Power efficient delay allocation in multihop wireless networks

Dinesh Rajan, Member, IEEE

Abstract—In this paper, we present delay allocation strategies that minimize the total transmit power in a multihop wireless network. The focus is on guaranteeing an end-to-end average delay bound for a single variable bit rate flow on a multihop fading channel. We first compute an analytical approximation for the transmit power required to send a variable bit rate source over a finite state fading channel. We then use this approximation to derive a low complexity and near optimal delay allocation method for multihop networks when the fading processes on the multiple hops are independent. Properties of the optimal delay allocation are also studied; in the special case of a Gaussian network, the optimal delay allocation strategy is completely characterized. The tradeoff between single hop and multihop transmission is studied under an end-to-end delay constraint.

Index Terms—Queuing, Scheduling, Multihop, Wireless network, Delay guarantees.

I. INTRODUCTION

The capacity of wireless ad-hoc networks has been an area of significant recent research and has been studied under different constraints on interference, maximum node range, node mobility, hybrid ad-hoc-infrastructure networks [1–8]; these results typically study the scaling of throughput with number of nodes. Scheduling and capacity in multihop networks with interference constraints has been considered by many recent researchers (see [9–12] and references therein).

The importance of incorporating traffic models in wireless physical layer design has been well recognized [13–15] and cross layer optimization has been widely investigated [16–22]. Numerous scheduling mechanisms have been proposed that explore the intricate relationship between packet delay, fairness, throughput and power. Scheduling in the power limited regime is considered in [23–26] using a linear model for the power consumption with increasing rate. A consequence of this linear model is that power savings (or equivalently throughput increases) with increasing delays are only possible in fading channels. On the other hand, a bandwidth limited regime and an exponential model (as suggested by Shannon’s capacity formulation or error rate analysis [27]) for power consumption with increasing number of packets transmitted has been used in [28–32]. A consequence of this exponential relationship is that power savings with increasing delays are possible both in fading channels and in Gaussian channels with bursty traffic sources.

The problem that we explore in this paper is the optimal allocation of total end-to-end delay at the different nodes of a multihop wireless network to optimize a desired function (e.g., summation) of the transmit powers at each of the nodes. A bandwidth limited system is studied using a standard distance-2 interference model. The proposed framework is also used to compare single-hop versus multi-hop transmissions under end-to-end delay constraints. One of the main distinguishing characteristics of this work over other related works [33] is the joint consideration of bursty source, fading channel, and end-to-end delay constraint.

The importance of the delay allocation problem is motivated with the following simple example. Consider the transmission of a time-varying source over a single hop fading channel. Intuitively, it is clear that increasing the queuing delay at the transmitter and waiting for good channel conditions results in improved performance (lower transmit power/higher throughput). Now, consider the same traffic being transmitted over a two-hop network with an end-to-end delay constraint. Clearly, increasing the delay at the first hop reduces the power at that hop but also reduces the available delay at the second hop thereby leading to increased power requirements at the second hop. One the other hand, increasing the delay at the second hop reduces the power requirement at the second hop at the cost of increased power at the first hop. Thus, there is a tradeoff between the delay that is allocated to the two hops and the power required at the transmit nodes; exploring this tradeoff is the main focus of this paper.

The importance of optimal delay allocation is also motivated by comparing the optimal multihop delay allocation scheme (defined in Section IV) with a scheme in which the available delay is uniformly allocated to each of the hops; This comparison is given in Table I for a simple 2-hop network. The sum of the transmit powers across the 2-hops is given for two different values of end-to-end delay bound. It can be seen that uniformly allocating the available delay to each hop results in over 1 dB loss in performance compared to the optimal allocation. We also show examples in Section IV where optimal allocation provides over 3 dB performance benefit.

The main contributions of this paper can be succinctly summarized as below:

- We construct an analytical approximation for the power required to transmit a bursty source over a finite state block fading channel under an average delay bound. This approximation, which is accurate for medium to high delays, clearly shows the dependence of the transmit power on the source burstiness, channel conditions and
delay.

- For a multihop Gaussian network, we characterize the optimal end-to-end delay allocation strategy that minimizes any linear, non-negative combination of the transmit power at each of the nodes. The optimal strategy allocates a delay of 1 time-slot to all but the first node; the remaining delay is given to the first node.

- For arbitrary fading ad-hoc networks, where each hop is modeled as a finite state block fading channel, we provide a low complexity delay allocation mechanism that has significantly better performance than a simple equal delay allocation strategy. This low complexity delay allocation is obtained by solving a convex optimization problem and the complexity of computing the optimal solution scales with the number of nodes.

- We study the tradeoff between single hop and multihop transmission under finite delay constraints. We consider a simple straight line topology and characterize the achievable power-delay performance for various hops between sender and receiver. Similar linear network topologies have been considered in [33–37]. The optimal number of hops between transmitter and receiver depends on the source distribution, channel fading statistics and the end-to-end delay constraint and is characterized in some special cases.

The results in this paper can be used for network capacity provisioning/planning, admission control or for designing rate-throttling mechanisms. Further, the proposed optimization framework can be easily modified to include upper bounds on provisioning/planning, admission control or for designing rate-scheduling in multihop networks. In Section IV-A the multihop hop fading channels and derive an analytical approximation on single hop scheduling and selected prior results are summarized.

With standard ad-hoc network routing algorithms. At each of the subsequent hops, instantaneous (1 or 2 bit information) channel knowledge at the transmitter is also required. Although considering routing protocols is beyond the scope of this paper, we briefly discuss the implementation of the proposed strategy in conjunction with standard ad-hoc network routing algorithms.

The rest of the paper is organized as follows. In Section II we describe the system under investigation. Section III focuses on single hop scheduling and selected prior results are summarized in Section III-A. In Section III-B, we analyze single hop fading channels and derive an analytical approximation between transmit power and delay. In Section IV, we explore scheduling in multihop networks. In Section IV-A the multihop problem is formulated and Sections IV-B and IV-C focus, respectively, on multihop Gaussian and fading networks. Finally, we summarize in Section V.

### II. Preliminaries and Problem Formulation

Consider an ad-hoc wireless network with nodes $N_1, N_2, \ldots, N_m$, which carries a single flow. Let nodes $N_1$ and $N_{m+1}$ be the source and destination, respectively, for the traffic (or flow). Further, assume that packets are routed via nodes $N_2, N_3, \ldots, N_m$. We assume that node $N_1$ is not capable of simultaneously transmitting data to node $N_{m+1}$ and receiving data from node $N_{m+1}$.

The traffic generated at source node $N_1$ is assumed to be bursty (time-varying) and is stored in a buffer of size $B_1$. Consider a time-slotted system and let the source produce packets at an average rate of $\lambda$ packets/time-slot. Between time $n-1$ and $n$, the source produces $a_n$ packets, each of fixed size $S$ bits, where $a_n$ has distribution $\mathcal{N}(a_n)$. At intermediate node $N_j$, the incoming packets are stored in a buffer of size $B_j$. In this paper, we assume that the buffers are large enough that buffer overflows do not occur.\(^1\) The number of packets in buffer $B_j$ at the beginning of the $n^{th}$ time-slot is denoted by $x_{n,j}$. The scheduler at node $N_j$ chooses $u_{n,j}$ packets for transmission at the beginning of the time-slot, and uses power $P_{n,j}$ for transmission. Since the length of the time-slot is fixed, the rate of transmission depends on the number of packets, $u_{n,j}$, selected for transmission. In this paper, we consider only distributed scheduling by which we mean that at each node $N_j$, the scheduler cannot choose $u_{n,j}, P_{n,j}$ based on instantaneous buffer/channel state at the other nodes.

#### A. Channel model

The received signal $Y_{n,j+1}$ at node $N_{j+1}$, depends on the transmitted signal $X_{n,j}$ at node $N_j$ and is given by,

$$Y_{n,j+1} = \frac{A_{n,j}}{d_{j,j+1}^\beta} X_{n,j} + \varepsilon_{n,j}, \quad (1)$$

where, $d_{j,j+1}$ is the distance between nodes $N_j$ and $N_{j+1}$, $\beta$ is the propagation loss coefficient (typical power loss factor with distance is between 2 and 4 and hence $\beta \in [1, 2]$), and $\varepsilon_{n,j}$ is the complex circularly symmetric additive white Gaussian noise with zero mean and variance $\sigma^2$. For simplicity, we set $\sigma^2 = 1$. We use two models for the channel between nodes $N_j$ and $N_{j+1}$: an AWGN channel and a

\(^1\) Analysis of packet outages or losses resulting for buffer overflows do not occur.
block fading channel. In the AWGN channel model, we set $A_{n,j} = A_0 n$. In the block fading channel model, the channel gain $A_{n,j} \in \{A_{ji}, i = 1, \ldots, n_{ch-states}\}$ is assumed to be constant over the period of a time-slot, i.e., the coherence interval of the channel is the same as the length of the time-slot. Further, the channel state at each hop is assumed to form a Markov chain with transition probabilities given by $q_{ji}$. The invariant distribution of the Markov chain, denoted $p_{ji}$, can be easily calculated from the transition probabilities $q_{ji}$. Recognize that $p_{ji}$ also represents the probability that the channel gain in the $j^{th}$ hop equals $A_{ji}$. We typically consider $n_{ch-states} = 2$, which corresponds to 1-bit of channel state information at the transmitter during every time-slot.

### B. Buffer and delay model

Recall that the output traffic at node $N_{j-1}$ during time-slot $n$ is the input traffic to node $N_j$. Thus, the buffer update is given by,

$$x_{n+1,j} = x_{n,j} + a_{n+1,j} - u_{n,j} = x_{n,j} + a_{n,j-1} - u_{n,j}, \forall j$$  \hspace{1cm} (2)

where, $a_{n,j}$ represents the input traffic at node $N_j$ during time-slot $n$ and for notational simplicity we let $u_{n,0} = a_n$. In this paper, the physical layer coding scheme is chosen to ensure that the selected packets are transmitted completely within one time-slot. At each node, the average queuing delay, $D_j$, is related to the average buffer length via Little’s theorem [42] and is given by,

$$D_j = \frac{1}{\lambda} E[x_{n,j}], \hspace{1cm} (3)$$

where $\lambda = E[a_n]$ is the average packet arrival rate. A natural constraint on $u_{n,j}$ is that $0 \leq u_{n,j} \leq x_{n,j}$. We assume that all packets that arrive at the beginning of time-slot $n$ can be transmitted in time-slot $(n)$ or later. At each hop, the smallest average delay of 1 time-slot is achieved when all arriving packets are transmitted in the same time-slot, i.e., $u_{n,j} = a_{n,j}$, which implies that $D_j = \frac{1}{\lambda} E[x_{n,j}] = 1$, assuming that the buffer is initially empty. The average end-to-end delay is given by $D_{te} = \sum_{j=1}^{m} D_j$. Using our convention, the minimum end-to-end delay, $D_{te}$, equals $m$, the number of hops.

### C. Interference model

We consider a standard distance-2 interference model in the network, i.e., while nodes $N_1$ and $N_2$ are in communication, nodes $N_3$ and $N_4$ cannot be in communication but nodes $N_4$ and $N_5$ can also be in simultaneous communication. The incorporation of a distance-2 interference ensures that the minimum end-to-end delay is actually greater than $m$. For instance, consider a simple 2-hop network with preassigned channel allocation as follows: node $N_1$ transmits during every even time-slot and node $N_2$ transmits during every odd time-slot. It can be easily shown that all packets arriving at node $N_1$ during even time-slots experience a minimum end-to-end delay of 2, whereas all packets that arrive at node $N_1$ during odd time-slots experience minimum end-to-end delay of 3. Thus, the minimum average end-to-end delay equals 2.5 time-slots. Clearly, such a preassigned channel allocation scheme only provides a lower bound on the channel contention delay.

### D. Scaled down equivalent discrete time model at each node

In this work, we consider a preassigned channel assignment similar to that described for the 2-hop network. In general, due to the distance-2 constraint, a node will acquire channel for transmission once every $k$ time-slots, where $k = \min(m, 3)$, in a periodic or round-robin manner. Consequently, we will consider an equivalent discrete time queuing model at each node, in which time-index is scaled down by a factor $k$. The $j^{th}$ time-slot in the equivalent queuing model at node $N_1$ corresponds to time-slot $kj$ in the original model. Similarly, the $j^{th}$ time-slot in the equivalent model at node $N_i$ corresponds to time-slot $kj + i - 1$ in the original model. This equivalence is depicted pictorially in Figure 1 for a simple 2-hop network.

Let $\tilde{x}_{i,j}$ denote the number of packets in the buffer at node $N_j$ during time-slot $i$. We use the notation that quantities with a $\tilde{}$ are the equivalents in the time-scaled model of the corresponding quantities in the original model.

The equivalent arrival process at node $N_1$ is now the sum of the original arrivals over $k$ consecutive time-slots, i.e. $\tilde{a}_{i,1} = \sum_{j=0}^{k-1} a_{k+i,j}$. The distribution of the effective arrival process $\{\tilde{a}\}$ is given by the $k \times k$ convolution of the distribution $r(a_n)$ and the average arrival rate equals $k\lambda$. The number of packet transmitted in time-slot $i$ is given by $\tilde{u}_{i,1} = u_{k+i,j}$. The effective delay $\tilde{D}_1$ can be easily computed as

$$D_1 = k(\tilde{D}_1 - 1) + 1 + (k - 1)/2, \hspace{1cm} (4)$$

where $\tilde{D}_1$ is the queuing delay at node $N_1$ given by $\tilde{D}_1 = \frac{\tilde{E}[\tilde{x}_{1,1}]}{k\lambda}$. The factor of $k$ in (4) accounts for the scaling in time and the additional term of $\frac{k-1}{2}$ is due to the fact that packet arrivals are possible even in the time-slots when node $N_1$ does not have access to the channel.

For all the subsequent nodes, the effective packet arrivals are given by $\tilde{a}_{i,j} = \tilde{a}_{i-j-1}$. The effective delay $\tilde{D}_i = k(\tilde{D}_i - 1) + 1$, where $\tilde{D}_i = \frac{\tilde{E}[\tilde{x}_{i,j}]}{k\lambda}$. Recognize that for nodes $N_i, i \neq 1$.

---

2. The careful reader will rightly point out that node $N_j$ might generate traffic on its own, which should be accounted for. However, as mentioned earlier for simplicity, we only consider a single flow.

3. The propagation delay in the wireless medium is assumed to be a constant and is not considered in this paper.

4. Such an interference model is commonly used to model 802.11 based networks which use a handshaking mechanism prior to data transfer.
packets arrive only once every \( k \) time-slots and hence there is no additive factor (like in (4)) in calculating \( D_{\text{ute}} \). The total end-to-end delay is thus given by

\[
D_{\text{ute}} = \sum_{i=1}^{m} D_i = k \sum_{i=1}^{m} \tilde{D}_i - (k - 1)m + \frac{k - 1}{2} \tag{5}
\]

With the new equivalent queuing model at each node, the effective fading channel transition probabilities are easily computed using the Chapman-Kolmogorov equations ([42], p.p. 259). Specifically,

\[
\tilde{q}_{j,i} = \frac{P_{r_i} \{ \tilde{A}_{k+1,j} = A_{j,i} \}}{ \tilde{A}_{j,i} } = \sum_{c=0}^{\infty} q_{j,i}^r \tilde{q}_{j,i,c}^{k-r}, r, k \geq 0. \tag{6}
\]

The corresponding invariant probability distribution \( \tilde{p}_{j,i} \) can be easily calculated using these transition probabilities.

At node \( N_i \), the scheduler now computes \( \tilde{u}_{j,i} \) based on \( \tilde{x}_{j,i} \) and \( \tilde{A}_{j,i} \) as explained in [31]. The transmit power at node \( N_i \) during time-slot \( n \) depends on the number of packets transmitted, the coding and modulation scheme used and the desired performance in terms of bit error rate (BER) or frame error rate (FER). In this paper, we assume reliable packet transmissions in the Shannon theoretic sense. Although such an assumption of reliability is only feasible asymptotically, it provides two benefits: a closed form expression can be used for required power and it also provides a lower bound on the power required by any practical transmission strategy. Moreover, advanced codes like LDPC and Turbo codes achieve performance very close to capacity with finite block lengths. Thus, we use the famous Gaussian capacity formulation [43] to derive the required power \( \tilde{P}_{n,j,i} \) at node \( N_j \) during time-slot \( n \) and channel state \( i \) as

\[
\tilde{P}_{n,j,i} = \frac{Q_{j,i}^{2\beta}}{|\tilde{A}_{j,i}|^2} \left( e^{S_{n,j,i}/T_c} - 1 \right), \tag{7}
\]

where \( T_c \) is the length of the time-slot. For simplicity, we set \( S/T_c = 1 \) in the rest of the paper. The average transmit power at node \( N_j \) is given by \( P_j = \frac{1}{k} E[\tilde{P}_{j,i}] \), since the channel is used by each transmitter only once every \( k \) time slots. The total network power is then given by

\[
\frac{1}{k} \sum_{j=1}^{m} \sum_{i=1}^{n_{\text{ch, states}}} \tilde{u}_{n,j,i} \tilde{P}_{n,j,i} \tag{8}
\]

where the expectation is over the distribution of \( u_{n,j,i} \).

III. SINGLE FLOW, ONE HOP CHANNEL

In this section, we focus on scheduling over a single hop wireless channel that forms the basis for the multihop analysis in the rest of the paper. We first summarize prior results on scheduling in Section III-A and then derive a new closed form approximate formula for transmit power in Section III-B. Numerical results are discussed in Section III-C.

![Fig. 2. Plot to demonstrate the accuracy of the proposed closed form approximation for power required versus delay in a single hop fading channel.](image-url)
B. Single hop fading channels: Analytical performance evaluation

In this section, we derive a simple closed form analytical approximation for the required transmit power at each hop as a function of the traffic, fading channel and delay bound. This approximation forms the basis for the proposed (Section IV) low complexity delay allocation strategy for multihop networks. Since all results in this section are derived for a single hop network, we use symbols from the original model rather than the equivalent model. The analytical approximation, which is derived in Appendix I based on a multivariate Gaussian distribution assumption on the traffic output of a scheduler, is given by,

\[
\hat{P}_j = \mathbb{E} \left[ \sum_{i=1}^{n_{ch \text{- states}}} p_{ji} P_{n,j,i} \right] \\
\approx \sum_{i=1}^{n_{ch \text{- states}}} \frac{d_{j,j+1}}{|A_{ji}|^2} \left( \frac{\sigma_{a_{ji}}(D_j)}{2} + \frac{\sigma_{u_{ji}}(D_j)}{2} - 1 \right) \tag{9}
\]

where, \(D_j\) is the delay, \(m_{ji}(D_j)\) is the mean transmission rate, \(\sigma_{a_{ji}}^2(D_j)\) is the variance of the transmission rates, \(|A_{ji}|^2\) represents the effective channel gain and \(p_{ji}\) is the probability of channel fading state \(i\) in the \(j\)th hop. In Appendix I, we also derive an approximate value for \(m_{ji}(D_j)\) and \(\sigma_{a_{ji}}^2(D_j)\) in terms of the mean arrival rate, variance of the arrival rate and the fading channel statistics. The accuracy of the approximation is numerically studied and compared with the power of the optimal scheduler in Figure 2. Recall that the optimal scheduler is computed using dynamic programming methods in [28, 31]. It is evident from the figure that the approximation is accurate for moderate to large delay values.

As noted before, the output traffic at node \(N_j\) is the input traffic to node \(N_{j+1}\). We define the effective channel variance, denoted by \(\sigma_{ch}^2\), as

\[
\sigma_{ch}^2 = \sum_{i=1}^{n_{ch \text{- states}}} p_{ji} m_{ji}^2(\infty) - \lambda^2. \quad \tag{10}
\]

This channel variance indicates how different the optimal transmission rates (equivalently gains) in the fading states are. The variance of the output traffic at node \(N_j\) can be derived (Appendix I) in terms of the variance of the input traffic to that node, as

\[
\sigma_{u_j}^2 = \frac{\sigma_{a_{j-1}}^2}{2\lambda D_j - 1} + \frac{2\lambda D_j}{2\lambda D_j - 1} \frac{(D_j - 1)^2}{D_j^2} \sigma_{ch}^2. \quad \tag{11}
\]

Thus, the variance of the output traffic is a convex combination of the variance of the input traffic and the effective channel variance scaled by \(\frac{(D_j - 1)^2}{D_j^2}\). Further, the variance of the output traffic could be lesser or greater than the variance of the input traffic. For the transmission of a constant rate source, the first term in (11) equals 0 and the output variance depends only on channel fading characteristics. In the case of an AWGN channel, the second term in (11) reduces to zero and the output variance only depends on input traffic variance and the delay. The following examples quantitatively illustrate the relationship between input and output traffic variances and delay.

C. Numerical results: Input vs Output traffic variances at a node

**Example 1:** Consider a two state fading channel with gains \(A_j \in \{1, 100\}\). Assume that the probabilities of being in each individual state is given by \(Pr(A_{n,j} = 1) = 0.75, Pr(A_{n,j} = 100) = 0.25\). In this case, the optimal transmission rates in the two fading states at infinite delay can be calculated using (36) as \(m_{j1}(\infty) = 2.74, m_{j2}(\infty) = 11.76\). For a sample traffic arrival with mean arrival rate, \(\lambda = 5\) and variance \(\sigma_a^2 = 10\), the ratio of the output variance to the input variance is given in Table II for various delays. Clearly the output traffic variance could be higher or lower than the input variance depending on the delay constraint. Asymptotically as \(D \to \infty\), the output traffic variance equals \(\sigma_a^2\), and thus the ratio of output to input traffic variance equals \(\frac{\sigma_b^2}{\sigma_a^2} = 1.52\).

**Example 2:** In this example, we let \(A_j \in \{1, 100\}\) and set \(Pr(A_{n,j} = 1) = 0.25, Pr(A_{n,j} = 100) = 0.75\). The optimal transmission rates in the two fading states at infinite delay are given by \(m_{j1}(\infty) = 0, m_{j2}(\infty) = 6.67\). For same traffic arrivals as in Example 1, the ratio of output variance to input variance is given in Table II. In this case, the output variance is strictly lesser than the input traffic variance but the output traffic variance increases with delay. Asymptotically as \(D \to \infty\), the ratio of output to input traffic variance equals \(\frac{\sigma_b^2}{\sigma_a^2} = 0.834\).
Example 3: In this example, we let $A_j = 1$, i.e., an AWGN channel, and assume the same traffic statistics as in the earlier examples. In this case too, the output variance is strictly less than the input traffic variance as illustrated in Table II. However, unlike Example 2, the output traffic variance decreases with delay and asymptotically reaches 0 as $D \to \infty$.

In summary, examples 1 and 2 show that the ratio of output to input variance increases with delay; in the first case, the supremum of the ratio is greater than one while in the second case the supremum is less than one. Example 3 shows that this ratio of output to input variance decreases with delay.

IV. MULTIHOP NETWORKS

In this section, we investigate the delay allocation problem in a multihop network.

A. Problem formulation

The optimization problem of interest is posed as follows:

$$\min \sum_{j=1}^{m} \frac{w_j}{k} \sum_{i=1}^{n_{ch-stat\_states}} \left( \tilde{\bar{p}}_{ji} \tilde{P}_{n,ji} \right)$$

(12)

where $\tilde{u}_{n,j,i}$, $\forall i,j,n$ are the variables of optimization, $\bar{D}$ is the bound on the end-to-end delay and $w_j$ is the weighting of the power at node $N_j$. A higher weight $w_j$ can be used for instance when battery resources at node $N_j$ are scarce. In this paper, for simplicity, we choose $w_j$ to be a constant for all $j$. In other words, we only consider the minimization of sum of transmit powers at the different nodes. Recognize that optimization problem (12) is extremely complex to solve due to the dependency of input traffic at one node on the output traffic at the previous node. The optimal results given in Table I are computed in the simple 2-hop case based on an exhaustive search and serves as a lower bound on the total transmit power required to support that traffic under delay constraint. For each possible value of $\bar{D}$ the optimal scheduler at node $N_1$ is computed; using the output traffic at node $N_1$ as the input to node $N_2$, the optimal scheduler at node $N_2$ is computed with delay $\bar{D} - \bar{D}_1$. The resulting sum power is compared with the sum power using all other possible delay allocation values and the minimum value is selected.

To make the problem tractable, we use a statistical characterization of the output traffic at each node. Specifically, we use the derived relationships (in (11)) for the mean and variance of output traffic at each node to determine the input traffic at each node. Such a characterization allows us to decouple the optimizing function into tractable functions. Using (9), the summation of the powers required in the entire network can be approximated as,

$$\tilde{P}_{net} = \sum_{j=1}^{m} \frac{w_j}{k} \tilde{P}_{j}$$

(13)

$$= \sum_{j=1}^{m} \sum_{i=1}^{n_{ch-stat\_states}} \frac{w_j}{k} A_{ji}^2 \left( e^{m_{ji}(\tilde{D}_j)} + \frac{\sigma_{ji}^2(\tilde{D}_j)}{2} \right) - 1$$

where we have used the quantities in the equivalent scaled down discrete model.

The optimization problem is then rewritten as follows:

$$\min \tilde{P}_{net}$$

(14)

$$\sum_{j=1}^{m} \tilde{D}_j - (k-1)m + \pi \leq \bar{D}, \tilde{D}_j \geq 1$$

Now, the variables of optimization are the delays $\tilde{D}_j$ allocated to the various nodes; conditioned on delay $\tilde{D}_j$ being allocated to node $N_j$, the optimal scheduler that minimizes the power at node $N_j$ is calculated using dynamic programming techniques [28, 31]. In Appendix III we shown that the optimizing function in (14) is convex. Further, the constraints in (14) are linear. Hence, optimization problem (14) is easily solved (even for large number of hops) using a numerical optimizer (e.g., the fmincon constrained optimization function in MATLAB). In some simple cases, analytical solutions to (14) are also possible and discussed in the next section. The performance of the end-to-end scheduler based on this approximation is given in Table I; clearly the proposed low complexity solution is near optimal.

It should be reiterated that the proposed allocation strategy requires knowledge of channel statistics at all hops and the input traffic distribution. One can conceive implementation strategies wherein such statistics can be gathered adaptively, for example using an on-demand routing algorithm like ad-hoc on demand distance vector (AODV) [47]. For instance, each node can append its channel statistics along with route request frames (RREQ) frames and the destination can then make the optimal delay allocation and notify the intermediate nodes of the delays at each hop using the route reply (RREP) packets.

To understand the relative impacts of the source and channel on delay allocation, we study in detail the following special cases: i) Gaussian network, bursty source, and ii) Fading network, bursty source.
B. Delay allocation: Gaussian Network and Bursty Source

It turns out that in a Gaussian network, i.e., channel between successive nodes $N_i$ and $N_{i+1}$ is a Gaussian channel, the optimal delay allocation strategy is quite simple and is characterized by the following Theorem.

Theorem 1: Consider a multihop wireless network with $m$-hops and let each hop of the network be modeled as a time-invariant additive Gaussian noise channel. The incoming traffic or traffic generated at node $N_1$ is bursty. This traffic is transmitted to destination node $N_{m+1}$ via nodes $N_2, N_3, \ldots, N_m$. There is an end-to-end delay constraint $D_{ste}$ for flow. Any linear non-negative combination of the transmit powers at nodes $N_1, N_2, \ldots, N_m$ is minimized by allocating delay $D_i = 1, i \neq 1$ and the remaining delay to node $N_1$.

Proof: The proof is by contradiction. The proof also uses the following property of scheduling over single hop Gaussian channels; the larger the delay allocated to a link, the smoother is the output traffic at that node and lower is the required transmit power [28, 31].

Let the “optimal” delay allocation at two successive hops be $\tilde{D}_i$ and $\tilde{D}_{i+1}$, where $\tilde{D}_{i+1} \neq 1$ ($\implies \tilde{D}_{i+1} > 1$) and let the corresponding transmit powers equal $\tilde{P}_i$ and $\tilde{P}_{i+1}$. We now construct a different delay allocation strategy, labeled ALT, at nodes $N_i$ and $N_{i+1}$ with delays $\tilde{D}'_i$ and $\tilde{D}'_{i+1}$ such that $\tilde{D}'_i + \tilde{D}'_{i+1} = \tilde{D}_i + \tilde{D}_{i+1}$. Further, we set $\tilde{D}'_{i+1} = 1$ and show that the total power of strategy ALT, is lower than the “optimal” delay allocation strategy.

Consider buffers $B_i$ and $B_{i+1}$ at nodes $N_i$ and $N_{i+1}$ respectively. Now, in strategy ALT, assume that both the buffers are at node $N_i$ with one preceding the other and the scheduler at node $N_{i+1}$ has delay of 1 time-slot as depicted in Figure 4. The ALT scheduler action at node $N_i$ includes the scheduler action of the “optimum” scheduler at nodes $N_i$ and $N_{i+1}$ with one minor difference; packets that arrive into the second buffer at node $N_i$ during time-slot $n$ using scheduler ALT, can be transmitted in time-slot $n$ or later. Clearly, with this new scheduling mechanism, the delay $\tilde{D}'_i$ at node $i$ equals $\tilde{D}_i + \tilde{D}_{i+1} - 1$ and delay $\tilde{D}'_{i+1} = 1$. Further, the output traffic at node $N_{i+1}$ using the new scheduler is the same as the “optimal” scheduler and hence the transmit power at node $N_{i+1}$ is unchanged, i.e., $\tilde{P}'_{i+1} = \tilde{P}_{i+1}$. However, the transmit power at node $N_i$ is reduced since it is allocated a larger delay $\tilde{D}'_i = \tilde{D}_i + \tilde{D}_{i+1} - 1 > \tilde{D}_i$, since $\tilde{D}_{i+1} > 1$. Thus, the total power of strategy ALT is lower than the power using the optimal strategy, which is a contradiction. Hence, the power is minimized by allocating a delay of 1-time slot to node $N_{i+1}$. By simple extension, we can prove that the optimal allocation of delay at all the hops after the first hop equals 1.

We now provide a qualitative justification for the theorem. For a given delay constraint, the average transmit power at each node increases with increasing burstiness (or variability) of incoming traffic. Moreover, in an AWGN channel, the variance of the output traffic strictly decreases with increasing delay [31]. Thus, the total power is minimized by allocating as much delay as possible to the first node without violating the end-to-end delay bound.

To further understand and characterize the behavior of the optimal delay allocation strategy, we consider the following cases. For simplicity, we consider a straight line topology with $m + 1$ uniformly spaced nodes and study the performance for various values of $m$. Let $d_0$ be the distance between nodes $N_1$ and $N_{m+1}$.

1) Equal delay versus optimal delay allocation: The sum power using optimal and equal delay allocation strategies is plotted in Figure 5a for a simple on-off traffic source. The input distribution is $r(a_n) = 0.5\delta(0) + 0.5\delta(10)$, i.e., in each time-slot either 0 or 10 packets are received. Using
both the strategies, the power decreases with increasing delay constraint, as expected. The significant savings in power using the optimal allocation scheme are clearly evident from the figure.

For the Gaussian network, we can also analytically approximate the gains of optimal delay allocation over uniform delay allocation strategy. It should be added that with both the delay allocation strategies, each node schedules its packets optimally to minimize the local transmission power given the local delay constraint at that node. Let $D_{eq,i}$ represent the delay at each of the nodes using the equal delay allocation strategy; then using (5) it is easy to show that $D_{eq,i} = 1 + \frac{1}{m} (D_{ete} - m - (k-1)/2) \forall i$, where $D_{ete}$ is the total end-to-end delay. In the optimal delay allocation case, $D_{opt,i} = 1, i > 1$ and thus, using (5) $D_{opt,1} = 1 + \frac{1}{m} (D_{ete} - m - (k-1)/2)$. The ratio of the power required using equal delay allocation to all hops versus optimal allocation to all hops can be calculated using (13) as,

$$\left(\frac{P_{eq}}{P_{opt}}\right) = \frac{\sum_{k=1}^{m} e^{\frac{x^2}{2(2kD_{eq,1}-1)^k}}}{\sum_{k=1}^{m} e^{\frac{x^2}{2(2kD_{opt,1}-1)(2k-1)(2k-1)}}}$$  \hspace{1cm} (15)$$

A plot of this ratio is given in Figure 6 for two different traffic arrival streams - one stream contains Ethernet traffic trace obtained from [48] and the other stream is an MPEG trace. Clearly substantial power savings are possible when the delay is allocated optimally across the multiple hops. It is also interesting to note that maximum gains arise only at medium delays, which is of high practical value. When $D_i = 1$ or when $\tilde{D} \rightarrow \infty$, the optimal allocation strategy only provides marginal gains over the equal delay allocation strategy.

2) Single hop versus Multihop transmission: The proposed framework can also be used to study the performance of varying number of hops between transmitter and receiver nodes. For a fixed $d_0$, we study the variation of $\tilde{P}_{net}$ versus $m$. The sum transmit power $\tilde{P}_{net}$, for various number of hops and delays is plotted in Figures 7a and b for two different arrival rates $\lambda$. In Figures 7a and b, the traffic arrivals have a uniform distribution over support sets $\{0, 1, 2\}$ and $\{0, \ldots, 8\}$, respectively. As expected for fixed $m$, the power decreases with delay. However, for a fixed end-to-end delay bound, the power versus $m$ exhibits interesting non-monotonic behavior. We now provide a qualitative explanation for behavior in Figures 7a and b.

Recognize that the power at each hop depends on the rate and distance to next hop in a non-linear way. First consider $m \leq 3$, $\Rightarrow k = m$. Now increasing $m$ from 1 to 3 has three consequences: i) The average rate of transmission in each node increases by $k$ and thus power increases as $e^{k\lambda}$, ii) The distance between nodes decreases and the power per hop behaves as $(\frac{d_0}{m})^{2\beta}$, and iii) The number of transmitters increases linearly with $m$ but at any given time only one transmitter is active. Thus, the third factor does not increase or decrease $\tilde{P}_{net}$ as $m$ goes from 1 to 3. For small $\lambda$, the reduction due to factor (ii) is larger than the increase due to factor (i) and hence the power decreases as $m$ increases from 1 to 3, a result confirmed by Figure 7a. For large $\lambda$ the reverse is true and the power decreases as $m$ increases from 1 to 3; Figure 7b confirms the result.

As the number of nodes continue to increase beyond 3, $k = 3$ stays constant. Again consider the three factors mentioned above: i) The average rate of transmission does not increase since $k$ is a constant and this term does not affect $\tilde{P}_{net}$. ii) The distance between nodes decreases as $m$ increases and the power decreases as $(\frac{d_0}{m})^{2\beta}$, iii) The number of transmitters increases linearly with $m$ but each transmitter is used only once every $k$ time-slots. Hence, the power at a node has to be scaled by a factor $\frac{m}{\lambda}$ to obtain the total network power. The third issue causes a linear increase in power with $m$ whereas the second issue causes a decrease in power that is of the order $m^{-2\beta}$. Since $\beta > 1$ the total power reduces with increasing number of hops when $m > 3$. Figures 7a and b confirm this qualitative analysis. A similar conclusion is stated in [33] where for large rates, single hop performance is superior to multihop performance without using any interference cancellation. Asymptotically as $m \rightarrow \infty$, the end-to-end delay $D_{ete} \rightarrow \infty$ and that case is studied next.

3) Large delay analysis: Asymptotically, for large delays, the scheduler completely smoothes the burstiness of the source and the output traffic depends only on the channel conditions. For a Gaussian network, each node transmits at constant rate $k\lambda$ and the transmit power at each hop is given by $\frac{1}{k} (\frac{d_0}{m})^{2\beta} (e^{k\lambda} - 1)$. The total power in the network is given by

$$\tilde{P}_{net} = m \frac{1}{k} \frac{1}{A_j^2} \frac{x}{2} \left(\frac{d_0}{m}\right)^{2\beta} (e^{k\lambda} - 1).$$  \hspace{1cm} (16)$$

The plot of sum power versus average arrival rate $\lambda$ is given in Figure 8 for various number of hops. It is clear from the plot that the number of hops that minimize the power

![Image of Figure 6](image-url)

Fig. 6. Plot of ratio of power in dB with equal delay allocation versus optimal delay allocation for two different traffic traces. A 4-hop wireless AWGN network is considered and hence smallest end-to-end delay equals 5.
depends on $\lambda$. The set of $\lambda$’s for which a 1-hop route results in lower power than a 2-hop route can be easily computed in closed form as: $\lambda < \log(2^{2\beta-1} - 1)$. Similar bounds can be computed to compare a $m$–hop and $m’$-hop route. Also, recognize that the total power monotonically decreases with $m$ for $m > 3$ and asymptotically reaches 0. However, the total delay increases monotonically with $m$ and thus the given data can be transmitted with arbitrarily small power $\epsilon \rightarrow 0$ only as $D \rightarrow \infty$.

C. Delay allocation: Fading network, bursty source

Analogous to the earlier case (Gaussian network, bursty source), even in this scenario, the transmit power decreases monotonically with increasing delay bound. However, the output traffic at a given node could have more or less burstiness than the input traffic at that node.\(^5\) Qualitatively, larger the variation in $|A_{ij}|^2$ at a given hop, the larger is the variance of the output traffic, i.e., the output is more bursty. As a result of this increased burstiness, the optimal delay allocation can be substantially different from case 1. For example, using a simple two state fading channel with gains 1 and 1000, and a two hop network, the optimal delay allocation strategy gives nearly 80% of the total delay to the second hop and the remaining 20% to the first hop.

Similar to the Gaussian network, we now study the following special cases in detail.

1) Equal vs optimal delay allocation: The plot of the power required using optimal and equal delay allocation strategies are given in Figure 5b. Simple on-off traffic is considered with either 10 packets or 0 packets arriving each time-slot. The average rate $\lambda$ equals 5 packets/time-slot and a 4-hop network is considered. The channel gains $A_j \in \{30\}, j > 1$ and $A_1 \in \{1, 300\}$, i.e. one of the hops is in deep fade and the other hops experience a Gaussian channel. The probability of being in deep fade equals 0.5. The figure clearly shows that additional power savings of around 1 dB are possible using the proposed delay allocation strategy as compared to the uniform delay allocation strategy. As in the Gaussian network, the maximum benefit due to delay allocation strategy is obtained for intermediate delay values; at the two extremal delays of $D_j = 1$ and $D_j \rightarrow \infty$, no additional gains are obtained beyond equal delay allocation.

2) Single hop vs multihop transmission: The plot of power versus delay for various number of hops is given in Figure 9. The average channel gain $A_j \in \{1, 300\}$, i.e., each of the

\(^5\) A similar observation on output entropy of a queue is made in [49].
hops experiences one deep fade state with probability 0.5. The traffic arrival distribution is considered to be uniform distribution over support set \( \{0 \ldots 4\} \) with average rate \( \lambda = 2 \). As expected, for a fixed \( m \), the power decreases monotonically as delay increases. However, the variation of power with \( m \) is non-monotonic for a fixed end-to-end delay bound. For large delays, data is only sent during the good fading state at each hop. In that case, using an analysis similar to the Gaussian case, one can see that larger \( m \) results in lower power. For smaller delays, the optimal number of hops depends on the actual \( D_{ete} \) and fading characteristics. No generalizations like Theorem 1 on the optimal allocation are possible.

For large \( \lambda \) as before, increasing number of hops from 1 to 3, results in increased power. Further, increase in \( m \) beyond 3 results in a decrease in power. Similarly for very small \( \lambda \), increasing \( m \) beyond 1 results in decreased power consumption. The plots in these cases are similar to Gaussian network and are not shown. The qualification of large and small \( \lambda \) depends on fading characteristics.

3) Large delay analysis: For a fading network, at asymptotically large delays, the output traffic at each node depends on the channel state \( A_j \) and equals \( m_{ji}(\infty) \) (36); note however that \( \lambda \) in (36) should be replaced by \( k\lambda \) since each hop is only used once every \( k \) time-slots. The total power in the network is given by

\[
P_{\text{net}} = m \frac{1}{k} \left( \frac{d_0}{m} \right) 2\beta \sum_{i=1}^{\eta_{\text{ch-states}}} \frac{(e^{m_{ji}(\infty)} - 1)}{|A_j|},
\]

assuming identical channel statistics across all hops. The behavior of power versus \( \lambda \) is similar to the Gaussian case (Figure 8) except for a different scaling factor \( \sum_{i} \frac{(e^{-m_{ji}(\infty)} - 1)}{|A_j|} \) instead of \( \frac{\lambda - 1}{\lambda} \) in (16).

4) Empirical properties: As discussed in the above cases, the optimal allocation of delay depends on the characteristics of the source and the fading channels at the different hops. Although no generalization like Theorem 1 can be made, we observed the following properties.

Empirically observed Property 1: Given an end-to-end delay \( D_{ete} \) across a \( m \)-hop wireless network, the variation of delay at each of the hops is an affine function of the total delay \( D_{ete} \). The slope and intercept of this affine function depends on the input traffic arrival statistics, and the fading characteristics of the different hops. This property is illustrated in Figure 10 for a simple 4-hop network. Since \( m = 4 \), the minimum \( D_{ete} = 5 \). The fading gains in the different hops are chosen at random and are given by \( A_1 \in \{300\}, A_2 \in \{3, 30\}, A_3 \in \{8, 20\}, A_4 \in \{5, 10\} \). Each of these states have equal probability. The input traffic has uniform distribution over support set \( \{0 \ldots 10\} \). This property could be used to derive closed form delay allocation strategies in future work.

Empirically observed property 2: We found that using the optimal allocation of delay at the various nodes, the spectra of the transmitted data rates across the various links were “matched” to one another. (e.g. Figure 11 shows the spectrum in the general case of fading channels and bursty source in which delay is split between the two links.) This critical observation leads us to conjecture that a deeper connection exists between scheduler design and spectral analysis. The ‘matching’ of sources and channels at the bit/signal level has been considered [50] to optimize performance. Our results indicate that similar ‘matchings’ between traffic arrivals and fading process occurs at the packet level. Unlocking this connection could lead to delay allocation strategies similar to the impedance matching approach and should be considered in future work.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a framework for studying the delay allocation problem in an ad-hoc wireless network. A closed form expression for the total required power in a
network is derived as a function of the delay allocation and approximation is exploited to find near optimal schedulers.

This work can be extended in many different ways. For instance, the effect of multiuser interference on delay allocation problem needs to considered in future work. The proposed framework can also be used to make routing decisions. Specifically, to compare between two different routes, the proposed framework can be used to compute the total power required in the two routes for a given end-to-end delay constraint. The total power can then be used as the metric to make routing decisions. Further, other channel allocation strategies beyond the temporal fair allocation used in this paper should be considered in future investigations.

APPENDIX I
DERIVATION OF APPROXIMATE CLOSED FORM EXPRESSION FOR POWER CONSUMPTION IN FADING CHANNELS WITH DELAY CONSTRAINTS
Recall that the buffer update is given by

\[ x_{n+1,j} = x_{n,j} + u_{n,j-1} - u_{n,j}, \tag{18} \]

Taking expectation on both sides of (18), we obtain,

\[ \mathbb{E}[u_{n,j-1}] = \mathbb{E}[u_{n,j}] = \mathbb{E}[a_n] = \lambda. \tag{19} \]

Squaring (18) and taking expectation on both sides, we get,

\[ \mathbb{E}[x_{n+1,j}^2] = \mathbb{E}[x_{n,j}^2] + \mathbb{E}[u_{n,1-j-1}^2] + \mathbb{E}[a_{n,j}^2] + 2 \lambda^2 + 2 \mathbb{E}[x_{n,j} u_{n,j}] \]

\[ \implies \sigma_{u,j}^2 + \sigma_{u,j}^2 + 2\lambda^2 D_j = 2\mathbb{E}[x_{n,j} u_{n,j}] \tag{20} \]

where the last step is obtained by using (3). We now approximate the scheduler action to be linear function of the buffer occupancy; a similar approximation is used for scheduling over an AWGN channel [31]. Denote by \( u_{n,j,i} \) the number of packets transmitted at time-slot \( n \) if the fading channel state in the \( j \)th hop equals \( A_{j,i} \). Then,

\[ u_{n,j,i} = \mu_{ji} x_{n,j} + \nu_{ji} \tag{21} \]

\[ \implies \mathbb{E}[u_{n,j,i}] = m_{ji}(D_j) = \rho_{ji} \lambda D_j + \nu_{ji}, \tag{22} \]

where, \( m_{ji}(D_j) \) represents the average number of packets transmitted in fading channel state \( A_{j,i} \) and \( \mu_{ji} \) and \( \nu_{ji} \) are constants which determine the number of packets to be transmitted in fading state \( i \) at node \( N_j \) [31]. Squaring (21) and taking expectation, we obtain,

\[ \mathbb{E}[u_{n,j,i}^2] = \mu_{ji}^2 \mathbb{E}[x_{n,j}^2] + \nu_{ji}^2 + 2\mu_{ji} \nu_{ji} \lambda D_j \tag{23} \]

Squaring (22) and subtracting from (23), the variance of the output traffic is given by,

\[ \sigma_{u,j,i}^2 = \mu_{ji}^2 \mathbb{E}[x_{n,j}^2] - \nu_{ji}^2 \lambda^2 D_j^2 = \mu_{ji}^2 \sigma_{x,j}^2 \tag{24} \]

Thus,

\[ \sigma_{u,j,i}^2 = \sigma_{x,j}^2 \nu_{ji}. \tag{25} \]

We now use the following heuristic functional form for selecting \( m_{ji}(D_j) \) as a function of delay,

\[ m_{ji}(D_j) = m_{ji}(\infty) + \lambda \frac{m_{ji}(\infty)}{D_j} = m_{ji}(\infty) \frac{D_j - 1}{D_j} + \frac{\lambda}{D_j}. \tag{26} \]

This heuristic is basically a linear interpolation for average transmission rate in each fading state, between \( \lambda \) and \( m_{ji}(\infty) \). The heuristic form is thus exact at the two extreme delays of \( 1 \) and \( \infty \). The optimal transmission rate, \( m_{ji}(\infty) \), in the various fading states, at infinite delay is calculated in Appendix II. Substituting from (21) into (20),

\[ \sigma_{u,j}^2 + \sigma_{u,j}^2 + 2\lambda^2 D_j = 2 \left\{ \sum_i p_{ji} \mu_{ji} \mathbb{E}[x_{n,j}^2] + \lambda D_j \sum_i p_{ji} \nu_{ji} \right\} \tag{27} \]

Now, we set \( \mu_{ji} = \frac{1}{\lambda D_j}, \forall i \) and \( \sigma_{u,j,i} = \sigma_{u,j,k}, \forall i, k \). Thus, \( \nu_{ji} = m_{ji}(D_j) - 1, \forall i \). Using these values for \( \mu_{ji} \) and \( \nu_{ji} \), along with (25), we obtain,

\[ \sigma_{u,j}^2 + \sigma_{u,j}^2 + 2\lambda^2 D_j = \ldots \]

\[ 2 \sum_i \left( \frac{p_{ji}}{\mu_{ji}} (\sigma_{u,j,i}^2 + \mu_{ji}^2 \lambda^2 D_j^2) + \lambda D_j p_{ji} m_{ji}(D_j) \right) - 2\lambda D_j \]

\[ = 2\lambda D_j (\sigma_{u,j,i}^2 + 1) + 2\lambda^2 D_j - 2\lambda D_j \quad \text{(for any } i) \]

\[ = 2\lambda D_j (\sigma_{u,j,i}^2 + \lambda) \tag{28} \]

Characterizing the output traffic variance
The variance of the traffic output at a node can be derived as,

\[ \mathbb{E}[u_{n,j}^2] = \sum_i p_{ji} \mathbb{E}[u_{n,j,i}^2] \tag{29} \]

\[ \implies \sigma_{u,j}^2 = \sum_i p_{ji} \left\{ \sigma_{u,j,i}^2 + m_{ji}(D_j)^2 \right\} - \lambda^2 \tag{30} \]

\[ = \sigma_{u,j,k}^2 + \sum_i p_{ji} m_{ji}(D_j)^2 - \lambda^2 \quad \text{(for any } k) \tag{31} \]

Substituting from (31) into (28), we obtain,

\[ \sigma_{u,j,k}^2 = \frac{1}{2\lambda D_j - 1} \left[ \sigma_{u,j}^2 - \lambda^2 + \sum_i p_{ji} m_{ji}(D_j)^2 \right] \]
\[
\begin{align*}
\frac{1}{2\lambda D_j} - 1 &= \left[ \frac{\sigma^2_{u_{j-1}}}{2 D_j} + \frac{(D_j - 1)^2}{D_j^2} \left( \sum_i p_{ji} m_{ji}(\infty)^2 - \lambda^2 \right) \right] \\
&= \frac{\sigma^2_{u_{j-1}}}{2 D_j} + \frac{2\lambda D_j}{2 D_j - 1} \left( 1 + \frac{\sigma^2_{c_h}}{2 D_j} \right) \\
\sigma^2_{u_j} &= \frac{\sigma^2_{u_{j-1}}}{2 D_j} + \frac{2\lambda D_j}{2 D_j - 1} \left( 1 + \frac{\sigma^2_{c_h}}{2 D_j} \right), \\
\end{align*}
\]  

Now, combining (31) and (32) we obtain the variance of the output traffic in terms of the variance of the input traffic as,

\[
\sigma^2_{u_j} = \frac{\sigma^2_{u_{j-1}}}{2 D_j} + \frac{2\lambda D_j}{2 D_j - 1} \left( 1 + \frac{\sigma^2_{c_h}}{2 D_j} \right) 
\]

where \(\sigma^2_{c_h}\) is defined in (10). The transmit power in each channel state is calculated by assuming that the traffic output in each channel state has a Gaussian distribution with mean \(m_{ji}(D_j)\) and variance \(\sigma^2_{n_{ji}}\). Thus,

\[
\mathbb{E}[P_{n_{ji}}] = P_{ji} = \sum_{i=1}^{n_{ch-states}} \frac{1}{\sqrt{2\pi \sigma^2_{u_{j-1}}}} 
\]

\[
\int_{-\infty}^{\infty} \frac{p_{ji} d^2 e^\beta}{|A_{ji}|^2} \left[ \left( e^{un_{ji}} - \lambda \right)^2 \left[ \sum_{i=1}^{n_{ch-states}} p_{ji} m_{ji}(\infty)^2 - \lambda \right] \right] du_{n_{ji}}, 
\]

\[
\int_{-\infty}^{\infty} \frac{p_{ji} d^2 e^\beta}{|A_{ji}|^2} \left[ \left( e^{un_{ji}} - \lambda \right)^2 \left[ \sum_{i=1}^{n_{ch-states}} p_{ji} m_{ji}(\infty)^2 - \lambda \right] \right] du_{n_{ji}}. 
\]

**APPENDIX II**

**OPTIMAL TRANSMISSION RATE IN EACH FAADING STATE AT INFINITE DELAY**

The average rate of transmission \(m_{ji}(\infty)\) in each channel fading state at infinite delay can be easily computed using standard water-filling techniques as follows. The total power over a fading channel is given by

\[
\sum_{i=1}^{n_{ch-states}} \left( e^{m_{ji}(\infty)} - 1 \right). 
\]

The rates \(m_{ji}(\infty)\) are computed to minimize the total power as follows:

\[
m_{ji}(\infty) = \arg \min_{\sum_{i=1}^{n_{ch-states}} p_{ji} m_{ji}(\infty)} \sum_{i=1}^{n_{ch-states}} p_{ji} m_{ji}(\infty)^2 - \lambda \left( e^{m_{ji}(\infty)} - 1 \right). 
\]

Using standard Lagrangian methods, we can easily show that

\[
m_{jk}(\infty) = \frac{1}{\sum_{i} p_{ji}} \left( \lambda + \sum_{i} 2p_{ji} \log \left( \frac{A_{jk}}{A_{ji}} \right) \right),
\]

where \(i\) is the summation over states for which \(0 \leq m_{ji}(\infty) \leq \sum_{i} p_{ji}\).

**APPENDIX III**

**CONVEXITY OF OPTIMIZING FUNCTION (14)**

We know that the sum of convex functions is a convex function and hence, it is sufficient to prove that each of the terms inside the summation in (14) is convex. Now consider the term

\[
\left( e^{m_{ji}(D_j)} + \frac{\sigma^2_{ji}(D_j)}{2} - 1 \right). 
\]

Since \(e^x\) is a convex nondecreasing function, it is sufficient to show that the term in exponent is a convex function (See [51] for rules on convexity of composite functions). Using (26) and (32) the term in the exponent can be equivalently rewritten in the form \(\sum D_j\), where \(a < 0\). Each of these terms is a convex function of \(D_j\) (see for e.g., [51]). Hence, the function is convex in \(D_j\)\forall j.


