Beyond Proportional Fairness: A Resource Biasing Framework for Shaping Throughput Profiles in Multihop Wireless Networks

Sumit Singh, Upamanyu Madhow and Elizabeth M. Belding
University of California, Santa Barbara
Email: {sumit@ece, madhow@ece, ebelding@cs}.ucsb.edu

Abstract—Throughput performance of multihop wireless networks is governed by how the network’s transport capacity (in bit-meters per second) is partitioned among different network flows. Max-min fair allocation leads to poor throughput performance for all flows because connections traversing a large number of hops consume a disproportionate share of resources. While proportional fair allocation provides a significant improvement, we point out here that there is a much richer space of resource allocation strategies for introducing a controlled bias against resource-intensive long connections in order to significantly improve the performance of shorter connections. We present an analytical model that gives insight into the impact of a particular resource allocation strategy on network performance, in a manner that captures the effect of finite network size and spatial traffic patterns. Our simulation results demonstrate that it is possible to provide significantly better performance to shorter connections than max-min fair or proportional fair resource allocations, with minimal impact on the performance of long connections, using mixed bias strategies blending “fair” allocations with a strong bias against long connections.

I. INTRODUCTION

We consider the problem of resource allocation in static multihop wireless networks. For a fixed link rate and uniform traffic distribution where each node randomly chooses a destination, the number of hops and hence the amount of network resources required per-connection scales with network size. This spatial traffic distribution results in obtainable per-connection throughput diminishing approximately as $O\left(\frac{1}{\sqrt{N}}\right)$ because the network transport capacity only increases as $O(\sqrt{AN})$ [1], where $N$ is the number of nodes and $A$ is the network deployment area. Implicit in this estimate of per-flow throughput, however, is the assumption that the resource allocation to each connection is throughput-fair across the network, which implies that connections that traverse a larger number of hops (longer connections) consume significantly more network resources than those traversing fewer hops (shorter connections) to achieve the same end-to-end throughput. If we sidestep this fundamental assumption by biasing network resource allocation against longer connections, then shorter connections can achieve significantly higher throughput at the cost of controlled performance degradation for longer connections. In this paper, we introduce a rich class of resource allocation strategies based on this design philosophy.

An existing strategy that falls within our framework is proportional fairness [2], in which, roughly speaking, a connection’s throughput is inversely proportional to the number of hops it traverses. As shown in our paper, however, it is possible to design the resource allocation strategy to choose from among a much larger class of flow throughput profiles (i.e., flow throughput versus the number of hops it traverses) while maintaining efficient network utilization. For example, it is possible to significantly improve the performance of shorter connections beyond that attained by proportional fairness, while minimal degradation in the performance of longer connections. Alternatively, the throughput for both short and long connections can be improved relative to proportional fairness, at the expense of slightly smaller throughput for flows traversing a moderate number of hops.

A key contribution of this paper, which enables the preceding flexibility in shaping throughput profiles, is the introduction of mixed-bias strategies, in which a fraction of the total available network resource is allocated via a resource allocation strategy that is strongly biased against long connections, with the remainder allocated in an unbiased (e.g., max-min fair) or mildly biased (e.g., proportional fair) manner. Note that such mixing is critical: a naïve approach that strongly biases against long connections can lead to poor performance in a large network with spatially uniform traffic.

Another contribution of this paper is an analytical model that provides quick estimates of the throughput profile achieved by a given resource allocation strategy. This provides a mechanism for the exploration of the rich design space of mixed bias strategies in trading off the performance seen by short and long connections. Underlying our analytical framework is a two-scale model for multihop wireless networks. At the global scale, we observe that wireless plays a fundamental role only in influencing the network topology: we are constrained to use short distance hops due to considerations of power and spatial reuse. At this scale, elementary flow calculations treating node bandwidth shares (or data transfer capacity) as bit pipes suffice to analyze the impact of various resource allocation strategies. While our approach is based on flow balance at a “typical” node in a large network, it is also refined to account for the effect of finite network size and spatial traffic patterns. The effective node data transfer capacities used in the preceding “wired-equivalent” global scale calculations are obtained from a detailed analysis of the wireless medium at the local scale, which takes into account contention among nodes.
Related Work: Most prior work on resource allocation is based on a utility function framework: a flow attaining rate $r$ obtains utility $U(r)$, where $U(\cdot)$ is concave, and the goal of resource allocation is to maximize the sum of the utilities over flows, subject to network resource constraints. Reference [3] proposes a class of utility functions that incorporates the commonly used resource allocation policies including max-min fairness [4], proportional fairness [2], potential delay minimization [5], and total throughput maximization. Our starting point is quite different, in that we explicitly assume that the resource allocation for a flow is proportional to a specific decreasing function of the number of hops traversed by a flow, and then compute the proportionality constant via flow balance. However, our biasing strategies appear to be consistent with a utility function framework, and we intend to explore this connection further in future work. Other relevant prior work includes a discussion of the tradeoff between network capacity and fairness for heterogeneous multihop wireless networks [6], which concludes with advocacy of proportional fairness as a reasonable compromise between rate efficiency and fairness, and also observes that max-min fairness yields poor performance because all flows obtain essentially the same rate as the minimum rate flow. Reference [7] presents a realization of proportional fairness based on end-to-end rate control for a connection based on the number of hops it traverses.

II. RESOURCE ALLOCATION FRAMEWORK

We first discuss the global scale analysis, which essentially involves flow balance in a wired-equivalent network whose capacities would be set by a local scale analysis. For a uniform traffic model, each node generates a persistent flow, choosing a destination at random from among all other nodes. Our framework also accommodates spatially localized traffic distributions in which the probability of choosing a destination depends on its number of hops from the source. The biased resource allocation problem is to assign bandwidth to each flow as a function of its resource requirements (e.g., number of hops it traverses) such that node data transfer capacity constraints are honored. Since it is intractable to identify bottleneck links for randomized traffic patterns, we apply an averaged analysis at the global scale to do flow balance for a “typical” node of the following form:

$$C \geq \sum_{h=1}^{h_{max}} n(h)c(h),$$  \hspace{1cm} (1)

where $C$ is the available data transfer capacity of a node (in bits per-second), $n(h)$ is the average number of $h$-hop connections traversing the node, $c(h)$ is the capacity assigned to $h$-hop flows, and $h_{max}$ is the maximum number of hops. For an $N$-node network, $h_{max} = O(N)$ for a linear topology, and $h_{max} = O(\sqrt{N})$ for a two-dimensional grid.

For a uniform traffic model, a typical node is more likely to see a longer flow (since such a flow traverses more hops), so that $n(h)$ increases with $h$. It therefore makes sense to make $c(h)$ a decreasing function of $h$, in order to ensure that the sum in (1) converges even as the network size (and hence $h_{max}$) becomes large. A throughput profile refers to the dependence of $c(h)$ with $h$. Our goal is to explore the design space of throughput profiles that are feasible, in terms of meeting the link balance condition (1). For simplicity of exposition, we first illustrate our framework with some quick global scale computations for a one-dimensional network in Section II-A. We then provide a more comprehensive analysis for a two-dimensional network which models the IEEE 802.11 network to be simulated.

A. Computations for a One-Dimensional Network

Consider a linear network with $N$ nodes (arranged as a ring to avoid edge effects) and $C = 1$. For uniform traffic, the probability of a connection initiated by any given node traversing $h$ hops is $p(h) = \frac{1}{N-1}$, $h = 1, ..., N - 1$, since any of the other $N - 1$ nodes in the network can be chosen with equal probability. The average number of $h$-hop connections traversing a node is $n(h) = hp(h)$, so that flow balance at a typical node becomes (assuming full utilization):

$$\sum_{h=1}^{N-1} hp(h)c(h) = 1$$  \hspace{1cm} (2)

Max-min fair allocation: For max-min fair allocation, the bandwidth allocated to an $h$-hop connection is independent of $h$, so that $c(h) = \lambda$. Solving (2), we obtain $c(h) = \lambda \approx \frac{2}{N}$. The sum throughput $T_{sum} = N \sum_{h} c(h)p(h)$, summed over all connections, is approximately 2.

Proportional fair allocation: Now consider the case when the bandwidth assigned to a connection is inversely proportional to the number of hops it traverses; i.e., $c(h) = \frac{\lambda}{h}$. Here $\lambda$ denotes the maximum allowed source application data rate (in bits per-second). This allocation can easily be shown to be proportional fair [7]. Plugging this into (2), we obtain that $\lambda \approx 1$, so that $c(h) \approx \frac{1}{h}$. The sum throughput $T_{sum} \approx \log N$. Thus, we have significantly improved both the sum throughput and the performance of shorter connections, with only a factor of two loss in the throughput obtained by the longest connections. One can show using typical link flow analysis that the throughput of a $h$-hop connection is approximately $\frac{1}{h}$, with the sum throughput of $\log N$, and $\lambda = O(1)$. Comparing with the max-min fair allocation, we realize that biasing against the long connections via proportional fairness results in a win-win strategy: it leads to far better performance for the shorter connections, while reducing the throughput attained by the long connections only by roughly a factor of two. The cumulative effect of this is the improvement of the total throughput by a factor of $\frac{\log N}{2}$ compared to the max-min fair allocation.

The pitfalls of naive biasing: Given the improved performance tradeoff offered by proportional fairness, it is natural to ask if we can get still larger improvements by biasing even more severely against the long connections. Consider a more severely biased allocation $c(h) = \frac{\lambda}{h^p}$. Substituting into (2), it can be shown that there is indeed a gain in sum throughput, but we require that $\lambda = O\left(\frac{N}{\log(N)}\right)$. 

322
That is, the maximum flow throughput must scale up with network size, which is clearly not compatible with the finite capacity of one. Basically, there are not enough short flows to take advantage of the transport capacity released by such a strong bias. If we now impose the constraint that $\lambda$ cannot scale up with $N$, we find (details are provided for a two-dimensional model later) that the network is poorly utilized.

**The promise of mixed biasing:** The problem with the preceding biasing strategy was that the short connections would have to send at too high a rate (more than the available link capacity) to take advantage of the bias. However, if we only have to send at too high a rate (more than the available link capacity) to take advantage of the bias, then we can eliminate this problem. For example, consider a mixed bias strategy of the form $c(h) = \lambda_1 + \frac{\lambda_2}{h^2}$; this mixes max-min fairness with a bias stronger than proportional fairness. Let us assume that we use a proportion $\alpha_1$ of the node data transfer capacity to implement max-min fairness, and a proportion $\alpha_2 = 1 - \alpha_1$ to implement the stronger bias such that flow balance decomposes into two separate equations:

$$\sum_{h=1}^{N-1} h p(h) \lambda_1 = \alpha_1, \quad \text{and} \quad \sum_{h=1}^{N-1} h p(h) \frac{\lambda_2}{h^2} = \alpha_2.$$ 

Now, by choosing $\alpha_2$ to scale down with $N$, we can prevent $\lambda_2$ from scaling up with $N$. Specifically, we can set $\alpha_2 = \frac{\log N}{N}$, which will result in $\lambda_2 \approx a$, which no longer scales up with $N$. The resulting throughput profile $c(h)$ is a convex combination of the throughput profiles from the two different strategies:

$$c(h) \approx \frac{a}{h^2} + \left( 1 - \frac{\log N}{N} \right) \frac{2}{N}.$$  

(3)

(Here we constrain $a \leq 1$ so that the maximum throughput does not exceed the node/link capacity of one). We note that (3) provides a very different throughput profile than either max-min fairness or proportional fairness. In particular, the throughput seen by long connections is almost as good as that of max-min fairness (and hence is a factor of two better than that obtained from proportional fairness), while the throughput seen by short connections does not scale down as a function of $N$ (and hence is better than max-min fairness). In short, mixed bias strategies open up a large design space not covered by existing resource allocation strategies. We shall explore this in more detail in our discussion of two-dimensional networks.

**B. Two-Dimensional Networks**

We now consider two-dimensional $N$ node regular grid topology networks where the inter-node distance is such that direct communication is only possible between immediate neighbors on the grid. The hop distance $h(i, j)$ between two nodes $i$ and $j$ equals the number of hops along the shortest path between the nodes. We consider a power-law traffic model [8], in which the probability that a node $i$ will communicate with a node $j$, hops away, is given by $p_{i,j} = \frac{a}{h^r}$, where $0 < a < 1$. Thus, the probability $p_H(h)$ that a pair of communicating nodes are $h$ hops apart is given by $p_H(h) = m(h)/(h^b \sum_{r=1}^{h_{\text{max}}} m(r))$, where $m(h)$ is the number of $h$-hop flows in the network. For $k > 0$, this traffic distribution models a spatially localized traffic pattern. We consider uniformly distributed traffic ($k = 0$) to study the effect of different resource biasing policies.

We focus on allocation of the available data transfer capacity $C$ of a typical network node using link flow analysis to obtain the expected per-connection throughput. The underlying medium access control (MAC) layer (e.g., IEEE 802.11) implements max-min fair bandwidth allocation among contending nodes. The bias weight function $w(h)$ is defined as $w(h) = \frac{1}{h^r}$, where the bias exponent $b \geq 0$ determines the degree of bias. Let $n(h)$ denote the average number of $h$-hop flows passing through or initiated by a typical network node. We have

$$C \geq \sum_{h=1}^{h_{\text{max}}} n(h) \frac{\lambda}{h^r}. \quad (4)$$

Thus, the required $\lambda$ for maximum utilization of the available capacity is given by $\lambda = C/\sum_{h=1}^{h_{\text{max}}} n(h) / h^r$. We first consider a symmetric, regular grid (i.e., a 2-d torus) topology to avoid edge effects, and focus on how the required $\lambda$ scales with increasing network size as a function of the traffic model and the bias weight function. For a symmetric regular grid topology, $p_H(h) = 1/(h(k-1) \sum_{r=1}^{h_{\text{max}}} \frac{1}{r^{k+1}})$. Using $n(h) = h p_H(h)$, we have

$$\lambda = \frac{C}{\sum_{h=1}^{h_{\text{max}}} h p_H(h) / h^r} = \frac{C \sum_{h=1}^{h_{\text{max}}} 1/h^{(k-1)}}{\sum_{h=1}^{h_{\text{max}}} 1/h^{(k+b-2)}}. \quad (5)$$

For uniform traffic distribution ($k = 0$), we infer that with unbiased resource allocation ($b = 0$), $\lambda = O(1/\sqrt{N})$; for proportional fairness ($b = 1$), we have $\lambda = O(1)$; and for $b = 2$, $\lambda = O(\sqrt{N})$. Further calculations show that the required $\lambda$ grows with the network size $N$ for $b > 1$. However, since $\lambda$ cannot exceed the maximum stable data rate $s$ of a node, this means that after the network size increases to a certain limit for $b > 1$, a node cannot fully utilize its fair share of bandwidth because the stronger bias prevents long flows from using the leftover capacity from the short flows. As in the linear network case, we infer that with proportional fair resource allocation, the per-flow throughput will be higher than the unbiased resource allocation for any $N$ since the required $\lambda$ is $O(1)$. The preceding analysis also applies to infinite regular grid topologies where edge effects can be neglected.

We now consider finite, regular grid topologies where the contention at the network edges is less than the core because there are fewer competing nodes at the edges; also, with a uniform traffic model where every node randomly picks a destination among other network nodes, it is easy to show that the traffic volume increases towards the core of the network. Therefore, the network core acts as a bottleneck for network flows in most cases. Hence, we focus on the network-core to avoid offsets from the edge effects. We expect that the flow throughput results obtained from analysis of the network core are pessimistic estimates of the actual flow throughputs, especially for smaller networks (e.g., $N < 50$) where the edge effect dominates. Based on these observations, we modify (4) as follows:

$$C \geq \sum_{h=1}^{h_{\text{max}}} n_c(h) \min\left( \frac{\lambda}{h^r}, s \right), \quad (6)$$
where \( n_c(h) \) denotes the average number of \( h \)-hop flows passing through or initiated by a typical network-core node. We evaluate \( n_c(h) \) by averaging the values obtained from MatLab simulations of traffic instances over the network for 5000 seed values. Thus, we can calculate the maximum \( \lambda \) that satisfies (6) for a given bias exponent \( b \). This \( \lambda \) value can be used to estimate the average per-flow throughput \( T \) of the network as a function of the network size, traffic distribution and the resource allocation policy as follows:

\[
T = \sum_{h=1}^{h_{max}} p_H(h) \min(\frac{\lambda}{h^b}, s).
\]  

(7)

**Mixed Bias Resource Allocation Policy:** The underlying idea is to allocate a portion of the total available capacity at a node via a strongly biased policy, and allocate the rest employing a fairer policy. Assume that, of the total available node capacity \( C \), we allocate \( \alpha C \) via a resource allocation policy determined by weight function \( w_1(h) = \frac{1}{h^b} \), and allocate the remaining bandwidth via weight function \( w_2(h) = \frac{1}{h^2} \). We calculate the maximum \( \lambda_1 \) and \( \lambda_2 \) that satisfy the following inequalities:

\[
\alpha C \geq \sum_{h=1}^{h_{max}} n_c(h) \min(\frac{\lambda_1}{h^b}, s),
\]

(8)

\[
C \geq \sum_{h=1}^{h_{max}} n_c(h) \min(\frac{\lambda_1}{h^b}, \frac{\lambda_2}{h^2}, s).
\]

(9)

The average per-flow throughput in this case is given by:

\[
T = \sum_{h=1}^{h_{max}} p_H(h) \min(\frac{\lambda_1}{h^b} + \frac{\lambda_2}{h^2}, s).
\]

(10)

The parameters \( \alpha \), \( b_1 \) and \( b_2 \) allow us to span a large design space of resource allocation strategies, trading off throughput performance versus fairness.

**III. APPLICATION TO AN IEEE 802.11 NETWORK**

We now apply our resource allocation framework to an IEEE 802.11 multihop network to illustrate how to determine global-scale parameters from local-scale analysis, which takes into account physical layer technology and medium access control (MAC). We then employ simulations to determine how well the design prescriptions obtained by analysis work.

For our local-scale analysis, Bianchi’s saturation throughput analysis [9] is used to determine the data transfer capacity \( S \) of a contention region in an IEEE 802.11 network. Based on the observation that the IEEE 802.11 MAC protocol implements an approximation to max-min fair bandwidth allocation among contending neighbors [6], [10], a typical node’s share of single hop capacity can be estimated as \( C = S/n_c \), where \( n_c \) is the average number of contending nodes in a contention region at the network core. This value of \( C \) is now used in (6) in the global-scale analysis to estimate \( \lambda \). This is then plugged into (7) to obtain the average per-flow throughput. Fig. 1 illustrates the predicted average per-flow throughput for an example mixed bias allocation with \( b_1 = 5, b_2 = 1 \) (proportional fairness) and \( \alpha = 0.2 \), for the individual allocation policies in the mixture and also for max-min fair allocation. We observe that the mixed bias allocation has a higher average flow throughput as compared to proportional fair and max-min fair allocations. Evaluation for different \( \alpha \) values shows that increasing \( \alpha \) to a higher value emphasizes strong bias whereas decreasing \( \alpha \) to a smaller value such as 0.2 sways the throughput more towards the fairer allocation in the mixture.

We now employ simulations to obtain flow throughput profiles, and to evaluate the benefits of mixed bias strategies relative to the individual strategies in the mixture, as well as to proportional fairness. We use the QualNet Network Simulator [11] over regular grid topology networks with IEEE 802.11 MAC in basic access mode. We present our results with static routing in order to isolate the effects of routing overhead on throughput. The results are averaged over 20 different seeds. Our application model comprises CBR flows of packet size of 1000 bytes, with data rates determined analytically for a given biasing strategy, as in Section II-B and enforced using source data rate control. Destinations are chosen at random based on the uniform traffic model.

While we fix the offered rate based on the analytical model in order to compare different resource allocation strategies, an important topic for future work is to allow nodes to dynamically tune their flow rates while implementing a biased resource allocation strategy. This can be used to alleviate congestion, to exploit spatial reuse opportunities, and more broadly, to react to specific network topologies and traffic patterns. Embedding mixed-bias resource allocation strategies within a utility function framework may provide the flexibility needed to arrive at decen-
centralized adaptive implementations that achieve the preceding goals.

Fig. 2 illustrates the per-flow throughput performance for the example mixed bias allocation considered in Section II-B ($b_1 = 5$, $b_2 = 1$, $\alpha = 0.2$) and compares it with that of the single bias policies in the mixture and also max-min fair allocation. The results are consistent with the analytical prediction that the mixed bias allocation provides higher average flow throughput than max-min fair or proportional fair allocations. However, the average flow throughput values are lower, and decay faster with network size, than the analytical predictions shown in Fig. 1. This is because flows incur higher packet loss probabilities as they traverse a larger number of hops in a larger network, which results in an increase in the amount of wasted network resources. Another factor causing performance degradation that is not explicitly modeled in the analysis is the waste of medium access time due to the interaction between the IEEE 802.11 backoff mechanism and multihop packet forwarding [12].

Fig. 3 presents the scatterplots of individual flow throughputs as functions of the number of hops for a 144-node network over all simulation seeds. We consider a mixed bias strategy, mixing a strong bias with proportional fairness. The strategy leads to significantly higher throughput for shorter connections than proportional fairness, while incurring virtually no degradation for the longer connections. Other desirable mixed bias strategies that, for example, provide better performance for long connections than proportional fairness, can be obtained by mixing a strongly biased strategy with max-min fairness (e.g., Fig. 3(d)). Note that the strongly biased strategy shuts out the long connections, and is therefore not a feasible choice on its own. The choice of the capacity fraction $\alpha$ and the strength $b$ of the bias are other parameters at the disposal of the system designer to craft a desirable throughput profile.

IV. CONCLUSIONS

Our resource biasing framework opens up a rich design space for sharing transport capacity in multihop wireless networks that goes beyond existing paradigms such as max-min fairness and proportional fairness. The simple two-scale model provides quick performance estimates for different biasing strategies, providing a tool that allows system designers to tune parameters so as to shape throughput profiles while maintaining network efficiency. For 802.11 networks with fixed link speeds, the global scale, wired-equivalent model provides predictions of throughput profiles that match trends obtained by simulation. Important topics for future research include generalization of the framework to allow for multiple raw link speeds (which may complicate the interaction between the local-scale and global-scale models), and the design of efficient protocols for implementing these resource allocation strategies in a manner that adapts to the network topology and traffic pattern. In particular, embedding our strategies within a utility function framework may lead to simple decentralized implementations (e.g., via end-to-end rate control), analogous to the approach of Kelly [2].

ACKNOWLEDGMENTS

This work was supported in part by NSF Grant CCF-0431205, ONR Grant N00014-06-1-0066, NSF Career Award CNS-0347886 and NSF NeTS Award CNS-0435527.

REFERENCES