Scheduling Reserved Traffic in Input-Queued Switches: New Delay Bounds via Probabilistic Techniques

Matthew Andrews and Milan Vojnović

Abstract— **We consider the problem of providing delay bounds to reserved traffic in high-speed input-queued switches. We assume that the matrix of bandwidth demands is known and we use the now standard approach of decomposing this matrix into a convex combination of permutation matrices. Our problem therefore reduces to the problem of constructing a schedule for these permutation matrices.**

In this paper we derive delay bounds for four algorithms that are based on probabilistic techniques. For each algorithm we first place tokens randomly in continuous time for each permutation matrix. If the nth token that appears corresponds to permutation matrix M_k then we schedule matrix M_k in the nth time slot. The algorithms differ in how the random token processes are defined. For two of the algorithms we are able to perform a derandomization so as to obtain deterministic schedules.

We show through numerical computation that in many situations the resulting delay bounds are smaller than the previously best-known delay bounds of Chang, Chen, and Huang [1].

Keywords—**Input-queued switches. Decomposition-based schedules. Delay bounds.**

I. Introduction

In recent years there has been a great deal of work on scheduling algorithms for input-queued switches. The key feature of an input-queued switch is that at each time step, each input can be connected to at most one output and each output can be connected to at most one input. The aim of the scheduler is to determine how to configure the switch at each time step so as to provide high throughput and low delays for the arriving packets.

Most of the previous work has concentrated on providing *stability* or *100% throughput* for the switch. A scheduler is said to be stable if the queues remain bounded as long as the load on each port is less than the capacity of that port. Algorithms that provide stability generally fall into two categories depending on whether or not we know the arrival rates for each input-output pair in advance. For the case in which we do know the arrival rates we can decompose the rate matrix into a convex combination of permutation matrices. If the scheduler configures the switch according to this decomposition then we have stability. We refer to these schedulers as *decomposition-based* schedulers (e.g. [1], [2], [3]). Specific algorithms for performing the decomposition can be derived from results of Birkhoff [4] and von Neumann [5]. Note that the schedule can be computed in advance of the arriving traffic and hence it is acceptable for the calculation to have significant complexity.

For the case of high-speed optical switches it is reasonable to assume that the arrival rates are known to us because such switches are likely to be deployed in the core of networks where traffic engineering using MPLS is becoming more prevalent. Switches that support MPLS must be able to provide *bandwidth guarantees* for certain input-output pairs. For each MPLS path that passes through input i and output j on a switch, the switch will be required to *reserve* bandwidth for the path between these two ports. Another justification that the input rates are known can be found with Expedited Forwarding (EF) in the context of differentiated services [6]. There it is commonplace to assume the network is engineered such that the load of EF traffic at each node is bounded by some configured value.

For the case in which we do not know the arrival rate matrix in advance then most of the stable schedulers set up a bipartite graph whose edge weights correspond to the queue sizes of the corresponding input queues. At each time step the scheduler finds an (approximate) maximum-weight matching in this graph and configures the switch accordingly. These schedulers are sometimes known as maximum weight matching (MWM) type schedulers (e.g. [7], [8], [9], [10]). Note that an MWM scheduler must operate in real-time since it needs access to queue information and hence it must have extremely low complexity.

More recently, attention has been paid to the problem of minimizing the *delay* experienced by packets passing through an input-queued switch. Leonardi *et al.* [11] analyzed the MWM algorithm and showed that the mean delay through an arbitrary pair of input-output ports of a switch with I input and I output ports, uniformly loaded to $\rho < 1$, is bounded by $(I-\rho)/(1-\rho)$. A related work is that of Shah and Kopikare [12] who observe that for uniform Bernoulli arrivals to the switch with the scheduling policy that at each time slot takes a matching uniformly at random from the entire set of $I!$ matchings, the expected delay is $(I-1)/(1-\rho)$. Note that this is smaller than the bound obtained in [11].

In [1] Chang, Chen, and Huang showed how to derive worst-case deterministic delay bounds for a Birkhoff-von Neumann schedule in which the permutation matrices are scheduled according to Packetized Generalized Processor Sharing (PGPS) [13] such that at an instant a matrix is served, it is placed as a new arrival into the PGPS system.

In this paper we consider delay bounds for decomposition-based schedules. We show that by using probabilistic techniques we are able to tighten the bounds of [1] for the worst case input-output pairs in many scenarios. We use the term "probabilistic techniques" rather than "randomized algorithms" since for two of our algorithms we are able to derandomize the random processes so as to obtain *deterministic algorithms*.

It is interesting to observe that for a node to support EF it needs to conform to a rigorous definition of the per-hop-

Author Affiliations: Matthew Andrews, Bell Laboratories, Lucent Technologies, Murray Hill, NJ <andrews@research.bell-labs.com>, Milan Vojnović, EPFL, CH-1015 Lausanne, Switzerland <milan.vojnovic@epfl.ch>. Work performed in part while the second author was visiting Bell Labs.

behavior, namely, Packet Scale Rate Guarantee with rate r and latency e; see [14]. Our work can be viewed as calculating what the value of the latency e would be for an input-queued switch. Before we present our algorithms and results in detail we must present some notation and our goals.

Assumptions and Notation

We consider an $I \times I$ switch. Let ρ_{ij} be the bandwidth that needs to be reserved between input i and output j, normalized by the link rate. Let M be the matrix whose ij entry is ρ_{ij}. We refer to M as the *rate matrix*. For the majority of this paper we consider the case in which M is a *doubly stochastic* matrix, i.e. $\sum_i \rho_{ij} = 1$ and $\sum_j \rho_{ij} = 1$. This corresponds to the case in which the entire bandwidth of the switch is reserved. However, we shall sometimes consider the substochastic case in which we only have $\sum_i \rho_{ij} \leq 1$ and $\sum_j \rho_{ij} \leq 1$. In this case the residual bandwidth of the switch could be used by best effort traffic.

By standard results of Birkhoff and von Neumann (see e.g. [1]) we can decompose the matrix M into a convex combination of permutation matrices,

$$M = \sum_{k=1}^{K} \varphi_k M_k,$$

where $K \leq I^2 - 2I + 2$. Here, M_k is a permutation matrix (a $0-1$ matrix with exactly one "1" in each row and column), φ_k is the *rate* of matrix M_k and $\sum_{k=1}^{K} \varphi_k = 1$. Let S_{ij} be the set of matrices in the decomposition that have a 1 in the ij position. Then $\rho_{ij} = \sum_{k \in S_{ij}} \varphi_k$. Our aim is to create a schedule in which exactly one of the permutation matrices is scheduled in each time slot. Input-output pair ij is served whenever a matrix from the set S_{ij} is scheduled. Hence we require that a matrix from S_{ij} is scheduled approximately once every $1/\rho_{ij}$ time slots.

As an example, suppose that,

$$M = \begin{pmatrix} 1/6 & 5/6 & 0 \\ 1/2 & 1/6 & 1/3 \\ 1/3 & 0 & 2/3 \end{pmatrix}.$$

Then,

$$M = \frac{1}{2} M_1 + \frac{1}{3} M_2 + \frac{1}{6} M_3,$$

where,

$$M_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, M_2 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix},$$
$$M_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

and a possible schedule is,

$$M_1, M_2, M_1, M_2, M_1, M_3,$$
$$M_1, M_2, M_1, M_2, M_1, M_3, \ldots$$

The class of the schedulers that we consider can be formulated by the following unifying framework. We first place *tokens* for each matrix M_k in continuous time. We schedule matrix M_k in time slot n if the nth token to appear corresponds to matrix M_k. (See Figure 1).

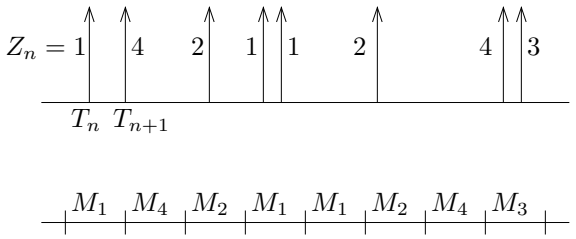

Fig. 1. (Top) The token process. T_n is the time at which the nth token appears. Z_n is the type of the nth token, i.e. if the nth token corresponds to permutation matrix M_k then $Z_n = k$. (Bottom) The corresponding schedule.

More formally, we associate with M_k a *counting process* N_k defined on $\mathbb{R}^+$. For any interval $\mathcal{I} \subseteq \mathbb{R}^+$, $N_k\mathcal{I}$ equals the number of tokens for M_k that land in interval $\mathcal{I}$. We require that N_k has *intensity* φ_k, i.e. $\lim_{t\to\infty} N_k[0,t)/t = \varphi_k$.

We define the superposition process $N\mathcal{I} = \sum_{k=1}^{K} N_k\mathcal{I}$ to which there is an associated point process $(T_n)_{n\geq 0}$ defined on $\mathbb{R}^+$. Next, let $(Z_n)_{n\geq 0}$ be the sequence of marks such that $Z_n = k$ if and only if the n-th point of the superposition point process, T_n, belongs to N_k. Let $N_{ij}\mathcal{I}$ be the number of tokens for input-output pair ij that land in the interval $\mathcal{I}$, i.e. $N_{ij}\mathcal{I} = \sum_{k\in S_{ij}} N_k\mathcal{I}$. Likewise, let $N_{\overline{ij}}\mathcal{I}$ be the number of tokens that land in $\mathcal{I}$, but do not belong to S_{ij}.

The schedule is given by the sequence $(Z_n)_{n\geq 0}$. If for any given n, $Z_n = k$, then the matrix M_k is scheduled in the nth slot. We say the nth token is of type k. Notice that by taking $(Z_n)_{n\geq 0}$ we in fact construct a *non-idle* schedule. A key feature of this schedule is,

Observation 1: The total number of slots in which input-output pair ij can be served during the time slots $n, n+1, \ldots, n+m-1$ is equal to $N_{ij}[T_n, T_{n+m})$.

Service Characterization

We give different characterizations of the service offered to an arbitrary input-output pair ij. Informally speaking, we would like $N_{ij}[T_n, T_{n+m})$ to be close to $\rho_{ij}m$. The following is the simplest, but weakest, service characterization: for any $n \geq 0$ and $m > 0$, and some fixed $E_1^{ij} \geq 0$,

$$\{N_{ij}[T_n, T_{n+m}) \geq \rho_{ij}(m - E_1^{ij})\}. \quad (1)$$

If the above event holds with probability $1-\varepsilon$, $\varepsilon \geq 0$, a probabilistic interpretation of the service offered is: for any fixed m one picks at random a slot n, then, the number of slots given to the input-output pair ij in the next m slots is at least $\rho_{ij}(m - E_1^{ij})$ with probability $1-\varepsilon$.

A natural extension of the above characterization is by requiring that for any $n \geq 0$ and some fixed $E_2^{ij} \geq 0$,

$$\{\forall_{m>0} :\ N_{ij}[T_n, T_{n+m}) \geq \rho_{ij}(m - E_2^{ij})\}. \quad (2)$$

The strongest guarantee is offered by requiring, for some fixed $E_3^{ij} \geq 0$,

$$\{\forall_{n\geq 0}\forall_{m>0} :\ N_{ij}[T_n, T_{n+m}) \geq \rho_{ij}(m - E_3^{ij})\}. \quad (3)$$

If this event holds then we lower bound the service offered to input-output pair ij over *any* interval of time slots. In particular, it can be seen that $f(m) = \rho_{ij} \max[m - E_3^{ij}, 0]$ is a strict minimum service curve offered to the ijth pair (see Proposition 1.3.6, Section 1.3.2, [15]). The service curve is "rate-latency" with rate ρ_{ij} and latency E_3^{ij}.

The service characterizations introduced so far bound how much the service offered is behind the service that would be offered by an idealistic fluid system (which would serve $\rho_{ij}m$ bits in m slots). Thus, these service characterizations bound the *lateness* of the scheduler. Analogous characterizations can be established to bound the *earliness*; one only needs to reverse the inequalities in the above definitions, and replace minus with plus in the rate-latency functions. Small earliness of the schedule is desirable to reduce burstiness at the output of the switch.

What Can We Compute from the Service Characterizations?

Consider an arbitrary input-output port pair ij. Let $A_{ij}[n]$ be number of the bits that arrive in $[0, n]$ at input port i and are destined for output port j. Then, by the result known for variable-capacity nodes (see [15], Sec. 1.3.2, also Sec. 4.3.2), we know that the number of bits in $[0, n]$ observed at the output port j that originate from i, $A^*_{ij}[n]$, satisfies,

$$A^*_{ij}[n] = \min_{1 \le m \le n} [A_{ij}[m] + N_{ij}[T_m, T_n)].$$

In particular, suppose that the arrivals are (σ_{ij}, ρ_{ij})-*bounded*, i.e., $A_{ij}[n] - A_{ij}[m] \le \rho_{ij}(n - m) + \sigma_{ij}$ for all $m \le n$. Then, the following is a classical network calculus result.

Fact 2: The backlog of ij packets waiting for service at the switch is at most $\sigma_{ij} + \rho_{ij} E_3^{ij}$. If FIFO scheduling is used *within* the aggregate of ij packets then the maximum delay for these packets is at most $\sigma_{ij}/\rho_{ij} + E_3^{ij}$.

If $\sigma_{ij} = 0$ (i.e. the arrivals are bounded by a idealized fluid system of rate ρ_{ij}) then the packet delay is bounded by E_3^{ij} (assuming FIFO scheduling within the aggregate). However, in a perfect schedule for ij, a matrix in S_{ij} would appear *exactly once* every $1/\rho_{ij}$ time slots. In this case the packet delay would be $1/\rho_{ij}$. Hence, if $\sigma_{ij} = 0$ we have,

$$\frac{\text{worst case packet delay}}{\text{optimal packet delay}} \le \rho_{ij} E_3^{ij}.$$

For these reasons our objective is to keep E_3^{ij} small.

Algorithms and results

Our algorithms will be divided into two types, *frame-based* and *non-frame-based*. Suppose that for some fixed integers ℓ_k and L, $\varphi_k = \ell_k/L$. (We note that this is always possible if the φ_k are rational.) We can compute a schedule for the interval $[0, L)$ that contains exactly ℓ_k occurrences of the permutation matrix m_k and then simply repeat this schedule for all subsequent intervals of length L. We call such a schedule a *frame-based* schedule of length L. Notice that the frame length L and number of the permutation matrices K are related as $K = L/\bar{\ell}$, where $\bar{\ell}$ is the arithmetic mean of ℓ_k, $k = 1, \ldots, K$. Since $\ell_k \ge 1$ for all k it follows that $L \ge K$, with equality if and only if $\ell_k = 1$ for all k.

If the schedule is not periodic in the above way then we say that it is *non-frame-based*. For a non-frame-based schedule we have to define it explicitly in the entire interval $[0, \infty)$.

In [1], Chang *et al.* propose a non-frame-based algorithm in which the permutation matrices are scheduled according to a PGPS [13] system that is fed with its own departures (which is initialized such that all tokens arrive at time 0). In our setting, this corresponds to placing the nth token for matrix M_k at time n/φ_k. More formally, for each $k = 1, \ldots, K$,

$$N_k[0, t) = \sum_{n \ge 0} 1_{[0,t)}(\frac{n}{\varphi_k}).$$

Chang *et al.* [1] show that for this algorithm,

$$E_3^{ij} \le \min[\frac{K}{\rho_{ij}}, \frac{|S_{ij}|}{\rho_{ij}} + (K - 1)]. \tag{4}$$

The aim of our work is to show that by using probabilistic techniques, it is possible to tighten this bound in many scenarios. Our results are as follows.

1. We begin in Section III-A with an extremely simple frame-based scheduler in which the tokens for the permutation matrices in a frame are randomly permuted. We call this the *Random Permutation* scheduler. We require that (3) holds with probability $1 - \varepsilon$ and we show that, as $L \to \infty$,

$$E_3^{ij} \to \sqrt{A\left(\frac{1}{\rho_{ij}} - 1\right) L}, \tag{5}$$

where A is a constant depending on ε specified in this paper. For E_2^{ij} the same expression holds, with $A = \frac{1}{2} \ln \varepsilon^{-1}$.

2. In Section III-B we present a *deterministic* frame-based algorithm. We require that (3) holds with probability 1 and we show that,

$$E_3^{ij} \le \frac{|S_{ij}|}{\rho_{ij}} + (2 + \sqrt{2K \ln(2L + 1)}). \tag{6}$$

We derive this algorithm from a randomized algorithm in which the nth token for matrix M_k is placed at time $U_k + n/\varphi_k$ where U_k is chosen uniformly at random in $[0, 1/\varphi_k)$. We call this the *Random-Phase Periodic Competition* scheduler. We then show how to *derandomize* this scheduler to obtain a deterministic algorithm using the method of conditional probabilities [16]. In Section IV we show that in many scenarios, (6) is significantly smaller than (4), largely due to the presence of the square-root in (6).

3. In Section III-C we present a deterministic non-frame-based algorithm. We require that (3) holds with probability 1 and we show that,

$$E_3^{ij} \le \frac{1}{\rho_{ij}} \sqrt{2|S_{ij}| \ln D} + (2 + \sqrt{2K \ln D}), \tag{7}$$

where $D = 1 + (4(2I^2 + 2)/\min_k \varphi_k)$. This algorithm is derived from a randomized algorithm in which the nth token for matrix M_k is placed uniformly at random in the interval

$[(n-1)/\varphi_k, n/\varphi_k)$. We call this the *Random-Distortion Periodic Competition* scheduler. By using the method of conditional probabilities we are able to *derandomize* this scheduler although the analysis is more complex than it was for the random-phase scheduler since we now have to consider the entire interval $[0,\infty)$ instead of a finite frame. In Section IV we show that in many scenarios, (7) is significantly smaller than (4), largely due to the presence of the square-roots in (7).

4. In Section III-D we analyze a non-frame-based scheduler in which the tokens for matrix M_k are placed according to a Poisson process. We call this the *Poisson Competition* scheduler. For this scheduler only we assume that the load ρ on each input and output is strictly less than 1. We show using the Brownian approximation [17] that,

$$E_2^{ij} \approx \frac{1}{2}\ln\varepsilon^{-1}\frac{\rho}{1-\rho}\left(\frac{1-\rho}{\rho_{ij}}+\rho\right). \tag{8}$$

The latency E_3^{ij} does not make much sense for this scheduler since the event in (3) would fail with probability 1 as we require that the inequality in (3) holds for all n.

Comparison with single server polling

We remark that our problem is significantly different from the single server polling problem (e.g. see [18] and references therein) in which a single server has to poll a set of clients at predetermined frequencies. Note that in our problem it is not sufficient for each matrix M_k to be served at evenly spaced intervals of $1/\varphi_k$ slots. This is because input-output pair ij is served whenever a matrix in S_{ij} is served. If $k,\ell \in S_{ij}$ and M_k and M_ℓ are served close together then the service to ij is bursty even though each permutation matrix might receive smooth service. Note however that we cannot in isolation change the schedule to improve service for one particular input-output pair since each permutation matrix is a member of S_{ij} for I different pairs ij.

Previous work

As mentioned earlier, papers that analyze schedulers based on decomposing the rate matrix include [1], [2], [3]. Analyses of MWM-type schedulers can be found in, for example, [7], [8], [9], [10], [11]. Some frame-based schedulers were presented in [19], [20]. If the switch fabric has an internal speedup of 2 then it is known that it can emulate output-queued switches (in which there is no contention at the inputs) [21], [22], [23]. In [24], an algorithm is presented whose aim is to "track" an idealized fluid policy.

If the switch is sufficiently underloaded then tight delay bounds can be achieved. In [25] it is shown that if the total load on any input or output is at most one quarter of the link rate, then it is possible to serve each ij pair at least once every $1/\rho_{ij}$ steps.

The remainder of the paper is organized as follows. In Section II we derive some useful facts about the event, $\{N_{ij}[T_n, T_{n+m}) \geq \rho_{ij}(m-E)\}$. In Section III we present our four schedulers and analyze them in detail. In Section IV we present some numerical results to evaluate our bounds for specific rate matrices. We defer the proofs of some of our results to the Appendix.

II. Preliminary Analysis

Note that the event (1) is equivalent to

$$\{\exists_{t>0} \quad : \quad N_{ij}[T_n, T_n+t) \geq \rho_{ij}(m-E_1^{ij}), \\ N[0, T_n+t) = n+m-1\}.$$

This can be further equivalently written as

$$\{\exists_{t>0}\exists_{s<t} \quad : \quad N_{ij}[s, s+t) \geq \rho_{ij}(m-E_1^{ij}), \\ N[0,s+t) = n+m-1, N[0,s) = n-1\}.$$

Unfortunately, it is hard in general to calculate the probability of the above event since there is too much dependence between the constituent events. It is however feasible for the case of point processes with independent increments. An example of this special case is the Poisson Competition scheduler that we analyze in Section III-D using a Geo/D/1 queue.

In the remainder of the section, we try to define a subevent of (1) whose probability is easier to bound in the general case. To that end, let $G_{n,m}$ be the *good* event $\{N_{ij}[T_n, T_{n+m}) \geq \rho_{ij}(m-E)\}$. Let $B_{n,m}$ be the *bad* event $B_{n,m} = \overline{G}_{n,m}$.

Let $\Delta_1, \Delta_2 \in \mathbb{Z}^+$ and $\Delta_3, \Delta_4 \in \mathbb{R}^+$ satisfy,

$$\Delta_1 + \Delta_2 + (\Delta_3+\Delta_4)/\rho_{ij} \leq E,$$

where $E = E_1^{ij}, E_2^{ij}$ or E_3^{ij}, depending on our calculation.

Let $t = n+m-\Delta_1$ and let $s = n+\Delta_2$. Note that s and t are integers.

Lemma 3: Suppose that $N_{ij}[s,t) \geq \rho_{ij}(t-s) - (\Delta_3+\Delta_4)$ and $[s,t) \subseteq [T_n, T_{n+m})$. Then, $N_{ij}[T_n, T_{n+m}) \geq \rho_{ij}(m-E)$.[1]

Proof: We have,

$$\begin{aligned} N_{ij}[T_n, T_{n+m}) &\geq N_{ij}[s,t) \\ &\geq \rho_{ij}(t-s) - (\Delta_3+\Delta_4) \\ &= \rho_{ij}((n+m-\Delta_1) - (n+\Delta_2) - \\ &\quad -(\Delta_3+\Delta_4)/\rho_{ij}) \\ &= \rho_{ij}(m-(\Delta_1+\Delta_2) - (\Delta_3+\Delta_4)/\rho_{ij}) \\ &\geq \rho_{ij}(m-E). \end{aligned}$$

The first two inequalities come from the assumptions of the lemma. The first equality comes from the definitions of s and t. The final inequality comes from our constraint on the Δ's. ■

From the definition of $G_{n,m}$, Lemma 3 implies,

$$\begin{aligned} G_{n,m} \quad \supseteq \quad & \{N_{ij}[s,t) \geq \rho_{ij}(t-s) - (\Delta_3+\Delta_4)\} \cap \\ & \cap \{[s,t) \subseteq [T_n, T_{n+m})\}. \end{aligned} \tag{9}$$

By the above results we can focus on the quantity $N_{ij}[s,t)$ and the relationship between the intervals $[s,t)$ and $[T_n, T_{n+m})$. However, for the random processes we consider, each interval $[s,t)$ will be dependent on too many other intervals. The way to solve this problem is to concentrate on intervals that have one of their endpoints fixed. For this purpose we must refine our results.

Lemma 4: If $N[0,t) < t+\Delta_1$ and $N[0,s) \geq s-\Delta_2$ then $[s,t) \subseteq [T_n, T_{n+m})$.

[1] Note that we are interested in the values of m such that $m \geq E$. This implies $m \geq \Delta_1+\Delta_2$, which is equivalent to $s \leq t$. For $m < E$ the inequalities in (1), (2), (3) do indeed hold.

Proof: Since $t = n+m-\Delta_1$, $N[0,t) < t+\Delta_1 \Rightarrow [0,t) \subseteq [0, T_{n+m})$. Similarly, since $s = n + \Delta_2$, $N[0,s) \geq s - \Delta_2 \Rightarrow [0,s) \supseteq [0, T_n)$. Therefore, $[s,t) \subseteq [0, T_{n+m}) \backslash [0, T_n) = [T_n, T_{n+m})$. ∎

Lemma 5: If $N_{ij}[0,t) \geq \rho_{ij} t - \Delta_3$ and $N_{ij}[0,s) \leq \rho_{ij} s + \Delta_4$ then $N_{ij}[s,t) \geq \rho_{ij}(t-s) - (\Delta_3 + \Delta_4)$.

Proof: We have, $N_{ij}[s,t) = N_{ij}[0,t) - N_{ij}[0,s) \geq (\rho_{ij} t - \Delta_3) - (\rho_{ij} s + \Delta_4) = \rho_{ij}(t-s) - (\Delta_3 + \Delta_4)$. ∎

Lemma 4, Lemma 5 and (9) imply,

$$\begin{aligned} G_{n,m} \quad \supseteq \quad & \{N[0,t) < t + \Delta_1\} \cap \\ & \cap \{N[0,s) \geq s - \Delta_2\} \cap \\ & \cap \{N_{ij}[s,t) \geq \rho_{ij}(t-s) - (\Delta_3 + \Delta_4)\}, \end{aligned} \tag{10}$$

and,

$$\begin{aligned} G_{n,m} \quad \supseteq \quad & \{N[0,t) < t + \Delta_1\} \cap \\ & \cap \{N[0,s) \geq s - \Delta_2\} \cap \\ & \cap \{N_{ij}[0,t) \geq \rho_{ij} t - \Delta_3\} \cap \\ & \cap \{N_{ij}[0,s) \leq \rho_{ij} s + \Delta_4\}. \end{aligned} \tag{11}$$

If we are interested in calculating E_1^{ij} then we only need to focus on some fixed n and m.

However, if we are interested in E_3^{ij} then we need to know whether $G_{n,m}$ for *all* n, m. For the latter case we have,

$$\begin{aligned} \textstyle\bigcap_{n,m} G_{n,m} \quad \supseteq \quad & \textstyle\bigcap_t \{N[0,t) < t + \Delta_1\} \cap \\ & \textstyle\bigcap_s \{N[0,s) \geq s - \Delta_2\} \cap \\ & \textstyle\bigcap_{s,t} \{N_{ij}[s,t) \geq \rho_{ij}(t-s) - (\Delta_3 + \Delta_4)\}. \end{aligned} \tag{12}$$

and,

$$\begin{aligned} \textstyle\bigcap_{n,m} G_{n,m} \quad \supseteq \quad & \textstyle\bigcap_t \{N[0,t) < t + \Delta_1\} \cap \\ & \textstyle\bigcap_s \{N[0,s) \geq s - \Delta_2\} \cap \\ & \textstyle\bigcap_t \{N_{ij}[0,t) \geq \rho_{ij} t - \Delta_3\} \cap \\ & \textstyle\bigcap_s \{N_{ij}[0,s) \leq \rho_{ij} s + \Delta_4\}. \end{aligned} \tag{13}$$

We note that since s and t are integers we only need to take the intersection over a discrete set of events.

III. Four Schedulers

A. Random Permutation

We consider a frame-based scheduler in which the permutation matrices in a frame are scheduled in random order. More formally, denote by $z = (z_1, z_2, \ldots, z_L)$ some fixed order of the token types such that there are exactly ℓ_k tokens of type k, $k = 1, \ldots, K$. Let $\pi = (\pi(1), \pi(2), \ldots, \pi(L))$ be a random permutation of the elements $(1, 2, \ldots, L)$.

For $n = 1, \ldots, L$ we define the schedule by randomly permuting the elements of z, i.e.,

$$Z_n = z_{\pi(n)},\ n = 1, \ldots, L.$$

The schedule is extended for $n > L$ by concatenating replicas of the schedule Z_n, $n = 1, \ldots, L$.

As an aside, note that the scheduler as defined above can be formulated in the framework of point processes. We can first construct the counting processes N_k, $k = 1, \ldots, K$, on $[0,1]$ by placing ℓ_k points uniformly at random in $[0,1]$. Then, N_k is extended to the whole positive real line by periodic extension of the points in $[0,1]$.

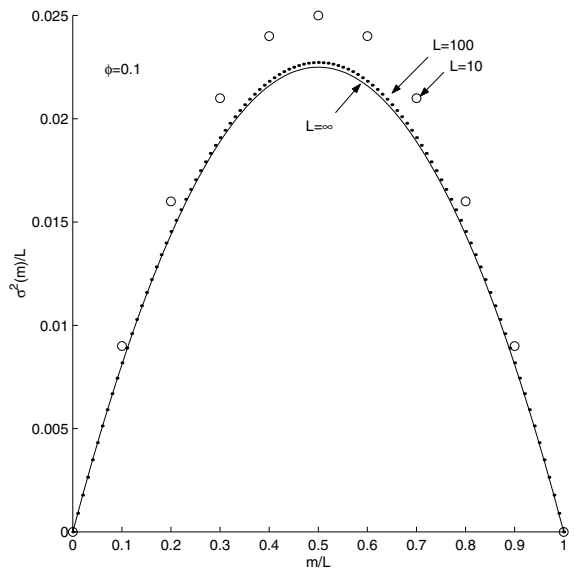

Fig. 2. Normalized variance of $N_{ij}[T_n, T_{n+m})$ for varying frame size L, and $\rho_{ij} = \phi = 1/10$.

We first discuss some elementary properties of the scheduler as described above, and then display the latencies. By a routine combinatorial argument we obtain, for $l = 1, \ldots, \min[\ell_{ij}, m]$,

$$\mathsf{P}(N_{ij}[T_n, T_{n+m}) = l) = \frac{\binom{m}{l}\binom{L-m}{\ell_{ij}-l}}{\binom{L}{\ell_{ij}}},$$

where $\ell_{ij} = \sum_{k \in S_{ij}} \ell_k$. Notice that the probability does not depend on n, which reflects the fact that the counting process is stationary. The above explicit expression enables us to compute the latency E_1^{ij} defined by (1).

As an aside, we remark that for any fixed m, $N_{ij}[T_n, T_{n+m})$ converges in distribution to Binomial random variable (m, ρ_{ij}), as $L \to \infty$. (This can be checked, for example, by Stirling's formula.)

One may check that the variance of $N_{ij}[T_n, T_{n+m})$ is

$$\sigma_{ij}^2(m) = \frac{L^2}{L-1} \rho_{ij}(1-\rho_{ij}) \frac{m}{L} \left(1 - \frac{m}{L}\right).$$

We omit this calculation for the benefit of space. Note that the variance forms a bridge ($\sigma_{ij}^2(0) = \sigma_{ij}^2(L) = 0$, with the global maximum at $m = L/2$); see Figure 2. Note also that, $\sigma_{ij}^2(m) \to L\rho_{ij}(1-\rho_{ij})\frac{m}{L}\left(1 - \frac{m}{L}\right)$, as $L \to \infty$.

We finally show the main results of this section; asymptotic expressions for the latencies E_2^{ij} and E_3^{ij}. The proofs are given in Appendices A and B.

Proposition 6: As $L \to \infty$,

$$E_2^{ij} \to \sqrt{\frac{1}{2} \ln \frac{1}{\varepsilon} \left(\frac{1}{\rho_{ij}} - 1\right) L}.$$

Note that E_2^{ij} scales with the frame length L as $O(\sqrt{L})$.

We have the following result for the latency E_3^{ij}.

Proposition 7: As $L \to \infty$,

$$E_3^{ij} \to \sqrt{A\left(\frac{1}{\rho_{ij}} - 1\right) L}, \tag{14}$$

where A is the positive solution of

$$\sum_{\ell=1}^{\infty} \left(4\ell^2 A - 1\right) e^{-2\ell^2 A} = \frac{1}{2}\varepsilon.$$

Note that E_3^{ij} also scales with the frame length L as $O(\sqrt{L})$, but with a different constant. In Figure 3, we plot values for E_2^{ij} and E_3^{ij} obtained empirically, together with the above limits, for different values of L.

B. Random-Phase Periodic Competition

Let U_k, $k = 1, \ldots, K$, be a collection of independent uniformly distributed random variables on $[0, 1]$. We define the scheduler as follows; for each $k = 1, \ldots, K$,

$$N_k[0, t) = \sum_{n \geq 0} 1_{[0,t)}(\frac{n-1}{\varphi_k} + \frac{1}{\varphi_k} U_k).$$

Thus the tokens for matrix M_k form a periodic stream of period $1/\varphi_k$ with the random phase shift U_k/φ_k.

We assume that the φ_k have the property that we can define a frame-based scheduler (see the Introduction). Therefore we only need to concentrate on the time interval L. For any interval $[s, t)$, $N_k[s, t) \geq \varphi_k(t - s) - 1$. This implies,

$$N_{ij}[s,t) \geq \sum_{k \in S_{ij}} (\varphi_k(t-s)-1) \Rightarrow N_{ij}[s,t) \geq \rho_{ij}(t-s) - |S_{ij}|.$$

We follow the method of Section II and set $t = n + m - \Delta_1$, $s = n + \Delta_2$, where Δ_1, Δ_2 are defined below. For any permutation matrix M_k, let $a_k = \lfloor \varphi_k t \rfloor$. Then $N_k[0, t) = a_k + X_k$ where X_k is a binary random variable with mean $\varphi_k t - a_k$. Let $\mu = \mathsf{E}[\sum_k X_k] = \sum_k (\varphi_k t - a_k) = t - \sum_k a_k$. We have,

$$\begin{aligned} N[0, t) \geq t + \Delta_1 \quad &\Leftrightarrow \quad \textstyle\sum_k (a_k + X_k) \geq t + \Delta_1 \\ &\Leftrightarrow \quad \textstyle\sum_k X_k \geq t + \Delta_1 - \sum_k a_k \\ &\Leftrightarrow \quad \textstyle\sum_k X_k \geq \mu + \Delta_1. \end{aligned}$$

Therefore, by Hoeffding's inequality [26], $\mathsf{P}(N[0, t) \geq t + \Delta_1) \leq \exp(-2(\Delta_1)^2/K)$. Similarly, $\mathsf{P}(N[0, s) < s - \Delta_2) \leq \exp(-2(\Delta_2)^2/K)$.

Let,

$$\begin{aligned} \Delta_1 = \Delta_2 &= \left\lceil \sqrt{(K/2) \cdot \ln(2L+1)} \right\rceil, \\ \Delta_3 &= \Delta_4 = |S_{ij}|/2, \\ E_3^{ij} &= (|S_{ij}|/\rho_{ij}) + (2 + \sqrt{2K \ln(2L+1)}). \end{aligned}$$

Then, by the above Hoeffding bounds and the containment (12) from Section II we have,

$$\begin{aligned} &\mathsf{P}(\textstyle\bigcap_{ij} \bigcap_{nm} \{N_{ij}[T_n, T_m) \geq \rho_{ij}(m - E_3^{ij})\}) \\ \geq\ & 1 - \textstyle\sum_{t=1}^{L} \mathsf{P}(N[0, t) \geq t + \Delta_1) - \\ & - \textstyle\sum_{s=1}^{L} \mathsf{P}(N[0, s) < s - \Delta_2) \\ \geq\ & 1 - \textstyle\sum_{t=1}^{L} \exp(\frac{-2(\Delta_1)^2}{K}) - \sum_{s=1}^{L} \exp(\frac{-2(\Delta_2)^2}{K}) \\ \geq\ & 1 - \frac{2L}{2L+1}. \end{aligned} \tag{15}$$

(Note that since we are considering a finite frame we only need sum over $s, t \in \{1, \ldots, L\}$.)

Hence with probability $1 - 2L/(2L + 1)$, the rate-latency condition (3) holds *for all* ij. We now show how to *derandomize* the algorithm so that condition (3) holds with probability 1.

Derandomization

We use the method of conditional probabilities (see e.g. [16]) that is motivated by the following lemma (which we prove in Appendix C).

Lemma 8: Let $Y_1, \ldots, Y_{n_1}$ be a set of random variables, let $X_1, \ldots, X_{n_2}$ be a set of independent binary random variables and let $\sigma_1, \ldots, \sigma_{n_3}$ be a set of events such that for some functions $f_{ij}(\cdot)$,

$$\mathsf{P}(\sigma_i | A) \leq \mathsf{E}[\prod_{j=1}^{n_2} f_{ij}(X_j) | A], \tag{16}$$

for any event A. Then there exists a set of values $y_1, \ldots, y_{n_1}$ such that,

$$\sum_i \mathsf{P}(\sigma_i | Y_1 = y_1, \ldots, Y_{n_1} = y_{n_1}) \leq \sum_i \mathsf{E}[\prod_{j=1}^{n_2} f_{ij}(X_j)].$$

In particular if $\sum_i \mathsf{E}[\prod_{j=1}^{n_2} f_{ij}(X_j)] < 1$ and σ_i is completely determined by $Y_1, \ldots, Y_{n_1}$ then,

$$\sum_i \mathsf{P}(\sigma_i | Y_1 = y_1, \ldots, Y_{n_1} = y_{n_1}) = 0.$$

To compute y_v given $y_1, \ldots, y_{v-1}$ we minimize,

$$\begin{aligned} &\textstyle\sum_i \mathsf{E}[\prod_{j=1}^{n_2} f_{ij}(X_j) | Y_1 = y_1, \ldots, Y_v = y_v] = \\ &= \textstyle\sum_i \prod_{j=1}^{n_2} \mathsf{E}[f_{ij}(X_j) | Y_1 = y_1, \ldots, Y_v = y_v], \end{aligned}$$

as we vary y_v over the full range of Y_v. (Recall that the random variables X_j are independent and so we can exchange the expectation with the product).

To apply this lemma in our setting we take $Y_1, \ldots, Y_{n_1}$ to be the random phase shifts $U_1, \ldots, U_K$; $X_1, \ldots, X_{n_2}$ to be binary random variables of the form $N_k[0, t) - \lfloor \varphi_k t \rfloor$ and $\sigma_1, \ldots, \sigma_{n_3}$ to be events of the form $\{N[0, t) \geq t + \Delta_1\}$ or $\{N[0, s) < s - \Delta_2\}$. The functions $f_{ij}(\cdot)$ are defined by,

$$\mathsf{P}(N[0, t) \geq t + \Delta_1) \quad \leq \quad e^{-\theta(t+\Delta_1)} \mathsf{E}[\textstyle\prod_{k=1}^{K} e^{\theta N_k[0,t)}], \tag{17}$$

where $\theta = \ln(b(1 - a)/a(1 - b))$, $a = (t - \sum_{k=1}^{K} \lfloor \varphi_k t \rfloor)/K$, $b = (t + \Delta_1 - \sum_{k=1}^{K} \lfloor \varphi_k t \rfloor)/K$. A similar inequality holds for $\mathsf{P}(N[0, s) < s - \Delta_2)$.

In the derivation of (15) we showed that,

$$\sum_{t=1}^{L} \mathsf{P}(N[0, t) \geq t + \Delta_1) + \sum_{s=1}^{L} \mathsf{P}(N[0, s) < s - \Delta_2) \leq \frac{2L}{2L+1},$$

using Hoeffding bounds that are derived from (17). Hence by Lemma 8 there exist fixed values $u_1, \ldots, u_K$ for the initial phase shifts such that $N[0, t) < t + \Delta_1$ for all t and $N[0, s) \geq s - \Delta_2$ for all s.

The one complication that arises in the calculation of the u_k is that the U_k are continuous random variables, they do not

take discrete values. However, as U_k is varied between 0 and 1, $N_k[0,t) - \lfloor \varphi_k t \rfloor$ changes from 0 to 1 at one discrete point. Hence it is sufficient to consider only $L+1$ values of U_k. The right-hand side of (17) may be computed in time polynomial in K and L, even if some of the phase shifts have already been fixed. Hence, we can fix the value of U_k in time polynomial in K and L.

Theorem 9: The resulting deterministic scheduler satisfies (3) with,

$$E_3^{ij} = \frac{|S_{ij}|}{\rho_{ij}} + (2 + \sqrt{2K \ln(2L+1)}).$$

C. Random-Distortion Periodic Competition

Let U_{kn}, $k = 1, \ldots, K$, $n \in \mathbb{Z}^+$, be a collection of independent uniformly distributed random variables on $[0,1]$. We define the scheduler as follows; for each $k = 1, \ldots, K$,

$$N_k[0,t) = n + 1_{U_{kn} < \varphi_k t - n},$$

where $n = \lfloor \varphi_k t \rfloor$. Another interpretation is: the nth point of the kth token type is placed uniformly at random in the interval $[n/\varphi_k, (n+1)/\varphi_k)$.

We make use of the containment (13) from Section II. We apply Hoeffding bounds in a similar manner to the previous subsection to obtain.

$$\begin{array}{rcl} \mathsf{P}(N[0,t) \geq t + \Delta_1) & \leq & \exp(-2(\Delta_1)^2/K), \\ \mathsf{P}(N[0,s) < s - \Delta_2) & \leq & \exp(-2(\Delta_2)^2/K), \\ \mathsf{P}(N_{ij}[0,t) \leq \rho_{ij} t - \Delta_3) & \leq & \exp(-2(\Delta_3)^2/|S_{ij}|), \\ \mathsf{P}(N_{ij}[0,s) \geq \rho_{ij} s + \Delta_4) & \leq & \exp(-2(\Delta_4)^2/|S_{ij}|). \end{array}$$

Let,

$$\begin{array}{rcl} \Gamma_1(t) & = & \mathsf{P}(N[0,t) \geq t + \Delta_1) + \\ & & + \sum_{ij} \mathsf{P}(N_{ij}[0,t) < \rho_{ij} t - \Delta_3), \\ \Gamma_2(s) & = & \mathsf{P}(N[0,s) < s - \Delta_2) + \\ & & + \sum_{ij} \mathsf{P}(N_{ij}[0,s) > \rho_{ij} s + \Delta_4), \\ D & = & 1 + (4(2I^2+2)/\min_k \varphi_k), \\ \Delta_1 = \Delta_2 & = & \left\lceil \sqrt{(K/2) \cdot \ln D} \right\rceil, \\ \Delta_3 = \Delta_4 & = & \sqrt{(|S_{ij}|/2) \cdot \ln D}, \\ E_3^{ij} & = & \Delta_1 + \Delta_2 + (\Delta_3 + \Delta_4)/\rho_{ij}. \end{array}$$

For fixed s and t we have,

$$\Gamma_1(t) + \Gamma_2(s) \leq \frac{(2I^2+2)}{D}.$$

Note that we cannot apply a union bound over s and t as we did in the previous subsection because s and t range over the entire interval $[0, \infty)$. However, note that if $\Gamma_1(t) = \Gamma_2(s) = 0$ for all s, t then from (13) we know that (3) holds for all i, j with probability 1. Hence we focus on derandomizing the algorithm.

Derandomization

Instead of placing the nth token for matrix M_k at random into the interval $[n/\varphi_k, (n+1)/\varphi_k)$, we now wish to place it deterministically. Let $P = \lceil 2/\min_k \varphi_k \rceil$. We divide time into intervals of length P, namely, $[0,P), [P,2P), \ldots$. Let A^ω be the set of tokens that fall into the interval $[\omega P, (\omega+1)P)$ with probability 1, i.e. the nth token for matrix M_k is in A^ω if and only if $[n/\varphi_k, (n+1)/\varphi_k) \subseteq [\omega P, (\omega+1)P)$. Let B^ω be the set of tokens that are not in $A^{\omega'}$ for any ω' and that fall into the interval $[(\omega - \frac{1}{2})P, (\omega + \frac{1}{2})P)$ with probability 1. We have chosen P sufficiently large so that all tokens are in $A^\omega \cup B^\omega$ for some ω.

Suppose inductively that for $\omega' < \omega$ we have fixed the positions of all the tokens in $A^{\omega'} \cup B^{\omega'}$. Since none of the tokens that have already been fixed affect the interval $[\omega P, (\omega+1)P)$, our previous analysis implies,

$$\textstyle\sum_{\omega P}^{(\omega+1)P} \Gamma_1(t) + \sum_{\omega P}^{(\omega+1)P} \Gamma_2(s) \leq \frac{P(2I^2+2)}{D}.$$

By applying the method of conditional probabilities in a similar manner to Section III-B, we can fix the positions of tokens in A^ω one after the other so that we still have,

$$\textstyle\sum_{\omega P}^{(\omega+1)P} \Gamma_1(t) + \sum_{\omega P}^{(\omega+1)P} \Gamma_2(s) \leq \frac{P(2I^2+2)}{D}.$$

Here, the constituent probabilities of $\Gamma_1(t)$ and $\Gamma_2(t)$ are now conditioned on the fact that the tokens in A^ω are fixed. We obtain,

$$\begin{array}{rl} & \sum_{(\omega-\frac{1}{2})P}^{(\omega+\frac{1}{2})P} \Gamma_1(t) + \sum_{(\omega-\frac{1}{2})P}^{(\omega+\frac{1}{2})P} \Gamma_2(s) \\ \leq & \sum_{(\omega-1)P}^{(\omega+1)P} \Gamma_1(t) + \sum_{(\omega-1)P}^{(\omega+1)P} \Gamma_2(s) \\ \leq & \frac{2P(2I^2+2)}{D} < 1. \end{array}$$

By the method of conditional probabilities we can fix the positions of tokens in B^ω so that we still have,

$$\textstyle\sum_{(\omega-\frac{1}{2})P}^{(\omega+\frac{1}{2})P} \Gamma_1(t) + \sum_{(\omega-\frac{1}{2})P}^{(\omega+\frac{1}{2})P} \Gamma_2(s) \leq \frac{2P(2I^2+2)}{D} < 1.$$

All tokens in $A^\omega \cup B^\omega$ are now fixed and so we have a deterministic schedule up to time $(\omega + \frac{1}{2})P$. Recall that $\Gamma_1(t)$ and $\Gamma_2(s)$ are sums of probabilities. Therefore $\Gamma_1(t) = \Gamma_2(s) = 0$ for all $s, t \in [(\omega - \frac{1}{2})P, (\omega + \frac{1}{2})P)$. This process can be repeated indefinitely.

Theorem 10: The resulting deterministic scheduler satisfies (3) with,

$$E_3^{ij} = \frac{1}{\rho_{ij}} \sqrt{2|S_{ij}| \ln D} + (2 + \sqrt{2K \ln D}).$$

Adaptation to the substochastic case

For the previous three schedulers, we have assumed that the rate matrix M is doubly stochastic, i.e. $\sum_i \rho_{ij} = 1$ and $\sum_j \rho_{ij} = 1$. For the case in which M is only substochastic, i.e. $\sum_i \rho_{ij} \leq 1$ and $\sum_j \rho_{ij} \leq 1$, it is known by a result of von Neumann (see e.g. [1]) that there exists a matrix M' with ij entry ρ'_{ij} such that $\rho_{ij} \leq \rho'_{ij}$ for all ij and M' is doubly stochastic. In this case, we can apply all the results of this paper to the matrix M' to obtain latencies E_1^{ij}, E_2^{ij} and E_3^{ij}. Note that the ij traffic might not be able to use all the service it is offered. In this case the residual bandwidth can be used for best-effort traffic.

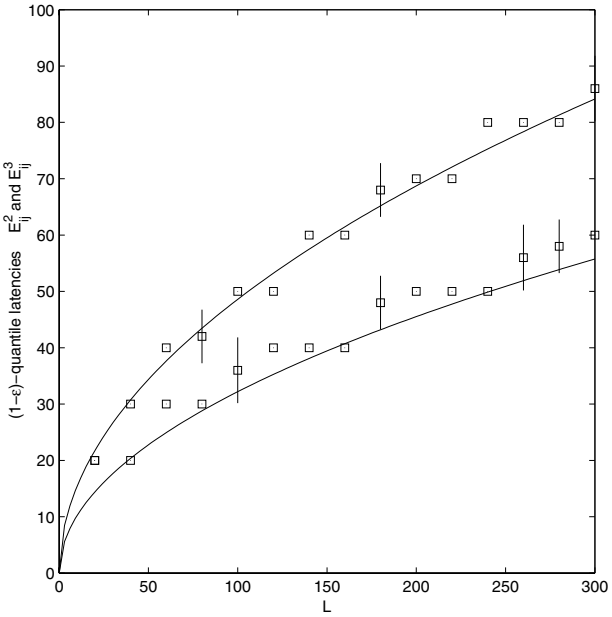

Fig. 3. 0.99-quantile empirical and theoretical limit latencies of Random Permutation Scheduler: (Bottom curve) E_2^{ij} and (Top curve) E_3^{ij}. The empirical quantiles are computed from 5 independent samples each of 500 samples of random permutations. The empirical quantiles are shown as averages over the 5 samples along with 0.95-confidence intervals.

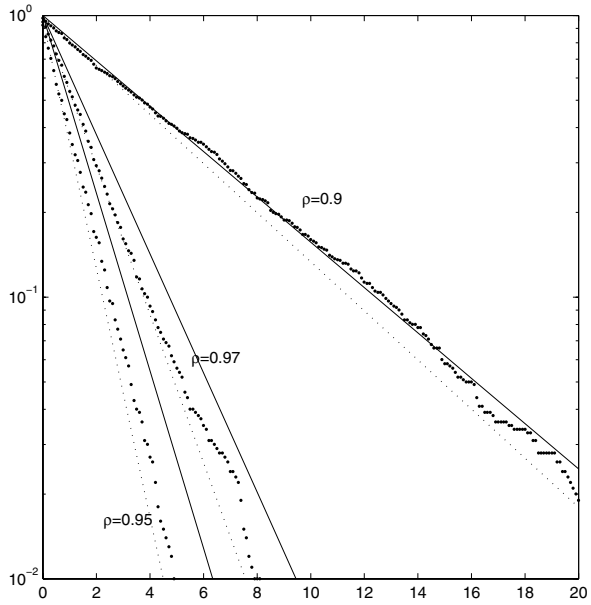

Fig. 4. Complementary distribution of $V_{\overline{ij}}$: (dots) empirical estimates, (dotted line) M/D/1, (solid line) Brownian approximation. $V_{\overline{ij}}$ is estimated by averaging over 1000 random samples of length 10000. $\rho_{ij} = 0.1$.

D. Poisson Competition

For our final scheduler we require that the load on each input and output is strictly less than 1. Let N_k be Poisson with intensity φ_k, all $k = 1, \ldots, K$. Then the following holds.

Lemma 11: For any ij, and $n, m \geq 0$, $l = 1, 2, \ldots, m$,

$$\mathsf{P}(N_{ij}[T_n, T_{n+m}) = l) = \left(\begin{array}{c} m \\ l \end{array} \right) \rho_{ij}^l (1 - \rho_{ij})^{m-l}.$$

The above result may be obvious to many; we give an elementary proof in Appendix D. We note that $(T_n, Z_n)_{n \geq 0}$ is a marked point process with independent identically distributed marks, where $Z_n = k$ with probability φ_k. Our naming of this scheduler is inspired by the Poisson competition theorem (Theorem 1.3, Chapter 8 [27]).

We continue further by observing the following queueing interpretation of the latencies defined in (2) and (3). Locally to this section, assume $\sum_{k=1}^K \varphi_k < 1$; we impose this condition to ensure stability. Moreover, for a fixed ij, let $\rho < 1$ be such that $\sum_{k \ni S_{ij}} \varphi_k = \rho(1 - \rho_{ij})$. We also assume that the counting processes N_k are extended to $\mathbb{R}$, the whole real line. Then, it is not difficult to observe that (2) is equivalent to

$$\{V_{\overline{ij}}[0] \leq \rho_{ij} E_2^{ij}\},$$

where $V_{\overline{ij}}[n]$, $n = 0, \pm 1, \pm 2, \ldots$, is *the unfinished work* of a slotted single server queueing system with infinite buffer capacity, service rate $(1 - \rho_{ij})$ and an arrival process that is 0 or 1 with the probability of an arrival equal to $\rho(1 - \rho_{ij})$. The above observation follows immediately by Reich's formula,

$$V_{\overline{ij}}[n] = \max_{m \geq 1} [N_{\overline{ij}}[T_n, T_{n+m}) - (1 - \rho_{ij})m].$$

From Lemma 11, it follows that the unfinished work is of a Geo/D/1 queue. The distribution of the unfinished work of a Geo/D/1 queue is known in closed form [28], which in our context amounts to

$$\mathsf{P}(V_{\overline{ij}}[0] \leq v) = \\ = \frac{1 - qD}{(1-q)^{v+1}} \sum_{l=0}^{j} [q(1-q)^{D-1}]^l (-1)^l \left(\begin{array}{c} v - (D-1)l \\ l \end{array} \right),$$

where j is the integer such that $jD \leq v \leq (j+1)D - 1$, $q := \rho(1 - \rho_{ij})$, and $D := 1/(1 - \rho_{ij})$ is implicitly assumed to be an integer. (If D would be a real, then one may redefine $D := \lceil 1/(1 - \rho_{ij}) \rceil$ to obtain a lower bound, provided that $\rho D < 1$.)

The last would enable one at least in theory to exactly compute the latency E_2^{ij} in (2). In practice one may expect numerical instability as $\rho D \to 1$. The following heuristic argument gives us an approximation that is numerically stable, but also brings us some insight about the latency. By appealing to the Brownian approximation (see [17], Sec. 5.7, Equation (7.16)) we claim

$$\mathsf{P}(V_{\overline{ij}}[0] \leq \rho_{ij} E_2^{ij}) \approx 1 - e^{-2 \frac{1-\rho}{\rho(1-\rho(1-\rho_{ij}))} \rho_{ij} E_2^{ij}}.$$

Hence, we have

$$E_2^{ij} \approx \frac{1}{2} \ln \varepsilon^{-1} \frac{\rho}{1-\rho} \left(\frac{1-\rho}{\rho_{ij}} + \rho \right).$$

Another approximation can be obtained by considering M/D/1 queue, a continuous time analog[2] of Geo/D/1. A simple exponential approximation is known for M/D/1 ([29], Equation

[2] A notable difference is that with Geo/D/1, in contrast to M/D/1, the number of arrivals over any interval of length m is bounded by m.

6.1.6, Section 6.1.2). In Figure 4 we show a numerical comparison of the approximations mentioned above with their empirical companions. We observe that the above approximation for E_2^{ij} should be good in the heavy-traffic limit as $\rho \to 1$. It is perhaps interesting to observe that in the heavy-traffic limit E_2^{ij} becomes insensitive to ρ_{ij}.

As mentioned in the Introduction, the latency E_3^{ij} does not make much sense for this scheduler since the event in (3) would fail with probability 1 as we require that the inequality in (3) holds for all n.

IV. Numerical Results

In this section we evaluate some of our bounds for specific rate matrices. Recall that the best possible latency for input-output pair ij is $1/\rho_{ij}$. Hence the ratio between the latency provided by the scheduler, E_3^{ij}, and the best possible latency is $\rho_{ij}E_3^{ij}$. For this reason we define $\max_{ij} \rho_{ij}E_3^{ij}$ to be the figure of merit for a scheduler.

We evaluate the bounds (6) and (7) for the deterministic algorithms derived from Random-Phase Periodic Competition and Random-Distortion Periodic Competition. We compare them with the bound (4) for PGPS. We would like to use matrices drawn uniformly from the set of doubly-stochastic matrices. However, we do not know of a method to generate such a matrix uniformly. Hence we use the following method to generate our example matrices. We start with a uniform matrix in which all entries are equal to I/L where $L = I \times I$. We then repeatedly choose parameters i_1, i_2, j_1, j_2 and δ uniformly at random such that $\delta \leq \min\{\rho_{i_1 j_1}, \rho_{i_2 j_2}\}$. We subtract δ from $\rho_{i_1 j_1}$ and $\rho_{i_2 j_2}$ and we add δ to $\rho_{i_1 j_2}$ and $\rho_{i_2 j_1}$. We carry out this operation 100000 times. Note that it preserves the doubly stochastic nature of the matrix. We also ensure that all entries of the rate matrix are integer multiples of $1/L$. Hence we can define frame-based schedulers with frame-length L.

In Figure 5 we plot the value of $\max_{ij} \rho_{ij}E_3^{ij}$ for different values of I, the switch size. We see that except for extremely small switches, the bound for the Random-Distortion scheduler is smaller than the bound for the Random-Phase scheduler which is in turn smaller than the bound for PGPS.

In Figure 6 we examine how $\rho_{ij}E_3^{ij}$ varies for different pairs ij. In particular we examine a 64×64 matrix for which $K = 2423$. For each value of x we plot the fraction of ij pairs for which $\rho_{ij}E_3^{ij} \leq x$. We see that the bound (6) for the Random-Phase based algorithm is consistently smaller than the bound (4) for PGPS. The bound (7) for the Random-Distortion based algorithm has a smaller range than the other two bounds. There are fewer pairs ij with large values of $\rho_{ij}E_3^{ij}$ but there are also fewer pairs ij with small values of $\rho_{ij}E_3^{ij}$. The reason for the latter phenomenon is that the bound (7) is typically larger than the bounds, (4), (6) when the value of $|S_{ij}|$ is small.

We remark that we cannot directly compare the expressions (5) and (8) for the Random Permutation and Poisson Competition schedulers with the bounds (4), (6) and (7) considered in this section. This is because the expression (5) is a limit and the expression (8) is for E_2^{ij}, not E_3^{ij}.

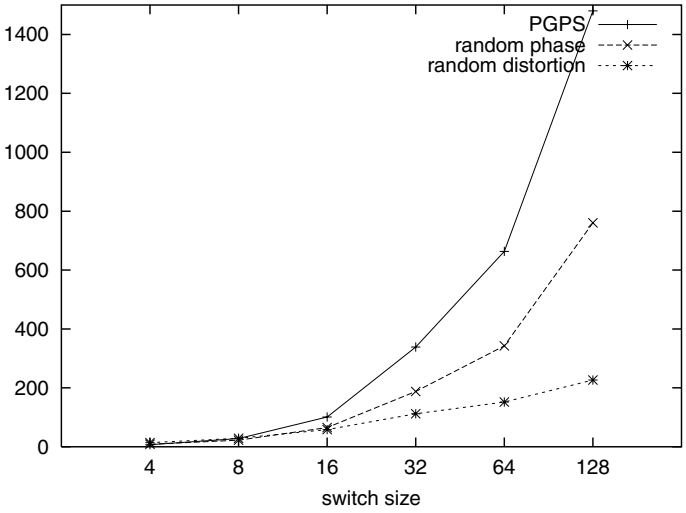

Fig. 5. The value of $\max_{ij} \rho_{ij}E_3^{ij}$ for switches of varying size.

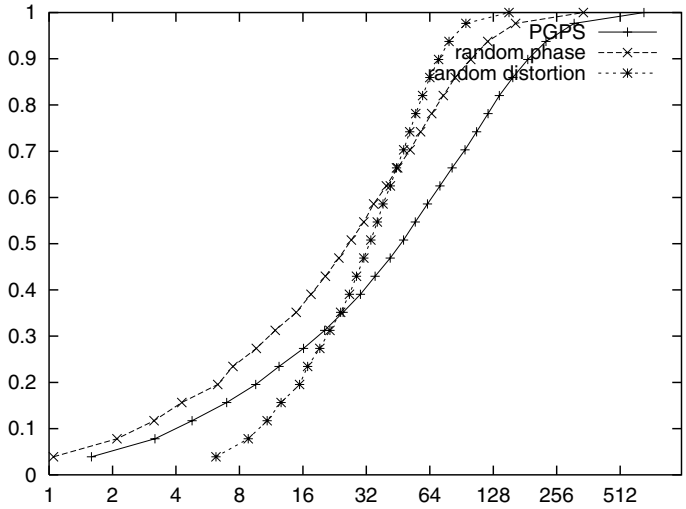

Fig. 6. Fraction of ij pairs for which $\rho_{ij}E_3^{ij} \leq x$. The matrix has $I = 64$, $L = 4096$ and $K = 2423$.

V. Conclusion

In this paper we have analyzed the latency of four decomposition-based schedulers for input-queued switches. We believe that there are a number of interesting open problems. First, it is possible that a tighter analysis of our framework of point processes could lead to better bounds for the existing schedulers, and may even motivate the construction of new ones. Second, we know of no non-trivial lower bounds on the best possible latency. It would be interesting to know what is the best value of $\max_{ij} \rho_{ij}E_3^{ij}$ that can be achieved.

VI. Acknowledgments

The authors would like to thank Sem Borst, Jean-Yves Le Boudec and Lisa Zhang for many helpful discussions. The second author would like to thank the members of the Math Center at Bell Labs for their hospitality.

REFERENCES

[1] C.S. Chang, W.J. Chen, and H.Y. Huang, "On service guarantees for input buffered crossbar switches: A capacity decomposition approach by Birkhoff and von Neumann," in *Proc. of IEEE IWQoS*, London, UK, 1999.

[2] C.S. Chang, W.J. Chen, and H.Y. Huang, "Birkhoff-von Neumann input buffered crossbar switches," in *Proc. of IEEE INFOCOM'00*, Tel-Aviv, Israel, March 2000.

[3] S. Li and N. Ansari, "Input-queued switching with QoS guarantees," in *Proc. of IEEE INFOCOM '99*, New York, NY, March 1999, pp. 1152 – 1159.

[4] G. Birkhoff, "Tres observaciones sobre el algebra lineal," *Univ. Nac. Tucumán Rev. Ser. A5*, pp. 147 – 150, 1946.

[5] J. von Neumann, "A certain zero-sum two-person game equivalent to the optimal assignment problem," *Contributions to the Theory of Games*, vol. 2, pp. 5 – 12, 1953.

[6] B. Davie (Editor), A. Charny, F. Baker, J. Bennett, K. Benson, J.-Y. Le Boudec, A. Chiu, W. Courtney, S. Davari, V. Firoiu, C. Kalmanek, K. K. Ramakrishnan, and D. Stiliadis, "An expedited forwarding PHB," April 2001.

[7] N. W. McKeown, V. Anantharam, and J. Walrand, "Achieving 100% throughput in an input-queued switch," in *Proc. of IEEE INFOCOM*, 1996.

[8] A. Mekkittikul and N. W. McKeown, "A practical algorithm to achieve 100% throughput in input-queued switches," in *Proc. of IEEE INFOCOM*, 1998.

[9] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," *IEEE Transactions on Automatic Control*, vol. 37, no. 12, pp. 1936 – 1948, December 1992.

[10] J. Dai and B. Prabhakar, "The throughput of data switches with and without speedup," in *Proc. of IEEE INFOCOM '00*, Tel Aviv, Israel, March 2000, pp. 556 – 564.

[11] E. Leonardi, M. Mellia, F. Neri, and M. Ajmone Marsan, "Bounds on average delays and queue size averages and variances in input-queued cell based switch," in *Proc. of IEEE Infocom 2001*, Anchorage, AK, 2001.

[12] Devavrat Shah and Milind Kopikare, "Delay bounds for approximate maximum weight matching algorithms for input queued switches," in *Proc. of IEEE Infocom 2002*, New York, USA, June 2002.

[13] A. K. Parekh and R. G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: The single node case," *IEEE/ACM Trans. on Networking*, vol. 1-3, pp. 344–357, June 1993.

[14] J. Bennett, K. Benson, A. Charny, W. Courtney, and J.-Y. Le Boudec, "Delay jitter bounds and packet scale rate guarantee for expedited forwarding," in *Proc. of IEEE INFOCOM'2001*, March 2001.

[15] J.-Y. Le Boudec and P. Thiran, *Network Calculus*, Springer-Verlag, 2001, (also available on-line at http://ica1www.epfl.ch/PS_files/NetCal.htm).

[16] R. Motwani and P. Raghavan, *Randomized algorithms*, Cambridge University Press, 1995.

[17] Ward Whitt, *Stochastic-Processes Limits*, Springer, 2002.

[18] H. Takagi, "Analysis and application of polling models," in *Performance Evaluation: Origins and Directions, Lecture Notes in Computer Science 1769*, G. Haring, C. Lindemann, and M. Reiser, Eds., 2000, pp. 423–442.

[19] A. Hung, G. Kesidis, and N. McKeown, "ATM input-buffered switches with the guaranteed rate property," in *Proc. of IEEE ISCC*, 1998, pp. 331–335.

[20] T. Lee and C. Lam, "Path switching – A quasi-static routing scheme for large scale ATM packet switches," *IEEE Journal on Selected Areas of Communications*, vol. 15, pp. 914 – 924, 1997.

[21] S. Chuang, A. Goel, N. McKeown, and B. Prabhakar, "Matching output queueing with combined input and output queueing," *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 6, pp. 1030–1039, 1999.

[22] A. Charny, P. Krishna, N. Patel, and R. Simcoe, "Algorithms for providing bandwidth and delay guarantees in input-buffered crossbars with speedup," in *Proc. of IEEE IWQoS*, Napa, CA, 1998.

[23] I. Stoica and H. Zhang, "Exact emulation of an output queueing switch by a combined input output queueing switch," in *Proc. of IEEE IWQoS*, Napa, CA, 1998.

[24] V. Tabatabaee, L. Georgiadis, and L. Tassiulas, "QoS provisioning and tracking fluid policies in input queuing switches," *IEEE/ACM Trans. on Networking*, vol. 9, no. 5, October 2001.

[25] J. Giles and B. Hajek, "Scheduling multirate periodic traffic in a packet switch," in *1997 Conference on Information Sciences and Systems at John Hopkins University*, 1997.

[26] Wassily Hoeffding, "Probability inequalities for sums of bounded random variables," *American Statistical Association Journal*, pp. 13–30, March 1963.

[27] Pierre Brémaud, *Markov Chains: Gibbs Fields, Monte Carlo Simulation and Queues*, Springer, 1999.

[28] Annie Gravey, Jean-Raymond Louvion, and Pierre Boyer, "On the Geo/D/1 and Geo/D/1/n queues," *Performance Evaluation, North-Holland*, vol. 11, pp. 117–125, 1990.

[29] J. W. Roberts (Editor), *COST 224: Performance evaluation and design of multiservice networks*, Commision of the European Communities, 1991.

[30] Jaroslav Hájek, Zbyněk Šidák, and Pranab K. Sen, *Theory of Rank Tests*, Academic Press, 1999.

[31] J. L. Doob, "Heuristic approach to the Kolmogorov-Smirnov theorems," *Ann. Math. Statist.*, vol. 20, pp. 293–403, 1949.

[32] Wim Vervaat, "A relation between brownian bridge and brownian excursion," *The Annals of Probability*, vol. 7, no. 1, pp. 143–149, 1979.

[33] Bruce Hajek, "A queue with periodic arrivals and constant service rate," *in Probability, Statistics and Optimization – a Tribute to Peter Whittle (F.P. Kelly ed., John Wiley and Sons*, pp. 147–158, 1994.

[34] Frank B. Knight, *Essentials of Brownian Motion and Diffusion*, vol. 18, Mathematical Surveys, American Mathematical Society, 1981.

APPENDIX

I. PROOF OF PROPOSITION 6

Proof: Note that by periodicity of the counting process N_{ij}, (2) is equivalent to

$$\left\{\max_{1\le m\le L}[\rho_{ij}m - N_{ij}[T_n, T_{n+m})] \le \rho_{ij}E_2^{ij}\right\} \Leftrightarrow \left\{\max_{1\le m\le L}[N_{\overline{ij}}[T_n, T_{n+m}) - (1-\rho_{ij})m] \le \rho_{ij}E_2^{ij}\right\}.$$

Now, it is a standard result (e.g. Theorem 1, Section 6.3.7 [30]) that as $L \to \infty$ we have the convergence in distribution

$$\max_{1\le m\le L}[N_{\overline{ij}}[T_n, T_{n+m}) - (1-\rho_{ij})m] \Rightarrow \Rightarrow \sqrt{\rho_{ij}(1-\rho_{ij})L}\sup_{0\le t\le 1} B_0(t),$$

where B_0 is Brownian bridge, the Gaussian process with $\mathsf{E}[B_0(t)] = 0$ and $\mathsf{cov}[B_0(s)B_0(t)] = s(t-s)$, $0 \le s \le t \le 1$. Another definition of Brownian bridge is given by $B_0(t) = B(t) - tB(1)$, $t \in [0,1]$, where $B(t)$, $t \ge 0$, is standard Brownian motion. Hence, Brownian bridge is a Brownian motion conditioned on hitting 0 at $t = 1$.

An exact expression for the complementary distribution of the maximum of Brownian bridge is known (Doob [31]),

$$\mathsf{P}(\sup_{0\le t\le 1} B_0(t) > b) = e^{-2b^2}.$$

From the above convergence and equating the last limit distribution with ε, we obtain the stated result. ■

II. PROOF OF PROPOSITION 7

Proof: Note that (3) is equivalent to

$$\{\max_{n\ge 0, m>0}[N_{\overline{ij}}[T_n, T_{n+m}) - (1-\rho_{ij})m] \le \rho_{ij}E_3^{ij}\}.$$

From the periodicity of $N_{\overline{ij}}$, it follows

$$\begin{aligned} Y &:= \max_{n\ge 0, m>0}[N_{\overline{ij}}[T_n, T_{n+m}) - (1-\rho_{ij})m] \\ &= \max_{1\le k\le m\le 2L}[N_{\overline{ij}}[T_{k-1}, T_{m-1}) - (1-\rho_{ij})(m-k)] \\ &= \max_{1\le k\le 2L} X_k - \min_{1\le k\le 2L} X_k, \end{aligned}$$

where $X_k := N_{\overline{ij}}[T_0, T_{k-1}) - (1-\rho_{ij})k$, $k = 1, 2, \ldots$. Now, similarly as in the proof of Proposition 6, we conclude that, as $L \to \infty$, $Y \Rightarrow \sqrt{\rho_{ij}(1-\rho_{ij})L}W$, where $W =_d \sup_{0\le t\le 1} B_0(t) - \inf_{0\le t\le 1} B_0(t)$, the range of the Brownian bridge. It is known that the range of the Brownian bridge

is equal in distribution to the maximum Brownian excursion (see Vervaat [32] and [33]). The Brownian excursion can be represented in terms of standard Brownian motion B as $(Z(t))_{t\in[0,1]} =_d ((\tau_+ - \tau_-)^{-1/2}|B((1-t)\tau_+ + t\tau_-)|)_{t\in[0,1]}$, where τ_- is the last zero of B before 1 and τ_+ the first zero after 1. It is known that ([34] Theorem 5.2.10), for $z > 0$,

$$\mathsf{P}(\sup_{0\le t\le 1} Z(t) > z) = 2\sum_{\ell=1}^{\infty}(4\ell^2 z^2 - 1)e^{-2\ell^2 z^2}.$$

Now let E_3^{ij} be equal to the right-hand side in (14) for some $A > 0$. It follows from the above convergence in distribution that

$$\mathsf{P}(\max_{n\ge 0, m>0}[N_{\overline{ij}}[T_n, T_{n+m}) - (1-\rho_{ij})m] > \rho_{ij}E_3^{ij}) \to$$
$$\to 2\sum_{\ell=1}^{\infty}\left(4\ell^2 A - 1\right)e^{-2\ell^2 A}, \text{ as } L\to\infty.$$

Lastly, we equate the limit in the last display to ε, for a fixed $\varepsilon \ge 0$, which completes the proof. ■

III. Proof of Lemma 8

We now prove Lemma 8.

Proof: Suppose inductively that we have already chosen $y_1, \ldots, y_{v-1}$ such that,

$$\begin{array}{rl} & \sum_i \mathsf{E}[\prod_{j=1}^{n_2} f_{ij}(X_j)|Y_1 = y_1, \ldots, Y_{v-1} = y_{v-1}] \\ \le & \sum_i \mathsf{E}[\prod_{j=1}^{n_2} f_{ij}(X_j)]. \end{array}$$

This can trivially be done for $v = 1$. Then,

$$\begin{array}{rl} & \sum_y \sum_i \mathsf{P}(Y_v = y|Y_1 = y_1, \ldots, Y_{v-1} = y_{v-1}) \cdot \\ & \cdot\mathsf{E}[\prod_{j=1}^{n_2} f_{ij}(X_j)|Y_1 = y_1, \ldots, Y_{v-1} = y_{v-1}, Y_v = y] \\ = & \sum_i \mathsf{E}[\prod_{j=1}^{n_2} f_{ij}(X_j)|Y_1 = y_1, \ldots, Y_{v-1} = y_{v-1}] \le \\ \le & \sum_i \mathsf{E}[\prod_{j=1}^{n_2} f_{ij}(X_j)]. \end{array}$$

Since $\sum_y \mathsf{P}(Y_v = y|Y_1 = y_1, \ldots, Y_{v-1} = y_{v-1}) = 1$, there exists a fixed value y_v such that,

$$\begin{array}{rl} & \sum_i \mathsf{E}[\prod_{j=1}^{n_2} f_{ij}(X_j)|Y_1 = y_1, \ldots, Y_{v-1} = y_{v-1}, Y_v = y_v] \\ \le & \sum_i \mathsf{E}[\prod_{j=1}^{n_2} f_{ij}(X_j)]. \end{array}$$

Then, by Inequality (16),

$$\begin{array}{rl} & \mathsf{P}(\sigma_i|Y_1 = y_1, \ldots, Y_{v-1} = y_{v-1}, Y_v = y_v) \\ \le & \sum_i \mathsf{E}[\prod_{j=1}^{n_2} f_{ij}(X_j)|Y_1 = y_1, \ldots, Y_{v-1} = y_{v-1}, Y_v = y_v] \\ \le & \sum_i \mathsf{E}[\prod_{j=1}^{n_2} f_{ij}(X_j)]. \end{array}$$

The proof follows by induction. ■

To find y_v we minimize,

$$\sum_i \mathsf{E}[\prod_{j=1}^{n_2} f_{ij}(X_j)|Y_1 = y_1, \ldots, Y_{v-1} = y_{v-1}, Y_v = y] =$$
$$= \sum_i \prod_{j=1}^{n_2} \mathsf{E}[f_{ij}(X_j)|Y_1 = y_1, \ldots, Y_{v-1} = y_{v-1}, Y_v = y],$$

over all possible values of y. We can exchange the expectation with the product due to the independence of the X_j. Hence we only need to be able to calculate the $\mathsf{E}[f_{ij}(X_j)|Y_1 = y_1, \ldots, Y_{v-1} = y_{v-1}, Y_v = y]$ in isolation. This is feasible in all our applications of Lemma 8.

IV. Proof of Lemma 11

Proof: Let us first recall definition of Poisson process. A counting process N on $\mathbb{R}$ is Poisson with intensity λ if for any two disjoint intervals $\mathcal{I}$ and $\mathcal{J}$ on $\mathbb{R}$, $N\mathcal{I}$ and $N\mathcal{J}$ are independent, and in addition, for any $\mathcal{I}$ on $\mathbb{R}$,

$$\mathsf{P}(N\mathcal{I} = m) = \frac{(\lambda|\mathcal{I}|)^m}{m!}e^{-\lambda|\mathcal{I}|},\ m = 0, 1, \ldots.$$

Now, it is an elementary result that, if N_k $k = 1, 2, \ldots, K$ are Poisson counting processes with respective finite intensities φ_k $k = 1, 2, \ldots, K$, then $N_{ij}\mathcal{I} = \sum_{k\in S_{ij}} N_k\mathcal{I}$, for any ij and $\mathcal{I} \in \mathbb{R}$, is Poisson with intensity $\rho_{ij} = \sum_{k\in S_{ij}} \varphi_k$.

Thus, we can write

$$\begin{array}{rl} & \mathsf{P}(N_{ij}[T_n, T_{n+m}) = l) \\ = & \mathsf{P}(N_{ij}\{T_n\} = 0, N_{ij}(T_n, T_{n+m}) = l)+ \\ & +\mathsf{P}(N_{ij}\{T_n\} = 1, N_{ij}(T_n, T_{n+m}) = l-1) \qquad (18) \\ = & (1-\rho_{ij})\mathsf{P}(N_{ij}(T_n, T_{n+m}) = l)+ \\ & +\rho_{ij}\mathsf{P}(N_{ij}(T_n, T_{n+m}) = l-1). \end{array}$$

We exercise a simple calculus, for any $l = 0, 1, \ldots, m-1$,

$$\begin{array}{rl} & \mathsf{P}(N_{ij}(T_n, T_{n+m}) = l) \\ = & \int_0^\infty \mathsf{P}(N_{ij}(T_n, T_n + t) = l, N(T_n, T_n + t) = m-1)dt \\ = & \int_0^\infty \mathsf{P}(N_{ij}(T_n, T_n + t) = l, N_{\overline{ij}}(T_n, T_n + t) = m-1-l)dt \\ = & \int_0^\infty \mathsf{P}(N_{ij}(0, t) = l, N_{\overline{ij}}(0, t) = m-1-l)dt \\ = & \int_0^\infty \mathsf{P}(N_{ij}(0, t) = l)\mathsf{P}(N_{\overline{ij}}(0, t) = m-1-l)dt \\ = & \frac{\rho_{ij}^k(1-\rho_{ij})^{m-1-l}}{k!(m-1-l)!}\int_0^\infty t^{m-1}e^{-t}dt \\ = & \frac{(m-1)!}{(l-1)!(m-l)!}\rho_{ij}^l(1-\rho_{ij})^{m-1-l}. \end{array}$$

The second equality is obtained by $N\mathcal{I} = N_{ij}\mathcal{I} + N_{\overline{ij}}\mathcal{I}$, any $\mathcal{I} \subset \mathbb{R}$; the third equality is by independence and stationarity of the increments of N_{ij} and $N_{\overline{ij}}$; the forth equality follows by the independence of N_{ij} and $N_{\overline{ij}}$; in the fifth equality we utilize the fact that for any fixed $\mathcal{I} \in \mathbb{R}$, $N_{ij}\mathcal{I}$ and $N_{\overline{ij}}\mathcal{I}$ are Poisson random variables; and finally, the last equality follows from $\int_0^\infty t^m e^{-t}dt = m!$, for m an integer.

The statement of the lemma follows by plugging the resulting identity in the last above display into (18). ■