A Power Efficient QoS Provisioning Architecture for Wireless Ad Hoc Networks

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

[1]Department of Electrical and Computer Engineering
New Jersey Institute of Technology
Newark, NJ, 07102 USA
Emails:{dg52,nirwan.ansari,jxy9918}@njit.edu

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

***Abstract*— The work presented in this paper[1] focuses on a new approach in provisioning Quality of Service (QoS) in ad hoc wireless networks, aiming at making the best use of ad hoc networking as a candidate technology for next generation wireless networks. Specifically, a cross-layer QoS provisioning architecture for wireless ad hoc networks is introduced and described, based on the integration of the recently proposed *service vector* concept at the network layer and a delay bounded power efficient scheduling at the data link layer. It is demonstrated through modeling and simulations that this novel architecture can provide considerable performance improvements in terms of both power savings and enhanced QoS granularity in wireless ad hoc networks. Furthermore, the performance of the proposed scheme under various traffic arrival rates and distributions is evaluated.**

I. Introduction

The evolution towards next generation wireless networks follows a path from wireless being simply a mode of access to a wired network to where the entire network architecture aims to provide wireless communications. Therefore, ad hoc wireless networking has enjoyed dramatic increase in popularity over the last few years. At the same time rapid growth of Internet and real-time multimedia communications necessitate Quality of Service (QoS) provisioning mechanisms. As a result, QoS provisioning in wireless networks becomes of high practical and research importance.

Ad hoc wireless networks are typically composed of equal nodes that communicate over wireless links without any central control. These networks inherit the traditional problems of wireless communications, such as power control, bandwidth allocation, and quality of service provisioning. In ad-hoc networks, power consumption is an even more critical factor than the conventional wireless networks. The lack of a fixed network infrastructure is often accompanied by non-existence of a power supply infrastructure. In such circumstances, energy needs to be considered as a limited network resource. At the same time the infrastructure-less nature of ad hoc wireless networks makes QoS provisioning a more challenging task.

Two service models have hitherto been proposed for QoS provisioning in Internet. IntServ [1] allocates network resources on a per-flow basis, and hence suffers from the scalability problem; whereas DiffServ [2] overcomes the scalability problem by implementing per-aggregate based resource allocation; however, it can only provide coarse QoS granularity. On the other hand, Explicit Endpoint Admission Control and Service Vector (EEAC-SV) paradigm has been recently introduced [3][4], as an efficient QoS provisioning architecture that preserves the scalability characteristic of DiffServ networks and improves its QoS granularity.

In this paper, this paradigm is adopted and extended to wireless ad hoc networks. A cross-layer architecture based on the combination of the service vector scheme and delay bounded link level scheduling has been designed. It is demonstrated through modeling and simulation that within this architecture the service vector scheme can provide significant power savings as well as better QoS granularity in wireless ad hoc networks.

The remaining of the paper is organized as follows: Section II provides some background information about the service vector paradigm. The corresponding problem in wireless ad hoc networks is formulated in Section III, whereas Section IV describes the delay bounded power efficient link layer scheduling in detail. Section V presents some simulation results that demonstrates the performance improvements, in terms of both power savings and service differentiation, that can be achieved by our proposed scheme. Finally, Section VI concludes the paper.

II. Related Works on Service Vector Paradigm

The Explicit Endpoint Admission Control with Service Vector (EEAC-SV) scheme consists of two stages, namely the probing phase and the data transfer phase. Initially, the end host sends probing packets along the path to the destination host, which then creates an acknowledgement packet and sends it back to the end host. Each router along the path attaches to the acknowledgement packets the QoS related information about each service class. The end host determines the best service vector by implementing an optimization procedure, as described in [4].

Assuming that there are m routers along the path and n service classes at each router, the set of n service classes

[1]This work has been supported in part by the National Science Foundation under grant no. 0435250.

1-4244-0355-3/06/$20.00 (c) 2006 IEEE

can be represented as $S = (S_0, S_1, ..., S_{n-1})$ and the service vector can be denoted as $s = (s_0, s_1, ..., s_{m-1})$, where s_i corresponds to the service class used at router i. One of the key principles of the EEAC-SV scheme is that the flow is allowed to choose different service classes at different routers.

With reference to this generic paradigm, the various end-to-end service provisioning mechanisms, i.e., static service mapping and dynamic service mapping schemes, can be categorized and incorporated into the service vector concept, as follows:

Scheme 1-Conventional Scheme (EAC-CS) (Static Service Mapping): The users' QoS requirements are statically mapped to a predetermined service class, and hence the service vector is a constant vector. The end host checks whether the measured QoS performance of the statically mapped service class meets the user's QoS requirements or not. Consequently, the resultant QoS granularity is $O(1)$.

Scheme 2-EEAC with Single Class of Service Scheme (EEAC-SCS) (Dynamic Service Mapping): The service vector is a constant vector; nevertheless, the flow is dynamically mapped to the available best service class. The resultant QoS granularity is *O(n).*

Scheme 3-EEAC with Combination of Service Classes Scheme (EEAC-CSC) (Combination of Service Classes via the Service Vector): Different service classes can be selected at different routers. Since the number of possible service vectors is n^m, the resultant QoS granularity is $O(n^m)$.

III. Problem Formulation

In principle, all nodes in wireless ad hoc networks can be regarded as routers, since they can transmit each others' packets in a multi-hop fashion, and hence in this paper we use the terms nodes and routers interchangeably. Assuming that there are m intermediate routers along the path of a data flow and n service classes are provisioned at each router, the resultant service vector determined by the end host can be denoted as $s = (s_0, s_1, ..., s_{m-1})$, where s_i corresponds to the service class used at router i. The average end-to-end delay is considered as the QoS parameter, where the average delay bound of service class s_i is represented by $delay(s_i)$. A time-slotted system is considered in the data transmission phase and the average end-to-end delay bound of a data flow is inelastic; i.e., the application does not care if better than required QoS is provided.

In order to minimize the total transmission power along the path after the service vector is determined, we need to solve the following problem:

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

where $\overline{P} = \lim_{n\to\infty} \sum_{i=0}^{m-1} P_{i,n}$, $P_{i,n}$ 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 chosen at router i. This problem decouples into the link layer multi-user scheduling problem of minimizing the average transmission power while satisfying the average delay constraints of all the service class buffers.

The key merit of the service vector scheme is that if a certain service class with a less stringent delay guarantee is unavailable at a certain node along the path, the data flow can still choose that particular service class in some other nodes where it is available. Therefore, the transmission rate can be decreased at the node where the service class is available, thus directly reducing the transmission power of that node. Consequently, our proposed method of combining the service vector scheme with a delay-bounded power-efficient multi-user scheduling mechanism, leads to significant power savings in wireless ad hoc networks.

IV. Delay Bounded Power Efficient Multi-User Scheduling

As mentioned before, the delay bounded power efficient multi-user scheduler constitutes an integral part of our proposed methodology. Authors in [5, 6] proposed optimal single user and multi-user schedulers, in which a dynamic programming technique referred to as Value Iteration Algorithm (VIA) is implemented. Furthermore, they also proposed suboptimum schedulers for single user, called log-linear scheduler [5], and multi-user cases [6]. The suboptimum multi-user scheduler basically consists of two phases. The flow choice is made in the first phase, and the optimum single user scheduler is used in the second phase to determine the number of packets to be transmitted.

The proposed optimum schedulers have three main drawbacks due to the impracticality and computational complexity of VIA technique. Firstly, the number of states in the algorithm increases exponentially as the buffer size and the number of queues increase. Secondly, the Lagrangian value ε, which is a parameter of the cost function, is mathematically intractable to obtain. Thirdly, implementation of this algorithm requires knowledge about the traffic arrival distribution at each router along the path. Measurement of the arrival distributions in real-time is impractical, firstly due to its high implementation complexity, and secondly because incorrect measurement results may cause flawed scheduler actions.

Due to the reasons mentioned above, even the suboptimal multi-user scheduler, which utilizes the optimum single user scheduler in its second stage, cannot be implemented for the service vector scheme. Therefore, in this paper, the suboptimal TDMA scheduler proposed in [6] is modified as follows: flow choice is made in the first stage, and the number of packets to be transmitted is determined in the second stage using the suboptimal log-linear scheduler rather than the optimum single user scheduler. Therefore, the following scheduler has been implemented:

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

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

2. *Number of packets:*

This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE ICC 2006 proceedings.

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

where 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. The first condition in determining the flow choice ensures zero buffer overflow, whereas the second condition chooses the flow that is closest to violating its delay bound.

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 $k = 1, 2, ..., K$, 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. On the other hand, the maximum number of packets that the scheduler can transmit in a certain time slot was set to be equal to $\sum_{k=1}^{K} M_k$.

V. Performance Evaluation

The performance of our proposed scheme is evaluated through modeling and simulation using the Optimized Network Engineering Tool (OPNET) framework. The wireless links are assumed to be AWGN and the route of a flow is assumed to be predetermined. Two different flows with different QoS requirements are considered. The performance of the three types of service provisioning schemes, i.e., EAC-CS, EEAC-SCS, and EEAC-CSC, are evaluated under both uniform and *On-Off* traffic arrivals. The impact of varying traffic arrival rates on the performance of the proposed mechanism has also been studied for both arrival distributions.

A. Models and Assumptions

Expedited Forwarding (EF), Assured Forwarding (AF), and Best Effort (BE) service classes are provisioned at each router, and the TDMA system is used. For all routers, the time-slot length is $T_s = 0.05s$, the maximum number of packets that can arrive at the class k buffer is $M_k = 6, \forall k = 1, 2, 3$, and the buffer size of service class k is $L_k = 170, \forall k = 1, 2, 3$.

The network topology under consideration is shown in Figure 1. During the probing phase, each router attaches the information about the availability of its service classes. The end host executes an optimization procedure and determines the best service vector among the available ones. Table 1 illustrates the average delay bounds for the various service classes under consideration.

The performance of a data flow originating from node A and destined to node E is evaluated in this study. Cross traffic is assumed to be uniformly distributed, and the maximum number of packets that can arrive in a time slot for the background traffic flows is summarized in Table 2.

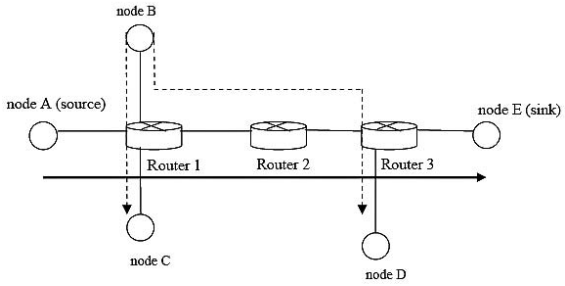

Fig. 1. The simulated network topology.

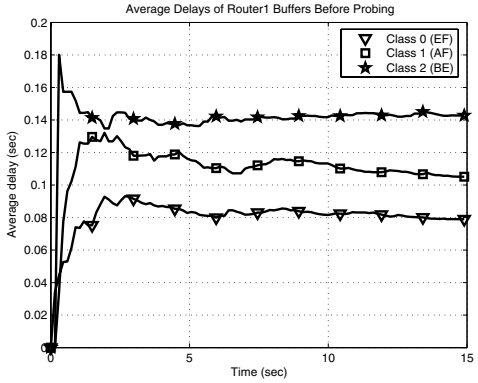

Fig. 2. Average packet delays of Router 1 buffers before probing.

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

Table 1. Service class definitions

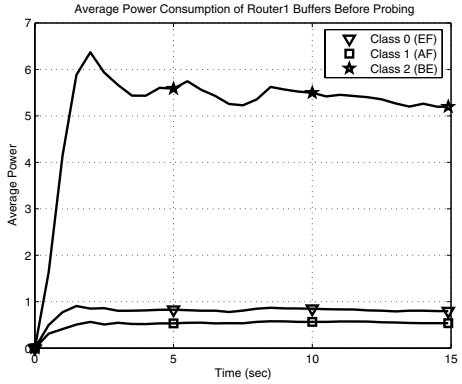

Fig. 3. Average power consumption of Router 1 buffers before probing.

This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE ICC 2006 proceedings.

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

Table 2. Summary of background traffic

The reason for checking the availability of the service classes at each router during the probing phase is demonstrated in Figure 2 and Figure 3, which illustrate respectively the average end-to-end delay and average power consumption of the various service classes at router 1, before the flow from node A to node E starts sending traffic. As observed from Figure 3, class 2 traffic at router 1 has higher power consumption than the others, although it is the service class with the least stringent delay bound requirement. Besides, the average packet delay of this traffic class is much smaller than its required value, as shown in Figure 2. The reason for this is attributed 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 actual average delay at class 2 buffer becomes smaller than its required value at the cost of larger power consumption. Furthermore, the scheduler can guarantee zero buffer overflow provided that the maximum number of packet arrivals per time slot is less than or equal to its upper bound. For instance, in our model, the maximum number of packets that can arrive at the class k buffer is $M_k = 6, \forall k = 1, 2, 3$. Therefore, it is crucial to determine the current maximum number of packets per time slot arriving at each buffer. Consequently, 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:

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

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. At each router, $\overline{r}_{S_j}$ is updated whenever a data packet of service class S_j is received. In the probing phase, if $\overline{r}_{S_j} > \frac{M_{S_j}-1}{2}$, S_j is marked as unavailable in the probe acknowledgement packet; otherwise, it is marked as available.

Selection of constant K affects the performance of the arrival rate estimation. More specifically, a small value of K enables the estimation process to track the variation in traffic appropriately; nevertheless it cannot filter out the transient changes in the data rate. On the other hand, a large value of K can filter out these changes, and hence provide stable network performance. However in this case, it cannot respond to the changes in the traffic arrival pattern quickly. The exponential moving average filter has the following unit sample response function:

$$h(a) = (1 - e^{-\tau_{S_j}^{\min}/K})(e^{-\tau_{S_j}^{\min}/K})^a U(a)$$

where a is the number of packet arrivals that determines the convergence time needed for the measurement results to converge to the actual packet arrival rate, and $\tau_{S_j}^{\min}$ is the minimum time between subsequent packet arrivals. Since the time slot length $T_s = 0.05s$, and the maximum number of packets arriving at a service class buffer is $M_{S_j} = 6, \forall j = 1, 2, 3$, $\tau_{S_j}^{\min} = 0.00833s$. Assume that a new flow will use service class S_j with probability p_{S_j}. In order to avoid buffer overflow due to slow convergence of the exponential moving average filter, the average convergence time should be less than $\sum_j p_{S_j} L_{S_j}$ packet arrivals. Since $L_{S_j} = 170, \forall j = 1, 2, 3$, the average convergence time should be less than 170 packet arrivals; i.e., $a < 170$. Let a_{stop} represent the convergence time where $h(a_{stop}) = -10db$ due to the reason that $h(a)$ will have little impact on the exponential moving average result when $a > a_{stop}$. Therefore, $a_{stop} = 170$, and hence $K \leq 12.3$. On the other hand, the smallest possible value of K stands for the case where the exponential moving average filter immediately converges to the actual measurement result, i.e., $a_{stop} = 1$, and hence $K \geq 0.0724$. Consequently, K should be in the range of $0.0724 \leq K \leq 12.3$. In our work, K was selected to be 0.35, which was found to be able to provide an accurate estimate of the actual arrival rate, while at the same time being within the above mentioned required bounds.

B. Simulation Results and Discussions

Two different types of flows are considered to be generated by the source (node A). *Type 1* has an average end-to-end delay bound of 950 ms and *Type 2* has an average end-to-end delay bound of 750 ms. 50% of the traffic generated by the source is *Type 1*, and the remaining is *Type 2*. The total number of packets generated by the source is assumed to be uniformly distributed with a maximum of 4 packets/time slot.

Figures 4 and 5 illustrate the average end-to-end delay and power consumption for the two types of flows under uniformly distributed arrival traffic. Figure 4 demonstrates that all the three service provisioning schemes can meet the inelastic average end-to-end delay bound requirement for both types of flows. Scheme 3 presents the highest average end-to-end delay, since it attempts to utilize all possible combinations of service classes. On the other hand, Figure 5 demonstrates that Scheme 3 results in the lowest power consumption for both types of flows, since it allows the use of higher delay and consequently less power consuming service classes. As it is further illustrated in these figures, EEAC-CSC scheme (Scheme 3) is the only scheme that can provide service differentiation, by providing different quality to each one of the two different flows according to their requirements. On the other hand, EEAC-SC and EAC-CS schemes are unable to provide this differentiation since they map these two flows to the same service vector. In other words, the cross-layer approach proposed in this paper not only enables significant

This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE ICC 2006 proceedings.

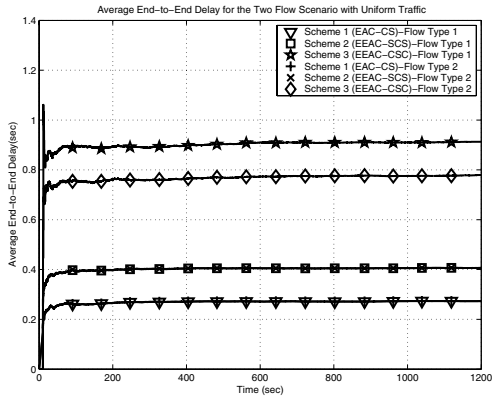

Fig. 4. Average end-to-end delay of the three schemes for the two flows with uniform traffic.

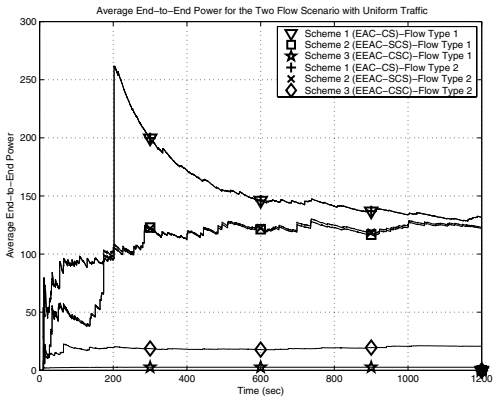

Fig. 5. Average end-to-end power consumption of the three schemes for the two flows with uniform traffic.

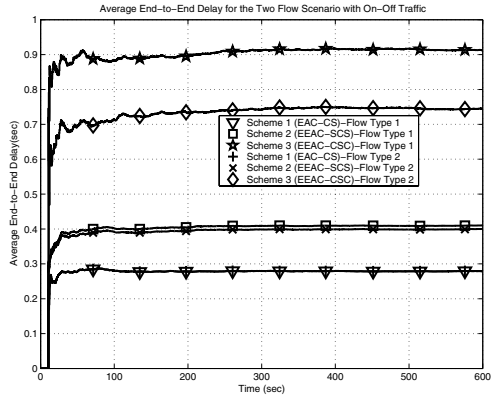

Fig. 6. Average end-to-end delay of the three schemes for the two flows with On-Off traffic.

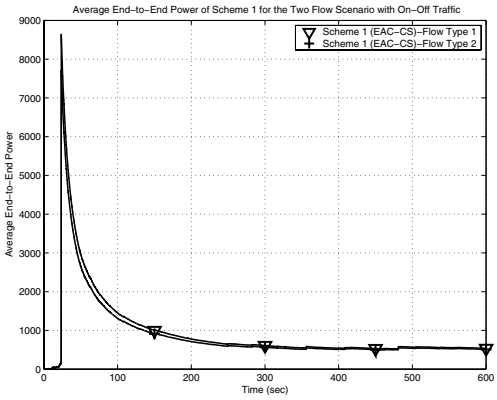

Fig. 7. Average end-to-end power consumption of Scheme 1 for the two flow scenario with On-Off traffic.

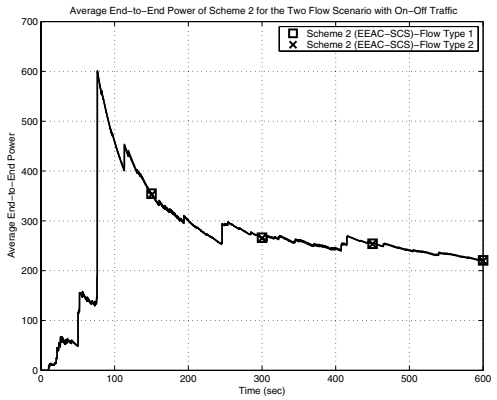

Fig. 8. Average end-to-end power consumption of Scheme 2 for the two flow scenario with On-Off traffic.

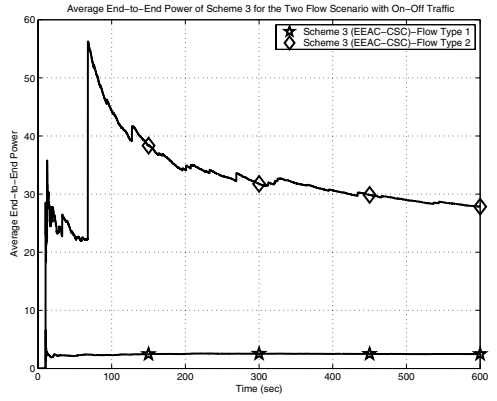

Fig. 9. Average end-to-end power consumption of Scheme 3 for the two flow scenario with On-Off traffic.

This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE ICC 2006 proceedings.

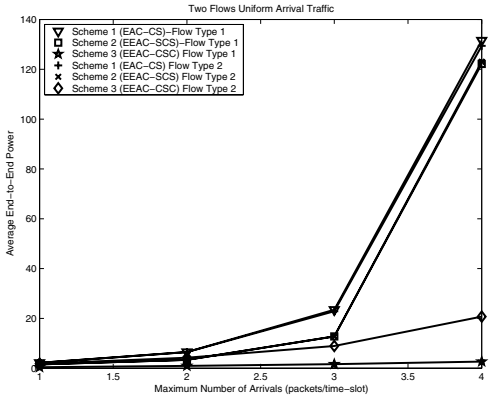

Fig. 10. Average end-to-end power consumption of the three schemes for the two flows with varying arrival rates and uniform traffic.

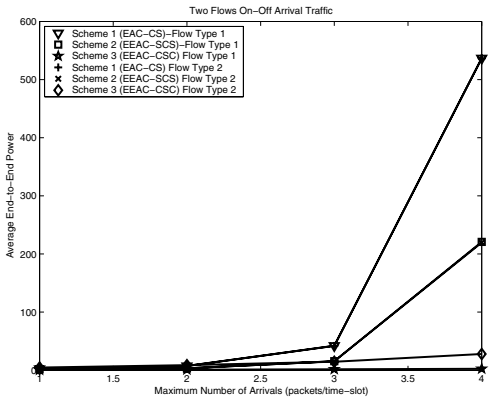

Fig. 11. Average end-to-end power consumption of the three schemes for the two flows with varying arrival rates and On-Off traffic.

power savings in wireless ad hoc networks, but also enhances the QoS granularity both in terms of the average end-to-end power consumption and delay.

Similarly, Figure 6 presents the end-to-end average delay for the two flows when *On-Off* traffic is generated by the source. Under this traffic pattern, the *On* state and *Off* state are assumed to be equally likely, while 4 packets are generated during the *On* state. Furthermore, Figures 7, 8 and 9 illustrate the average power consumption for the two flows, under each one of the three QoS provisioning schemes when *On-Off* traffic is considered. These results again confirm the capability of our proposed approach in providing finer QoS granularity both in terms of end-to-end delay and end-to end power consumption. 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 due to the fact that the *On-Off* arrival process requires the highest transmit power at any delay in an AWGN channel among all arrival processes with the same average and finite maximum arrival rate [5].

Figures 10 and 11 present 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 other two schemes for all of the arrival rates. Moreover, the performance improvement in power savings as well as QoS granularity increases as the arrival rate increases. The exponential shape of the plots is attributed to the exponential relation between transmission rate and power.

VI. Conclusions and Future Work

The rapid growth of Internet and real-time multimedia communications, along with the continuous expansion of wireless networks and services, have made the QoS provisioning in wireless ad hoc networks, more a need rather than a desire. Therefore, a cross layer QoS provisioning architecture for wireless ad hoc networks is introduced in this paper. The proposed methodology integrates the *service vector* scheme at the network layer and a power minimizing delay bounded multi user scheduling approach at the data link layer. It has been demonstrated that this integrated approach facilitates considerable power savings, while at the same time achieves enhanced QoS granularity in wireless ad hoc networks.

Due to the impracticality and implementation complexity of the optimal schedulers, suboptimal multi-user scheduling, which can only operate in AWGN channels, has been employed in this work. Our future work will investigate the implications of fading on the performance of our proposed scheme. Therefore, this suboptimal scheduler will be extended by considering fading, which would be of high practical importance. Furthermore, the probing process will be utilized to gather information regarding the performance of the wireless channel, such as the fading coefficients, which are usually unknown to the end user device.

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.

This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE ICC 2006 proceedings.