Resource Allocation for QoS Provisioning in Wireless Ad Hoc Networks

Mung Chiang[1], Daniel ONeill, David Julian[2] and Stephen Boyd
Electrical Engineering Department
Stanford University, CA 94305, USA

***Abstract-* For wireless ad hoc networks with multihop transmissions and Rayleigh fading, this paper maximizes the overall system throughput subject to QoS constraints on power, probability of outage, and data rates. Formulations are also given which minimize delay and optimize network resources in a wireless ad hoc network, where each link is shared by multiple streams of traffic from different QoS classes, and each traffic traverses many links. Although these optimal resource allocation problems are non-linear, they can be posed as geometric programs, which are transformed into convex optimizations, and can be solved globally and efficiently through interior-point methods.**

I. Introduction

Quality of service (QoS) has become an important issue in various kinds of data networks as some users are no longer satisfied with resource allocation based on service provisioning. Three major considerations of QoS support are bandwidth, delay and delivery guarantee. Voice, data, image, and video have different bandwidth requirements. Some classes of traffic, such as voice, are also much more sensitive to delay than background classes, such as data. QoS provisioning in a wireless network is a particularly difficult issue because physical layer problems; such as path loss, fading, and multipath; can make the communication links unreliable. This makes delivery guarantee a necessary feature in wireless ad hoc network QoS provisioning.

The challenge is to first prescribe a feasible QoS scheme for different classes of traffic, and then to optimize the use of network resources, mainly link capacities and transmitter powers, to satisfy QoS requirements for all classes while maximizing either the total network performance, or the QoS for the premier class. Within the wireless arena, ad hoc wireless networks pose additional technical challenges for QoS support. Unlike cellular wireless networks, ad hoc networks have no fixed infrastructure, and long range communications require multihop transmissions where a packet is routed through the network by other transceivers that act as relay nodes.

In sections 3 and 4, the following resource allocation problems for QoS provisioning in wireless ad hoc networks are solved:

$P1$ As a special case of resource allocation, the power control of user nodes are optimized to maximize the overall system throughput.

$P2$ Turning to the general cases, feasibility of service level agreement (SLA) terms are determined under network resource constraints.

$P3$ Taking delay into consideration, the total delay for the most time sensitive class of traffic is minimized by optimizing over powers, capacities, and SLA terms.

$P4$ Optimizing over powers, capacities, and SLA terms; the unused capacity of the network is maximized.

Because the mobile radio channel is fast varying and the number of user nodes is large, a fast and robust decision making algorithm is needed that accommodates a large number of variables for dynamic resource allocation to be feasible. Several ad hoc heuristics have been proposed to tackle the above problem, but they cannot meet all of the following criteria: optimality, speed, and the ability to accommodate a variety of constraints and a large number of variables. Solutions for wireless cellular networks have been proposed in [1]. This paper tackles the more difficult problems of resource allocation for wireless ad hoc networks.

A global solution to non-linear problems $P1$ to $P4$ is found by transforming the problems into convex optimization problems. Solution methods for these problems not only produce globally optimal solutions as efficiently as for linear programs, but also unambiguously determine feasibility. This second property is used to determine the feasibility and pricing scheme of admitting a new user with a defined QoS requirement.

II. Convex optimization and geometric programming

An efficient algorithm is needed in order to find the optimal solution to the above nonlinear problems in a high speed dynamic network with a large number of links and nodes. Formulations for $P1$ to $P4$ are provided that can be turned into convex optimizations, which have fast algorithms, such as the interior point method and the primal dual method, that make them as easy to solve as linear programs.

Convex optimization refers to minimizing a convex objective function over convex constraint sets. The particular type of convex optimization used in this paper is in the form of geometric programming [2]. First consider the following definition

Definition 1 *A* monomial *is a function $h : \mathcal{R}^n \rightarrow \mathcal{R}$, where the domain contains all real vectors with positive components:*

[1]This work was supported by Hertz Foundation Fellowship and Stanford Graduate Fellowship.
[2]This work was supported by NSF Grant CCR-9973134 and the Stanford Network Research Center.

$$h(\mathbf{x}) = cx_1^{a_1}x_2^{a_2}\cdots x_n^{a_n}, \quad c \geq 0 \; and \; a_i \in \mathcal{R} \tag{1}$$

A sum of monomials $f(\mathbf{x}) = \sum_k c_k x_1^{a_{1k}} x_2^{a_{2k}} \cdots x_n^{a_{nk}}$ is called a *posynomial.* Geometric programming is an optimization problem with the following form:

$$\begin{array}{lll} \text{minimize} & f_0(\mathbf{x}) & \\ \text{subject to} & f_i(\mathbf{x}) & \leq \; 1 \\ & h_j(\mathbf{x}) & = \; 1 \end{array} \tag{2}$$

where f_0 and f_i are posynomials and h_j are monomials. Geometric programming in the above form is not a convex optimization problem. However, with a change of variables: $y_i = \log x_i$ and $b_{ik} = \log c_{ik}$, the geometric programming form is put into convex form:

$$\begin{array}{llllll} \text{minimize} & p_0(\mathbf{y}) & = & \log \sum_k exp(\mathbf{a}_{0k}^T\mathbf{y} + b_{0k}) & & \\ \text{subject to} & p_i(\mathbf{y}) & = & \log \sum_k exp(\mathbf{a}_{ik}^T\mathbf{y} + b_{ik}) & \leq & 0 \\ & q_j(\mathbf{y}) & = & \mathbf{a}_j^T\mathbf{y} + b_j & = & 0 \end{array} \tag{3}$$

It can be verified that the log of sum of exponentials is a convex function [2]. Therefore p_i are convex functions and q_j are affine functions, and the problem is a convex optimization problem. Note that if all posynomials are in fact monomials, geometric programming becomes linear programming.

Convex optimization problems can be solved globally and efficiently through interior point and primal dual methods [3], with running times that usually scale to the square root of the problem size. These methods also offer duality interpretations, stability analyses and accommodate a variety of constraints. This paper shows how geometric programming can solve many versions of QoS provisioning problems in wireless ad hoc networks.

III. Power control for throughput optimization

First a special case of resource allocation optimization is formulated and solved in this section. The variables are user node powers and the objective is to maximize the overall system throughput in bps. The method in this paper explicitly takes into account the statistical variation of the received signal and the interference power.

A. Multi-hop network model and Rayleigh fading

Consider a wireless ad hoc network with n transmitter/receiver pairs, labeled $1, \ldots, n$, which transmit at powers $P_1 \ldots, P_n$. The power received from transmitter j, at receiver i is given by

$$G_{ij}F_{ij}P_j \tag{4}$$

The nonnegative number G_{ij} represents the path gain in the absence of fading from the j^{th} transmitter to the i^{th} receiver. G_{ij} can encompass path loss, shadowing, antenna gain, coding gain, and other factors.

The Rayleigh fading between each transmitter j and receiver i is given by F_{ij}. The F_{ij}'s are assumed to be independent and have unit mean. The G_{ij}'s are appropriately scaled to reflect variations from this assumption. The distribution of the received power between any pair of transmitter j and receiver i is exponential with mean value,

$$E\left[G_{ij}F_{ij}P_j\right] = G_{ij}P_j \tag{5}$$

The signal to interference ratio (SIR) for user i determines the quality of the received signal and is defined as

$$SIR_i = \frac{P_i G_{ii}}{\sum_{j\neq i}^{N} P_j G_{ij} + n_i} \tag{6}$$

Unlike SNR, SIR cannot be increased by simply increasing all users' transmitting powers since that would raise both the signal level and the interference level. This introduces a bit error floor and a QoS bottleneck.

B. Outage probability and system throughput

An outage is declared when the received SIR falls below a given threshold defined as SIR_{th}, often computed from a BER requirement. The outage probability associated with the i^{th} hop is given by

$$\begin{array}{lll} O_i & = & Pr(SIR_i \leq SIR_{th}) \\ & = & Pr(G_{ii}F_{ii}P_i \leq SIR_{th}\sum_{k\neq i} G_{ik}F_{ik}P_k) \end{array} \tag{7}$$

The outage probability can be expressed as [4]

$$O_i = 1 - \prod_{k\neq i} \frac{1}{1+\frac{SIR_{th}G_{ik}P_k}{G_{ii}P_i}} \tag{8}$$

Outage probability over a hop induces an outage probability over a path S

$$\begin{array}{lll} O_{pathS} & = & \text{Prob(outage along the path } S) \\ & = & 1 - \prod_{s\in S}(1 - O_i) \\ & = & 1 - \prod_{s\in S}\prod_{k\neq s} \frac{1}{(1+\frac{SIR_{th}G_{ik}P_k}{G_{ii}P_i})}. \end{array} \tag{9}$$

The constellation size M used by a hop can be closely approximated for MQAM modulation as follows

$$M = 1 + \frac{-1.5}{\ln(5BER)} SIR \tag{10}$$

where BER is the bit error rate. Defining $K = \frac{-1.5}{\ln(5BER)}$ leads to a monotonic expression for the data rate of the i^{th} hop as a function of the received SIR:

$$R_i = (1/T)\log_2(1 + KSIR_i) \tag{11}$$

The aggregate data rate for the system can then be written simply as the sum of terms of this form.

$$R_{system} = \sum_i R_i = (1/T)\log_2 \prod_i (1 + KSIR_i) \tag{12}$$

Overall system throughput is defined as the maximum aggregate data rate supportable by the system given a set of users with defined QoS.

C. Throughput optimization

Theorem 1 *(Optimize power for throughput maximization) The following problem of optimizing user node powers to maximize total network throughput is a convex optimization problem.*

$$\begin{aligned}
&\text{maximize} \quad R_{system} \\
&\text{subject to} \\
&R_i \geq R_{i,LB}, \quad \forall i \\
&1 - \prod_{k \neq i} \frac{1}{1+\frac{SIR_{th} G_{ik} P_k}{G_{ii} P_i}} \leq Pr_{out_i} \quad \forall i \\
&1 - \prod_{s \in S} \prod_{k \neq s} \frac{1}{(1+\frac{SIR_{th} G_{ik} P_k}{G_{ii} P_i})} \leq Pr_{out_path_s} \quad \forall S \\
&P_i \leq P_{max}
\end{aligned} \tag{13}$$

The objective function is the overall system throughput. It is optimized over the set of all feasible powers P_i. The first set of constraints are the data rates demanded by existing system users. The second set of constraints are the outage probability limitations demanded by users using single hops. The third set of constraints are the outage probability limitations for users using a multi-hop path. Lastly, the forth set of constraints are regulatory or system limitations on transmitter powers.

D. Simulation

A simple four node multi-hop network is considered in the following simulation. As shown in figure 1, the network consists of 4 nodes A, B, C, and D, and 4 links $1, 2, 3$, and 4. On link 1 node A is the transmitter and node B is the receiver; similarly, the transmitter and receiver nodes for each link are shown in the figure. Note that node A is also the transmitter on link 3, illustrating that a node can be a transmitter and/or receiver on many links. Nodes A and D as well as B and C are separated by a distance of 20m. By geometry the distance of each transmit path is $10\sqrt{2}$m.

For this simulation each link has a maximum transmit power of 1W. Alternatively, the power constraint could be placed on each node instead of each link by adding a constraint that $P_1 + P_3 \leq 1$W. All nodes are using MQAM modulation. The baseband bandwidth for each link is 10kHz, the minimum data rate for each link is 100bps for maintenance data, and the target BER is 10^{-3}. For the Rayleigh fading a probability of outage of $P_{out} = 0.1$ is required for an SIR threshold of 10dB. The gains for each link are computed as $G_{ij} = \frac{1}{200}\left[\frac{1}{d}\right]^4$ for $i \neq j$, and $G_{ii} = \left[\frac{1}{d}\right]^4$, with the exception of G_{12} and G_{34} which are set equal to 0 since it is assumed that a node does not transmit and receive at the same time. The factor of $\frac{1}{200}$ can be viewed as the spreading gain in a CDMA system, or the power falloff with frequency in a FDMA system. This gives the following gain matrix:

$$G = 10^{-4} \cdot \begin{bmatrix} 0.2500 & 0.0003 & 0.0012 & 0.0003 \\ 0 & 0.2500 & 0.0003 & 0.0012 \\ 0.0012 & 0.0003 & 0.2500 & 0.0003 \\ 0.0003 & 0.0012 & 0 & 0.2500 \end{bmatrix} \tag{14}$$

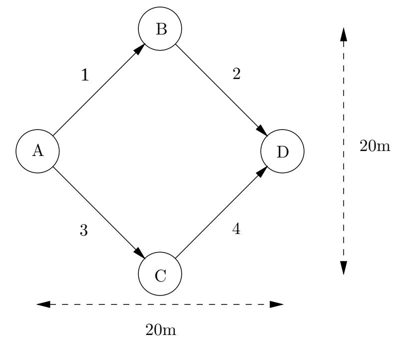

Fig. 1: Network Topology for Simulation

The maximum aggregate data rate, found using the geometric programming optimization method, is $R = 216.8$kbps, with $M = 42.8$QAM modulation for each link, $R_i = 54.2$kbps for each link, and $P_1 = P_3 = 0.709$W and $P_2 = P_4 = 1$W link transmit powers. The resulting $SIR = 21.7$dB on each link. The symmetry in modulation levels and SIR is due to the symmetries in the network topology, and not due to any explicit optimization constraint.

E. Admission control and pricing

The methodology used in this section can also be extended for other resource management purposes in a wireless ad hoc network. For example, a new user is admissible if his QoS requirements can be supported by the system without disturbing the existing QoS requirements of current users. In this model a user is admissible if a feasible solution of the problem in Theorem 1 exists after the new user's QoS constraints have been added. An infeasible solution is a definitive statement that this new user may not be added to the system without a change in required QoS.

Our model also supports several service-pricing approaches. The decrease in the overall system throughput associated with a new user can be used to estimate the incremental cost of supporting that new user, and should be proportional to the price that might be charged. A second approach is to calculate the effect of different levels of QoS for a given data rate. The effect on overall system throughput for each of the chosen different levels of QoS can then be used to set relative pricing for these levels. One example is to set pricing as a function of probability of outage for a given data rate.

IV. Resource allocation for delay and efficiency optimization

The more general cases of resource allocation are formulated and solved in this section. Optimization variables include powers, the number of packets in each traffic, bandwidth, delay and delivery guarantee required for each QoS class, and capacity for each link. Potential objective criteria include delay, unused capacity and SLA feasibility, in addition to the overall system throughput in the last section.

A. Problem formulations

Consider a network with J links with capacity of C_j packets per second for each link j. There are K classes of traffic with different QoS requirements to be transported over the network. For each QoS class k, the bandwidth required is b_k Hz, and the delay guarantee in the service level agreement (SLA) is $d_{k,UB}$ seconds. Also, a minimum probability of delivering the packet across the unreliable network is required in the SLA, denoted by $p_{k,LB}$. In this problem formulation the delay is the delay due to transmission time; propagation delay is ignored because it is constant for the optimization parameters. The more refined model with queueing delay is treated in the extension.

Similar to the last section, each stream of traffic from source s to destination d will traverse certain specific links as dictated by the particular routing protocol used for the network. Denote by K_j the set of traffic using link j and by J_k the set of links traversed by QoS class k. Denote by n_k the number of packets dynamically admitted in the k^{th} class of traffic.

In an ad hoc network each link may fail due to either power shut down of a user or deep fading that causes an outage. Therefore p_j, a real number between 0 and 1, is attached to each link as the probability that this link will be maintained during the transmission. By increasing transmitter power over a link j while keeping other parameters of the network constant, SIR of link j and therefore p_j can be increased.

As will be shown in the following problem formulations, with the above constraints on link capacity, bandwidth requirement, delay, and delivery probability guarantees, the problem is not a linear programming problem. However, these nonlinear optimizations can be turned into geometric programming problems and solved as efficiently as linear programs.

The first formulation is the following.

Theorem 2 *(SLA feasibility under network constraints) The following problem of testing SLA feasibility is a convex optimization problem.*

$$\begin{array}{lllll}
\text{minimize} & \text{No Objective Function} & & \\
\text{subject to} & \sum_{k\in K_j} b_k n_k & \leq & C_j,\ \forall j \\
& \sum_{j\in J_k}\left(\frac{\sum_{i\in K_j} n_i}{C_j}\right) & \leq & d_{k,UB},\ \forall k \\
& \prod_{j\in J_k} p_j & \geq & p_{k,LB},\ \forall k \\
& b_k n_k & \geq & R_k,\ \forall k \\
& b_{k^*} n_{k^*} & = & C_{j^*} \\
& \frac{n_{k^*}}{C_{j^*}} & = & d^*_{k,j} \\
& p_j & \leq & p_{j,UB} \\
& b_k, C_j, p_j, d_{k,UB}, p_{k,LB} & \geq & 0
\end{array} \quad (15)$$

No objective function is necessary since in this formulation only the feasibility of the SLA terms $p_j, d_{k,UB}$ and $p_{k,LB}$ is being tested. Alternatively, a cost function as the objective function could be used for relative pricing during admission control. Note that the first constraint is the link capacity constraint, the second one is the delay guarantee constraint and the third one the delivery probability constraint. The fourth constraint delivers a guaranteed data rate to each class of traffic. The fifth constraint makes room for SLA terms that give a class of traffic the sole right to traverse a link j^*. This could be for bandwidth requirements or for security reasons. The sixth constraint allows for SLA terms that specify not just an end to end total delay guarantee, but also an exact delay requirement for a particular traffic class k^* on a link j^*. The other constraints are positivity constraints on the variables, and upper bound constraints on p_j.

The following parameters are all potential optimization variables: $b_k, n_k, p_j, C_j, d_{k,UB}$ and $p_{k,LB}$. Variables $b_k, d_{k,UB}$ and $p_{k,LB}$ are terms in the SLA. The link capacities C_j and probability of maintaining a link p_j are network resources to be optimized over. Admission control is reflected in n_k.

The formulation is a non-linear optimization problem because optimization variables are multiplied together, such as $b_k n_k$ in the first constraint or the product of p_j in the third constraint, and appear in the denominator, such as C_j in the second and sixth constraints. However, all the inequality constraints are in posynomial form and all the equality constraints are in monomial form.

In the second formulation, the unused capacity of a particular link j_0 is maximized. The link may be a bottleneck link, or the most often traversed link in the network. [5].

Theorem 3 *(Unused capacity maximization) The following problem of maximizing the unused capacity under SLA and network constraints is a convex optimization problem.*

$$\begin{array}{lllll}
\text{maximize} & C_{j0} - \sum_{k\in K_{j_0}} b_k n_k & & \\
\text{subject to} & \sum_{k\in K_j} b_k n_k & \leq & C_j,\ \forall j \\
& \sum_{j\in J_k}\left(\frac{\sum_{i\in K_j} n_i}{C_j}\right) & \leq & d_{k,UB},\ \forall k \\
& \prod_{j\in J_k} p_j & \geq & p_{k,LB},\ \forall k \\
& b_k n_k & \geq & R_k,\ \forall k \\
& b_{k^*} n_{k^*} & = & C_{j^*} \\
& \frac{n_{k^*}}{C_{j^*}} & = & d^*_{k,j} \\
& p_j & \leq & p_{j,UB} \\
& b_k, C_j, p_j, d_{k,UB}, p_{k,LB} & \geq & 0
\end{array} \quad (16)$$

The objective function is to maximize unused capacity of a link j_0 by keeping the used capacity to the minimum under all network and QoS constraints. The constraints are the same as in Theorem 2.

In the third formulation, the total delay for a particular class of traffic is minimized.

Theorem 4 *(Delay Minimization) Delay minimization under SLA and network constraints is a convex optimization problem.*

$$
\begin{array}{lll}
\text{minimize} & \sum_{j \in J_{k_0}} \frac{\sum_{i \in K_j} n_i}{C_j} + \alpha \left(\sum_j C_j \right) & \\
\text{subject to} & \sum_{k \in K_j} b_k n_k & \leq \; C_j, \; \forall j \\
& \sum_{j \in J_k} \left(\frac{\sum_{i \in K_j} n_i}{C_j} \right) & \leq \; d_{k,UB}, \; \forall k \\
& \prod_{j \in J_k} p_j & \geq \; p_{k,LB}, \; \forall k \\
& b_k n_k & \geq \; R_k, \; \forall k \\
& b_{k^*} n_{k^*} & = \; C_{j^*} \\
& \frac{n_{k^*}}{C_{j^*}} & = \; d^*_{k,j} \\
& p_j & \leq \; p_{j,UB} \\
& b_k, C_j, p_j, d_{k,UB}, p_{k,LB} & \geq \; 0
\end{array}
\tag{17}
$$

Where α is the marginal tradeoff of capacity for delay. By increasing capacities available on each link at the relative cost α through bandwidth allocation or bandwidth leasing, delay of the most time sensitive QoS class can be decreased. Therefore, a weighted sum, parametrized by α, of the premier QoS class's delay and the cost of capacity provisioning is maximized.

B. Simulation

The following is a simulation for Theorem 4, which investigates the tradeoff between delay and cost of capacity.

For the network in Fig. 2 there are three classes of traffic. The first class is audio data traffic sent along path ABCD requiring a rate of 50 packets/second and a maximum delay of 0.2 seconds. The second class is also audio data traffic sent along path DFEA with the same rate and delay requirements. The third class of traffic is video data sent along path ABFD with a rate requirement of 250 packets/second. The goal is to minimize both the delay of the video data and the cost of capacity that must be provisioned or leased. This is accomplished by minimizing a weighted sum of the video data delay and the total capacity used subject to meeting the rate constraints on all traffic classes, and the delay constraints on the audio data traffic. For each α, the marginal tradeoff value between delay and capacity, Fig. 3 shows the minimum delay achievable for the video traffic given. The x-axis uses a log scale. The tradeoff curve shows that the minimum delay increases rapidly with increasing cost of capacity until it reaches the delay associated with the minimum capacity required to support the video signal; from that point onwards the tradeoff curve is flat.

C. Extensions

Some extensions to Theorems 2 to 4 include minimizing the maximum delay and accounting for queueing delay at the nodes. Minimizing the maximum delay can be accomplished by minimizing a dummy variable subject to all delays less than that dummy variable. Queueing delay in a store and forward network can also be accounted for in the geometric programming construction [5].

V. Summary

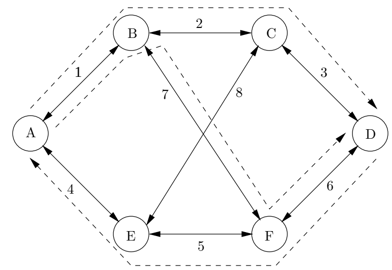

Fig. 2: Network Topology for Simulation

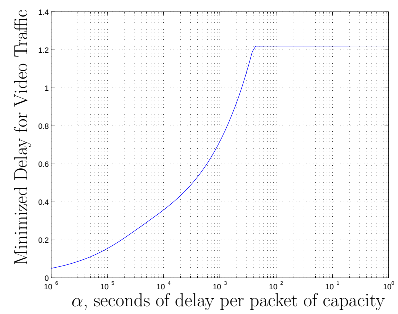

Fig. 3: Trade off between video traffic delay and capacity cost

Important performance metrics are optimized, such as throughput and delay, subject to a variety of realistic constraints on both SLA terms and network resources. This is done for wireless ad hoc networks with multihop transmissions, mutual interference, and intrinsic unreliability of links and nodes. Although these resource allocation optimizations are non-linear problems, they can be solved efficiently using fast convex optimization algorithms.

Acknowledgments

We would like to thank Nicholas Bambos and Andrea Goldsmith for discussions of wireless network problems.

References

[1] D. Julian, M. Chiang and D. O'Neill, "Constrained power control optimization for wireless cellular networks," to appear in *IEEE Vehicular Technology Conference,* Oct. 2001.

[2] S. Boyd and L. Vandenberghe, *Convex Optimization and Its Engineering Applications*, Stanford University EE 364 Course Reader 1999.

[3] Yu. Nesterov and A. Nemirovsky, *Interior Point Polynomial Methods in Convex Programming,* SIAM 1994.

[4] S. Kandukuri and S. Boyd, "Optimal power control in interference limited fading wireless channels with outage probability specifications." *To appear IEEE JSAC*, 2001.

[5] D. ONeill, D. Julian, M. Chiang and S. Boyd, "Resource allocation optimization for QoS provisioning in wireless cellular and ad hoc networks," *Submitted to IEEE Transaction on Communication*, 2001.