u,s^{(j)})] \tag{19}$$

According to Lemma 3, at least one transition in $\mathcal{R}$ is critical for $t^{(j+\varepsilon)}$ under any delay assignment. Particularly, if $v \in \mathcal{R}$ is critical for $t^{(j+\varepsilon)}$, we have $\tau(v) + \delta^*(v,t^{(j+\varepsilon)}) = \tau(t^{(j+\varepsilon)})$ and hence,

$$\gamma^{(j)}(s,t,\varepsilon) \le \delta^*(v,t^{(j+\varepsilon)}) - \delta^*(v,s^{(k)}).$$

Considering the possibility for each of the events in $\mathcal{R}$ to be critical for $t^{(j+\varepsilon)}$, we conclude,

$$\gamma^{(j)}(s,t,\varepsilon) \le \max_{v\in\mathcal{R}}[\delta^*(v,t^{(j+\varepsilon)}) - \delta^*(v,s^{(k)})]$$

and the desired result of the lemma follows. ■

Before we prove Theorem 2, we must first show following two simple lemmas.

Lemma 9 *Let (G, M_0) be a LSMG and H its unfolding segment. Let $j,k \in \mathbf{N}$ such that $k > j$. There is a path $\rho \in H(j,k)$ such that $t^{(j)} \overset{\rho}{\rightsquigarrow} t^{(k)}$ for every $t \in T$.*

Proof It is sufficient to show that the lemma holds for every k s.t. $k = j+1$ and $j > 0$. However, since G is live and safe, Lemma 7 ensures t belong to some circuit $\xi \in C(G)$ such that $\mu_I(\xi) = 1$. Let p be the only marked place in ξ under initial marking. According to Lemma 1, we must have $p^{(j)} \rightsquigarrow t^{(j)} \rightsquigarrow p^{(j+1)} \rightsquigarrow t^{(j+1)}$ after unfolding G for $(j+1)$-th time. ■

Lemma 10 *For any transitions s,t of a LS marked graph G, there is a path ρ in $H(j,k)$ such that $s^{(j)} \overset{\rho}{\rightsquigarrow} t^{(k)}$ iff $k \ge j + \phi(s,t)$.*

Proof Let $\rho \in G$ such that $s \overset{\rho}{\rightsquigarrow} t$ and $\mu_{M_0}(\rho) = \phi(s,t)$. Let $p_1, p_2, \ldots, p_{\phi(s,t)}$ be the marked places of ρ under initial marking such that $p_m \rightsquigarrow p_{m+1}$ for $1 \le m < \phi(s,t)$. With similar argument in the proof of Lemma 9, we verify that there is a path in $H(j, j+\phi(s,t))$ along which we have $s^{(j)} \rightsquigarrow {p_1}^{(j+1)} \rightsquigarrow \cdots \rightsquigarrow {p_m}^{(j+m)} \rightsquigarrow \cdots \rightsquigarrow {p_{\phi(s,t)}}^{(j+\phi(s,t))} \rightsquigarrow t^{(j+\phi(s,t))}$. However, since $k > j + \phi(s,t)$, we have $t^{(j+\phi(s,t))} \rightsquigarrow t^{(k)}$ according to Lemma 9. Verifying the reverse part of the lemma is trivial. ■

Theorem 2 *Let s, t be arbitrary transitions of G. For any $j \geq \max(\Phi(s), \Phi(s) - \varepsilon, 1 - \varepsilon)$, there is a subset of transitions A in $G^{(0)}$ which is a reference set for pair $(s^{(j)}, t^{(j+\varepsilon)})$. Moreover, $\Delta_k A$ is a reference set for the pair $(s^{(j+k)}, t^{(j+\varepsilon)})$ for every $k > 0$.*

Proof First, since $j + \varepsilon \geq 1$ as assumed by the theorem, any path leading a source node of $H(0, \infty)$ to transition $t^{(j+\varepsilon)}$ must pass through at least one transition in $G^{(0)}$. Moreover, since $j + \varepsilon \geq \Phi(t)$, every transition in $G^{(0)}$ must have path leading to $t^{(j+\varepsilon)}$. Therefore, there is at least one subset of transitions in $G^{(0)}$ which is a cut set for $t^{(j+\varepsilon)}$. Let this cut set be denoted by A. Next, for any transition v in A, $v \rightsquigarrow s^{(j)}$ according to Lemma 10 because $j = \Phi(s) \geq \phi(v, s)$. Therefore, A is reference set for $(s^{(j)}, t^{(j+\varepsilon)})$.

For the remaining of the theorem, suppose first $\Delta_k A$ is not a cut set for $t^{(j+\varepsilon+k)}$. Then, there exists a path ρ and a source node v of $H(0, \infty)$ such that $v \overset{\rho}{\rightsquigarrow} t^{(j+\varepsilon+k)}$. Further, any path must pass through the set of source nodes of $H(k, \infty)$ as implied by Lemma 8. Let ρ visit a source node $u^{(k)}$ of $H(k, \infty)$. Thus, we find a path $\rho' \subseteq \rho$ such that $\rho' \cap A = \emptyset$ but $u^{(k)} \overset{\rho'}{\rightsquigarrow} t^{(j+\varepsilon+k)}$. According to Lemma 2, we have $\Delta_{-k}\rho' \in H(0, \infty)$ such that $\Delta_{-k}\rho' \cap A = \emptyset$ but $u^{(0)} = \Delta_{-k} u^{(k)} \overset{\Delta_{-k}\rho}{\rightsquigarrow} \Delta_{-k} t^{(j+\varepsilon+k)} = t^{(j+\varepsilon)}$. This contradicts the fact that A is a cut set of for $t^{(j+\varepsilon)}$. Finally, Lemma 10 ensures every transition in $\Delta_k A$ leads to $s^{(j+k)}$. ■

Finally, we complete this section by proving Lemma 5 and the final theorem, Theorem 3.

Lemma 5 *Let $R \subseteq G^{(0)}$ be a reference set for pair $(s^{(j)}, t^{(j+\varepsilon)})$ where $j \geq \pi(s, t, \varepsilon)$. For any $m > 0$, we have*

$$U^{(j+m)}(\Delta_m R, s, t, \varepsilon) = U^{(j)}(R, s, t, \varepsilon). \tag{20}$$

Proof With the given assumption, we have $\Delta_m R$ as a valid reference set for $(s^{(j+m)}, t^{(j+m+\varepsilon)})$ according to Theorem 2.

Consider an arbitrary transition $v^{(m)} \in \Delta_m R$. For any path $\rho \in \mathcal{P}(v^{(j+m)}, t^{(j+\varepsilon+m)})$, there is a path $\rho' \in \mathcal{P}(v^{(j)}, t^{(j+\varepsilon)})$ which is identical to ρ under operator Δ according to Lemma 2. That is, $\rho' = \Delta_{-m}\rho$ and, of course, $v^{(j)} \in R$. In addition, the delays experience by tokens flowing into a given place is assumed to independent identically distributed. Therefore, $\delta(\rho) = \sum_{p \in \rho} X(\ell(p)) \overset{L}{\sim} \sum_{p \in \rho'} X(\ell(p)) = \delta(\rho')$. Conversely, if $u^{(0)}$ is any transition in R and ρ' is a path in $\mathcal{P}(u^{(0)}, t^{(j+\varepsilon)})$, there is a path ρ in $\mathcal{P}(u^{(m)}, t^{(j+\varepsilon+m)})$ such that the sums of place delays along ρ' and ρ share the same probability law. By similar arguments, we claim the same invariance property property transitions $s^{(j)}$ and $s^{(j+m)}$. Thus, for any $v^{(m)} \in \Delta_m R$, and hence, $v^{(0)} \in R$, we have $\delta^*(v^{(m)}, t^{(j+\varepsilon+m)}) \overset{L}{\sim} \delta^*(v^{(0)}, t^{(j+\varepsilon)})$. The lemma is thus immediate by combining this observation with definition of U in (7). ■

Theorem 3 *Let $R \subseteq G^{(0)}$ be a reference set for the pair $(s^{(\pi(s,t,\varepsilon))}, t^{(\pi(s,t,\varepsilon)+\varepsilon)})$. Then,*

$$\overline{\gamma}(s, t, \varepsilon) \leq \mathbf{E} \max_{u \in R} [\delta^*(u, t^{(\pi(s,t,\varepsilon)+\varepsilon)}) - \delta^*(u, s^{(\pi(s,t,\varepsilon))})]. \tag{21}$$

Proof By Theorem 1, we know that the sequence of time separations $\{\gamma^{(k)}(s, t, \varepsilon) : k > 0\}$ converges almost surely and in mean to $\overline{\gamma}(s, t, \varepsilon)$. That is, $\overline{\gamma}(s, t, \varepsilon) = \lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} \mathbf{E}\gamma^{(k)}(s, t, \varepsilon)$. Since $\pi(s, t, \varepsilon)$ is finite, we have $\overline{\gamma}(s, t, \varepsilon) = \lim_{n \to \infty} \frac{1}{n} \sum_{k=\pi(s,t,\varepsilon)}^{n} \mathbf{E}\gamma^{(k)}(s, t, \varepsilon)$. By Lemma 4, it is upper bounded as

$\overline{\gamma}(s,t,\varepsilon) \leq \lim_{n\to\infty} \frac{1}{n} \sum_{k=\pi(s,t,\varepsilon)}^{n} \mathbf{E}[\delta^*(u,t^{(k+\varepsilon)}) - \delta^*(u,s^{(k)})]$. Finally, Lemma 5 ensures that for all $k \geq \pi(s,t,\varepsilon)$, $\mathbf{E}\max_{u\in calR}[\delta^*(u,t^{(k+\varepsilon)}) - \delta^*(u,s^{(k)})] = \mathbf{E}\max_{u\in\mathcal{R}}[\delta^*(u,t^{(\pi(s,t,\varepsilon)+\varepsilon)}) - \delta^*(u,s^{(\pi(s,t,\varepsilon))})]$. This concludes the proof. ■