A Cross-Layer Architecture for End-to-End QoS Provisioning in Wireless Ad Hoc Networks

Didem Gozupek[1], Symeon Papavassiliou[2], and Nirwan Ansari[1]

[1]Department of Electrical and Computer Engineering
New Jersey Institute of Technology
Newark, NJ 07102 USA
Emails: {dg52, nirwan.ansari}@njit.edu

[2]School of Electrical and Computer Engineering
National Technical University of Athens
Zografou, 15780 Athens Greece
Email: papavass@mail.ntua.gr

***Abstract*-In this paper, a novel cross layer architecture is proposed to provide enhanced Quality of Service (QoS) in wireless ad hoc networks. The key feature of the proposed architecture is the integration of the service vector paradigm at the network layer with a delay-bounded power efficient scheduling approach at the data link layer. It has been demonstrated through modeling and simulations that this cross-layer architecture can provide substantial power savings and at the same time meet the delay requirements in wireless ad hoc networks.**

I. INTRODUCTION

Due to the rapid growth of Internet and real-time multimedia communications as well as the evolution of the networking environment towards wireless domain, Quality of Service (QoS) provisioning in wireless networks has become an issue of high practical and research importance. However, the power limited and time varying nature of the wireless medium complicates the design and use of QoS provisioning mechanisms. Ad hoc wireless networks introduce additional challenges due to their infrastructure-less and multi-hop nature. For the same reasons, power consumption becomes a more significant design constraint in ad hoc wireless networks as compared to their conventional wired and/or wireless counterparts.

Two fundamental frameworks have been proposed for QoS provisioning in the Internet: IntServ [1] and DiffServ [2]. IntServ suffers from the scalability problem because it requires per-flow based service provisioning and resource allocation. DiffServ is considered as a more feasible solution since it overcomes the scalability problem by aggregating individual flows and providing only a certain number of services to the aggregated data flows. However, it can only provide coarse QoS granularity. A concept, referred to as *service vector*, has recently been introduced in the literature [3, 4]. This service provisioning model enhances the QoS granularity of the DiffServ architecture while retaining its scalability feature.

The service vector scheme was originally designed for wireline networks. In our work, we have extended this scheme to wireless ad hoc networks and proposed a cross-layer architecture based on the integration of the service vector scheme at the network layer and delay bounded power efficient scheduling at the link layer. Our proposed scheme provides significant power savings, and hence improves the end-to-end QoS provisioning in wireless ad hoc networks.

The rest of the paper is organized as follows. Section II provides an overview of the service vector paradigm, while section III defines and presents the corresponding framework within the wireless ad hoc realm. Section IV describes in detail the delay bounded power efficient link layer scheduling, integrated within the service vector paradigm. Section V presents some initial simulation results that demonstrate the performance improvements in terms of power savings that can be achieved by our proposed scheme. Finally, Section VI concludes the paper.

II. OVERVIEW OF THE SERVICE VECTOR PARADIGM

The Explicit Endpoint Admission Control with Service Vector (EEAC-SV) scheme essentially consists of two phases. In the probing phase, the end host sends a probe request packet to the destination, and the destination host responds by sending a probe acknowledgement packet in the reverse direction. Each router along the path explicitly provides information about the performance of each service class and attaches this information to the acknowledgement packet. This way, the end host gathers the QoS related information about each service class at all routers along the path, and uses this information in determining the optimum service vector. In the data transfer phase, the service vector is attached to the data packets, which are allowed to utilize different service classes at different routers along the path [4].

Assume that there are m routers along the path and n service classes at each router. The set of n service classes can be denoted as $S=(S_0, S_1, ..., S_{n-1})$ and the service vector can be represented as $s=(s_0, s_1, ..., s_{m-1})$, where s_i denotes the service class used at router i. Allowing the flow to choose different service classes at different routers is the key principle of the EEAC-SV scheme.

The existing end-to-end service provisioning mechanisms of static service mapping and dynamic service mapping schemes can be included in the service vector concept and classified as follows:

Scheme 1-Conventional Scheme (EAC-CS) (Static Service Mapping): The end host maps the users' QoS requirements to a certain service class, measures the performance of this constant service vector, and determines whether it satisfies the QoS requirements of the data flow. If the requirements are met, the flow is

accepted; otherwise, it is rejected. As a result, the end-to-end QoS granularity is *O(1)*.

Scheme 2-EEAC with Single Class of Service Scheme (EEAC-SCS) (Dynamic Service Mapping): The end host dynamically maps the data flow's QoS requirements to the available best service class. The service vector is a constant vector as in EAC-CS; however, since the mapping is dynamic rather than static, the resultant end-to-end QoS granularity is *O(n)*.

Scheme 3-EEAC with Combination of Service Classes Scheme (EEAC-CSC) (Combination of Service Classes via the Service Vector): In this scheme, the flow is allowed to choose different service classes at different routers along the data path. Consequently, the end-to-end QoS granularity is $O(n^m)$.

III. Service Vector Realization In Wireless Ad Hoc Domain

In the following, we use the terms nodes and routers interchangeably because all the nodes in a wireless ad hoc network can transmit each others' packets in a multi-hop manner, and therefore can be treated as routers. Let us also assume that a flow going from its source to the destination passes through *m* intermediate routers, the set of available service classes at each router is $S=(S_0, S_1, ..., S_{n-1})$, and the service vector determined after the probing phase is denoted as $s=(s_0, s_1, ..., s_{m-1})$, where s_i denotes the service class used at router *i*. A time-slotted system is considered and the QoS parameter is the average end-to-end delay bound, which is inelastic; i.e., the user's level of satisfaction with the perceived QoS is the same as long as the provided QoS performance satisfies the requirements.

To minimize the total average transmission power along the path after the service vector is determined, the following problem needs to be considered:

$$\min(E\{\bar{P}\})$$
$$s.t. \quad E\{D_i\} \le delay(s_i) \quad \forall i \in (0,1,...,m-1)$$

where $E\{\bar{P}\} = \lim_{n\to\infty} \sum_{n=0}^{m-1} P_{i,n}$ $P_{i,n}$, is the power in time slot *n* at router *i*, D_i is the delay experienced at router *i*, and s_i is the service class selected at router *i*. Apparently, the above problem is decoupled into the link level scheduling problem of minimizing the average transmission power subject to the average delay constraints of all the service class buffers.

The fundamental achievement of the EEAC-CSC scheme is that even though a service class with the less strict delay guarantee is unavailable in some part of the network, the data flow is permitted to utilize that service class in some other part of the network where it is available. This way, the transmission rate and consequently the power consumption of the node where the service class is available can be diminished. For instance, suppose that the service classes 0, 1, and 2 correspond to average delay bounds of 100 ms, 200 ms, and 300 ms, respectively. Furthermore, assume that there are three nodes along the data path, and the end-to-end delay bound of the data flow is 750 ms. EAC-CS scheme results in the usage of Class 0 along the entire path and hence 300 ms average end-to-end delay, whereas EEAC-SCS scheme results in the usage of Class 1 along the data path and hence 600 ms average end-to-end delay. On the other hand, EEAC-CSC scheme results in the usage of Class 2, Class 2, and Class 3 along the path and consequently an average end-to-end delay of 700 ms (which still meets the end-to-end delay bound of the data flow under consideration). Therefore, EEAC-CSC scheme results in the least power consumption, since larger delay corresponds to smaller transmission rate and hence less power consumption. Accordingly, our proposed methodology of integrating the network layer service vector concept with the link layer delay bounded power efficient scheduling facilitates considerable power savings in ad hoc wireless networks, while still meeting the average end-to-end delay requirements.

IV. Delay Bounded Power Efficient Multi-User Scheduling

An essential part of our proposed scheme is the delay bounded power efficient multi-user scheduling. Authors in [5, 6] introduced optimal and suboptimal multi-user schedulers, and employ a dynamic programming technique called Value Iteration Algorithm (VIA) in the optimum schedulers. Their proposed suboptimum multi-user TDMA scheduler initially determines the flow that is permitted to transmit in a specific time slot, and then decides on the number of packets to be transmitted in that time slot, where the optimum single user scheduler is utilized in this decision. Moreover, a suboptimum scheduler, referred to as log-linear scheduler, for the single user case has been proposed.

The optimum schedulers in [5, 6] have three major shortcomings. First of all, the number of possible states in VIA grows exponentially as the buffer sizes and the number of queues at the router increase. Second, it is difficult, if not impossible, to analytically and numerically obtain the Lagrangian value ε, which is a cost function parameter in VIA. Lastly, information about the actual (real-time) traffic arrival distribution at each router is required for the implementation of VIA.

Owing to the above mentioned drawbacks of the optimum schedulers, even the suboptimum multi-user TDMA scheduler in [6] is impractical to implement. Consequently, we modify the TDMA scheduler to tailor for our work:

1. Flow Choice: Index *k* of the flow chosen to transmit:

$$k = \begin{cases} l & if \ x_l > L_l - M_l \\ \arg\max_l \dfrac{x_l}{\lambda_l D_{l,0}} & else \end{cases}$$

2. Number of packets:

$$u_n = \min(x_n, \lfloor \log(\kappa x_n) \rfloor)$$

The first stage of the algorithm determines the flow choice basically by choosing the most “desperate flow”,

which corresponds to the one that is closest to violating its delay bound while at the same time ensuring zero buffer overflow. The second stage determines the number of packets to be transmitted in that time slot, by utilizing the log-linear scheduler. Here, x_l denotes the number of packets in buffer l at the beginning of the time slot, L_l represents the size of buffer l , M_l denotes the maximum number of packets that can arrive at buffer l, λ_l represents the average arrival rate to buffer l, $D_{l,0}$ corresponds to the average delay bound of buffer l, u_n denotes the number of packets chosen for transmission from the selected buffer at the beginning of time slot n, x_n denotes the number of packets at the selected buffer at the beginning of time slot n, and κ is a parameter that is chosen so that the average delay bound is satisfied.

Moreover, our scheduler implementation guarantees zero outage conditions in which packets are not dropped at the transmitter; zero buffer overflow is ensured by guaranteeing that $x_k \geq L_k - M_k$ for at most one $k=1,2, \ldots, K$, where x_k is the number of packets at buffer k, L_k is the size of buffer k, M_k is the maximum number of packets that can arrive at buffer k in a time slot, and K is the total number of buffers at the router.

V. Performance Evaluation

The performance of our proposed scheme is evaluated using the Optimized Network Engineering Tool (OPNET). The route of a data flow is assumed to be predetermined and the wireless links are assumed to be AWGN channels. The performance of the three types of service provisioning schemes, i.e., EAC-CS, EEAC-SCS, and EEAC-CSC, are evaluated for a single flow going from the source to the sink under both uniform and *On-Off* traffic arrival patterns. Furthermore, the influence of different arrival rates on our proposed scheme has been studied for both traffic arrival processes.

A. Models and Assumptions

Each router is assumed to provision three different service classes, namely, Expedited Forwarding (EF), Assured Forwarding (AF), and Best Effort (BE) classes. TDMA is used as the multiple access scheme and the time slot length is $T_s=0.05s$ in the entire system. The buffer size of service class k is L_k=170, $\forall$ k=1, 2, 3 and the maximum number of packets that can arrive at the class k buffer is M_k=6, $\forall$ k=1, 2, 3.

The simulated network topology is depicted in Figure 1. Each router provides the information about the availability of its service classes in the probing phase and the end host determines the best service vector among the available ones. The average delay bounds for the service classes under consideration are shown in Table 1.

The performance of a data flow from node A to node E is evaluated, while cross traffic (as shown in Figure 1) is assumed to be uniformly distributed. Table 2 summarizes the maximum number of packets that can arrive in a time slot for the background traffic flows.

Figures 2 and 3 illustrate the average delay and power consumption of the service class buffers at router 1 before the flow from node A to node E starts sending traffic. Although class 2 is the service class with the least stringent average delay bound requirement, Figure 3 shows that it has higher power consumption than the others. Furthermore, Figure 2 illustrates that its average packet delay is much smaller than its required value. This situation is due to the increase in the rate of transmission from class 2 buffer in order to both meet the average delay requirement and prevent buffer overflow. As a result, the

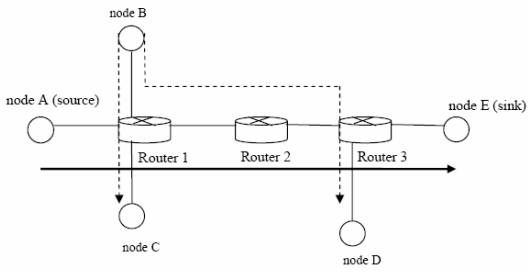

Figure 1. The simulated network topology

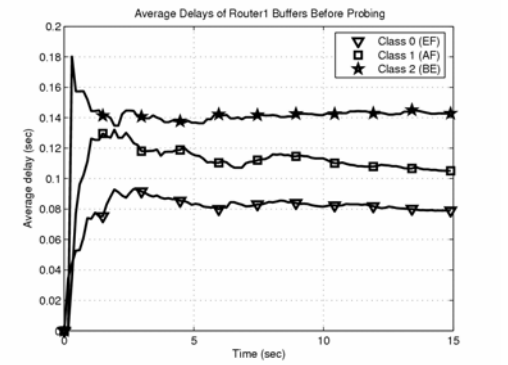

Figure 2. Average packet delays of Router 1 buffers before probing.

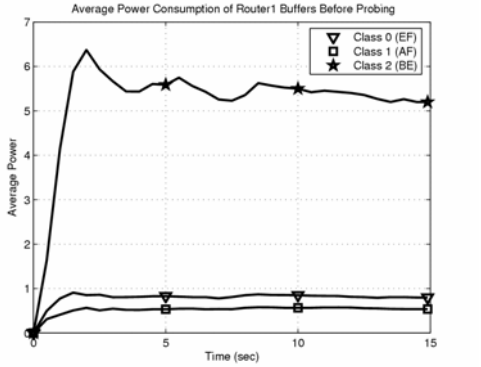

Figure 3. Average power consumption of Router 1 buffers before probing.

actual average delay at class 2 buffer becomes much smaller than its required value at the expense of immense power consumption. This is why the availability of the service classes at each router should be determined in the probing phase. Moreover, zero buffer overflow can be guaranteed as long as the maximum number of packet arrivals to each service class buffer is less than or equal to their corresponding upper bounds. Consequently, determining the current maximum number of packets arriving at each buffer is vital to ensure zero buffer overflow. Therefore, estimation of the packet arrival rate to the service class buffer is used as the parameter to check the availability of the service classes. Exponential moving average filter is used to estimate the packet arrival rate [7], which is measured in packets per time slot, as follows:

$$\overline{r}_{S_j}(t) = (1 - e^{-\tau_{S_j}(t)/K}) \frac{T_s}{\tau_{S_j(t)}} + e^{-\tau_{S_j}(t)/K} \overline{r}_{S_{j,old}}(t)$$

where T_s is the time slot length in seconds, $\overline{r}_{S_j}(t)$ is the estimated value of the packet arrival rate for service class S_j at time t, $\tau_{S_j}(t)$ is the interval between the arrival of the previous received packet of service class S_j and the current time t, and K is a constant. Each router updates $\overline{r}_{S_j}$ whenever it receives a data packet of service class S_j. If $\overline{r}_{S_j} > \frac{M_{S_j} - 1}{2}$ in the probing phase, S_j is marked as unavailable in the probe acknowledgement packet; otherwise, it is marked as available.

TABLE I
SERVICE CLASS DEFINITIONS

Service Class	Average Delay Bound
Class 0 (EF)	100 ms
Class 1 (AF)	150 ms
Class 2 (BE)	350 ms

TABLE II
SUMMARY OF BACKGROUND TRAFFIC

Source	Destination	Class 0 (EF)	Class 1 (AF)	Class 2 (BE)
Node B	Node D	2 packets / slot	2 packets / slot	2 packets / slot
Node B	Node C	0 packets/slot	0 packets/slot	4 packets / slot

B. Simulation Results

A single flow originating from node A and destined to node E, having an inelastic average end-to-end delay bound of 950 ms, is considered. The total number of packets generated by the source is initially assumed to be uniformly distributed with a maximum of 4 packets per time-slot.

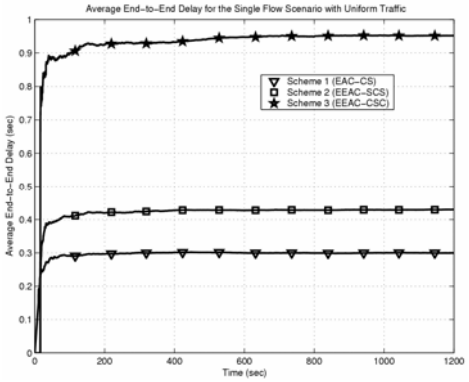

Figure 4. Average end-to-end delay of the three schemes for the single flow with uniform traffic.

Figures 4 and 5 illustrate the average end-to-end delay and power consumption for the single flow under uniformly distributed arrival traffic. Figure 4 shows that all the three service provisioning schemes can meet the inelastic average end-to-end delay bound requirement. Scheme 3 attempts to utilize all possible combinations of service classes; therefore, it leads to the highest average end-to-end delay. However, as shown in Figure 5, Scheme 3 achieves the lowest power consumption. This is attributed to the fact that this scheme permits the use of higher delay and consequently less power consuming service classes.

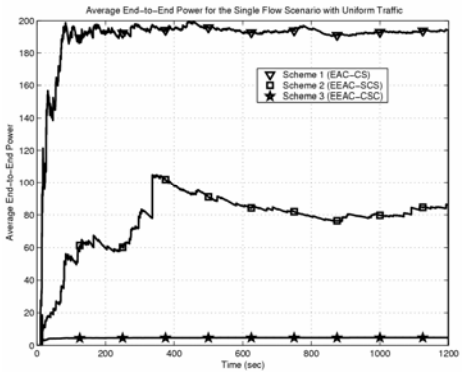

Figure 5. Average end-to-end power consumption of the three schemes for the single flow with uniform traffic.

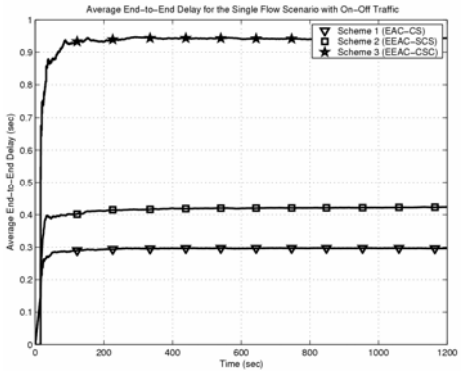

Figure 6. Average end-to-end delay of the three schemes for the single flow with *On-Off* traffic.

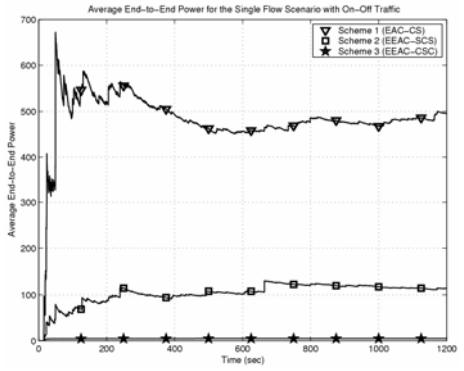

Figure 7. Average end-to-end power consumption of the three schemes for the single flow with *On-Off* traffic.

Figures 6 and 7 illustrate the average end-to-end delay and power consumption for the single flow under the *On-Off* arrival traffic, where the *On* state and *Off* state are assumed to be equally likely and 4 packets are generated in the *On* state. The results again confirm the power savings enabled by our proposed scheme. Furthermore, it should be noted that the power consumption for each scheme under the *On-Off* traffic arrival pattern is higher than the corresponding ones under the uniform arrival distribution counterparts. This is attributed to the fact that the *On-Off* arrival process requires the highest transmission power at any delay in an AWGN channel

among all arrival processes with the same average and finite maximum arrival rate [5].

Figures 8 and 9 illustrate the average end-to-end power consumption of the three service provisioning schemes for uniform and *On-Off* traffic, respectively, under different traffic loads (i.e., the maximum number of packets generated by the source is varied from 1 to 4). Under both traffic patterns, our proposed scheme outperforms the

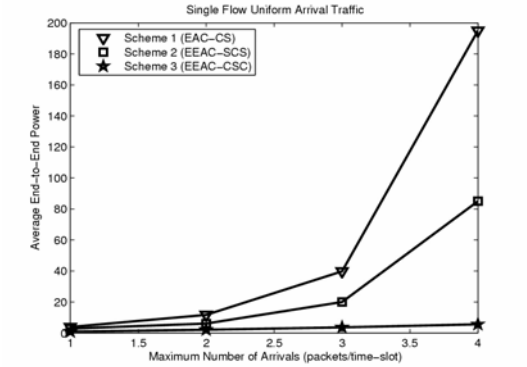

Figure 8. Average end-to-end power consumption of the three schemes for the single flow with varying arrival rates and uniform traffic.

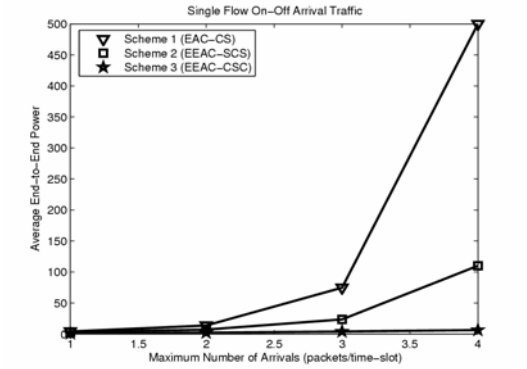

Figure 9. Average end-to-end power consumption of the three schemes for the single flow with varying arrival rates and *On-Off* traffic.

other two schemes for all of the arrival rates. Furthermore, the performance improvement in power savings increases as the arrival rate increases. On the other hand, the exponential shape of the plots is due to the exponential relation between transmission rate and power.

VI. CONCLUSIONS

QoS provisioning in wireless ad hoc networks has become a crucial issue due to the proliferation of Internet applications and services, as well as the emergence and deployment of wireless ad hoc networks. Power consumption is a vital QoS constraint in wireless ad hoc networks. Consequently, a power efficient cross-layer QoS provisioning architecture for wireless ad hoc networks is proposed in this paper. Our proposed scheme capitalizes on the network layer *service vector* concept and the link layer delay bounded multi-user scheduling. It has been demonstrated through modeling and simulation that this approach enables significant power savings in wireless ad hoc networks. Since optimum scheduling is not feasible, suboptimal multi-user scheduling, which can only function in AWGN channels, has been utilized. Extending this suboptimum scheduler by considering fading would be of high practical and research importance.

REFERENCES

[1] R. Braden, D. Clark, and S. Shenker, "Integrated Services in the Internet Architecture: An Overview," RFC1633, June 1994.

[2] S. Blake, D. Black, M. Calson, E. Davies, Z. Wang, and W. Weiss, "An Architecture for Differentiated Services", RFC2475, December 1998.

[3] J. Yang, J. Ye, S. Papavassiliou, and N. Ansari, "A Flexible and Distributed Architecture for Adaptive End-to-End QoS Provisioning in Next Generation Networks", *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 2, pp. 321-333, February 2005.

[4] J. Yang, J. Ye, and S. Papavassiliou, "Enhancing End-to-End QoS Granularity in Diffserv Networks via Service Vector and Explicit Endpoint Admission Control," *IEE Proceedings on Communications*, vol. 151, no. 1, pp. 77-81, February 2004.

[5] D. Rajan, A. Sabharwal, and B. Aazhang, "Delay-Bounded Packet Scheduling of Bursty Traffic over Wireless Channels," *IEEE Transactions on Information Theory*, vol. 50, no. 1, pp. 125-144, January 2004.

[6] D. Rajan, "Power Efficient Transmission Policies for Multimedia Traffic over Wireless Channels," Ph.D. thesis, Rice University, April 2002.

[7] S. Floyd, V. Jacobson, "Random Early Detection for Congestion Avoidance", *IEEE/ACM Transactions on Networking*, vol. 1, no. 4, pp. 397-413, July 1993.