CROSS-LAYER DESIGN FOR VIDEO STREAMING OVER WIRELESS AD HOC NETWORKS

Taesang Yoo, Eric Setton, Xiaoqing Zhu, Andrea Goldsmith and Bernd Girod

Department of Electrical Engineering, Stanford University,
{yoots, esetton, zhuxq, andrea, bgirod}@stanford.edu

Invited paper

ABSTRACT

We propose a cross-layer design framework for supporting delay-critical traffic over ad hoc wireless networks and analyze its benefits for video streaming. In this framework, link capacities and traffic flows are jointly allocated to minimize the congestion experienced by video packets. The optimal solution, calculated via time sharing among different transmission schemes, concentrates resources only on active links. Experimental results on a simulated network illustrate the advantages of cross-layer design over another method based on oblivious layers. With one path, the cross-layer approach yields a 10-fold gain in supported data rate or equivalently 8.5 dB improvement in PSNR of achievable received video quality. Using 3 paths, the gain is 3-fold in rate or 5 dB in video quality. While multipath routing is essential to high data rate in oblivious-layered design, cross-layer design achieves efficient resource utilization regardless of the number of routes.

1. INTRODUCTION

In multi-hop ad hoc wireless networks, nodes establish connections without the help of a fixed infrastructure. Each node can be a source, a destination or a relay for traffic. The flexibility promised by such networks could be leveraged in a variety of contexts. For example, in disaster areas, the ability to rapidly deploy a network supporting low-latency voice and video transmission is very appealing. Media streaming applications, however, typically have high data rates and hard delay constraints. This imposes great challenges on the underlying ad hoc network, where node mobility may cause link failures, and interference among nodes limits the transmission rate over a single link. In contrast to the traditional paradigm where each layer in the network is independently designed, a potentially more powerful approach is to perform joint optimization across the layers, hence allowing more efficient resource utilization. This new paradigm is called *cross-layer design*.

Some recent research has studied joint optimization of physical layer power allocation, MAC layer link scheduling, and network layer flow assignment [1, 2, 3]. In our work, we propose a more general framework that encompasses the entire protocol stack, as illustrated in Fig. 1. At the physical layer, adaptive link layer techniques adjust transmission rates over individual wireless links and provide this information for dynamic capacity assignment to the MAC layer, where resource allocations are determined to different traffic flows according to the solutions from congestion-optimized multipath routing in the network layer. Based on the joint optimization between the network and the MAC layers, intelligent packet scheduling and error-resilient source coding can be applied at the transport and the application layers to further enhance the performance for low-latency delivery of media over ad hoc wireless networks.

In this paper, we focus on the joint optimization of capacity and flow assignment for live video streaming. The optimal capacity assignment lies on the edge of the capacity region, and is determined by time sharing among different transmission schemes. Network congestion, a quantity reflecting the amount of delay experienced by the video packets, is chosen to be the cost function. This cross-layer approach extends further the benefits from congestion-optimized routing alone shown in [4].

The rest of the paper is organized as follows. We describe our ad hoc wireless network model in the next section and explain how to compute a capacity region. In Section 3, we formulate the problem of minimizing the network congestion by jointly allocating capacity and flow and illustrate how cross-layer design affects the form of the solution compared to an oblivious design where each layer is independently optimized. The benefits of cross-layer design for live video streaming are discussed in Section 4.

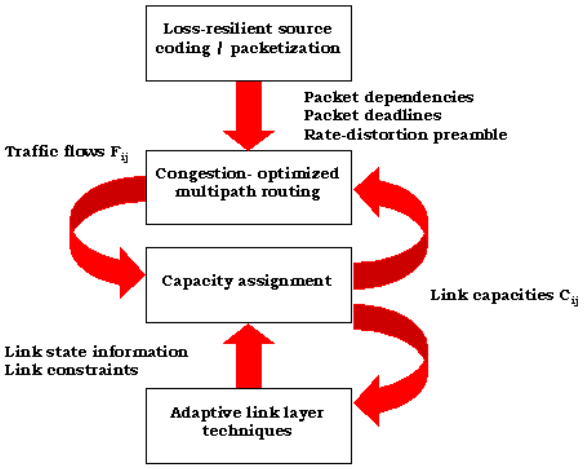

Fig. 1. Cross-layer design framework for low-latency media streaming over ad hoc wireless network.

This work was funded in part by NSF grant CCR-0325639.

2. WIRELESS NETWORK MODEL

Consider an ad hoc wireless network model with N nodes and single hop transmissions among the neighbors, where multi-hop transmissions will be handled explicitly by routing in the network layer. We assume that a node can act as a transmitter or a receiver at each time instance, but cannot do both simultaneously. Due to the broadcast nature of wireless networks, the achievable rate between two nodes depends on the power of the transmitter and on the level of interference caused by other simultaneous transmissions. In this section, we develop the concept of a *capacity region*, defined as the entire set of rates simultaneously achievable among pairs of nodes [1]. Our approach is similar to [5], where this notion of a capacity region is first introduced.

2.1. Capacity region

Conceptually, we define a *transmission scheme* as the state of operation of the nodes in the network. More precisely, consider a set of transmitting nodes $\mathcal{T} = \{t_1, \cdots, t_n\} \in \{1, \cdots, N\}$ and the set of their intended receivers $\mathcal{R} = \{r_1, \cdots, r_n\} \in \{1, \cdots, N\}$. Let the transmit powers of the nodes be $\mathbf{P} = [P_1, \cdots, P_N]$, ($P_i = 0$ for $i \notin \mathcal{T}$). A transmission scheme $\mathcal{S}$ is defined as a triplet $\mathcal{S} = \{\mathcal{T}, \mathcal{R}, \mathbf{P}\}$. Together with channel gain values, this determines achievable rates between the transmitter and receiver pairs.

For a specific transmission scheme $\mathcal{S}_k = \{\mathcal{T}_k, \mathcal{R}_k, \mathbf{P}_k\}$, the signal to interference and noise ratio (SINR) at node r_j is given by:

$$SINR_{k,t_j,r_j} = \frac{P_{k,t_j} G_{t_j,r_j}}{N_0 W + \sum_{i:t_i \in T_k, t_i \neq t_j} P_{k,t_i} G_{t_i,r_j}} \qquad (1)$$

where $G_{i,j}$ is the channel gain between nodes i and j, N_0 is the background noise spectral density, and W is the system bandwidth. We assume that each transmitter and receiver pair can adapt its rate to this SINR value by using adaptive link layer techniques such as MQAM [6] [7]. The achievable data rate between nodes i and j is given by

$$R_{k,ij} = \begin{cases} W \log_2(1 + \frac{SINR_{k,i,j}}{\Gamma}), & (i,j) \in (\mathcal{T}_k, \mathcal{R}_k) \\ 0 & \text{otherwise} \end{cases} \qquad (2)$$

where Γ is a constant determined by the BER requirement and the coding scheme used at the physical layer. The rate matrix $\mathbf{R}_k = \{R_{k,ij}\}$ represents simultaneously achievable rates between the given transmitter and receiver pairs using the transmission scheme $\mathcal{S}_k$.

Finally, we define the *capacity region* $\mathcal{C}$ of the ad-hoc wireless network as the set of rates simultaneously achievable by time sharing the different elementary transmission schemes:

$$\mathcal{C} = \left\{ \sum_{k=1}^{L} \lambda_k \mathbf{R}_k : \lambda_k \geq 0, \sum_{k=1}^{L} \lambda_i \leq 1 \right\}. \qquad (3)$$

2.2. Reducing the complexity

When all the possible combinations of transmitters and receivers are considered, and assuming that each node can transmit at one of p possible power levels, it can be shown that the total number of transmission schemes is $L = \sum_{n=1}^{\lfloor N/2 \rfloor} \frac{N!}{n!(N-2n)!} p^n$, which is a large number even for a small network size. To reduce the complexity, in the current work, we use a fixed transmit power, i.e., $P_{k,i} = P$ if $i \in \mathcal{T}_k$ and $P_{k,i} = 0$ otherwise. The complexity can be further reduced by the following methods.

(1) *Pruning useless transmission schemes*: If a rate matrix $\mathbf{R}_k$ can be generated by an appropriate time division of the other rate matrices, the scheme $\mathcal{S}_k$ can be removed. More precisely, if nonnegative λ_ks exist such that $\mathbf{R}_k = \sum_{i=1, i\neq k}^{L} \lambda_i \mathbf{R}_i$ and $\sum_{i=1, i\neq k}^{L} \lambda_i \leq 1$, then removing the scheme $\mathcal{S}_k$ does not change the capacity region. This effectively prevents spatial reuse within close neighbors.

(2) *Pruning inefficient transmission schemes*: Edges with link gains below a certain threshold can be removed to form a partially connected network. This has the effect of preventing long range communications, forcing the upper layer routing protocol to use multi-hop routing to reach the destination.

3. JOINT ALLOCATION OF CAPACITY AND FLOW

3.1. Problem formulation

To achieve the best media streaming performance, resources should be allocated to support the maximum data rates and yield minimum end-to-end delay. For general queuing systems, this does not lead to a tractable problem formulation. As an alternative, we propose to consider the allocation of capacity and flow that minimizes the network congestion Δ, while allowing communication between a sender and a receiver at a given data rate. In this work, we define congestion as the maximum link utilization over the links of the network:

$$\Delta(\mathbf{C}, \mathbf{f}) = \max_{(i,j)} \frac{f_{ij}}{C_{ij}} \qquad (4)$$

In (4), the capacity and traffic matrices are denoted by $\mathbf{C}$ and $\mathbf{f}$, and their coefficients C_{ij} and f_{ij} represent the capacity and traffic flow on the directional link (i, j). Optimal capacity and flow are determined by minimizing Δ over all feasible $\mathbf{C}$'s and $\mathbf{f}$'s. These particular feasibility conditions may be expressed as a set of linear constraints. Specifically, the capacity matrix $\mathbf{C}$ should lie in the capacity region defined by (3). In addition, the flow over a link should be positive and smaller than the capacity of this link, flow conservation equations need to be satisfied as well as rate constraints at the source and destination. The latter ensures that the traffic rate between the sender and the receiver sums up to a given rate.

The objective function $\Delta(\mathbf{C}, \mathbf{f})$ is not the classic network congestion measure analyzed in [8] derived from the M/M/1 formula for average queuing delay. However, as pointed out in [9], $\Delta(\mathbf{C}, \mathbf{f})$ has comparable properties to the classic congestion measure and has the important advantage of being a quasi-convex function of both flow and capacity [2]. This is indicated by the following equality which shows that the sub-level sets of the function are convex

[1] Note that this is not the information theoretic capacity region, which is still an open problem, but an achievable rate region under a suboptimal transmission strategy

[2] A function is quasi-convex if the domain over which its value is below any particular threshold (called sub-level set) is convex.

polyhedra:

$$\begin{aligned} & \{(\mathbf{C},\mathbf{f}) \quad | \quad \Delta(\mathbf{C},\mathbf{f}) \leq \alpha\} \\ = \; & \{(\mathbf{C},\mathbf{f}) \quad | \quad \max_{(i,j)} \frac{f_{ij}}{C_{ij}} \leq \alpha\} \\ = \; & \{(\mathbf{C},\mathbf{f}) \quad | \quad f_{ij} - \alpha C_{ij} \leq 0, \forall (i,j)\} \end{aligned}$$

Hence, as the feasibility set is also a convex polyhedra, the optimal solution can be obtained by finding the smallest α for which the intersection of these two sets is not empty. This is done by bisection, through a sequence of linear programs.

The optimal flow assignment given by the solution does not directly indicate a set of paths between the source and the destination. More importantly, the number of links used by the solution may be undesirably high. To convert the solution to a more practical form, we recursively extract from the solution the k paths carrying the most traffic as in [4]. We then re-optimize the capacity and flow assignment and constrain traffic to only flow along the extracted paths.

3.2. Solution example

A solution to the optimization problem is shown in the top part of Fig. 2. Flow and capacity are assigned jointly to minimize the congestion Δ for a 10-node network. Rather than pooling all the resources to support a high data rate path between the source and the destination and time sharing the links of this path, the optimal solution makes use of diversity at several layers of the protocol stack. In particular, the solution comprises spatial reuse of the available bandwidth as links (3,2) and (4,5) are scheduled simultaneously. The links picked by the optimization predominantly connect neighboring nodes to favor the most energy-efficient routes. At the network layer, 3 paths between the source and the destination are established and allocated comparable traffic. The efficiency of the solution resides in the exclusive allocation of resources to the directional links that compose these paths[3].

As a comparison, the bottom part of Fig. 2 shows another possible allocation of capacity and flow obtained in a network with oblivious layers. In this architecture, bidirectional links are established between nodes depending on distance. Nodes within a given transmission range are connected, and the minimum capacity of the resulting set of links is maximized through a linear optimization program. Not surprisingly, this results in assigning resources to many idle links even though they do not support any traffic. Flow assignment and path extraction are then performed by minimizing (4) for this fixed capacity map. In this case, multiple routes are necessary to increase the aggregate data rate available between the source and the destination. Given the fixed capacity map, routes are selected regardless of energy-efficiency of the links.

4. SIMULATION RESULTS

We use the network simulator NS-2 [10] to evaluate the performance of the proposed cross-layer joint optimization scheme and of the layered optimization approach, both described in Section 3. The simulated network consists of 10 mobile nodes within a 100m-by-100m square as in Fig. 2. Each node follows the random

[3]In a more practical setup, some residual capacity could easily be reserved to signal additional traffic requests from the nodes, which in turn would trigger a new optimization.

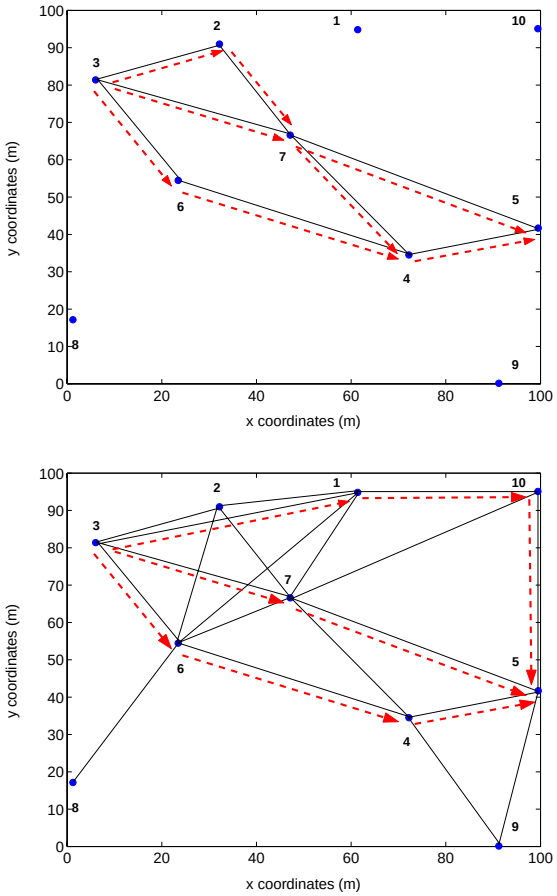

Fig. 2. Routes and capacity assigned when routing 240 kbps of traffic between nodes 3 and 5 along 3 paths. Capacity and flow are determined jointly (top) or independently (bottom).

walk mobility model with average speed 5 m/s. Capacity assignment and routing are updated once every 1.0 s in both schemes. The entire simulation duration is 100 s.

A video streaming session is set up from Node 3 to Node 5, using single-path or 3-path routing. The *Foreman* CIF video sequence is encoded by the H.264 codec at 30 frames per second, using a typical IBBP... coding structure with GOP length 16. The playout deadline is 150 ms, typical for live streaming scenarios. In the experiments, packet losses are caused by link failures or overflow of transmission queues due to congestion. Packets arriving at the receiver after their deadlines are discarded. Error concealment is performed by replacing an undecodable frame with the nearest correctly decoded frame.

Figure 3 shows the rate-distortion performance of the transmitted video using cross-layer and oblivious-layer designs, with single-path and 3-path routing. The encoder rate-distortion curve is also plotted for reference. With one route, the cross-layer design concentrates all available resources on the particular path, and can support a data rate of up to 1.9 Mbps without significant packet loss or delay; the design with oblivious layers, by contrast, can only hold 150 kbps of traffic. Consequently, the best achievable video quality with joint optimization is 39.2dB in PSNR, 8.5 dB higher than that with independent optimization. When three paths

are used for routing, the cross-layered network supports 3 times more data rate and offers 5dB improvement in PSNR of received video quality. It is interesting to note that while multipath routing increases the aggregate data rate for the design with oblivious layers, this is not the case with cross-layer design, which offers more flexibility by concentrating resources on the active links.

Video quality over time at the highest supportable rates is plotted in Fig. 4 for both designs using 3-path routing. The packet drops observed in the experiments are caused by route switches due to the mobility of the nodes, which is not captured in the theoretical analysis. Due to topology changes, packets transmitted shortly before a new capacity assignment may be traveling on a broken link in the next time instant. The design with oblivious layers is less susceptible to this phenomena as it reserves extra resources for unused links; the performance of cross-layer design is slightly affected at low rates. For the same reason, in the cross-layer design, while the theoretical objective (Eq. 4) achieved by 3-path routing is slightly lower than that with one path, more route switches and packet drops are likely to occur in the experiments, therefore the overall performance is compromised. In future work, the impact of these packet drops are expected to be reduced by error-resilience techniques at the transport and the application layers.

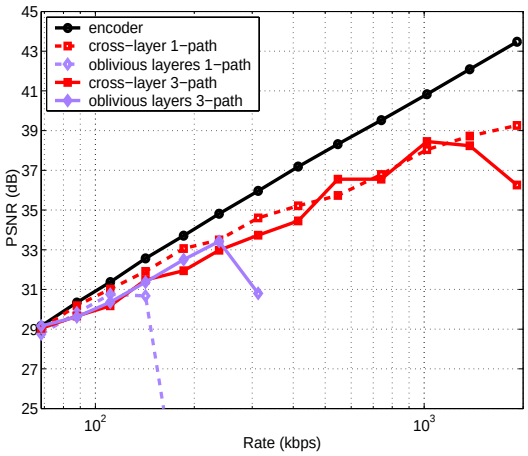

Fig. 3. Rate-PSNR performance of video streaming for two different network architectures. The rate is shown in logarithmic scale.

5. CONCLUSION

We analyze the benefits of cross-layer design on live video streaming. By determining a set of transmissions schemes and combining them through time sharing, the capacity region is computed for an ad hoc wireless network. When transmitting traffic between two nodes, congestion is minimized by jointly determining a point on the edge of this region and a feasible flow limited to a fixed number of paths. This cross-layer optimization makes efficient use of the network resources by allocating them only to links supporting traffic. In the experimental results with single-path and 3-path routing, cross-layer optimization results in 4-10$\times$ increase in the maximum supported data rate and gains of 5-8.5 dB in PSNR for received video quality transmitted between two mobile nodes.

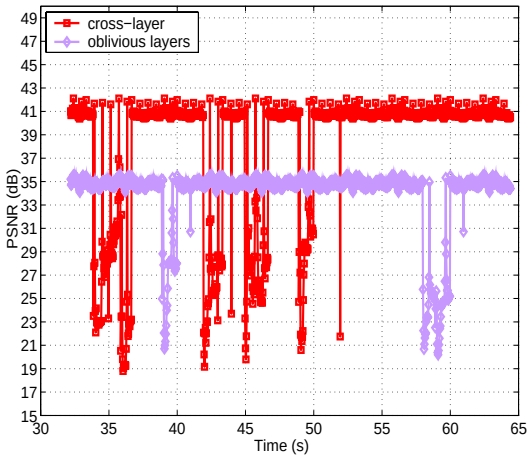

Fig. 4. Maximum video quality for cross-layer design and design with oblivious layers, using 3-path routing. Packet drop rates are around 1% in both traces. This plot shows one third of the entire simulation session.

6. REFERENCES

[1] L. Xiao, M. Johansson, and S. Boyd, "Simultaneous routing and resource allocation via dual decomposition," *to appear in IEEE Transactions on Communications*.

[2] R. L. Cruz and A. Santhanam, "Optimal routing, link scheduling and power control in multi-hop wireless networks," *Proc. INFOCOM, San Francisco, USA*, pp. 702–711, Mar. 2003.

[3] Y. Wu, P. Chou, Q. Zhang, W. Zhu K. Jain, and S-Y. Kung, "Network planning in wireless ad hoc networks: a cross-layer approach," *IEEE Journal on Selected Areas on Communications, submitted October 2003*.

[4] E. Setton, X. Zhu, and B. Girod, "Congestion-optimized multipath streaming of video over ad hoc wireless networks," *Proc. IEEE Internation Conference on Multimedia and Expo 2004, Taipei, Taiwan*, July 2004.

[5] S. Toumpis and A. Goldsmith, "Capacity Regions for Wireless Ad Hoc Networks," *Wiley Interscience, NY*, vol. 2, pp. 736–748, July 2003.

[6] A. Goldsmith and S. Chua, "Variable-rate variable-power M-QAM for fading channels," *IEEE Trans. Communications*, Oct. 1997.

[7] T. Yoo and A. Goldsmith, "Throughput Optimization Using Adaptive Techniques," *submitted to GlobeCom 2004*.

[8] L. Kleinrock, *Queuing Systems, Volume II: Computer Applications*, Wiley Interscience, NY, 1976.

[9] D. Bertsekas and R. Gallager, *Data Networks*, Prentice Hall, NJ, 1987.

[10] "The Network Simulator - ns-2," *www.isi.edu/nsnam/ns/*.