Delay Analysis of Scheduling Policies in Wireless Networks

Gagan Raj Gupta
School of Electrical and Computer Engineering
Purdue University, West Lafayette, IN, USA
Email: grgupta@purdue.edu

Ness B. Shroff
Departments of ECE and CSE
The Ohio State University, Columbus, OH, USA
Email: shroff@ece.osu.edu

***Abstract*—We consider a class of wireless networks with general interference constraints and heterogeneous transmission rates under single-hop traffic. The delay analysis of throughput optimal (queue length based) scheduling policies in such systems is extremely difficult due to complex correlations arising between the arrival, service and the queue length process. We use the underlying interference constraints to obtain a fundamental lower bound on the delay performance of any scheduling scheme for this system. We also present upper bounds for the performance of these networks operating under the well-known Maximum Weighted Matching (MWM) scheduling policy.**

I. Introduction

A large number of studies on wireless networks have been devoted to system stability and throughput maximization. These schemes are often called throughput-optimal scheduling schemes. The delay performance of these systems, however, has largely been an open problem. Our focus in this paper is to analyze the expected delay for this system. To that end, we will establish fundamental lower bound on the expected delay of any scheduling policy. We also derive an upper bound on the expected delay of a well-known and extensively-studied (e.g., [1], [3], [6], [7]) throughput-optimal scheme called the Maximum Weighted Matching (MWM).

We model the wireless network as a arbitrary graph $G = (V, E)$ with time varying links. Packets arrive in the system from independent exogenous sources and are stored in separate queues awaiting transmission. The links can have different capacities depending on the distance between the nodes, interference etc. Channel conditions on each link very independently every slot according to ON/OFF Bernoulli processes, so that a link l can transmit C_l packets when it is in the ON state and cannot transmit in the OFF state. Such ON/OFF channel states might arise from channel fluctuations or fading. Every time-slot, a network controller views the channel conditions and schedules a set of non-interfering links (matching). The set of all valid matchings in the system can be arbitrary, allowing for any interference model. Under the MWM scheme, the weights of each link is chosen as the product of its backlog with the capacity of the link. With this choice of weights, the maximum weighted matching is scheduled at every time slot.

The design of a delay optimal policy that achieves minimum possible average delay of packets in the network for a given routing matrix has proved to be very challenging. Except for a delay optimal scheduling scheme for the tandem queue under the node exclusive interference model derived in [14], no result is known for more general systems.

The analysis of scheduling policies is difficult because of correlations among mutually interfering queues. Moreover, throughput optimal algorithms like MWM use the queue length information while making the scheduling decisions. This results in complex interactions of arrival, service, and backlog processes and significantly complicates the analysis. The general research on the delay analysis of scheduling policies has progressed in the following main directions:

- *Heavy traffic regime using fluid models:* Fluid models have typically been used to either establish stability of the system or to study the workload process in the heavy traffic regime. It has been shown in [2], [13] that the MWM policy minimizes the workload process for a stochastic processing network in the heavy traffic regime.
- *Stochastic Bounds using Lyapunov drifts:* This method is developed in [4], [7], [11], [12] and is used to derive upper bounds on the average queue length for these systems. However, these results are order results and provide only a limited characterization of the delay of the system. For example, it has been shown in [12] that the maximal matching policies achieve $\mathcal{O}(1)$ delay for networks with single-hop traffic when the input load is in the reduced capacity region.
- *Large Deviations:* Large deviation results for cellular systems have been obtained in [8], [15], [17] to calculate queue-overflow probability. The analysis is much harder for the wireless network considered here, due to the complex interactions of the arrival, service, and backlog process.

In this paper we develop lower and upper bound on the average delay of a packet in a wireless network with single-hop traffic under a throughput-optimal scheme. A throughput optimal scheme can stabilize the system whenever there exists any other scheduling scheme which can stabilize the system.

The delay performance of any scheduling policy is primarily limited by the interference, which causes many bottlenecks to be formed in the network. We generalize the typical notion of a bottleneck. In our terminology, we define a bottleneck to be a set of links X such that no more than one of them can simultaneously transmit. We develop an efficient technique to reduce such bottlenecks to a single queue system fed by appropriate arrival processes which are simple functions of the exogenous arrival processes of the original network. The lower bound on the system-wide average delay of a packet is then computed by the analysis of these reduced systems and

978-1-4244-2941-7/08/$25.00 ©2008 IEEE

requires only the statistics of the exogenous arrival processes. Our idea of bottlenecks is similar to [5], which uses cliques in the conflict graph to characterize the capacity region of a wireless network.

We then construct a upper bound on the delay of a variant of MWM using the method of Lyapunov drifts which has been developed in [4], [7], [11], [12]. By designing an appropriate Lyapunov (potential) function we are able to ensure that the contribution of each queue to the drift (expected decrease in the potential) is proportional to the size of the queue. Using the fact that the matching computed by the MWM algorithm has the largest weight among the set of all possible matchings, we are able to establish an upper bound on the expected delay of the system. In the rest of the paper, we define the system model and subsequently develop the lower bound and the upper bound.

II. System Model

We consider a wireless network, G with N links denoted by set L. The capacity (maximum number of packets that can be transmitted in one slot) of link l is given by C_l. Let $s_l(t) \in \{ON, OFF\}$ represent the channel state of link l during time slot t. Assume that these channel states are i.i.d. over time-slots and independent across channels, and let p_l represent the probability of link l to be ON.

Each link has its own exogenous arrival stream $\{A_l(t)\}_{t=1}^{\infty}$. Each arrival stream is *i.i.d.* in time. Time is slotted. The distribution of the number of packets, $A_l(t)$, arriving to a link l in any given time slot t may be arbitrary but time invariant. Each packet has deterministic service time equal to one unit. Assume that the second moments, $\mathbf{E}[A_l^2]$, of the arrival processes are finite. Different input streams may be correlated with each other. Let $\mathbf{A}(t) = (A_1(t), \ldots, A_N(t))$ represent the vector of exogenous arrivals, where $A_l(t)$ is the number of packets that arrive to link l during time slot t (for $l \in 1, \ldots, N$). Let $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_N)$ represent the corresponding arrival rate vector.

The packets arriving at each link are queued. Let $Q_l(t)$ denote the queue length at link l. The queue length vector is denoted by $\mathbf{Q}(t) = (Q_l(t) : l = 1, 2, \ldots, N)$. A link can be activated in a time slot t only if the queue is non empty. We use the term activation (scheduling) of a link or a queue interchangeably in the paper. After service, each packet leaves the system. There is a slotted service structure. For each link l, the indicator function $I_l(t)$ indicates whether or not link l received service at time slot t. Note that

$$I_l(t) = \begin{cases} 1 & \text{if } Q_l(t) > 0 \text{ and } l \text{ is scheduled} \\ 0 & \text{otherwise} \end{cases} \tag{II.1}$$

The evolution of the queue is as follows,

$$\begin{aligned} Q_l(t+1) =& (Q_l(t) - I_l(t)C_l(t)\mathbf{1}_{\{s_l(t)=ON\}})^+ \\ &+ A_l(t), l = 1, .., N \end{aligned} \tag{II.2}$$

where

$$(x)^+ = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Define residual capacity r_l as follows.

$$r_l(t) = \begin{cases} C_l(t) - Q_l(t) & \text{if } Q_l(t) < C_l(t) \\ 0 & \text{otherwise} \end{cases}$$

Then, the queue evolution can be written as

$$\begin{aligned} Q_l(t+1) =& Q_l(t) - I_l(t)(C_l(t) - r_l(t))\mathbf{1}_{\{s_l(t)=ON\}} \\ &+ A_l(t), l = 1, .., N \end{aligned} \tag{II.3}$$

The vector of the scheduled queues is denoted by $\mathbf{I}(t) = (I_n(t)) : n = 1, ..N$. Because of interference, there are constraints on the combination of links that can be activated simultaneously. We allow these constraints to be arbitrary. $\mathbf{I}(t)$ is a valid activation vector if it satisfies these constraints. Let S be the collection of all activation vectors, $\mathbf{I}^j$. At each time-slot an activation vector $\mathbf{I}(t)$ is scheduled. A scheduling policy decides which activation vector is used in every time slot.

Let $\|\mathbf{Y}\|$ denote the Euclidean norm of vector $\mathbf{Y}$. The system is considered to be stable if $\lim_{t \to +\infty} \mathbf{E}[\sup \|\mathbf{Q}(t)\|]$ is bounded. If the system is stable then the throughput is the same as the arrival rates. A throughput vector $\boldsymbol{\lambda}$ is admissible if there is some scheduling policy under which the system is stable when the arrival rate vector is $\boldsymbol{\lambda}$. Let us denote by Λ the closure of the convex hull of the set of activation vectors, $\mathbf{I}^j$ and by C the interior of the convex hull. Note that Λ is a closed convex set. It has been shown in [6] that if each arrival process is *i.i.d.* in time, and the first two moments of all the arrival streams $\{A_l(t)\}_{t=1}^{\infty}$ are finite, then $\boldsymbol{\lambda} \in C$ is a necessary condition for a stabilizing scheduling policy to exist. It is also shown that the MWM policy, that chooses the maximum weighted activation vector (matching), stabilizes the system for any arrival rate satisfying the preceding condition.

MWM Scheduling Policy

$$\mathbf{I}(t) = \operatorname*{argmax}_{\mathbf{I}^j \in S} \sum_{i=1}^{N} Q_i C_i \mathbf{1}_{\{s_i=ON\}} I_i^j \tag{II.4}$$

where I_i^j is the i^{th} component of the j^{th} activation vector, $\mathbf{I}^j$, in set S.

Fig. 1. MWM Scheduling Policy

III. Lower Bound Analysis

In this section, we present our methodology to derive lower bounds on the average packet delay for a given multi-hop wireless network. The first step is to identify the bottlenecks in the system. We then explain how to lower bound the average delay of the packets in a given bottleneck. Finally, we present a greedy algorithm which takes as input, a system with possibly multiple bottlenecks, and returns a lower bound on the system-wide average packet delay.

Link interference causes certain bottlenecks to be formed in the system. Define a bottleneck to be be a set of links $X \subset L$ such that no more than one of its links can be scheduled simultaneously. Our idea of bottleneck is equivalent to identifying cliques in the conflict graph which was used by

[5] to estimate the capacity region of a given wireless network. We call these sets of links, *exclusive sets*.

We demonstrate our methodology to derive lower bounds on the average size of the queues corresponding to the links that belong to an exclusive set. Then by definition,

$$\sum_{i \in X} I_i \leq 1 \tag{III.5}$$

We define the weighted sum of arrival rates, λ_X corresponding to X as follows,

$$\lambda_X = \sum_{i \in X} \frac{\lambda_i}{C_i} \tag{III.6}$$

Similarly, A_X and $\boldsymbol{\mathcal{S}}_X$ are defined as follows,

$$A_X(t) = \sum_{i \in X} \frac{A_i}{C_i}(t) \tag{III.7}$$

$$\boldsymbol{\mathcal{S}}_X(t) = \sum_{i \in X} \frac{Q_i}{C_i}(t) \tag{III.8}$$

Reduced System: Consider a system with a single server and $A_X(t)$ as the input. Note that we allow the inputs to this system to be fractional. The server serves at most one packet from the queue whenever it is non-empty and there is at least one link state that is ON a time t. We denote the latter by $\mathbf{1}_{\{s_X(t)=ON\}}$. Let $\boldsymbol{\mathcal{Q}}_X(t)$ be the queue length of this system at time t. We define $R_X(t)$, the residual capacity of the reduced system as follows,

$$R_X(t) = \begin{cases} 1 - \boldsymbol{\mathcal{Q}}_X(t) & \text{if } Q_X(t) < 1 \\ 0 & \text{otherwise} \end{cases}$$

The queue evolution of the reduced system is given by the following equation.

$$\boldsymbol{\mathcal{Q}}_X(t+1) = \boldsymbol{\mathcal{Q}}_X(t) - \mathbf{1}_{\{\boldsymbol{\mathcal{Q}}_X(t)>0\}}\mathbf{1}_{\{s_X(t)=ON\}} + A_X(t) + R_X(t) \tag{III.9}$$

where $\mathbf{1}$ is the indicator function.

We now establish that at all times t, $\boldsymbol{\mathcal{Q}}_X(t)$ is smaller than $\boldsymbol{\mathcal{S}}_X(t)$.

Theorem 3.1: For an exclusive set X in the system, at any time T, $\boldsymbol{\mathcal{S}}_X$ under any scheduling policy is no smaller than that of the reduced system, i.e., $\boldsymbol{\mathcal{Q}}_X(T) \leq \boldsymbol{\mathcal{S}}_X(T)$.

Proof: Omitted for brevity.

$\blacksquare$

- The above analysis captures the combinatorial interference constraints and reduces the bottleneck to a G/D/1 system with appropriate inputs for the purpose of establishing lower bounds.
- We emphasize that $A_X(t)$ can be computed from Eq. (III.7) and considers only the exogenous inputs to the system. Furthermore, the lower bound on the expected delay can be computed using only the statistics of the exogenous arrival process and not their sample paths.

Using the above theorem it follows that,

$$\mathbf{E}[\boldsymbol{\mathcal{S}}_X] \geq \mathbf{E}[\boldsymbol{\mathcal{Q}}_X] \tag{III.10}$$

We now compute a lower bound on $\mathbf{E}[\boldsymbol{\mathcal{Q}}_X]$. Due to the lack of space, we omit the proof of the theorem.

Theorem 3.2: The expected queue length of the reduced system, $\mathbf{E}[\boldsymbol{\mathcal{Q}}_X] \geq \dfrac{\mathbf{E}[A_X^2] - 2\lambda_X^2 + \lambda_X}{2(1 - \prod_{i \in X}(1 - p_i)) - \lambda_X}$

We now present a greedy algorithm, Algorithm 1, which computes a lower bound on the average delay for a system containing multiple bottlenecks. The exclusive sets correspond to cliques in the conflict graph [5]. Let M be the largest number of links that interfere with a link $l \in L$. The time complexity to compute all the exclusive sets is exponential in M in the worst case.

The Algorithm 1 maintains a table $T(i)$ which indicates the number of times link i has been used in the bottleneck. The value of $T(i)$ is initialized to C_i. The algorithm proceeds by greedily searching for a bottleneck that yields the maximum lower bound. For each link in the chosen bottleneck, the value of $T(i)$ is decremented by 1 and the process is repeated until the table T has a non zero entry. Thus it decomposes the wireless network into several single queue systems. The average delay of the system can then be easily computed. Note that the decomposition obtained by the greedy algorithm is not the optimal decomposition. The optimal decomposition can alternately be obtained by using a dynamic programming approach with the cost of increased computation complexity.

Algorithm 1 Computing the Lower Bound

```
for $i = 1$ to $N$ do
    $T(i) \leftarrow C_i$
end for
$BOUND \leftarrow 0$
repeat
    Find the bottleneck which maximizes $\mathbf{E}[\boldsymbol{\mathcal{Q}}_X]$
    $BOUND \leftarrow BOUND + \mathbf{E}[\boldsymbol{\mathcal{Q}}_X]$
    for all $i \in X$ do
        $T(i) \leftarrow T(i) - 1$
    end for
until $\forall i, \quad T(i) = 0$
return $BOUND$
```

The lower bound may be loose on account of the following. We assume that the queueing in the bottlenecks is independent of each other, which may not be possible because of interference. Moreover, in the derivation of the lower bound by the reduction technique, we have neglected the non-empty queue constraints by grouping the arrivals into a single queue, and hence we underestimate the delay. Since the exclusive sets do not completely characterize the capacity region of the network, it may also be expected that if the input load is close to a boundary of the capacity region C, which is different from the boundaries generated by the exclusive sets, the lower bound may perform poorly. Thus, in certain cases, the delay of the system under MWM policy may be close to infinity while the lower bound is much smaller. This motivates the development of an upper bound for the system, which is tight in the sense that whenever the upper bound goes to infinity, the delay of the system under a throughput optimal policy also becomes infinite.

IV. Development of an Upper Bound

In this section, we analyze a class of Generalized Maximum Weighted Matching (GMWM($\mathbf{w}$)) policies, parameterized by weights w_i which is described in Figure 2. The MWM policy is a special case, where all the weights w_i are unity. We establish the following bounds on the sum of the expected queue lengths and the expected delay in the system.

GMWM Scheduling Policy

$$\mathbf{I}(t) = \operatorname*{argmax}_{\mathbf{I}^j \in S} \sum_{i=1}^{N} (w_i Q_i) C_i \mathbf{1}_{\{s_i = ON\}} I_i^j \qquad \text{(IV.11)}$$

where I_i^j is the i^{th} component of the j^{th} activation vector, $\mathbf{I}^j$, in set S and $w_i > 0$ are fixed constants.

Fig. 2. GMWM Scheduling Policy

Theorem 4.1: Given any input load vector $\boldsymbol{\lambda} \in C$ and any vector $\boldsymbol{\mu} \in C : \quad \forall i, \quad \mu_i > \lambda_i$, the following bound on the expectation of the sum of lengths of queues holds true in a system operating under the GMWM policy where the weights w_i are chosen as $w_i = \frac{1}{(\mu_i - \lambda_i)}$:

$$\sum_{i=1}^{N} \mathbf{E}[Q_i] \leq \sum_{i=1}^{N} \frac{(2C_i\lambda_i + \mathbf{Var}[A_i] - \lambda_i^2)}{2(\mu_i - \lambda_i)} \qquad \text{(IV.12)}$$

The total expected network delay, $\bar{D}$, satisfies:

$$\bar{D} \leq \sum_{i=1}^{N} \frac{(2C_i\lambda_i + \mathbf{Var}[A_i] - \lambda_i^2)}{2(\sum_{i=1}^{N} \lambda i)(\mu_i - \lambda_i)} \qquad \text{(IV.13)}$$

Proof: We prove the bound on sum of expected queue lengths in the appendix and the bound on delay follows by Little's law. ■

The above analysis naturally leads us to the question of which $\boldsymbol{\mu} > \boldsymbol{\lambda}$ should be selected in the capacity region C such that the upper bound is minimized. Intuitively this means that the distance between the load vector and the service process should be as large as possible. This can be formulated as an optimization problem to compute the value of $\boldsymbol{\mu}$ that minimizes the upper bound.

Upper Bounding Expected Delay

$$\text{Minimize} \sum_{i=1}^{N} \frac{(2C_i\lambda_i + \mathbf{Var}[A_i] - \lambda_i^2)}{2(\mu_i - \lambda_i)}$$

subject to $\boldsymbol{\mu} \in C$

Fig. 3. Optimization Problem for Minimizing the Upper Bound

The optimization problem in Figure 3 is convex because the objective function is convex and the capacity region is also convex, being a convex hull of the activation vectors. The formulation of the problem is very similar to the network utility maximization using convex optimization techniques (see [9], [10], [16]).

V. Conclusions

We have established a fundamental lower bound on the performance of a wireless system with single-hop traffic and general interference constraints. This result can be used to study the relative performance of any scheduling policy. We have presented analysis of the GMWM type of scheduling policies on the expected queue lengths and expected delay in the system. The GMWM policy analyzed in the paper, uses the information of the arrival rates to the links to achieve load balancing by assigning higher weights w_i to more congested links. Thus, we are able to obtain a sharper upper bound on the delay performance.

References

[1] M. Andrews and L. Zhang. Scheduling algorithms for multi-carrier wireless data systems. In *MOBICOM*, pages 3–14. ACM, 2007.

[2] D.Shah and D.J.Wischik. Optimal scheduling algorithms for input-queued switches. In *INFOCOM*, 2006.

[3] A. Eryilmaz, R. Srikant, and J. R. Perkins. Stable scheduling policies for fading wireless channels. *IEEE/ACM Trans. Netw.*, 13(2):411–424, 2005.

[4] L. Georgiadis, M. J. Neely, and L. Tassiulas. *Resource Allocation and Cross-Layer Control in Wireless Networks, Foundations and Trends in Networking*, volume 1. Now Publishers, 2006.

[5] K. Jain, J. Padhye, V. Padmanabhan, and L. Qiu. Impact of interference on multi-hop wireless network performance. In *MOBICOM*, 2003.

[6] T. Leandros and A. Ephremides. Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. *IEEE Trans. Aut. Contr.37*, 37(12):1936–1948, 1992.

[7] E. Leonardi, M. Mellia, F. Neri, and M. A. Marsan. On the stability of input-queued switches with speed-up. *IEEE/ACM Transactions on Networking*, 9(1), Feburary 2001.

[8] X. Lin. On characterizing the delay performance of wireless scheduling algorithms. In *44th Annual Allerton Conference on Communication, Control and Computing*, September 2006.

[9] X. Lin and N. B. Shroff. The impact of imperfect scheduling on cross-layer congestion control in wireless networks. *IEEE/ACM Transactions on Networking*, 14(2):302–315, April 2006.

[10] S. H. Low and D. E. Lapsley. Optimization flow control: basic algorithm and convergence. *IEEE/ACM Trans. Netw.*, 7(6):861–874, 1999.

[11] M. J. Neely. Order optimal delay for opportunistic scheduling in multi-user wireless uplinks and downlinks. In *44th Annual Allerton Conference on Communication, Control, and Computing*, September 2006.

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

[13] A. L. Stolyar. Maxweight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic. *Annals of Applied Probability*, Vol.14(No.1):pp.1–53., 2004.

[14] L. Tassiulas and A. Ephremides. Dynamic scheduling for minimum delay in tandem and parallel constrained queueing models. *nnals of Operation Research, Vol. 48, 333-355*, 1993.

[15] V. J. Venkataramanan and X. Lin. Structural properties of ldp for queue-length based wireless scheduling algorithms. In *45th Annual Allerton Conference on Communication, Control and Computing*, September 2007.

[16] X. Wang and K. Kar. Cross-layer rate control for end-to-end proportional fairness in wireless networks with random access. In *MOBIHOC '05*, pages 157–168, New York, NY, USA, 2005. ACM.

[17] L. Ying, R. Srikant, A. Eryilmaz, and G. E. Dullerud. A large deviations analysis of scheduling in wireless networks. *IEEE Transactions on Information Theory*, 52:5088–5098, 2006.

Appendix

Proof of Theorem 4.1 We first design an appropriate Lyapunov function.

$$V(\mathbf{Q}(t)) = \frac{1}{2}\sum_{i=1}^{N} w_i Q_i^2(t) \qquad \text{(A.14)}$$

Note that if all the weights w_i are chosen to be 1, this is exactly the quadratic Lyapunov function used in [6]. We begin with the calculation of the drift for any state **Q**(t).

$$\begin{aligned}
&\Delta(\mathbf{Q}(t)) \\
&= \frac{1}{2}\sum_{i=1}^{N} w_i \mathbf{E}[(Q_i(t+1) - Q_i(t))(Q_i(t+1) + Q_i(t))|\mathbf{Q}(t)] \\
&= \frac{1}{2}\sum_{i=1}^{N} w_i \mathbf{E}[(A_i(t) - (C_i - r_i(t))\mathbf{1}_{\{s_i(t)=ON\}} I_i(t)) \\
&\quad (2Q_i(t) + A_i(t) - (C_i - r_i(t))\mathbf{1}_{\{s_i(t)=ON\}} I_i(t))|\mathbf{Q}(t)] \\
&= \sum_{i=1}^{N} w_i \mathbf{E}[Q_i(t)(A_i(t) - (C_i - r_i(t))\mathbf{1}_{\{s_i(t)=ON\}} I_i(t))|\mathbf{Q}(t)] \\
&+ \frac{1}{2}\sum_{i=1}^{N} w_i \mathbf{E}[(A_i(t) - (C_i - r_i(t))\mathbf{1}_{\{s_i(t)=ON\}} I_i(t))^2|\mathbf{Q}(t)] \\
&= \sum_{i=1}^{N} w_i \mathbf{E}[Q_i(t)(A_i(t) - C_i\mathbf{1}_{\{s_i(t)=ON\}} I_i(t))|\mathbf{Q}(t)] \\
&+ \frac{1}{2}\sum_{i=1}^{N} w_i (\mathbf{E}[A_i^2(t) + (C_i^2 + r_i^2(t))\mathbf{1}_{\{s_i(t)=ON\}} I_i(t)|\mathbf{Q}(t)]) \\
&+ \sum_{i=1}^{N} w_i (\mathbf{E}[(Q_i(t) - C_i\mathbf{1}_{\{s_i(t)=ON\}} I_i(t)) \\
&\qquad\qquad r_i(t)\mathbf{1}_{\{s_i(t)=ON\}} I_i(t)|\mathbf{Q}(t)]) \\
&+ \sum_{i=1}^{N} w_i (\mathbf{E}[A_i]\mathbf{E}[(C_i - r_i(t))\mathbf{1}_{\{s_i(t)=ON\}} I_i(t)|\mathbf{Q}(t)]) \\
&= \sum_{i=1}^{N} w_i \mathbf{E}[Q_i(t)(A_i(t) - C_i\mathbf{1}_{\{s_i(t)=ON\}} I_i(t))|\mathbf{Q}(t)] \\
&+ \frac{1}{2}\sum_{i=1}^{N} w_i (\mathbf{E}[A_i^2(t) + (C_i^2 + r_i^2(t))\mathbf{1}_{\{s_i(t)=ON\}} I_i(t)|\mathbf{Q}(t)]) \\
&+ \sum_{i=1}^{N} w_i (-\mathbf{E}[r_i^2(t)\mathbf{1}_{\{s_i(t)=ON\}} I_i(t)|\mathbf{Q}(t)] + \mathbf{E}[A_i^2(t)]) \\
&= \sum_{i=1}^{N} w_i \mathbf{E}[Q_i(t)(A_i(t) - C_i\mathbf{1}_{\{s_i(t)=ON\}} I_i(t))|\mathbf{Q}(t)] \\
&+ \frac{1}{2}\sum_{i=1}^{N} w_i (\mathbf{E}[A_i^2(t)] - 2\mathbf{E}[A_i^2(t)]) \\
&+ \frac{1}{2}\sum_{i=1}^{N} w_i (\mathbf{E}[(C_i^2 - r_i^2(t))\mathbf{1}_{\{s_i(t)=ON\}} I_i(t)|\mathbf{Q}(t)])
\end{aligned} \qquad \text{(A.15)}$$

Note that $\mathbf{I}(t)$ is the activation vector chosen by the GMWM scheme at time-slot t. For any other activation vector $\mathbf{I}^* \in S$, the following holds true:

$$\begin{aligned}
&\sum_{i=1}^{N} w_i \mathbf{E}[C_i\mathbf{1}_{\{s_i(t)=ON\}} I_i^*(t) Q_i(t)|\mathbf{Q}(t)] \\
&\le \sum_{i=1}^{N} w_i \mathbf{E}[C_i\mathbf{1}_{\{s_i(t)=ON\}} I_i(t) Q_i(t)|\mathbf{Q}(t)].
\end{aligned}$$

Now, for a stationary randomized policy with rates $\boldsymbol{\mu} > \boldsymbol{\lambda}$, suppose the activation vector picked at time t is $\mathbf{M}(t)$. We define another scheduling policy $\mathbf{I}^*$ which schedules at time t, all the queues scheduled by $\mathbf{M}(t)$ except for those whose queues are empty. We define $\mathbf{I}^*$ as follows:

$$I_i^*(t) = \begin{cases} M_i(t) & \text{if } Q_i(t) > 0 \\ 0 & \text{if } Q_i(t) = 0 \end{cases}$$

Moreover, M_i is a stationary randomized policy and we have

$$\begin{aligned}
&\mathbf{E}[M_i] = \mu_i, \mu_i \ge \lambda_i \\
&\mathbf{E}[M_i(t)Q_i(t)|\mathbf{Q}(t)] = \mu_i Q_i(t).
\end{aligned}$$

By definition of $\mathbf{I}^*$,

$$\mathbf{E}[C_i\mathbf{1}_{\{s_i(t)=ON\}} I_i^*(t) Q_i(t)|\mathbf{Q}(t)] = \mathbf{E}[C_i\mathbf{1}_{\{s_i(t)=ON\}} M_i(t) Q_i(t)|\mathbf{Q}(t)].$$

Therefore,

$$\begin{aligned}
\Delta(\mathbf{Q}(t)) &\le \sum_{i=1}^{N} w_i(\lambda_i - \mu_i) Q_i(t) \\
&+ \frac{1}{2}\sum_{i=1}^{N} w_i (\mathbf{E}[A_i^2(t)] - 2\mathbf{E}[A_i^2(t)]) \\
&+ \frac{1}{2}\sum_{i=1}^{N} w_i (\mathbf{E}[(C_i^2 - r_i^2(t))\mathbf{1}_{\{s_i(t)=ON\}} I_i(t)|\mathbf{Q}(t)]).
\end{aligned}$$

Now we choose $w_i = \frac{1}{(\mu_i - \lambda_i)}$. We now use the Lyapunov Drift technique from [7] to obtain the following:

$$\begin{aligned}
&\limsup_{t\to\infty} \frac{1}{t}\sum_{\tau=0}^{t-1} \mathbf{E}[\sum_{i=1}^{N} Q_i(\tau)] \le \frac{1}{2}\sum_{i=1}^{N} w_i (\mathbf{E}[A_i^2(t)] - 2\mathbf{E}[A_i^2(t)]) \\
&+ \frac{1}{2}\sum_{i=1}^{N} w_i (\mathbf{E}[(C_i^2 - r_i^2(t))\mathbf{1}_{\{s_i(t)=ON\}} I_i(t)|\mathbf{Q}(t)])
\end{aligned} \qquad \text{(A.16)}$$

Finally, we simplify the last term in the above equation as follows,

$$\begin{aligned}
&\mathbf{E}[(C_i^2 - r_i^2(t))\mathbf{1}_{\{s_i(t)=ON\}} I_i(t)|\mathbf{Q}(t)] \\
&= \mathbf{E}[(C_i + r_i(t))(C_i - r_i(t))\mathbf{1}_{\{s_i(t)=ON\}} I_i(t)|\mathbf{Q}(t)] \\
&\le 2C_i\mathbf{E}[(C_i - r_i(t))\mathbf{1}_{\{s_i(t)=ON\}} I_i(t)|\mathbf{Q}(t)] = 2C_i\lambda_i.
\end{aligned} \qquad \text{(A.17)}$$

Thus we obtain the following bound on the sum of expected queue lengths in the system:

$$\sum_{i=1}^{N} \mathbf{E}[Q_i] \le \sum_{i=1}^{N} \frac{(2C_i\lambda_i + \mathbf{Var}[A_i] - \lambda_i^2)}{2(\mu_i - \lambda_i)}. \qquad \text{(A.18)}$$