Delay Analysis for Maximal Scheduling with Flow Control in Wireless Networks with Bursty Traffic

Michael J. Neely

***Abstract*—We consider the delay properties of one-hop networks with general interference constraints and multiple traffic streams with time-correlated arrivals. We first treat the case when arrivals are modulated by independent finite state Markov chains. We show that the well known maximal scheduling algorithm achieves average delay that grows at most logarithmically in the largest number of interferers at any link. Further, in the important special case when each Markov process has at most two states (such as bursty ON/OFF sources), we prove that average delay is independent of the number of nodes and links in the network, and hence is *order-optimal*. We provide tight delay bounds in terms of the individual auto-correlation parameters of the traffic sources. These are perhaps the first order-optimal delay results for controlled queueing networks that explicitly account for such statistical information. Our analysis treats cases both with and without flow control.**

***Index Terms*—Queueing analysis, Markov chains, Flow Control**

I. Introduction

This paper derives average delay bounds for one-hop wireless networks that use maximal scheduling subject to a general set of interference constraints. It is known that maximal scheduling algorithms are simple to implement and can support throughput within a constant factor of optimality. Our analysis shows that this type of scheduling also yields tight delay guarantees. In particular, when arrival processes are modulated by independent Markov processes, we show that average delay grows at most logarithmically in the number of nodes in the network. We then obtain an improved delay bound in the important special case when the individual Markov chains have at most two states (such as bursty ON/OFF sources). Average delay in this case is shown to be independent of the network size, and hence is *order-optimal*.

Specifially, we consider a network with N nodes and L links. Let $\mathcal{N}$ and $\mathcal{L}$ denote the node and link sets:

$$\mathcal{N}\triangleq\{1,2,\ldots,N\} \quad , \quad \mathcal{L}\triangleq\{1,2,\ldots,L\}$$

Each link $l\in\mathcal{L}$ represents a directed communication channel from one node to another, and we define $tran(l)$ and $rec(l)$ to be the corresponding transmitter and receiver nodes for link l (where $tran(l)\in\mathcal{N}$ and $rec(l)\in\mathcal{N}$). The network operates in slotted time with unit timeslots $t\in\{0,1,2,\ldots\}$. Every timeslot a decision is made about which links to activate for transmission. If a link l is activated during a particular slot, it sends exactly one packet from $tran(l)$ to $rec(l)$. However, due to scheduling and/or interference constraints, not all links can be simultaneously active during the same slot. These constraints are defined according to the general interference model of [2][3]: Each link $l\in\mathcal{L}$ is allowed to be active if and only if no other links within an *interference set* $\mathcal{S}_l$ are simultaneously active. For convenience, it is useful to define the set $\mathcal{S}_l$ to additionally include link l itself. That is, each set $\mathcal{S}_l$ consists of link l together with all possible interferers of link l. Note that these interference sets have the following pairwise symmetry property: For any two links $\omega, l\in\mathcal{L}$, we have that $\omega\in\mathcal{S}_l$ if and only if $l\in\mathcal{S}_\omega$.

The link sets $\mathcal{S}_l$ can be chosen to impose a variety of constraint models. For example, setting $\mathcal{S}_l$ to include all links adjacent to either the transmitter or receiver of link l imposes *matching constraints*. Matching constraint models arise naturally in scheduling problems for packet switches. They are also important for wireless networks where individual nodes can transmit or receive over at most one adjacent link, and cannot simultaneously transmit and receive (called the *node exclusive interference* model in [4]). More general sets $\mathcal{S}_l$ can be used to model topology-dependent interference constraints for wireless networks, such as the constraint that no additional transmitting nodes can be activated within a specified distance of a node that is actively receiving.

Every timeslot, new data randomly arrives to the network. Let $A_l(t)$ represent the (integer) number of packets that arrive during slot t that are intended for transmission over link l. Packets are stored in separate queues according to their corresponding link. This is a one-hop network, so that packets exit the network once they are transmitted over their intended link. A network controller observes the current queue backlog and makes link activation decisions every slot subject to the transmission constraints.

It is well known that generalized max-weight scheduling can be used to achieve maximum throughput in such networks [5] [6]. However, this type of scheduling is difficult to implement in wireless networks with general interference constraints. In this paper we consider a simpler class of *maximal scheduling algorithms*. Maximal scheduling is of recent interest due to its low complexity and ease of distributed implementation. For $N\times N$ packet switches, maximal scheduling is known to support throughput that is within a factor of 2 of optimality, and to also have nice delay properties for i.i.d. inputs [7] [8]. Related constant factor throughput results have also been shown for wireless networks, including factor of 2 results for networks under matching constraints [4] and constant factor results for more general interference models [2] [3]. However, the work on wireless scheduling in [4][2][3] considers only throughput results and does not provide a delay analysis.

This work was presented in part at the IEEE INFOCOM conference, Phoenix, AZ, April 2008.

Michael J. Neely is with the Electrical Engineering department at the University of Southern California, Los Angles, CA (web: http://www-rcf.usc.edu/∼mjneely).

This material is supported in part by one or more of the following: the DARPA IT-MANET program grant W911NF-07-0028, the NSF grant OCE 0520324, the NSF Career grant CCF-0747525.

Further, while the work in [8] considers delay analysis for i.i.d. arrivals and a $N \times N$ packet switch, no existing work provides explicitly computable and order-optimal delay bounds for time-correlated arrivals.

Our work addresses the issues of general interference constraints and time-correlated "bursty" traffic simultaneously. We treat the general interference model of [2][3], but use the concept of *queue grouping* to derive the order-optimal delay results. Queue grouping techniques have been used in [9][10][7][3] to reduce scheduling complexity in switches and wireless networks. They are used in [11] to provide order-optimal delay for opportunistic scheduling in a single-server downlink. Near order-optimal delay is established for $N \times N$ packet switches in [12]. The analysis in this paper particularly treats delay in wireless networks with general constraint sets and time-correlated arrivals. Previous work in [13][14][6] treats time-correlated arrivals, but does not obtain order optimal delay. Asymptotic delay optimality is studied in [15] for a heavy traffic regime where input rates are scaled so that they are very close to the capacity region boundary. Here, we focus on the case when input rates are a fixed fraction away from the boundary. We obtain tight delay bounds in terms of the individual auto-correlation parameters of the traffic sources. These are perhaps the first delay bounds for controlled queueing networks that explicitly incorporate such statistical information. This allows delay to be understood in terms of general models for network traffic. Our analysis includes the important special case of Markov modulated ON/OFF traffic sources, allowing for explicit delay bounds in terms of the parameters of each source.

We first treat the case of general Markovian arrivals and prove a *structural result* about average delay, showing that average delay grows at most logarithmically in the worst case number of interferers of a given link, and hence is at most $O(\log(N))$. We then provide an explicit and tighter delay analysis for the special case when all Markov chains have at most two states. In this case, we prove average delay is *independent of* N. Our analysis first assumes arrival rates are inside the stability region associated with maximal scheduling, and is next generalized to treat flow control when traffic rates are either inside or outside of this region. The time-correlated scenarios treated here are quite challenging to analyze, and we introduce a simple technique of *delayed Lyapunov analysis* to ensure the arrival processes couple sufficiently fast to a stationary distribution. We note that this delayed Lyapunov analysis is different from the related T-slot drift technique developed in [13] for queue stability with Markovian channels (also used in [14][6] for both stability and delay analysis), as the T-slot drift approach cannot achieve order-optimal delay results.

In the next section we present the network model. In Section III we provide the drift analysis, and in Section IV we present the logarithmic delay result for general time-correlated arrivals. The order-optimal delay analysis for 2-state chains is provided in Section V. Flow control mechanisms for systems with arbitrary traffic rates are considered in Section VI.

II. Network Model

Recall that $\mathcal{L}$ denotes the set of network links, and that transmission over each link is constrained by the general interference sets defined in the previous section.

A. Traffic Model

Suppose the arrival process $A_l(t)$ is modulated by a discrete time, stationary, ergodic Markov chain $Z_l(t)$ for each link $l \in \mathcal{L}$. Specifically, $Z_l(t)$ has finite state space $\mathcal{Z}_l = \{1, 2, \ldots, M_l\}$. For each link $l \in \mathcal{L}$ and state $m \in \mathcal{Z}_l$, arrivals $A_l(t)$ are conditionally independent and identically distributed according to mass function $p_l^{(m)}(a)$, where:

$$p_l^{(m)}(a) = Pr[A_l(t) = a \mid Z_l(t) = m] \quad \text{for } a \in \{0, 1, 2, \ldots\}$$

Define the conditional arrival rates $\lambda_l^{(m)}$ as follows:

$$\lambda_l^{(m)} = \mathbb{E}\{A_l(t) \mid Z_l(t) = m\}$$

We assume conditional second moments of arrivals are finite, so that $\mathbb{E}\left\{(A_l(t))^2 \mid Z_l(t) = m\right\} < \infty$ for all $l \in \mathcal{L}$ and all $m \in \mathcal{Z}_l$. Let $\pi_l^{(m)}$ represent the steady state probability that $Z_l(t) = m$. Define λ_l as the average arrival rate to link l:

$$\lambda_l = \sum_{m \in \mathcal{Z}_l} \pi_l^{(m)} \lambda_l^{(m)} \tag{1}$$

We assume that all Markov chains are in their steady state distribution at time 0, so that each $A_l(t)$ process is stationary and for all slots $t \geq 0$ and all links $l \in \mathcal{L}$ we have:

$$\mathbb{E}\{A_l(t)\} = \lambda_l$$

The Markov chains $Z_l(t)$ themselves may be correlated over different links $l \in \mathcal{L}$, although we mainly focus on the case when chains are independent. More detailed statistical information, such as the auto-correlation for individual inputs (and the spatial correlation between multiple inputs if they are not independent), is also important for delay analysis and shall be defined when needed. Note that this traffic model is quite general and includes the following important special cases:

- Case 1: $Z_l(t)$ has only one state and so arrivals $A_l(t)$ are i.i.d. over slots with some given distribution.
- Case 2: $Z_l(t)$ is a 2-state ON/OFF process where $A_l(t) = 1$ whenever $Z_l(t) = ON$ and $A_l(t) = 0$ whenever $Z_l(t) = OFF$.

Let $\sigma_l^2 \triangleq \mathbb{E}\left\{A_l(t)^2\right\} - \lambda_l^2$ represent the steady state arrival variance for link l. While our results hold for general traffic with finite variance, it is important to note that the order-optimal delay analysis we achieve requires the following assumption:

$$\frac{1}{\lambda_{tot}} \sum_{l \in \mathcal{L}} \sigma_l^2 = O(1) \tag{2}$$

where $\lambda_{tot} \triangleq \sum_{l \in \mathcal{L}} \lambda_l$. This is a mild assumption that typically holds for any inputs with a finite variance. For example, if arrivals are Poisson then we have $\sigma_l^2 = \lambda_l$, and so $\frac{1}{\lambda_{tot}} \sum_{l \in \mathcal{L}} \sigma_l^2 = 1$. Similarly, it is easy to show that if there is a finite constant A_{max} such that $A_l(t) \leq A_{max}$ for all l and all t, then $\frac{1}{\lambda_{tot}} \sum_{l \in \mathcal{L}} \sigma_l^2 \leq A_{max}$.

B. Queueing

Define $Q_l(t)$ as the number of queued packets waiting for transmission over link l during slot t. Let $\boldsymbol{Q}(t) = (Q_l(t))_{l\in\mathcal{L}}$ be the vector of queue backlogs. Define $\mu_l(t) \in \{0, 1\}$ as the *transmission rate* offered to the link during slot t (in units of packets/slot). That is, $\mu_l(t) = 1$ if link l is scheduled for transmission on slot t, and $\mu_l(t) = 0$ otherwise. We assume the scheduler only schedules a link l that does not violate the interference constraints and that has a packet ready for transmission (so that $Q_l(t) > 0$). Let $\boldsymbol{\mu}(t) = (\mu_l(t))_{l\in\mathcal{L}}$ represent the transmission rate vector for slot t. Define $\mathcal{X}(t)$ as the set of *feasible transmission vectors* for slot t, representing all $\boldsymbol{\mu}(t)$ rate vectors that conform to the constraints defined by the interference sets $\mathcal{S}_l$ and the additional constraint that $\mu_l(t) = 1$ only if $Q_l(t) > 0$ (for each $l \in \mathcal{L}$). The queueing dynamics thus proceed as follows:

$$Q_l(t + 1) = Q_l(t) - \mu_l(t) + A_l(t) \tag{3}$$

The goal is to observe the queue backlogs every slot and make scheduling decisions $\boldsymbol{\mu}(t) \in \mathcal{X}(t)$ so as to support all incoming traffic with average delay as small as possible.

C. Maximal Scheduling

Define the *network capacity region* Λ as the closure of the set of all arrival rate vectors $(\lambda_l)_{l\in\mathcal{L}}$ that can be stably supported, considering all possible scheduling algorithms that conform to the above constraints (see [6] for a discussion of capacity regions and stability). It is well known that scheduling according to a generalized *max-weight* rule every timeslot ensures stability and maximum throughput whenever arrival rates are interior to the capacity region [5] [6].[1] However, the max-weight rule involves an integer optimization that may be difficult to implement, and has delay properties that are difficult to analyze. Here, we assume scheduling is done according to a simpler *maximal scheduling* algorithm. Specifically, given a queue backlog vector $\boldsymbol{Q}(t)$, a transmission vector $\boldsymbol{\mu}(t)$ is *maximal* if it satisfies the interference constraints and is such that for all links $l \in \mathcal{L}$, if $Q_l(t) > 0$ then $\mu_\omega(t) = 1$ for at least one link $\omega \in \mathcal{S}_l$. In words, this means that if link l has a packet, then either link l is selected for transmission, or some other link within the interference set $\mathcal{S}_l$ is selected. There is much recent interest in maximal scheduling because of its implementation simplicity (described briefly below) and its ability to support input rates within a constant factor of the capacity region for wireless networks [4] [2] [3] and for $N \times N$ packet switches [7].

One way to achieve a maximal scheduling is as follows: First select any non-empty link $l \in \mathcal{L}$ and label it "active." Then select any other non-empty link that does not conflict with the active link l (i.e., that is not within $\mathcal{S}_l$). Label this second link "active." Continue in the same way, selecting new non-empty links that do not conflict with any previously selected links, until no more links can be added. It is not difficult to see that this final set of links labeled "active" has the desired maximal property. Maximal link selections are not unique, and can alternatively be found in a distributed manner, where multiple nodes attempt to activate their non-conflicting, non-empty links simultaneously, and contentions are resolved locally. This distributed implementation also requires multiple iterations before the set of selected links becomes maximal.

All maximal link selections have the following important mathematical property.

Lemma 1: Under any maximal link scheduling for $\boldsymbol{\mu}(t)$, for all links $l \in \mathcal{L}$ we have:

$$Q_l(t) \sum_{\omega\in\mathcal{S}_l} \mu_\omega(t) \geq Q_l(t) \tag{4}$$

Proof: Consider any link $l \in \mathcal{L}$. If $Q_l(t) = 0$, then (4) reduces to $0 \geq 0$ and is trivially true. Else, if $Q_l(t) > 0$ then $\mu_\omega(t) = 1$ for at least one link ω within $\mathcal{S}_l$ (by definition of a maximal link selection), and so $\sum_{\omega\in\mathcal{S}_l} \mu_\omega(t) \geq 1$, which proves (4). ■

In this paper, we assume transmission decisions are made every slot according to any maximal scheduling. For convenience, we further assume that the maximal scheduling has a well defined probabilistic structure given the queue backlog vector, so that the entire queueing system can be viewed as an ergodic Markov chain with a countably infinite state space. The inequality (4) is the only additional property of maximal scheduling required in our analysis.

D. The Reduced Throughput Region

Define Λ^* as the set of all rate vectors $(\lambda_l)_{l\in\mathcal{L}}$ that satisfy the following:

$$\sum_{\omega\in\mathcal{S}_l} \lambda_\omega \leq 1 \quad \text{for all } l \in \mathcal{L}$$

The set Λ^* is overly restrictive, as it is possible to have more than one simultaneously active link within a given set $\mathcal{S}_l$ (provided that link l is idle). However, the set Λ^* is typically within a constant factor of the capacity region Λ. For example, in networks with matching constraints only, it is not difficult to show that $\frac{1}{2}\Lambda \subset \Lambda^*$, so that the throughput region Λ^* is within a factor of 2 of optimality. Further, in networks with general inteference sets $\mathcal{S}_l$ where each set $\mathcal{S}_l$ can support at most K simultaneously active links (called the *interference degree* of the network), it can be shown that $\frac{1}{K}\Lambda \subset \Lambda^*$ [2]. It is easy to see in this case that the integer K is strictly less than the largest cardinality of any interference set. The work in [2] also constructs a network for which the set $\frac{1}{K}\Lambda$ shares a common boundary point with the set of all rates supportable through maximal scheduling.

Throughout this paper, we assume input rates $(\lambda_l)_{l\in\mathcal{L}}$ are interior to the set Λ^*. Specifically, we assume there exists a value ρ^* such that $0 < \rho^* < 1$, where:

$$\sum_{\omega\in\mathcal{S}_l} \lambda_\omega \leq \rho^* \quad \text{for all } l \in \mathcal{L} \tag{5}$$

The value ρ^* represents the *relative network loading*, as it can be viewed as a loading factor relative to the reduced throughput region Λ^*.

[1] Specifically, the generalized max-weight rule in this case schedules to maximize $\sum_{l\in\mathcal{L}} Q_l(t)\mu_l(t)$ subject to $\boldsymbol{\mu}(t) \in \mathcal{X}(t)$.

E. An Example for $N \times N$ Packet Switches

For an illuminating example of the reduced throughput region Λ^* and its comparison to the capacity region Λ, consider a $N \times N$ packet switch operating under the crossbar constraint with input rate matrix (λ_{ij}) for input ports $i \in \{1, \ldots, N\}$ and output ports $j \in \{1, \ldots, N\}$ (for a more detailed discussion of packet switches and the crossbar constraint, see, for example, [16] [17] [12]). There are N^2 links, each labeled (i, j) according to its corresponding input/output pair. Link activation sets correspond to *matchings* on the resulting bi-partite graph. Specifically, if link (i, j) is activated for transmission, then no other link (i, b) or (a, j) can be activated (for $b \neq j$, $a \neq i$). Therefore, the interference set S_{ij} for each link (i, j) is:

$$S_{ij} = \{(a, j) \mid a \in \{1, \ldots, N\}\} \cup \{(i, b) \mid b \in \{1, \ldots, N\}\}$$

It is well known that the capacity region Λ for this $N \times N$ switch is given by the set of all (λ_{ij}) rate matrices that satsify the following $2N$ inequality constraints:

$$\begin{array}{l} \sum_{b=1}^{N} \lambda_{ib} \leq 1 \quad \text{for all } i \in \{1, \ldots, N\} \\ \sum_{a=1}^{N} \lambda_{aj} \leq 1 \quad \text{for all } j \in \{1, \ldots, N\} \end{array}$$

On the other hand, the reduced throughput region Λ^* is given by the set of all non-negative rate matrices (λ_{ij}) that satisfy the following N^2 inequality constraints:

$$\sum_{a=1}^{N} \lambda_{aj} + \sum_{b \in \{1, \ldots, N\}, b \neq j} \lambda_{ib} \leq 1 \quad \text{for all links } (i, j)$$

That is, $(\lambda_{ij}) \in \Lambda^*$ if and only if all rates are non-negative, and for any crosspoint (i, j), the sum of the rates in row i and column j (including crosspoint (i, j) only once) is less than or equal to 1. It is not difficult to see that if a matrix in Λ has all entries halved, then it will be in Λ^*, and hence:

$$\frac{1}{2}\Lambda \subset \Lambda^*$$

However, the reduced throughput region Λ^* is strictly larger than $\frac{1}{2}\Lambda$. For example, consider the following rate matrices for a 2×2 switch and a 3×3 switch, respectively:

$$(\lambda_{ij}) = \left(\begin{array}{cc} .7 & .1 \\ .1 & .1 \end{array} \right) \ , \ (\lambda_{ij}) = \left(\begin{array}{ccc} .1 & .2 & 0 \\ .3 & .2 & 0 \\ .3 & 0 & .2 \end{array} \right)$$

The first matrix is outside the set $\frac{1}{2}\Lambda$ because the first column sums to 0.8 (larger than 0.5). The second matrix is also outside of the set $\frac{1}{2}\Lambda$ (for the corresponding 3×3 capacity region Λ) because the first column sums to 0.7. However, both matrices are inside their respective reduced throughput regions Λ^*, and both have relative loading $\rho^* = 0.9$.

III. Drift Analysis

Recall that $\boldsymbol{Q}(t) = (Q_l(t))$. Our technique relies on the concept of *queue grouping*. For each link $l \in \mathcal{L}$, define:

$$\hat{Q}_{\mathcal{S}_l}(t) \triangleq \sum_{\omega \in \mathcal{S}_l} Q_\omega(t) \tag{6}$$

Thus, $\hat{Q}_{\mathcal{S}_l}(t)$ is the sum of all queue backlogs of links within the interference set $\mathcal{S}_l$ of link l. Define the Lyapunov function:

$$L(\boldsymbol{Q}(t)) \triangleq \frac{1}{2} \sum_{l \in \mathcal{L}} Q_l(t) \hat{Q}_{\mathcal{S}_l}(t) \tag{7}$$

The queue-grouped structure of this Lyapunov function is similar to the functions used in [3][9] to prove queue stability when input rates are a fixed fraction away from the capacity boundary. An alternate proof of rate-stability is given in [2]. However, the prior work in this area does not directly consider delay performance. Below we provide a more detailed drift analysis that yields explicit and tight delay bounds.

For each link $l \in \mathcal{L}$, define the *group departures* $\hat{\mu}_{\mathcal{S}_l}(t)$ and *group arrivals* $\hat{A}_{\mathcal{S}_l}(t)$ as follows:

$$\hat{\mu}_{\mathcal{S}_l}(t) \triangleq \sum_{\omega \in \mathcal{S}_l} \mu_\omega(t) \quad , \quad \hat{A}_{\mathcal{S}_l}(t) \triangleq \sum_{\omega \in \mathcal{S}_l} A_\omega(t)$$

Thus:

$$\hat{Q}_{\mathcal{S}_l}(t+1) = \hat{Q}_{\mathcal{S}_l}(t) - \hat{\mu}_{\mathcal{S}_l}(t) + \hat{A}_{\mathcal{S}_l}(t) \tag{8}$$

Define the 1-step *unconditional Lyapunov drift* as follows:

$$\Delta(t) \triangleq \mathbb{E}\left\{L(\boldsymbol{Q}(t+1)) - L(\boldsymbol{Q}(t))\right\} \tag{9}$$

where the expectation is over the randomness of $\boldsymbol{Q}(t)$ and the randomness of the system dynamics given the value of $\boldsymbol{Q}(t)$.

Lemma 2: (Drift Under Maximal Scheduling) If maximal scheduling is implemented every timeslot (using any maximal scheduling algorithm), the resulting unconditional Lyapunov drift $\Delta(t)$ satisfies the following for all slots t:

$$\Delta(t) \leq \mathbb{E}\left\{B(t)\right\} - \sum_{l \in \mathcal{L}} \mathbb{E}\left\{Q_l(t)(1 - \hat{A}_{\mathcal{S}_l}(t))\right\} \tag{10}$$

where

$$B(t) \triangleq \frac{1}{2} \sum_{l \in \mathcal{L}} \left[A_l(t)\hat{A}_{\mathcal{S}_l}(t) - 2\hat{A}_{\mathcal{S}_l}(t)\mu_l(t) + \mu_l(t)\right] \tag{11}$$

Proof: See Appendix A. ■

A. Lyapunov Drift Theorem

The drift expression (10) can be used to prove stability and delay properties of maximal matching via the following theorem:

Theorem 1: (Lyapunov drift [6]) Let $\boldsymbol{Q}(t)$ be a vector process of queue backlogs that evolve according to some probability law, and let $L(\boldsymbol{Q}(t))$ be a non-negative function of $\boldsymbol{Q}(t)$. If there exist processes $f(t)$ and $g(t)$ such that the following is satisfied for all time t:

$$\Delta(t) \leq \mathbb{E}\left\{g(t)\right\} - \mathbb{E}\left\{f(t)\right\}$$

then:

$$\limsup_{t \to \infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\left\{f(\tau)\right\} \leq \limsup_{t \to \infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\left\{g(t)\right\} \quad \square$$

The proof of Theorem 1 involves summing the telescoping series (see [6]).

B. Analysis for i.i.d. Arrivals

Define $\boldsymbol{A}(t)\triangleq(A_l(t))_{l\in\mathcal{L}}$ as the vector of new arrivals. Consider first the case when all arrival vectors $\boldsymbol{A}(t)$ are i.i.d. over timeslots, with rate vector $\boldsymbol{\lambda}=(\lambda_l)_{l\in\mathcal{L}}$ (the arrivals over different links in the same slot may be correlated). For each link l, define $\hat{\lambda}_{\mathcal{S}_l}$ as the sum of arrival rates over all input streams corresponding to links within the set $\mathcal{S}_l$. That is:

$$\hat{\lambda}_{\mathcal{S}_l}\triangleq\sum_{\omega\in\mathcal{S}_l}\lambda_\omega$$

Note by the loading assumption (5) that $\hat{\lambda}_{\mathcal{S}_l}\leq\rho^*$, where ρ^* is a value such that $0<\rho^*<1$. By independence of arrivals every slot, we have for all t:

$$\begin{aligned}\mathbb{E}\left\{Q_l(t)(1-\hat{A}_{\mathcal{S}_l}(t))\right\} &= \mathbb{E}\left\{Q_l(t)\right\}(1-\hat{\lambda}_{\mathcal{S}_l})\\ &\geq \mathbb{E}\left\{Q_l(t)\right\}(1-\rho^*)\end{aligned}$$

Using this inequality directly in the Lyapunov drift expression (10) yields (for all slots t):

$$\Delta(t)\leq\mathbb{E}\left\{B(t)\right\}-(1-\rho^*)\sum_{l\in\mathcal{L}}\mathbb{E}\left\{Q_l(t)\right\}$$

Plugging the above drift inequality into the Lyapunov drift theorem (Theorem 1) (using $g(t)\triangleq B(t)$ and $f(t)\triangleq(1-\rho^*)\sum_{l\in\mathcal{L}}Q_l(t)$) yields:

$$\limsup_{t\rightarrow\infty}\frac{1}{t}\sum_{\tau=0}^{t-1}\sum_{l\in\mathcal{L}}\mathbb{E}\left\{Q_l(\tau)\right\}\leq\frac{\overline{B}}{1-\rho^*}\tag{12}$$

where:

$$\overline{B}\triangleq\limsup_{t\rightarrow\infty}\frac{1}{t}\sum_{\tau=0}^{t-1}\mathbb{E}\left\{B(\tau)\right\}$$

Note from (11) that $\overline{B}<\infty$, and hence the queueing network is strongly stable with finite time average queue backlogs. Because it evolves according to an ergodic Markov chain with a countably infinite state space, it can be shown that limiting time averages exist and are equal to the steady state averages. Thus, the left hand side of (12) represents the time average total queue backlog in the system (summed over all queues). Let $\overline{Q}_l$ be the average queue backlog in link l (for each $l\in\mathcal{L}$). We thus have:

Theorem 2: (Time-Independent Arrivals) If the arrival vector process $\boldsymbol{A}(t)$ is i.i.d. over slots with a relative network loading $\rho^*<1$ (defined in (5)), then:

(a) The average total network congestion satisfies:

$$\sum_{l\in\mathcal{L}}\overline{Q}_l\leq\frac{\sum_{l\in\mathcal{L}}\left[\mathbb{E}\left\{A_l(t)\hat{A}_{\mathcal{S}_l}(t)\right\}-2\lambda_l\hat{\lambda}_{\mathcal{S}_l}+\lambda_l\right]}{2(1-\rho^*)}$$

(b) If arrival streams $A_l(t)$ are i.i.d. over slots and are also independent of each other, then $\overline{W}$, the expected delay averaged over all packets in the network, satisfies:

$$\overline{W}\leq\frac{1+\frac{1}{\lambda_{tot}}\sum_{l\in\mathcal{L}}\left[\sigma_l^2-\lambda_l\hat{\lambda}_{\mathcal{S}_l}\right]}{2(1-\rho^*)}\tag{13}$$

where $\lambda_{tot}\triangleq\sum_{l\in\mathcal{L}}\lambda_l$, and where $\sigma_l^2\triangleq\mathbb{E}\left\{(A_l(t))^2\right\}-\lambda_l^2$ and represents the variance of $A_l(t)$.

The average congestion bound in part (a) of the above theorem is found by computing $\overline{B}$ using (11) and noting that the system is stable, has a steady state, and that the time average limit of $\mathbb{E}\left\{\mu_l(t)\right\}$ is equal to λ_l. Part (b) is proven by the fact that total average congestion is equal to $\lambda_{tot}\overline{W}$ (by Little's Theorem). This also uses inter-link independence for the identity $\mathbb{E}\left\{A_l(t)\hat{A}_{\mathcal{S}_l}(t)\right\}=\sigma_l^2+\lambda_l\hat{\lambda}_{\mathcal{S}_l}$. Note that the numerator in (13) is $O(1)$ if (2) holds.

C. Delay under Poisson and Bernoulli Inputs

Note that if all arrival processes $A_l(t)$ are independent and Poisson with rate λ_l, we have that $\sigma_l^2=\lambda_l$. The average delay bound (13) in this case reduces to:

$$\overline{W}_{Poisson}\leq\frac{1-\frac{1}{2\lambda_{tot}}\sum_{l\in\mathcal{L}}\lambda_l\hat{\lambda}_{\mathcal{S}_l}}{(1-\rho^*)}\tag{14}$$

This demonstrates that average delay is $O(1)$, that is, it is *independent of the network size* N. Hence, maximal scheduling achieves *order optimal delay* with respect to N, provided that the arrival rates are interior to the reduced throughput region Λ^*, as described by the constraints (5). This is in contrast to the $O(N)$ average delay bounds derived for the throughput-optimal max-weight scheduling for $N\times N$ packet switches in [17] and for wireless networks in [14] [18].[2] The expression in the right hand side of (14) also provides an upper bound on delay in the case of independent Bernoulli arrivals, because σ_l^2 for a Bernoulli variable is less than that of a Poisson variable.

IV. LOGARITHMIC DELAY FOR TIME-CORRELATED ARRIVALS

Consider the case of finite-state ergodic Markov modulated arrivals, as described in Section II-A. Assume that all traffic rates satisfy the loading constraints (5) with relative network loading ρ^*. Let $\boldsymbol{H}(t)$ represent the past history of all actual arrivals (of all processes) up to but not including time t. For a given link $l\in\mathcal{L}$, suppose there exists a non-negative function $\epsilon_l(T)$ (for $T\in\{0,1,2,\ldots\}$) such that, regardless of past history $\boldsymbol{H}(t)$, we have:

$$\mathbb{E}\left\{A_l(t)\,|\,\boldsymbol{H}(t-T)\right\}\leq\lambda_l+\epsilon_l(T)\tag{15}$$

and such that:

$$\lim_{T\rightarrow\infty}\epsilon_l(T)=0$$

That is, $\epsilon_l(T)$ characterizes the time required for the process $A_l(t)$ to converge to its stationary mean, regardless of the initial condition. It can be shown that all finite state ergodic Markov processes converge exponentially fast to their steady state (see, for example, [19]). Hence for each $l\in\mathcal{L}$ we can define $\epsilon_l(T)$ as follows:

$$\epsilon_l(T)=\nu_l\gamma_l^{T+1}\tag{16}$$

for some constant ν_l and some decay factor γ_l such that $0<\gamma_l<1$. The ν_l and γ_l constants can in principle be determined

[2] We emphasize that max-weight scheduling is a special case of maximal scheduling, and so our delay bounds also apply to max-weight in the case when arrival rates are inside the reduced throughput region Λ^*.

as parameters from the Markov chain $Z_l(t)$. Here, we prove a *structural result* concerning logarithmic delay in terms of these parameters. Define $\hat{\rho}(T)$ as follows:

$$\hat{\rho}(T) \triangleq \max_{l \in \mathcal{L}} \left[\hat{\lambda}_{\mathcal{S}_l} + \sum_{\omega \in \mathcal{S}_l} \epsilon_\omega(T) \right] \tag{17}$$

Note that $\hat{\lambda}_{\mathcal{S}_l} \leq \rho^* < 1$ for all $l \in \mathcal{L}$, and so there exist integers T such that $\hat{\rho}(T) < 1$.

Theorem 3: (General Time-Correlated Arrivals) If arrival processes have rates $(\lambda_l)_{l \in \mathcal{L}}$ in the interior of Λ^*, then for any integer $T \geq 0$ such that $\hat{\rho}(T) < 1$, we have:

(a) The average total network congestion satisfies:

$$\sum_{l \in \mathcal{L}} \overline{Q}_l \leq \frac{\tilde{B} + \tilde{F}_T}{1 - \hat{\rho}(T)}$$

where:

$$\tilde{B} \triangleq \frac{1}{2} \sum_{l \in \mathcal{L}} \left[\lambda_l + \mathbb{E}\left\{ A_l(t) \hat{A}_{\mathcal{S}_l}(t) \right\} \right] \tag{18}$$

$$\tilde{F}_T \triangleq \sum_{l \in \mathcal{L}} \sum_{k=1}^{T} \mathbb{E}\left\{ \hat{A}_{\mathcal{S}_l}(k) A_l(0) \right\} \tag{19}$$

(b) If arrival processes $A_l(t)$ for different links l are additionally independent of each other, then the constants $\tilde{B}$ and $\tilde{F}_T$ satisfy:

$$\tilde{B} = \frac{1}{2} \sum_{l \in \mathcal{L}} \left[\lambda_l + \lambda_l \hat{\lambda}_{\mathcal{S}_l} + \sigma_l^2 \right] \tag{20}$$

$$\tilde{F}_T = \sum_{l \in \mathcal{L}} \sum_{k=1}^{T} [\hat{\lambda}_{\mathcal{S}_l} \lambda_l + \theta_l(k)] \tag{21}$$

where $\sigma_l^2 \triangleq \mathbb{E}\left\{ (A_l(t))^2 \right\} - \lambda_l^2$ is the variance of $A_l(t)$, and $\theta_l(k)$ is the auto-correlation in $A_l(t)$ and is defined:

$$\theta_l(k) \triangleq \mathbb{E}\left\{ A_l(t+k) A_l(t) \right\} - \lambda_l^2$$

Proof: The proof is given in Appendix B. ■

A. Discussion of the Delay Result

By Little's Theorem, if the conditions of Theorem 3 are satisfied, then average network delay $\overline{W}$ satisfies:

$$\overline{W} \leq \frac{\frac{1}{\lambda_{tot}} (\tilde{B} + \tilde{F}_T)}{1 - \hat{\rho}(T)}$$

The parameter T only affects the delay bound and does not affect the maximal scheduling algorithm. Thus, the bound can be optimized over all integers T such that $\hat{\rho}(T) < 1$. Here we show how the resulting bound grows as a function of the network size. First note that in the case when arrival processes are independent of each other, the constant $\tilde{B}$ in (20) satisfies:

$$\tilde{B} \leq \lambda_{tot} \left[1 + \frac{1}{2\lambda_{tot}} \sum_{l \in \mathcal{L}} \sigma_l^2 \right]$$

where $\lambda_{tot} \triangleq \sum_{l \in \mathcal{L}} \lambda_l$. This is because $\hat{\lambda}_{\mathcal{S}_l} \leq 1$ for all $l \in \mathcal{L}$. Likewise, the constant $\tilde{F}_T$ in (21) satisfies:

$$\tilde{F}_T \leq \lambda_{tot} T + \sum_{l \in \mathcal{L}} \sum_{k=1}^{T} \theta_l(k)$$

Therefore, when arrival processes $A_l(t)$ are independent of each other, average network delay satisfies:

$$\overline{W} \leq \frac{T + 1 + \frac{1}{2\lambda_{tot}} \sum_{l \in \mathcal{L}} \sigma_l^2 + \frac{1}{\lambda_{tot}} \sum_{l \in \mathcal{L}} \sum_{k=1}^{T} \theta_l(k)}{1 - \hat{\rho}(T)}$$

The values of $\frac{1}{2\lambda_{tot}} \sum_{l \in \mathcal{L}} \sigma_l^2$ and $\frac{1}{\lambda_{tot}} \sum_{l \in \mathcal{L}} \theta_l(k)$ are typically independent of N (recall (2)), and so the numerator is roughly linear in the T value. Because we have finite state ergodic Markov chains, from (16) we see the function $\hat{\rho}(T)$ has the form:

$$\hat{\rho}(T) \leq \rho^* + |\mathcal{S}| \nu \gamma^{T+1}$$

where $|\mathcal{S}|$ is the cardinality of the largest interference set $\mathcal{S}_l$, and ν and γ are the largest values of ν_l and γ_l, respectively, over all links $l \in \mathcal{L}$. In this case, we have $1 - \hat{\rho}(T) \geq (1 - \rho^*)/2$ whenever:

$$|\mathcal{S}| \nu \gamma^{T+1} \leq (1 - \rho^*)/2$$

which holds when T is chosen as the smallest integer that satisfies:

$$\frac{\log(2\nu |\mathcal{S}| / (1 - \rho^*))}{\log(1/\gamma)} - 1 \leq T \leq \frac{\log(2\nu |\mathcal{S}| / (1 - \rho^*))}{\log(1/\gamma)}$$

Thus, the above delay bound grows at most logarithmically in $|\mathcal{S}|$. A more explicit and *order-optimal* delay analysis is provided in the next section, where the special case of 2-state Markov chains is considered and average delay is shown to be independent of the network size.

V. Time-Correlated Arrivals with Two States

Consider the case of Markov modulated arrivals, as described in Section II-A, where all Markov chains $Z_l(t)$ have at most two states (labeled "1" and "2"). Note that this model includes the important special case of ON/OFF inputs, where $A_l(t)$ has a single packet arrival when in the ON state and has no arrivals when in the OFF state.[3] Let $\tilde{\mathcal{L}}$ be the set of all links $l \in \mathcal{L}$ that have exactly two states with different conditional rates $\lambda_l^{(1)}$ and $\lambda_l^{(2)}$. The transition probabilities are given by β_l and δ_l for each two-state chain $Z_l(t)$, as shown in Fig. 1. Assume that $0 < \delta_l < 1$ and $0 < \beta_l < 1$ for all $l \in \tilde{\mathcal{L}}$, and define $\pi_l^{(1)}$ and $\pi_l^{(2)}$ as the steady state probabilities for each two-state chain $Z_l(t)$:

$$\pi_l^{(1)} = \frac{\delta_l}{\beta_l + \delta_l} \quad , \quad \pi_l^{(2)} = \frac{\beta_l}{\beta_l + \delta_l}$$

[3] The conditional rates for such an ON/OFF example are given by $\lambda_l^{(1)} = 1$, $\lambda_l^{(2)} = 0$, where states 1 and 2 are associated with the ON and OFF states, respectively, as shown in Fig. 1.

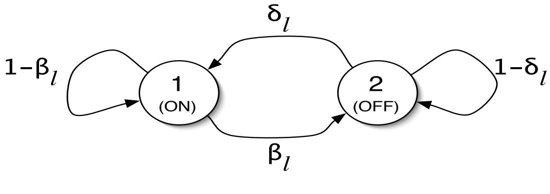

Fig. 1. The 2-state Markov chain $Z_l(t)$ for link l.

Arrival processes $A_l(t)$ for the remaining links $l \notin \tilde{\mathcal{L}}$ are either i.i.d. over slots (effectively "one-state" chains), or have two states but with $\lambda_l^{(1)} = \lambda_l^{(2)}$ (in the latter case, the different states may correspond to different conditional second moments).

The Markov chains $Z_l(t)$ are possibly correlated over different links $l \in \mathcal{L}$, although we focus primarily on the case when all chains are independent. The chains are assumed to be *stationary*, so that for each link $l \in \mathcal{L}$ we have $\mathbb{E}\{A_l(t)\} = \lambda_l$ for all time t. The time average rates λ_l for all 2-state arrival processes are given by:

$$\lambda_l = \pi_l^{(1)}\lambda_l^{(1)} + \pi_l^{(2)}\lambda_l^{(2)}$$

The traffic rates $(\lambda_l)_{l\in\mathcal{L}}$ are assumed to satisfy the loading constraints (5) with relative network loading $\rho^* < 1$. In the previous section we showed that the system is strongly stable and ergodic with finite steady state queue backlogs. Here we provide a tighter delay analysis using a combination of Markov chain theory and Lyapunov drift theory.

A. 2-State Drift Analysis

Recall the definition $\hat{A}_{\mathcal{S}_l}(t) \triangleq \sum_{\omega\in\mathcal{S}_l} A_\omega(t)$. Thus, from the drift inequality (10) we have that the unconditional Lyapunov drift $\Delta(t)$ satisfies the following every slot t:

$$\begin{aligned}\Delta(t) \quad \leq \quad & \mathbb{E}\{B(t)\} - \sum_{l\in\mathcal{L}} \mathbb{E}\{Q_l(t)\} \\ & + \sum_{l\in\mathcal{L}}\sum_{\omega\in\mathcal{S}_l} \mathbb{E}\{Q_l(t)A_\omega(t)\} \qquad (22)\end{aligned}$$

Now assume that the system is in *steady state* at time t, and is also in steady state at time $t-1$. Fix a link $l \in \mathcal{L}$ and an arrival process $A_\omega(t)$. We shall derive a relationship between $\mathbb{E}\{Q_l(t)A_\omega(t)\}$ and $\mathbb{E}\{Q_l(t)\}$. To this end, first note that for any link $l \in \mathcal{L}$ we have:

$$\mathbb{E}\{Q_l(t)A_\omega(t)\} = \mathbb{E}\{Q_l(t)\}\lambda_\omega \text{ if } \omega \notin \tilde{\mathcal{L}} \qquad (23)$$

That is, the expectation has a simple product form in the special case when $\omega \notin \tilde{\mathcal{L}}$, because $A_\omega(t)$ is either i.i.d. over slots or has the same conditional rate $\lambda_\omega^{(1)} = \lambda_\omega^{(2)} = \lambda_\omega$.

For the rest of this subsection we consider the opposite and more challenging case when $\omega \in \tilde{\mathcal{L}}$. In this case, we have:

$$\mathbb{E}\{Q_l(t)\} = \sum_{m=1}^{2} \pi_\omega^{(m)} \mathbb{E}\{Q_l(t) \mid Z_\omega(t) = m\} \qquad (24)$$

However, we also have:

$$\begin{aligned}&\mathbb{E}\{Q_l(t)A_\omega(t)\} \\ &\quad = \sum_{m=1}^{2} \pi_\omega^{(m)} \mathbb{E}\{Q_l(t)A_\omega(t) \mid Z_\omega(t) = m\} \\ &\quad = \sum_{m=1}^{2} \pi_\omega^{(m)}\lambda_\omega^{(m)} \mathbb{E}\{Q_l(t) \mid Z_\omega(t) = m\} \quad (25)\end{aligned}$$

This final equality holds because the expectation of $Q_l(t)$ is conditionally independent of $A_\omega(t)$ given $Z_\omega(t)$. Because the system is in steady state, the quantities $\mathbb{E}\{Q_l(t)\}$, $\mathbb{E}\{Q_l(t)A_\omega(t)\}$, and $\mathbb{E}\{Q_l(t) \mid Z_\omega(t) = m\}$ do not depend on t, and we define:

$$\begin{aligned}\mathbb{E}\{Q_l\} & \triangleq \mathbb{E}\{Q_l(t)\} \\ \mathbb{E}\{Q_lA_\omega\} & \triangleq \mathbb{E}\{Q_l(t)A_\omega(t)\} \\ x_{l,\omega}^{(m)} & \triangleq \pi_\omega^{(m)}\mathbb{E}\{Q_l(t) \mid Z_\omega(t) = m\}\end{aligned}$$

The equalities (24) and (25) can be re-written:

$$\mathbb{E}\{Q_l\} = x_{l,\omega}^{(1)} + x_{l,\omega}^{(2)} \qquad (26)$$

$$\mathbb{E}\{Q_lA_\omega\} = \lambda_\omega^{(1)}x_{l,\omega}^{(1)} + \lambda_\omega^{(2)}x_{l,\omega}^{(2)} \qquad (27)$$

Equations (26) and (27) are two linear equations that express a relationship between 4 unknowns (where the unknowns are $\mathbb{E}\{Q_l\}, \mathbb{E}\{Q_lA_\omega\}$, $x_{l,\omega}^{(1)}$ and $x_{l,\omega}^{(2)}$). To express a direct linear relationship between $\mathbb{E}\{Q_l\}$ and $\mathbb{E}\{Q_lA_\omega\}$, we require an additional equation. To this end, note that:

$$Q_l(t) = Q_l(t-1) - \mu_l(t-1) + A_l(t-1)$$

Therefore:

$$\mathbb{E}\{Q_l(t)A_\omega(t)\} = \mathbb{E}\{Q_l(t-1)A_\omega(t)\} - D_{l,\omega} + C_{l,\omega} \quad (28)$$

where $C_{l,\omega}$ and $D_{l,\omega}$ are defined as:

$$C_{l,\omega} \triangleq \mathbb{E}\{A_l(t-1)A_\omega(t)\} \qquad (29)$$

$$D_{l,\omega} \triangleq \mathbb{E}\{\mu_l(t-1)A_\omega(t)\} \qquad (30)$$

Now:

$$\begin{aligned}&\mathbb{E}\{Q_l(t-1)A_\omega(t)\} \\ &\quad = \sum_{m=1}^{2} \pi_\omega^{(m)}\mathbb{E}\{Q_l(t-1)A_\omega(t) \mid Z_\omega(t-1) = m\} \\ &\quad = \sum_{m=1}^{2} \pi_\omega^{(m)}h_\omega^{(m)}\mathbb{E}\{Q_l(t-1) \mid Z_\omega(t-1) = m\} \quad (31)\end{aligned}$$

where $h_\omega^{(m)}$ is defined:

$$h_\omega^{(m)} \triangleq \mathbb{E}\{A_\omega(t) \mid Z_\omega(t-1) = m\} \qquad (32)$$

The last equality follows again because $Q_l(t-1)$ is conditionally independent of $A_\omega(t)$ given $Z_\omega(t-1)$. However, because the system is in steady state at time t and also at time $t-1$, it follows that:

$$\pi_\omega^{(m)}\mathbb{E}\{Q_l(t-1) \mid Z_\omega(t-1) = m\} = x_{l,\omega}^{(m)} \qquad (33)$$

Therefore, using (33) and (31), equation (28) becomes the following:

$$\mathbb{E}\{Q_lA_\omega\} = C_{l,\omega} - D_{l,\omega} + h_\omega^{(1)}x_{l,\omega}^{(1)} + h_\omega^{(2)}x_{l,\omega}^{(2)} \qquad (34)$$

The constants $h_\omega^{(m)}$, defined in (32), can be computed directly from the transition probabilities for chain $Z_\omega(t)$:

$$h_\omega^{(1)} = (1-\beta_\omega)\lambda_\omega^{(1)} + \beta_\omega\lambda_\omega^{(2)} \qquad (35)$$

$$h_\omega^{(2)} = \delta_\omega\lambda_\omega^{(1)} + (1-\delta_\omega)\lambda_\omega^{(2)} \qquad (36)$$

The linear equations (26), (27), and (34) involve three equations and four unknowns, and can be shown to be linearly independent whenever $\lambda_\omega^{(1)} \neq \lambda_\omega^{(2)}$ (which holds for all

$\omega \in \tilde{\mathcal{L}}$). This directly leads to the following lemma that relates $\mathbb{E}\{Q_l\}$ and $\mathbb{E}\{Q_l A_\omega\}$.

Lemma 3: For all links $l \in \mathcal{L}$, we have:

$$\mathbb{E}\{Q_l A_\omega\} = \mathbb{E}\{Q_l\} \lambda_\omega + \frac{C_{l,\omega} - D_{l,\omega}}{\beta_\omega + \delta_\omega} \quad \text{if } \omega \in \tilde{\mathcal{L}} \tag{37}$$

Proof: The result follows by eliminating the $x_{l,\omega}^{(1)}$ and $x_{l,\omega}^{(2)}$ variables from (26), (27), and (34), and the computation is given in Appendix C. ■

Note from the definitions (29) and (30) that $C_{l,\omega} \geq 0$, $D_{l,\omega} \geq 0$ for all $l \in \mathcal{L}$ and all $\omega \in \tilde{\mathcal{L}}$. Define $C_{l,\omega} = 0$ for $\omega \notin \tilde{\mathcal{L}}$. Using (37) and (23) we find that for all link pairs $l, \omega \in \mathcal{L}$, we have:

$$\mathbb{E}\{Q_l A_\omega\} \leq \mathbb{E}\{Q_l\} \lambda_\omega + \frac{C_{l,\omega}}{\beta_\omega + \delta_\omega}$$

where the inequality comes because we have neglected the $D_{l,\omega}$ constant. Using this expression directly in the Lyapunov drift inequality (22) yields the following drift expression that holds at any time t at which the system is in steady state:

$$\begin{aligned} \Delta(t) \quad &\leq \quad \mathbb{E}\{B(t)\} - \sum_{l \in \mathcal{L}} \mathbb{E}\{Q_l(t)\}(1 - \rho^*) \\ &+ \sum_{l \in \mathcal{L}} \sum_{\omega \in \mathcal{S}_l \cap \tilde{\mathcal{L}}} \frac{C_{l,\omega}}{\beta_\omega + \delta_\omega} \end{aligned} \tag{38}$$

where we have used the fact that $\sum_{\omega \in \mathcal{S}_l} \lambda_\omega \leq \rho^*$.

B. The Delay Bound for 2-State Markov Modulated Arrivals

Theorem 4: If the input rates $(\lambda_l)_{l \in \mathcal{L}}$ satisfy the loading constraints (5) for a given relative network loading $\rho^* < 1$, then:

(a) The system is stable with steady state average congestion that satisfies:

$$\sum_{l \in \mathcal{L}} \overline{Q}_l \leq \frac{\tilde{B} + \tilde{C}}{1 - \rho^*}$$

where:

$$\begin{aligned} \tilde{B} \quad &\triangleq \quad \frac{1}{2} \sum_{l \in \mathcal{L}} \left[\mathbb{E}\left\{ A_l(t) \hat{A}_{\mathcal{S}_l}(t) \right\} + \lambda_l \right] \\ \tilde{C} \quad &\triangleq \quad \sum_{l \in \mathcal{L}} \sum_{\omega \in \mathcal{S}_l \cap \tilde{\mathcal{L}}} \frac{\mathbb{E}\{A_l(t-1) A_\omega(t)\}}{\beta_\omega + \delta_\omega} \end{aligned}$$

(b) If all Markov chains $Z_l(t)$ are additionally independent of each other, then:

$$\tilde{B} = \frac{1}{2} \sum_{l \in \mathcal{L}} \left[\lambda_l \hat{\lambda}_{\mathcal{S}_l} + \sigma_l^2 + \lambda_l \right]$$

$$\mathbb{E}\{A_l(t-1) A_\omega(t)\} = \lambda_l \lambda_\omega \;\; \text{if } l \neq \omega$$

Hence, average delay $\overline{W}$ satisfies:

$$\begin{aligned} \overline{W} \quad &\leq \quad \frac{1 + \frac{1}{\lambda_{tot}} \sum_{l \in \mathcal{L}} [\sigma_l^2 + \lambda_l \hat{\lambda}_{\mathcal{S}_l}]}{2(1 - \rho^*)} \\ &+ \frac{\frac{1}{\lambda_{tot}} \left[\sum_{l \in \tilde{\mathcal{L}}} \frac{\theta_l[1]}{(\beta_l + \delta_l)} + \sum_l \sum_{\omega \in \mathcal{S}_l \cap \tilde{\mathcal{L}}} \frac{\lambda_l \lambda_\omega}{(\beta_\omega + \delta_\omega)} \right]}{1 - \rho^*} \end{aligned}$$

where $\lambda_{tot} \triangleq \sum_{l \in \mathcal{L}} \lambda_l$, and $\theta_l[1] \triangleq \mathbb{E}\{A_l(t-1) A_l(t)\} - (\lambda_l)^2$ is the 1-slot auto-correlation for process $A_l(t)$, given by:

$$\theta_l[1] = \frac{\beta_l \delta_l (\lambda_l^{(1)} - \lambda_l^{(2)})^2 (1 - \beta_l - \delta_l)}{(\beta_l + \delta_l)^2}$$

Proof: The result follows directly from (38) via the Lyapunov drift theorem (Theorem 1). Details omitted for brevity. ■

Using the fact that $\hat{\lambda}_{\mathcal{S}_l} \leq \rho^* < 1$, the average delay bound $\overline{W}$ in part (b) of the above theorem can be simplified as follows:

$$\begin{aligned} \overline{W} \quad &\leq \quad \frac{1 + \rho^* + \frac{1}{\lambda_{tot}} \sum_{l \in \mathcal{L}} \sigma_l^2}{2(1 - \rho^*)} \\ &+ \max_{\omega \in \tilde{\mathcal{L}}} \left[\frac{1}{\beta_\omega + \delta_\omega} \right] \left(\frac{\rho^* + \frac{1}{\lambda_{tot}} \sum_{l \in \tilde{\mathcal{L}}} \theta_l[1]}{1 - \rho^*} \right) \end{aligned}$$

Note that $\theta_l[1] \leq \lambda_l \lambda^{(max)}$, where $\lambda^{(max)}$ is the largest conditional rate over all links and states. Thus, the numerator in the final term in the above delay bound satisfies: $\frac{1}{\lambda_{tot}} \sum_{l \in \tilde{\mathcal{L}}} \theta_l[1] \leq \lambda^{max}$. Therefore, the above bound is $O(1)$ (independent of the network size N). In the special case of ON/OFF sources, where a single packet arrives from stream l when $Z_l(t) = ON$ and no packet arrives when $Z_l(t) = OFF$, we have $\lambda_l^{(1)} = 1$ and $\lambda_l^{(2)} = 0$, and $\lambda_l = \pi_l^{(1)}$. Further: $\sigma_l^2 \leq \lambda_l$, $\theta_l[1] \leq \lambda_l$. Thus, average delay in this ON/OFF example satisfies:

$$\overline{W}_{ON/OFF} \leq \frac{1 + \frac{\rho^*}{2} + \max_{\omega \in \tilde{\mathcal{L}}} [(\rho^* + 1)/(\beta_\omega + \delta_\omega)]}{(1 - \rho^*)}$$

Note that $1/\beta_l$ is the average burst size (i.e., the average time spent in the ON state), and so the numerator roughly grows linearly in the largest average burst size over any input.

VI. Delay Analysis with Flow Control

The previous sections assume that the steady state input rate vector $(\lambda_l)_{l \in \mathcal{L}}$ satisfies the loading contstraints (5), and hence is interior to the reduced throughput region Λ^*. Here we consider arbitrary input rate vectors (possibly inside or outside of Λ^*), and develop a flow control mechanism that works together with maximal link scheduling. We restrict attention to arrival processes $A_l(t)$ that are Markov modulated and have at most two Markov states, as in the previous section.

We use the flow control framework of [6]. Specifically, every slot the flow controller at link l observes the current number of new packets $A_l(t)$, and decides how many of these new arrivals to place into the network. Let $R_l(t)$ be this flow control decision, where $R_l(t)$ is an integer that satisfies:

$$0 \leq R_l(t) \leq A_l(t) \quad \text{for all } l \in \mathcal{L} \text{ and all } t$$

Any packets that are not immediately admitted by the flow controller are dropped. The queueing dynamics are thus given by (compare with (3)):

$$Q_l(t+1) = Q_l(t) - \mu_l(t) + R_l(t) \tag{39}$$

Let $(r_l)_{l \in \mathcal{L}}$ denote the vector of time average admission rates into the network links (for $l \in \mathcal{L}$), which corresponds to some particular flow control algorithm. We define network fairness in terms of concave *utility functions* $g_l(r)$. Each utility function $g_l(r)$ is continuous, concave, non-negative, and non-decreasing. The goal is to design a flow control algorithm that ensures order optimal delay while maximizing the sum of utilities over all network links, where the maximum is defined with respect to the reduced throughput region Λ^*. Specifically, define g^* as the maximum sum utility associated with the following optimization problem:

$$\begin{aligned} &\text{Maximize:} && \textstyle\sum_{l\in\mathcal{L}} g_l(r_l) && (40)\\ &\text{Subject to:} && (r_l) \in \Lambda^* && (41)\\ &&& 0 \leq r_l \leq \lambda_l \;\text{ for all } l \in \mathcal{L} && (42)\end{aligned}$$

We desire our flow control algorithm to yield a sum utility that is close to (or larger than) the value of g^*. We note that a similar utility-based flow control problem with maximal scheduling is considered in [4], and a token based technique for max-min fairness is considered in [2]. Our algorithm is quite different from [4] [2], and our analysis is unique in that it demonstrates order-optimal delay for bursty traffic arrivals.

For technical reasons, we define $\mathcal{I}_\delta$ as the finite set of real numbers between 0 and 1, inclusive, that are uniformly spaced apart by some (arbitrarily small) quantization $\delta > 0$. For example, $\mathcal{I}_\delta$ might represent numbers as viewed by a computer that rounds to a finite decimal place. This quantization is convenient for limit theorems, as we will show the resulting system has only a finite number of possible states. Define g_δ^* as the maximum value of the problem (40)-(42) subject to the additional constraint:

$$r_l \in \mathcal{I}_\delta \quad \text{for all } l \in \mathcal{L} \tag{43}$$

It is clear that g_δ^* is very close to g^* when δ is small. We shall modify our goal to achieve a target utility of g_δ^*, rather than g^*.

A. The Flow Control Algorithm with Maximal Scheduling

Note that because the utility functions are non-decreasing, the optimization problem (40)-(43) is equivalent to:

$$\begin{aligned} &\text{Maximize:} && \textstyle\sum_{l\in\mathcal{L}} g_l(\gamma_l)\\ &\text{Subject to:} && (r_l) \in \Lambda^*\\ &&& r_l \leq \lambda_l \;\text{ for all } l \in \mathcal{L}\\ &&& 0 \leq \gamma_l \leq r_l \;,\; \gamma_l \in \mathcal{I}_\delta \;\text{ for all } l \in \mathcal{L}\end{aligned}$$

where we have introduced an auxiliary variable γ_l for each link l. Using our flow control framework of [6] [20], for each link l we define an *auxiliary process* $\gamma_l(t)$ and a *flow state queue* $Y_l(t)$. The flow state queue $Y_l(t)$ for each link l is implemented purely in software at link l, is initialized so that $Y_l(0) = 0$, and has update equation:

$$Y_l(t+1) = \max[Y_l(t) - R_l(t), 0] + \gamma_l(t) \tag{44}$$

where $R_l(t)$ is the admission decision made by link l on slot t and $\gamma_l(t)$ is an auxiliary variable chosen by link l on slot t according to the following algorithm. The algorithm is defined in terms of a control parameter $V > 0$ that affects a tradeoff between utility and delay.

Flow Control with Maximal Scheduling (FLOW-MAXIMAL): Every slot t, the flow controller at each link $l \in \mathcal{L}$ observes the queue backlog $\hat{Q}_{\mathcal{S}_l}(t)$ (defined in (6)), the flow state queue $Y_l(t)$, and the new arrivals $A_l(t)$, and performs the following:

- (*Admission Control*) Choose $R_l(t)$ as follows:

$$R_l(t) = \begin{cases} A_l(t) & \text{if } Y_l(t) \geq \hat{Q}_{\mathcal{S}_l}(t)\\ 0 & \text{otherwise}\end{cases}$$

- (*Auxiliary Variables*) Choose $\gamma_l(t)$ as the solution to:

$$\begin{aligned} &\text{Maximize:} && V g_l(\gamma_l(t)) - Y_l(t)\gamma_l(t) && (45)\\ &\text{Subject to:} && 0 \leq \gamma_l(t) \leq 1 \;,\; \gamma_l(t) \in \mathcal{I}_\delta && (46)\end{aligned}$$

  Note that this is a simple maximization of a concave function of a single variable γ over an interval. The additional constraint $\gamma_l(t) \in \mathcal{I}_\delta$ can be enforced by first calculating the non-quantized optimal $\gamma_l(t)$ value, and then choosing the quantized value either to the right or left according to which one maximizes (45).
- (*Virtual Queue Update*) Update $Y_l(t)$ according to (44), using the above chosen values of $R_l(t)$ and $\gamma_l(t)$.

The network then schedules links as before, using any maximal scheduling algorithm. Specifically, every slot t the following action is taken:

- (*Maximal Scheduling*) Choose $(\mu_l(t))_{l\in\mathcal{L}}$ by performing maximal link scheduling, so that queue backlog is updated according to (39).

B. Algorithm Performance

Suppose arrivals $A_l(t)$ are Markov modulated (possibly correlated over the different links l), and satisfy the two-state assumptions of Section V with steady state traffic rate vector $(\lambda_l)_{l\in\mathcal{L}}$. For simplicity, we additionally assume the maximum number of packets that can arrive to any link during a slot is bounded by a constant A_{max}, so that:

$$A_l(t) \leq A_{max} \text{ for all } t$$

Consider any utility functions $g_l(r)$ that are concave, non-decreasing, and non-negative. We assume that $g_l(0) = 0$ for all $l \in \mathcal{L}$ (so that zero throughput yields zero utility). Further define η_l as the maximum right-derivative of the utility function $g_l(r)$.[4] We assume that $\eta_l < \infty$ for all $l \in \mathcal{L}$, and define η as the largest η_l value: $\eta \triangleq \max_{l\in\mathcal{L}} \eta_l$. For example, consider the following two types of utility functions defined in terms of non-negative constants α_l and β_l:

- *Linear:* $g_l(r) = \alpha_l r$
- *Logarithmic:* $g_l(r) = \alpha_l \log(1 + \beta_l r)$

In the linear case above, we have $\eta_l = \alpha_l$. In the logarithmic case, we have $\eta_l = \alpha_l\beta_l$.

Our first result shows that the above algorithm yields bounded queue sizes for all time.

Theorem 5: (Worst Case Queue Bounds) Suppose all queues are initially empty and the above FLOW-MAXIMAL algorithm is implemented. Then for arbitrary arrival processes

[4] All concave functions of one variable have well defined right-derivatives.

$A_l(t)$ that satisfy the maximum arrival bound $A_l(t) \leq A_{max}$ for all t, we have the following:

$$\begin{aligned} 0 \leq \quad Y_l(t) \quad &\leq V\eta_l + 1 \quad \text{for all } t &(47)\\ 0 \leq \quad Q_l(t) \quad &\leq V\eta_l + 1 + A_{max} \quad \text{for all } t &(48)\end{aligned}$$

Proof: First note that because $g_l(\gamma)$ is concave, has maximum derivative η_l, and satisfies $g_l(0) = 0$, we have:

$$g_l(\gamma) \leq \gamma\eta_l \quad \text{for all } \gamma \geq 0$$

Now suppose that $Y_l(t) > V\eta_l$ for some slot t. The $\gamma_l(t)$ variable is chosen to maximize (45). However, for any possible value $\gamma \geq 0$, we have:

$$Vg_l(\gamma) - Y_l(t)\gamma \leq V\gamma\eta_l - Y_l(t)\gamma \leq V\gamma\eta_l - V\eta_l\gamma = 0$$

where the second inequality is an equality only if $\gamma = 0$. It follows that the controller chooses $\gamma_l(t) = 0$ on any slot in which $Y_l(t) > V\eta_l$. It follows from (44) that $Y_l(t)$ cannot increase on the next slot, and so it can only increase when it is currently less than or equal to $V\eta_l$. Because $\gamma_l(t) \leq 1$ for all t, the maximum amount of increase is 1, and so $Y_l(t) \leq V\eta_l + 1$ for all t, proving (47).

Similarly, note that if $Q_l(t) > V\eta_l + 1$, then $\hat{Q}_{\mathcal{S}_l}(t) > V\eta_l + 1 \geq Y_l(t)$, and so the admission control algorithm sets $R_l(t) = 0$. It follows that $Q_l(t)$ can only increase when it is less than or equal to $V\eta_l + 1$. Because A_{max} is the maximum amount it can increase, we have that (48) holds for all time. ■

Note from (44) and the constraint $\gamma_l(t) \in \mathcal{I}_\delta$ that $Y_l(t)$ can only take values that are integers plus an element in $\mathcal{I}_\delta$. Because $Y_l(t)$ is bounded, it can only take a finite number of values. Thus, there are only a finite number of system states $[\boldsymbol{Z}(t); \boldsymbol{Q}(t); \boldsymbol{Y}(t)]$ (recall that $Q_l(t)$ is a bounded integer and $\boldsymbol{Z}(t)$ has a finite state space). It follows that time average expectations are well defined for any fixed initial condition.

Let $(r_l^*)_{l\in\mathcal{L}}$ represent an optimal solution to (40)-(43) with optimal sum utility g_δ^*, so that $g_\delta^* = \sum_{l\in\mathcal{L}} g_l(r_l^*)$.

Theorem 6: (Performance of Flow Control with Maximal Scheduling) Suppose the above FLOW-MAXIMAL algorithm is implemented. Suppose arrivals are modulated by ergodic Markov processes with at most two-states as described in Section V, with steady state rates $(\lambda_l)_{l\in\mathcal{L}}$ (possibly inside or outside of Λ^*). Then queue backlog and sum utility satisfy:

$$\begin{aligned}\sum_{l\in\mathcal{L}} \overline{Q}_l \quad &\leq \quad \overline{r}_{tot}\left[1 + \frac{A_{max}}{2} + V\eta\right]\\ &\qquad + \frac{1}{2}\sum_{l\in\mathcal{L}} \mathbb{E}\left\{A_l(t)\hat{A}_{\mathcal{S}_l}(t)\right\} &(49)\\ \sum_{l\in\mathcal{L}} g_l(\overline{r}) \quad &\geq \quad g_\delta^* - F/V &(50)\end{aligned}$$

where $\overline{Q}_l$ is the time average expected backlog in link l, and $\overline{r}_l$ is the time average expected admitted rate to link l:

$$\overline{Q}_l \triangleq \lim_{t\to\infty} \frac{1}{t}\sum_{\tau=0}^{t-1} \mathbb{E}\left\{Q(\tau)\right\} \ , \ \overline{r}_l \triangleq \lim_{t\to\infty} \frac{1}{t}\sum_{\tau=0}^{t-1} \mathbb{E}\left\{R_l(\tau)\right\}$$

and $\overline{r}_{tot} = \sum_{l\in\mathcal{L}} \overline{r}_l$. Finally, the constant F is given by:

$$\begin{aligned} F \quad \triangleq \quad & \overline{r}_{tot}\left[1 + \frac{A_{max}}{2}\right] + \frac{1}{2}\sum_{l\in\mathcal{L}} \mathbb{E}\left\{A_l(t)\hat{A}_{\mathcal{S}_l}(t)\right\}\\ & + \sum_{l\in\mathcal{L}}\sum_{\omega\in\mathcal{S}_l\cap\tilde{\mathcal{L}}} \frac{(r_\omega^*/\lambda_\omega)\mathbb{E}\left\{A_l(t-1)A_\omega(t)\right\}}{\beta_\omega + \delta_\omega}\\ & + \sum_{l\in\tilde{\mathcal{L}}} \frac{(r_l^*/\lambda_l)\mathbb{E}\left\{A_l(t-1)A_l(t)\right\}}{\beta_l + \delta_l} \qquad (51)\end{aligned}$$

where $r_l^*/\lambda_l \leq 1$ due to the constraints of (40)-(42). Recall that the set $\tilde{\mathcal{L}}$ contains only links that have exactly two states and satisfy $\lambda_l^{(1)} \neq \lambda_l^{(2)}$. If the set $\tilde{\mathcal{L}}$ is empty (such as when arrivals are i.i.d. over slots), then the last two summation terms in the expression for F are equal to zero.

We emphasize that the $\mathbb{E}\left\{A_l(t)\hat{A}_{\mathcal{S}_l}(t)\right\}$ value in Theorem 6 is taken with respect to the steady state distribution for arrivals. Further, because the total input rate into the system is $\overline{r}_{tot}$, we have by Little's Theorem that average delay satisfies:

$$\overline{W} \leq \left[1 + \frac{A_{max}}{2} + V\eta + \frac{1}{2\overline{r}_{tot}}\sum_{l\in\mathcal{L}} \mathbb{E}\left\{A_l(t)\hat{A}_{\mathcal{S}_l}(t)\right\}\right]$$

It follows that we again have an explicit tradeoff between utility and average delay, as determined by the V parameter. Further, if exogenous inputs are independent of each other and $\lambda_{tot}/\overline{r}_{tot} = O(1)$, we have:

$$\frac{1}{2\overline{r}_{tot}}\sum_{l\in\mathcal{L}} \mathbb{E}\left\{A_l(t)\hat{A}_{\mathcal{S}_l}(t)\right\} = O(1)$$

In this case, average delay is $O(1)$ (independent of the network size). Theorem 6 is proven in Appendix D using a novel argument that combines Lyapunov optimization with the steady state analysis for bursty traffic as in Section V.

C. Discussion of "Periodic" Maximal Scheduling

Because any maximal scheduling algorithm can be used, one can consider algorithms where the initial collection of links that are activated as part of the maximal set are determined periodically, so that each link is chosen for this initial activation every M slots (where M is a function of the network size). Links in this initial set that contain packets are scheduled for transmission, and additional links are then chosen to ensure the maximal property is satisfied in the network on every slot. For example, in $N \times N$ bi-partite graphs with matching constraints, one can consider choosing links according to a round-robin schedule that selects one link every N slots. This ensures a service opportunity every N slots, and hence, because queue backlog $Q_l(t)$ is bounded by $V\eta_l + 1 + A_{max}$ for all time, we also have a worst-case delay bound of $N[V\eta_l + 1 + A_{max}]$ slots. This worst-case delay bound holds in addition to all of our other analytical guarantees, including our *average* delay bound that is independent of N when $\lambda_{tot}/\overline{r}_{tot} = O(1)$.

VII. Conclusion

We have developed order-optimal delay results for one-hop networks with general interference set constraints and bursty (time-correlated) traffic. Our results hold for cases when traffic is within the reduced throughput region Λ^*, a region that is typically within a constant factor of the network capacity region. Futher, we have developed a flow control technique that works together with maximal scheduling and yields an explicit utility-delay tradeoff.

Appendix A — Proof of Lemma 2

To compute $\Delta(t)$, note that using (8) and (3) yields:

$$\begin{aligned} Q_l(t+1)\hat{Q}_{\mathcal{S}_l}(t+1) &= Q_l(t)\hat{Q}_{\mathcal{S}_l}(t) \\ &\quad +(A_l(t)-\mu_l(t))(\hat{A}_{\mathcal{S}_l}(t)-\hat{\mu}_{\mathcal{S}_l}(t)) \\ &\quad -Q_l(t)(\hat{\mu}_{\mathcal{S}_l}(t)-\hat{A}_{\mathcal{S}_l}(t)) \\ &\quad -\hat{Q}_{\mathcal{S}_l}(t)(\mu_l(t)-A_l(t)) \end{aligned}$$

Thus, the 1-step unconditional Lyapunov drift is given by:

$$\begin{aligned} \Delta(t) &= \mathbb{E}\left\{B(t)\right\} \\ &\quad -\frac{1}{2}\sum_{l\in\mathcal{L}}\mathbb{E}\left\{Q_l(t)(\hat{\mu}_{\mathcal{S}_l}(t)-\hat{A}_{\mathcal{S}_l}(t))\right\} \\ &\quad -\frac{1}{2}\sum_{l\in\mathcal{L}}\mathbb{E}\left\{\hat{Q}_{\mathcal{S}_l}(t)(\mu_l(t)-A_l(t))\right\} \end{aligned} \tag{52}$$

where

$$B(t)\triangleq\frac{1}{2}\sum_{l\in\mathcal{L}}\left[(A_l(t)-\mu_l(t))(\hat{A}_{\mathcal{S}_l}(t)-\hat{\mu}_{\mathcal{S}_l}(t))\right] \tag{53}$$

We now use the following important structural property of the interference sets.

Lemma 4: (Sum Switching) For any function $f(l,\omega)$ (where $l\in\mathcal{L}$, $\omega\in\mathcal{L}$), we have:

$$\sum_{l\in\mathcal{L}}\sum_{\omega\in\mathcal{S}_l}f(l,\omega)=\sum_{\omega\in\mathcal{L}}\sum_{l\in\mathcal{S}_\omega}f(l,\omega)\ \Box$$

The above lemma follows directly from the *pairwise symmetry property* of the interference sets: For any two links $l,\omega\in\mathcal{L}$, we have that $\omega\in\mathcal{S}_l$ if and only if $l\in\mathcal{S}_\omega$, and hence $\{(l,\omega)\mid l\in\mathcal{L},\omega\in\mathcal{S}_l\}=\{(l,\omega)\mid\omega\in\mathcal{L},l\in\mathcal{S}_\omega\}$. Using this lemma we can re-write the final term in (52):

$$\sum_{l\in\mathcal{L}}\hat{Q}_{\mathcal{S}_l}(t)(\mu_l(t)-A_l(t))$$

$$\begin{aligned} &= \sum_{l\in\mathcal{L}}\sum_{\omega\in\mathcal{S}_l}Q_\omega(t)(\mu_l(t)-A_l(t)) && (54)\\ &= \sum_{\omega\in\mathcal{L}}\sum_{l\in\mathcal{S}_\omega}Q_\omega(t)(\mu_l(t)-A_l(t)) && (55)\\ &= \sum_{\omega\in\mathcal{L}}Q_\omega(t)(\hat{\mu}_{\mathcal{S}_\omega}(t)-\hat{A}_{\mathcal{S}_\omega}(t)) && (56)\\ &= \sum_{l\in\mathcal{L}}Q_l(t)(\hat{\mu}_{\mathcal{S}_l}(t)-\hat{A}_{\mathcal{S}_l}(t)) && (57) \end{aligned}$$

where (54) follows by the definition of $\hat{Q}_{\mathcal{S}_l}(t)$ given in (6), (55) follows by the Sum Switching Lemma (Lemma 4), and (57) follows by re-labeling the indices. Plugging the equality (57) directly into the drift expression (52) yields:

$$\Delta(t)=\mathbb{E}\left\{B(t)\right\}-\sum_{l\in\mathcal{L}}\mathbb{E}\left\{Q_l(t)(\hat{\mu}_{\mathcal{S}_l}(t)-\hat{A}_{\mathcal{S}_l}(t))\right\} \tag{58}$$

Further, we note that the expression for $B(t)$ in (53) is equivalent to that given in (11). This can be seen by using a sum-switching argument similar to (54)-(57) on the summation $\sum_{l\in\mathcal{L}}\hat{\mu}_{\mathcal{S}_l}(t)A_l(t)$, and by noting that $\mu_l(t)\hat{\mu}_{\mathcal{S}_l}(t)=\mu_l(t)$. The latter equality holds because if $\mu_l(t)=0$ we have $0=0$ which is trivially true, while if $\mu_l(t)=1$ then no other links within $\mathcal{S}_l$ can be active, and so $\hat{\mu}_{\mathcal{S}_l}(t)=1$.

We now use the fact that maximal scheduling is performed every timeslot. Specifically, we recall that any maximal scheduling algorithm satisfies (4) every timeslot. Using the definition of $\hat{\mu}_{\mathcal{S}_l}(t)$, we note that (4) is equivalent to:

$$Q_l(t)\hat{\mu}_{\mathcal{S}_l}(t)\geq Q_l(t)\quad\text{for all } l\in\mathcal{L}\text{ and all } t$$

Plugging the above inequality directly into (58) yields the expression (10) for $\Delta(t)$ under maximal scheduling.

Appendix B – Proof of Theorem 3

To prove Theorem 3, we introduce an artificial delay in the final term of the drift expression (10) to decouple correlations between queue state and arrivals. This is similar to the T-slot technique of [13][14][6], although, unlike [13][14][6], it allows a tight logarithmic delay result. To begin, fix an integer $T\geq 0$, and note that for $t\in\{0,1,2,\ldots,\}$ we have:

$$Q_l(t)\leq Q_l(t-T)+\sum_{v=0}^{T-1}A_l(t-T+v)$$

where we define $Q_l(t)=0$, $A_l(t)=0$ for $t<0$. Using the above inequality in (10) yields

$$\begin{aligned} \Delta(t) &\leq \mathbb{E}\left\{B(t)+F_T(t)\right\}-\sum_{l\in\mathcal{L}}\mathbb{E}\left\{Q_l(t)\right\} \\ &\quad +\sum_{l\in\mathcal{L}}\mathbb{E}\left\{Q_l(t-T)\hat{A}_{\mathcal{S}_l}(t)\right\} \end{aligned} \tag{59}$$

where $F_T(t)$ is defined:

$$F_T(t)\triangleq\sum_{l\in\mathcal{L}}\mathbb{E}\left\{\hat{A}_{\mathcal{S}_l}(t)\sum_{v=0}^{T-1}A_l(t-T+v)\right\}$$

We now use the $\hat{\rho}(T)$ function to modify the final term on the right hand side of (59):

$$\begin{aligned} &\mathbb{E}\left\{Q_l(t-T)\hat{A}_{\mathcal{S}_l}(t)\right\} \\ &\quad= \mathbb{E}\left\{Q_l(t-T)\mathbb{E}\left\{\hat{A}_{\mathcal{S}_l}(t)\mid\boldsymbol{Q}(t-T)\right\}\right\} \\ &\quad\leq \mathbb{E}\left\{Q_l(t-T)\left(\hat{\lambda}_{\mathcal{S}_l}+\sum_{\omega\in\mathcal{S}_l}\epsilon_\omega(T)\right)\right\} && (60)\\ &\quad\leq \mathbb{E}\left\{Q_l(t-T)\right\}\hat{\rho}(T) && (61) \end{aligned}$$

where (60) follows from (15) and (61) follows from (17). Using (61) in (59), it follows that unconditional Lyapunov drift satisfies:

$$\begin{aligned} \Delta(t) &\leq \mathbb{E}\left\{B(t)+F_T(t)\right\}-\sum_{l\in\mathcal{L}}\mathbb{E}\left\{Q_l(t)\right\} \\ &\quad +\hat{\rho}(T)\sum_{l\in\mathcal{L}}\mathbb{E}\left\{Q_l(t-T)\right\} \end{aligned} \tag{62}$$

Fixing the integer T and using the Lyapunov drift theorem (Theorem 1) in the drift expression (62) yields:

$$\limsup_{t\rightarrow\infty}\frac{1}{t}\sum_{\tau=0}^{t-1}\sum_{l\in\mathcal{L}}\mathbb{E}\left\{Q_l(\tau)\right\}\leq\left(\overline{B}+\overline{F_T}\right)$$
$$+\hat{\rho}(T)\limsup_{t\rightarrow\infty}\frac{1}{t}\sum_{\tau=0}^{t-1}\sum_{l\in\mathcal{L}}\mathbb{E}\left\{Q_l(\tau-T)\right\} \quad (63)$$

where

$$\overline{B}\triangleq\limsup_{t\rightarrow\infty}\frac{1}{t}\sum_{\tau=0}^{t-1}\mathbb{E}\left\{B(\tau)\right\}\ ,\ \overline{F_T}\triangleq\limsup_{t\rightarrow\infty}\frac{1}{t}\sum_{\tau=0}^{t-1}\mathbb{E}\left\{F_T(\tau)\right\}$$

However, it is not difficult to see that:

$$\limsup_{t\rightarrow\infty}\frac{1}{t}\sum_{\tau=0}^{t-1}\sum_{l\in\mathcal{L}}\mathbb{E}\left\{Q_l(\tau-T)\right\}$$
$$=\limsup_{t\rightarrow\infty}\frac{1}{t}\sum_{\tau=0}^{t-1}\sum_{l\in\mathcal{L}}\mathbb{E}\left\{Q_l(\tau)\right\} \quad (64)$$

Indeed, the equality (64) follows by noting the time-delayed version of the limit on the left hand side does not affect the overall time average, as any contribution to the sum over the extra T slots is finite and becomes negligible as $t\rightarrow\infty$. Therefore, because it is assumed that $\hat{\rho}(T)<1$, the inequality (63) simplifies to:

$$\limsup_{t\rightarrow\infty}\frac{1}{t}\sum_{\tau=0}^{t-1}\sum_{l\in\mathcal{L}}\mathbb{E}\left\{Q_l(\tau)\right\}\leq\frac{\overline{B}+\overline{F_T}}{1-\hat{\rho}(T)} \quad (65)$$

It is easy to show that:

$$\overline{B}\leq\frac{1}{2}\sum_{l\in\mathcal{L}}\left[\lambda_l+\mathbb{E}\left\{A_l(t)\hat{A}_{\mathcal{S}_l}(t)\right\}\right]$$

$$\overline{F_T}=\sum_{l\in\mathcal{L}}\mathbb{E}\left\{\hat{A}_{\mathcal{S}_l}(t)\sum_{v=0}^{T-1}A_l(t-T+v)\right\}$$

It follows by stationarity that $\overline{B}\leq\tilde{B}$ and $\overline{F_T}=\tilde{F}_T$, where $\tilde{B}$ and $\tilde{F}_T$ are defined in (18) and (19). Finally, we note that because the queueing dynamics are described by a countable state space, irreducible Markov chain, the $\limsup$ in the left hand side of (65) can be replaced by a regular limit, which proves part (a) of Theorem 3. Part (b) of Theorem 3 follows by computing $\tilde{B}$ and $\tilde{F}_T$ for the case when arrival processes $A_l(t)$ are independent of each other.

Appendix C — Proof of Lemma 3

Choose any two links l,ω such that $l\in\mathcal{L}$ and $\omega\in\tilde{\mathcal{L}}$. The linear equations (27) and (34) can be re-written in matrix form:

$$\mathbb{E}\left\{Q_lA_\omega\right\}\begin{bmatrix}1\\1\end{bmatrix}=(C_{l,\omega}-D_{l,\omega})\begin{bmatrix}1\\0\end{bmatrix}+\boldsymbol{H}_\omega\begin{bmatrix}x_{l,\omega}^{(1)}\\x_{l,\omega}^{(2)}\end{bmatrix} \quad (66)$$

where $\boldsymbol{H}_\omega$ is defined:

$$\boldsymbol{H}_\omega\triangleq\begin{bmatrix}h_\omega^{(1)} & h_\omega^{(2)}\\ \lambda_\omega^{(1)} & \lambda_\omega^{(2)}\end{bmatrix}$$

Using definitions of $h_\omega^{(1)}$ and $h_\omega^{(2)}$ from (35) and (36), it is not difficult to show that the determinant of $\boldsymbol{H}_\omega$ is given by:

$$\det(\boldsymbol{H}_\omega)=(\lambda_\omega^{(2)}-\lambda_\omega^{(1)})\lambda_\omega(\beta_\omega+\delta_\omega)$$

This determinant is non-zero if and only if $\lambda_\omega^{(1)}\neq\lambda_\omega^{(2)}$, which is true because $\omega\in\tilde{\mathcal{L}}$. Thus, $\boldsymbol{H}_\omega^{-1}$ exists and is given by:

$$\boldsymbol{H}_\omega^{-1}=\frac{1}{(\lambda_\omega^{(2)}-\lambda_\omega^{(1)})\lambda_\omega(\beta_\omega+\delta_\omega)}\begin{bmatrix}\lambda_\omega^{(2)} & -h_\omega^{(2)}\\ -\lambda_\omega^{(1)} & h_\omega^{(1)}\end{bmatrix}$$

However, note from (26) that $x_{l,\omega}^{(1)}+x_{l,\omega}^{(2)}=\mathbb{E}\left\{Q_l\right\}$, and hence:

$$\begin{aligned}\mathbb{E}\left\{Q_l\right\} &= \begin{bmatrix}1 & 1\end{bmatrix}\boldsymbol{H}_\omega^{-1}\boldsymbol{H}_\omega\begin{bmatrix}x_{l,\omega}^{(1)}\\x_{l,\omega}^{(2)}\end{bmatrix}\\ &= \mathbb{E}\left\{Q_lA_\omega\right\}\begin{bmatrix}1 & 1\end{bmatrix}\boldsymbol{H}_\omega^{-1}\begin{bmatrix}1\\1\end{bmatrix}\\ &\quad -(C_{l,\omega}-D_{l,\omega})\begin{bmatrix}1 & 1\end{bmatrix}\boldsymbol{H}_\omega^{-1}\begin{bmatrix}1\\0\end{bmatrix}\end{aligned}$$

where the last line follows from (66). However, using definitions of $h_\omega^{(1)}$ and $h_\omega^{(2)}$ from (35) and (36), it is not difficult to show that:

$$\begin{aligned}\begin{bmatrix}1 & 1\end{bmatrix}\boldsymbol{H}_\omega^{-1}\begin{bmatrix}1\\1\end{bmatrix} &= \frac{(\lambda_\omega^{(2)}-\lambda_\omega^{(1)})+(h_\omega^{(1)}-h_\omega^{(2)})}{\det(\boldsymbol{H}_\omega)}\\ &= \frac{1}{\lambda_\omega}\end{aligned}$$

Similarly:

$$\begin{bmatrix}1 & 1\end{bmatrix}\boldsymbol{H}_\omega^{-1}\begin{bmatrix}1\\0\end{bmatrix}=\frac{1}{\lambda_\omega(\beta_\omega+\delta_\omega)}$$

Using these identities yields:

$$\mathbb{E}\left\{Q_l\right\}=\frac{\mathbb{E}\left\{Q_lA_\omega\right\}}{\lambda_\omega}-\frac{C_{l,\omega}-D_{l,\omega}}{\lambda_\omega(\beta_\omega+\delta_\omega)}$$

This proves Lemma 3.

Appendix D — Proof of Theorem 6

A. Time Averages $\overline{\gamma}_l$ and $\overline{r}_l$

From the queueing equation for $Y_l(t)$ (given in (44)), it is clear that for all t we have:

$$\frac{1}{t}\sum_{\tau=0}^{t-1}\gamma_l(\tau)-\frac{1}{t}\sum_{\tau=0}^{t-1}R_l(\tau)\leq\frac{Y_l(t)}{t}$$

However, we know from (47) that $Y_l(t)\leq V\eta_l+1$ for all t. Define $\overline{r}_l$ and $\overline{\gamma}_l$ as time average expectations of the $R_l(t)$ and $\gamma_l(t)$ processes, respectively. Thus:

$$\overline{\gamma}_l\leq\overline{r}_l\ \text{ for all } l\in\mathcal{L} \quad (67)$$

B. Computing The Lyapunov Drift

Define the Lyapunov function $L(\boldsymbol{Q}(t))$ as in (7). Define $\boldsymbol{Y}(t)$ as the vector of flow state (virtual queues), and define the combined Lyapunov function $\Psi(\boldsymbol{Q}(t), \boldsymbol{Y}(t))$ as follows:

$$\Psi(\boldsymbol{Q}(t), \boldsymbol{Y}(t)) = L(\boldsymbol{Q}(t)) + \frac{1}{2}\sum_{l\in\mathcal{L}} Y_l(t)^2$$

The combined Lyapunov drift $\Delta(t)$ is the sum of drift for the $\boldsymbol{Q}(t)$ and $\boldsymbol{Y}(t)$ terms. The drift for the $\boldsymbol{Q}(t)$ terms is the same as (22) with the $A_l(t)$ values replaced by $R_l(t)$, and the drift for $\boldsymbol{Y}(t)$ is found from (44) using a standard quadratic drift argument (see, for example, [6]). Following the framework of [6] for joint Lyapunov stability and performance optimization, our goal is to make decisions that minimize a bound for the expression: $\Delta(t) - V\sum_{l\in\mathcal{L}}\mathbb{E}\{g_l(\gamma_l(t))\}$. Omitting the drift computation details for brevity, we have:

$$\begin{aligned}\Delta(t) - V\sum_{l\in\mathcal{L}}\mathbb{E}\{g_l(\gamma_l(t))\} &\leq \mathbb{E}\{D_Q(t) + D_Y(t)\}\\ &\quad - \sum_{l\in\mathcal{L}}\mathbb{E}\{Q_l(t)\}\\ + \sum_{l\in\mathcal{L}}\mathbb{E}\left\{\mathbb{E}\left\{R_l(t)[\hat{Q}_{\mathcal{S}_l}(t) - Y_l(t)] \mid \boldsymbol{Q}(t), \boldsymbol{Y}(t)\right\}\right\}&\\ + \sum_{l\in\mathcal{L}}\mathbb{E}\{\mathbb{E}\{Y_l(t)\gamma_l(t) - Vg_l(\gamma_l(t)) \mid \boldsymbol{Q}(t), \boldsymbol{Y}(t)\}\}& \quad (68)\end{aligned}$$

where $\hat{R}_{\mathcal{S}_l}(t) \triangleq \sum_{\omega\in\mathcal{S}_l} R_\omega(t)$ and Lemma 4 is used, and where $D_Q(t)$ and $D_Y(t)$ are defined:

$$\begin{aligned} D_Q(t) &\triangleq \frac{1}{2}\sum_{l\in\mathcal{L}}\left[R_l(t)\hat{R}_{\mathcal{S}_l}(t) - 2\hat{R}_{\mathcal{S}_l}(t)\mu_l(t) + \mu_l(t)\right]\\ D_Y(t) &\triangleq \frac{1}{2}\sum_{l\in\mathcal{L}}\left[R_l(t)^2 + \gamma_l(t)^2\right]\end{aligned}$$

It is evident that the FLOW-MAXIMAL algorithm observes the queue values $\boldsymbol{Q}(t)$ and $\boldsymbol{Y}(t)$ every slot t, and chooses $R_l(t)$ and $\gamma_l(t)$ to minimize the inside (conditional) expectations in the last two terms on the right hand side of (68) over all alternative options. It follows that the above expression is less than or equal to the corresponding expression when the decisions $R_l(t)$ and $\gamma_l(t)$ (for slot t only) are replaced by $R_l^*(t)$ and $\gamma_l^*(t)$, being any alternative decisions that can be made on slot t that satisfy the constraints:

$$0 \leq R_l^*(t) \leq A_l(t) \quad , \quad 0 \leq \gamma_l^*(t) \leq 1 \ , \gamma_l^*(t) \in \mathcal{I}_\delta \qquad (69)$$

Plugging these expressions back into the final four terms on the right hand side of (68) and rearranging terms yields:

$$\begin{aligned}\Delta(t) - V\sum_{l\in\mathcal{L}}\mathbb{E}\{g_l(\gamma_l(t))\} &\leq \mathbb{E}\{D_Q(t) + D_Y(t)\}\\ - \sum_{l\in\mathcal{L}}\mathbb{E}\{Q_l(t)\} &+ \sum_{l\in\mathcal{L}}\sum_{\omega\in\mathcal{S}_l}\mathbb{E}\{Q_l(t)R_\omega^*(t)\}\\ - \sum_{l\in\mathcal{L}}\mathbb{E}\{Y_l(t)R_l^*(t)\} &+ \sum_{l\in\mathcal{L}}\mathbb{E}\{Y_l(t)\gamma_l^*(t)\}\\ &- V\sum_{l\in\mathcal{L}}\mathbb{E}\{g_l(\gamma_l^*(t))\} \qquad (70)\end{aligned}$$

We emphasize a subtle but important point in the above expression: The expectations on the right hand side involve queue backlogs $Q_l(t)$ and $Y_l(t)$ that have distributions that arise from having implemented the FLOW-MAXIMAL algorithm on every timeslot up to time t (so that these are queue backlogs that arise in the actual algorithm). However, the $R_l^*(t)$ and $\gamma_l^*(t)$ values are not actually used in the algorithm implementation, are potentially *different* from the FLOW-MAXIMAL decisions, and are considered as being implemented only on slot t for computation of the expectation.

C. Establishing the Congestion Bound

Consider now the particular decisions $R_l^*(t) = \gamma_l^*(t) = 0$ for all $l \in \mathcal{L}$, and note that these indeed satisfy the required constraints (69). The inequality (70) thus becomes:

$$\begin{aligned}\Delta(t) - V\sum_{l\in\mathcal{L}}\mathbb{E}\{g_l(\gamma_l(t))\} &\leq \mathbb{E}\{D_Q(t) + D_Y(t)\}\\ &\quad - \sum_{l\in\mathcal{L}}\mathbb{E}\{Q_l(t)\}\end{aligned}$$

Using this inequality for the drift $\Delta(t)$ directly in the Lyapunov drift Theorem (Theorem 1), for $f(t) = \sum_{l\in\mathcal{L}} Q_l(t)$ and $g(t) = V\sum_{l\in\mathcal{L}} g_l(\gamma_l(t)) + D_Q(t) + D_Y(t)$, yields:

$$\begin{aligned}\limsup_{t\to\infty}\frac{1}{t}\sum_{\tau=0}^{t-1}\sum_{l\in\mathcal{L}}\mathbb{E}\{Q_l(\tau)\} &\leq \overline{D}_Q + \overline{D}_Y\\ + V\limsup_{t\to\infty}\frac{1}{t}\sum_{\tau=0}^{t-1}\sum_{l\in\mathcal{L}}&\mathbb{E}\{g_l(\gamma_l(\tau))\}\end{aligned}$$

where $\overline{D}_Q$ and $\overline{D}_Y$ represent time average expectation of the $D_Q(t)$ and $D_Y(t)$ processes. Using the fact that $g_l(0) = 0$, and that η is the largest right derivative of any utility function, we have $g_l(\gamma) \leq \eta\gamma$ for all l. Hence, using the fact that steady state exists, we have:

$$\sum_{l\in\mathcal{L}}\mathbb{E}\{Q_l\} \leq \overline{D}_Q + \overline{D}_Y + V\eta\sum_{l\in\mathcal{L}}\overline{\gamma}_l \qquad (71)$$

where $\overline{\gamma}_l$ represents the time average expectation of the $\gamma_l(t)$ process. Using the fact that $\overline{\gamma}_l \leq \overline{r}_l$ (from (67)), it follows that $\sum_{l\in\mathcal{L}}\overline{\gamma}_l \leq \overline{r}_{tot}$. Further, the $\overline{D}_Q$ and $\overline{D}_Y$ averages can be bounded as follows:

$$\begin{aligned}\overline{D}_Y &\leq \frac{\overline{r}_{tot}}{2}[A_{max} + 1]\\ \overline{D}_Q &\leq \frac{1}{2}\sum_{l\in\mathcal{L}}\mathbb{E}\left\{A_l(t)\hat{A}_{\mathcal{S}_l}(t)\right\} + \frac{\overline{r}_{tot}}{2}\end{aligned}$$

Using these in (71) proves the average queue bound (49).

D. Establishing the Utility Bound

Without loss of generality, assume that $\lambda_l > 0$ for all $l \in \mathcal{L}$ (else, such a link with $\lambda_l = 0$ can be effectively removed from the set of links $\mathcal{L}$, as no data ever arrives for transmission over this link). Define (r_l^*) as the solution to the optimization problem (40)-(43), so that $\sum_{l\in\mathcal{L}} g_l(r_l^*) = g_\delta^*$. Note also that the constraints in (40)-(43) ensure that $r_l^* \leq \lambda_l$. For each link $l \in \mathcal{L}$, consider the particular decisions for $R_l^*(t)$ and $\gamma_l^*(t)$:

$$\begin{aligned}\gamma_l^*(t) &= r_l^*\\ R_l^*(t) &= \begin{cases} A_l(t) & \text{with probability } r_l^*/\lambda_l\\ 0 & \text{otherwise}\end{cases}\end{aligned}$$

Note that the policy $R_l^*(t)$ is randomized, and the randomized decision is made independently of everything else. These decisions indeed satisfy the required constraints (69). Hence, plugging into (70) yields:

$$\begin{aligned}\Delta(t) - V\sum_{l\in\mathcal{L}}\mathbb{E}\left\{g_l(\gamma_l(t))\right\} &\leq \mathbb{E}\left\{D_Q(t) + D_Y(t)\right\} \\ &- \sum_{l\in\mathcal{L}}\mathbb{E}\left\{Q_l(t)\right\} + \sum_{l\in\mathcal{L}}\sum_{\omega\in\mathcal{S}_l}\mathbb{E}\left\{Q_l(t)R_\omega^*(t)\right\} \\ &- \sum_{l\in\mathcal{L}}\mathbb{E}\left\{Y_l(t)R_l^*(t)\right\} + \sum_{l\in\mathcal{L}} r_l^*\mathbb{E}\left\{Y_l(t)\right\} - Vg_\delta^* \qquad (72)\end{aligned}$$

Note that:

$$\mathbb{E}\left\{Q_l(t)R_\omega^*(t) \mid Z_\omega(t) = m\right\} = \frac{r_\omega^*}{\lambda_\omega}\lambda_\omega^{(m)}\mathbb{E}\left\{Q_l(t) \mid Z_\omega(t) = m\right\}$$

$$\mathbb{E}\left\{Y_l(t)R_l^*(t) \mid Z_l(t) = m\right\} = \frac{r_l^*}{\lambda_l}\lambda_l^{(m)}\mathbb{E}\left\{Y_l(t) \mid Z_l(t) = m\right\}$$

Hence, similar to the linear-algebraic derivation in Section V, we can show that for any time t at which the system is in steady state (omitting details for brevity):

$$\mathbb{E}\left\{Q_l(t)R_\omega^*(t)\right\} \leq \mathbb{E}\left\{Q_l(t)\right\} r_\omega^* + \frac{\tilde{C}_{l,\omega}}{\beta_\omega + \delta_\omega} \qquad (73)$$

$$\mathbb{E}\left\{Y_l(t)R_l^*(t)\right\} \geq \mathbb{E}\left\{Y_l(t)\right\} r_l^* - \frac{\tilde{C}_{l,l}}{\beta_l + \delta_l} \qquad (74)$$

where $\tilde{C}_{l,\omega}$ is defined: $\tilde{C}_{l,\omega}\triangleq\mathbb{E}\left\{R_l(t-1)R_\omega^*(t)\right\}$ if $\omega \in \tilde{\mathcal{L}}$, and 0 else. Note that $\tilde{C}_{l,\omega}$ is defined using the actual decision $R_l(t-1)$ implemented on slot $t-1$ and the alternative decision $R_\omega^*(t)$ (not implemented) for slot t, and satisfies:

$$\tilde{C}_{l,\omega} \leq \mathbb{E}\left\{A_l(t-1)R_\omega^*(t)\right\} = (r_\omega^*/\lambda_\omega)\mathbb{E}\left\{A_l(t-1)A_\omega(t)\right\}$$

Plugging (74) and (73) into (72) yields:

$$\begin{aligned}\Delta(t) - V\sum_{l\in\mathcal{L}}\mathbb{E}\left\{g_l(\gamma_l(t))\right\} &\leq \mathbb{E}\left\{D_Q(t) + D_Y(t)\right\} \\ &+ \sum_{l\in\mathcal{L}}\sum_{\omega\in\mathcal{S}_l\cap\tilde{\mathcal{L}}}\frac{\tilde{C}_{l,\omega}}{\beta_\omega + \delta_\omega} + \sum_{l\in\tilde{\mathcal{L}}}\frac{\tilde{C}_{l,l}}{\beta_l + \delta_l} - Vg_\delta^*\end{aligned}$$

where we have used the fact that $\sum_{\omega\in\mathcal{S}_l} r_\omega^* \leq 1$. Using the above drift expression in the Lyapunov drift Theorem (Theorem 1) yields:

$$\limsup_{t\rightarrow\infty}\frac{1}{t}\sum_{\tau=0}^{t-1}\sum_{l\in\mathcal{L}}\mathbb{E}\left\{g_l(\gamma_l(\tau))\right\} \geq g_\delta^* - F/V \qquad (75)$$

where F is defined as in (51). By concavity of $g_l(r)$, Jensen's inequality yields:

$$\frac{1}{t}\sum_{\tau=0}^{t-1}\mathbb{E}\left\{g_l(\gamma_l(\tau))\right\} \leq g_l\left(\frac{1}{t}\sum_{\tau=0}^{t-1}\mathbb{E}\left\{\gamma_l(\tau)\right\}\right)$$

Taking limits of the above inequality and noting that the time average expectation $\overline{\gamma}_l$ exists and satisfies $\overline{\gamma}_l \leq \overline{r}_l$ for each $l \in \mathcal{L}$ (by (67)), we have:

$$\limsup_{t\rightarrow\infty}\frac{1}{t}\sum_{\tau=0}^{t-1}\sum_{l\in\mathcal{L}}\mathbb{E}\left\{g_l(\gamma_l(\tau))\right\} \leq \sum_{l\in\mathcal{L}} g_l(\overline{r}_l)$$

Using this in (75) proves Theorem 6.

References

[1] M. J. Neely. Delay analysis for maximal scheduling in wireless networks with bursty traffic. *Proc. IEEE INFOCOM*, April 2008.

[2] P. Chaporkar, K. Kar, X. Luo, and S. Sarkar. Throughput and fairness guarantees through maximal scheduling in wireless networks. *IEEE Trans. on Information Theory*, vol. 54, no. 2, pp. 572-594, Feb. 2008.

[3] X. Wu, R. Srikant, and J. R. Perkins. Scheduling efficiency of distributed greedy scheduling algorithms in wireless networks. *IEEE Transactions on Mobile Computing*, June 2007.

[4] X. Lin and N. B. Shroff. The impact of imperfect scheduling on cross-layer rate control in wireless networks. *Proc. IEEE INFOCOM*, 2005.

[5] L. Tassiulas and A. Ephremides. Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. *IEEE Transacations on Automatic Control*, vol. 37, no. 12, pp. 1936-1949, Dec. 1992.

[6] L. Georgiadis, M. J. Neely, and L. Tassiulas. Resource allocation and cross-layer control in wireless networks. *Foundations and Trends in Networking*, vol. 1, no. 1, pp. 1-149, 2006.

[7] D. Shah. Maximal matching scheduling is good enough. *Proc. IEEE Globecom*, Dec. 2003.

[8] S. Deb, D. Shah, and S. Shakkottai. Fast matching algorithms for repetitive optimization: An application to switch scheduling. *Proc. of 40th Annual Conference on Information Sciences and Systems (CISS), Princeton, NJ*, March 2006.

[9] J. G. Dai and B. Prabhakar. The throughput of data switches with and without speedup. *Proc. IEEE INFOCOM*, 2000.

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

[11] M. J. Neely. Order optimal delay for opportunistic scheduling in multi-user wireless uplinks and downlinks. *Proc. of Allerton Conf. on Communication, Control, and Computing (invited paper)*, Sept. 2006.

[12] M. J. Neely, E. Modiano, and Y.-S. Cheng. Logarithmic delay for $n \times n$ packet switches under the crossbar constraint. *IEEE Transactions on Networking*, vol. 15, no. 3, pp. 657-668, June 2007.

[13] L. Tassiulas. Scheduling and performance limits of networks with constantly changing topology. *IEEE Trans. on Inf. Theory*, May 1997.

[14] M. J. Neely, E. Modiano, and C. E Rohrs. Dynamic power allocation and routing for time varying wireless networks. *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 1, pp. 89-103, January 2005.

[15] S. Shakkottai, R. Srikant, and A. Stolyar. Pathwise optimality of the exponential scheduling rule for wireless channels. *Advances in Applied Probability*, vol. 36, no. 4, pp. 1021-1045, Dec. 2004.

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

[17] 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 switches. *Proc. IEEE INFOCOM*, 2001.

[18] M. J. Neely, E. Modiano, and C. E. Rohrs. Power allocation and routing in multi-beam satellites with time varying channels. *IEEE Transactions on Networking*, vol. 11, no. 1, pp. 138-152, Feb. 2003.

[19] S. Ross. *Stochastic Processes*. John Wiley & Sons, Inc., New York, 1996.

[20] M. J. Neely, E. Modiano, and C. Li. Fairness and optimal stochastic control for heterogeneous networks. *Proc. IEEE INFOCOM*, March 2005.