Quality of Service in Tactical Ad Hoc Networks by Priority Queuing

Otto Tronarp

Swedish Defence Research Agency (FOI),
Dep. of Communication Systems, Linköping, Sweden
e-mail: ottot@foi.se

***Abstract*: One key component in the Network Based Defence (NBD) concept is a robust high capacity radio network. Several different types of services will be provided through the network, such as group calls and situation awareness services. As all services place specific demands on packet delays and packet losses in order to be fully functional, there is a need for a Quality of Service (QoS) mechanism in the network. In this paper we examine the effect of employing priority queues in the MAC layer in a TDMA network. We compare the performance of two different priority queues, namely fixed priority queuing and weighted fair queuing.**

Our simulations show that fixed priority queuing provides a sharp delay differentiation between service classes, whereas weighted fair queuing provides the ability to control the delay differentiation. One of those queuing schemes alone might not be the best solution for providing QoS. Instead we suggest that a combination of them be used.

1. Introduction

Since the early 1980s there have been an increasing adoption of information technology in military affairs. This innovative development of warfare is often referred to as *Revolution in Military Affairs* (RMA). As a part of this ongoing RMA the Swedish Armed Forces are beginning to adopt and develop the concept of Network Based Defence (NBD) [1]. One requirement for achieving NBD is the availability of high capacity networks that can distribute information between all entities in the network. The nature of military operations places greater demands than high capacity on such a network: it must also be robust and mobile.

One technology that realizes a network that might fulfill those demands is mobile ad hoc networks. The term mobile ad hoc networks refers to wireless networks that are created dynamically through cooperation between the participating wireless nodes, i.e. without the aid of a central administrative node or fixed infrastructure [8].

Several different types of services will be provided through the network including group calls, positional services, and situation awareness. As all services place specific demands on packet delays and packet losses in order to be fully functional. For example, as the human ear is very sensitive to delays, voice transmission demands low delays. File transfer and e-mail, on the other hand, have a much higher delay tolerance. Those types of demands are commonly referred to as Quality-of-Service (QoS) demands. Traffic can also have QoS demands for non-technical reasons such as the importance of the information it carries, e.g. *flash messages* should take precedence over all other messages in a military network.

Much research is being done on the provision of QoS guarantees on the Internet. Such as the Differentiated Services (DiffServ) architecture [4], a proposed standard from the Internet Engineering Task Force (IETF) for service differentiation on the Internet. In DiffServ individual traffic flows with similar QoS demands are tagged as members of the same service class, and the service classes are given differentiated treatment in the form of different *per-hop* behavior, i.e. traffic from different service classes is given different forwarding treatment when relayed through the network. One of the suggested per-hop behaviors is *assured forwarding*, where service classes are guaranteed a minimum forwarding rate and are also guaranteed a minimum buffer capacity. Another suggested per-hop behavior is *relative service*, where the network simply guarantees that higher classes will be provided with better QoS than lower classes [6].

When a node generates traffic, or receives relay traffic, at a higher rate than it can transmit, a *queue* is formed. Usually the queued traffic is scheduled for transmission according to the *first-come-first-serve* (FCFS) discipline, where the packets are served in the order of their arrival. However, there exists more elaborated queuing schemes. An entire family of queuing disciplines are the priority queues, where the packets are given differentiated treatment according to which service class they belong to. In priority queues the packets are assigned a priority that is a function of service class, and then the packets are served in decreasing order of priority.

The fixed priority queuing (FPQ) discipline [9] is the simplest form of priority queues. As the name suggests the packets are assigned a fixed priority according to service class membership. Another form of priority queue is weighted fair queuing (WFQ) [3], [2], where each service class can be guaranteed a minimum service rate.

The major contribution to packet delays in multihop radio networks comes from the queuing time in the individual nodes when the packets are routed through the network. Hence packet delays are to a high degree a *local* or *per-hop* problem. Therefore it is a natural strategy to change the per-hop behavior in an attempt to overcome the problems.

The purpose of this work is to investigate the possibility of providing such differentiated per-hop behaviors among service classes by employing priority queues in the MAC protocol. More specifically we will study the effects of fixed priority queuing and weighted fair queuing in TDMA-networks. Since the focus is on aggregated traffic flows (service classes) and not individual traffic flows, we will not be able to give absolute QoS guarantees, but rather a relative service differentiation.

This paper is organized as follows. First we give a brief introduction of the two queuing disciplines, fixed priority queuing and weighted fair queuing, that we use. We then build up the network model that we use. After that we describe the simulations and the results and finally we draw our conclusions.

2. Queuing systems

2.1. Fixed Priority Queuing

The fixed priority queuing (FPQ) discipline [9] is probably the simplest form of priority queues. As the name suggests the packets are assigned a fixed priority according to service class membership, i.e. if the packet belongs to service class c, it is assigned the priority q_c.

There is no sense of fairness in this strategy since packets that belong to the service class with the highest priority are always served first. So packets that do not belong to that service class are not guaranteed any service at all; they are merely given what is left after the highest priority class has been served.

2.2. Weighted Fair Queuing

Weighted fair queuing (WFQ) is packet approximation of the generalized processor-sharing scheme (GPS). It was developed in parallel in [3] and, under the name packet-by-packet generalized processor sharing (PGPS), in [2]. The GPS queuing scheme has a very attractive property. One can allocate a specific percentage, ϕ_c, of the total system capacity to a service class, c. Further, if some service classes do not utilize their full share, the excess capacity is fairly shared between those classes that need it. Thus, every service class is guaranteed a minimum service rate, but they may experience a better service rate if the system is not fully utilized.

3. Network Model

3.1. Data Link Layer

We let the directed graph $G = (\mathcal{V}, \mathcal{E})$ represent the network, where $\mathcal{V}$ is the set of *vertices* and $\mathcal{E}$ is the set of directed *edges*. The vertices represent the nodes in the network, and the edges represent the links between nodes.

Carrier sense multiple access (CSMA) is one of the most frequently used MAC protocols in ad hoc networks. As most contention-based protocols it inherently have problems with providing QoS, however there are efforts to provide QoS in CSMA [7].

We will instead use time division multiple access (TDMA) [10], which is a MAC protocol that is more suitable from a QoS perspective. Here, the time is divided into time slots and each node is assigned one or several time slots where it is allowed to use the channel. Because each node has a fixed resource allocation we can make delay bound guarantees for bounded network loads. In a TDMA network, nodes with greater communication needs can be assigned more time slots than other nodes and thus increase their capacity. To maximize the network capacity we will use perfect traffic adaption, which means that the nodes are assigned time slots corresponding to the average traffic load that they are exposed to. Further, to minimize the network delay the time slots for each node should be evenly spaced in the TDMA frame. That, however, is a tough problem. To circumvent this, we will permutate the slot allocation at the start of each new frame. In that way we obtain the evenly spaced property on average over time.

The only deviation we make from the standard TDMA protocols is that we will not use the common first-come-first-serve queuing discipline. Instead we will use WFQ and FPQ.

For simplicity we make the following assumptions:

- Perfect slot synchronization, i.e. every node has access to a perfectly synchronized time reference.
- All packets are of equal length and it takes a whole time slot to transmit a packet.

3.2. Transport and Network Layer

We will only consider unicast traffic, i.e. traffic with a single source and destination. Unicast traffic can be modeled as a stream of packets where each packet enters the network at a *source node* $i \in \mathcal{V}$ and leaves the network at a *destination node* $j \in \mathcal{V}$. The source and destination nodes are chosen according to a uniform distribution. The packets from each service class c arrive at the network according to a Poisson distribution with intensity λ_N^c.

For routing we will use the shortest-path algorithm, i.e. a packet will be routed along the route that traverses the least number of nodes. This algorithm minimizes the channel utilization, i.e. it requires the least number of retransmissions of a packet for it to reach its destination. If more than one shortest route exists between two nodes, then all traffic between those two nodes always uses the same route. The routing table can be calculated with Dijkstra's algorithm [11], for example. Denote this routing table R where the table entry $R(k,l)$ is a route r_{kl} from node k to node l. We will assume that the graph $(\mathcal{V}, \mathcal{E})$ forms a connected graph, i.e. there exists a route between every node pair. Hence the number of routes in the network is given by $|R| = N(N-1)$.

3.3. Performance Measures

Since we are interested in QoS from a delay perspective, we will use the *end-to-end* packet delay as a performance measure for the different service classes. More specifically we will look at the *network delay*. We define the network delay as the expected value of the average end-to-end packet delay over all routes. We let the stochastic variable d_i^c denote the node delay, i.e. the delay that a packet from service class c experiences when it passes node i. Further, let D_{kl}^c denote the end-to-end packet delay for route r_{kl}. D_{kl} can then be written as the sum of all node delays along the route r_{kl}. We get

$$D_{kl}^c = \sum_{i:(i,j)\in r_{kl}} d_i^c. \tag{1}$$

Since there is N possible start nodes for a route and for each start node there is $N-1$ possible end nodes, there is a total of $N(N-1)$ different routes in the network. With that we get the average end-to-end packet delay over all routes as

$$D^c = \frac{1}{N(N-1)} \sum_{k\in\mathcal{V}} \sum_{l\in\mathcal{V}\setminus k} D_{kl}^c. \tag{2}$$

The network delay for service class c is the expected value of Eq. (2)

$$\overline{D}^c = E[D^c]. \tag{3}$$

4. Simulations

To evaluate the performance of the queuing systems we will perform simulations of a network consisting of 40 nodes with the topology shown in Fig. 1.

The network was generated by placing 40 nodes randomly within a quadratic area, with the sides $1\ \mathrm{km}$, in the neighborhood of Skara, Sweden. Then the link gain between nodes was calculated with Detvag-90® [5], a two dimensional deterministic wave propagation model. With the link gain known, the transmission power was chosen to be the smallest possible value such that the graph $(\mathcal{V}, \mathcal{E})$ is a connected graph, i.e. there exists a way through the network between all node pairs.

We assume that the queue in each node can hold 1000 packets and if the queue becomes full the node starts to drop packets. Furthermore, to get a sufficiently low variance in the simulation results we choose a simulation length of $1.5 \cdot 10^6$ time slots.

Here we introduce a special service class: the best effort (BE) class. Packets that belong to the BE class are not queued with the same queuing scheme as packets from other service classes. Instead they end up in their own FCFS queue. Packets in this special queue are transmitted only if the other queuing system does not have any queued packets.

We will use two different simulation setups to evaluate the two queuing systems. The first has two service classes: class 1 and class 2. Each has an average arrival rate of $\lambda_N/2$ packets/time slot, where λ_N is the total network load. The second has three service classes: class 1, class 2 and class BE. Each has an average arrival rate of $\lambda_N/3$ packets/time slot.

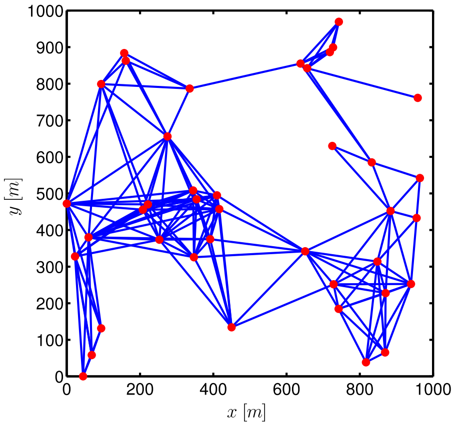


Figure 1: Network topology for the 40-node network used in the simulations.

5. Results

To see the effect the resource allocation parameters, ϕ_i, in WFQ have on the network delay we use the first simulation setup. In this setup both service classes has an average arrival rate of $\lambda_N/2$ packets/time slot and we let the resource allocations be $\phi_1 = \alpha$ for class 1 and $\phi_2 = \alpha - 1$ for class 2. In this simulation we keep the total network load, λ_N, fixed and let the parameter α vary.

The result for a moderate network load ($\lambda_N = 0.1$ packets/time slot) is shown in Fig. 2, where the network delay is viewed as a function of the parameter α. There we see that the ϕ_is give us the means to control the resource allocation. As one would expect, the network delay for the two classes is equal when they have 50% each of the resources. Then as α increases, and consequently more resources are allocated to class 1, the network delay for class 1 decreases, whereas it increases for class 2.

For comparison of FPQ against WFQ we use the second simulation setup, where each service class has an average arrival rate of $\lambda_N/3$ packets/time slot. Further, we keep the resource allocation parameters in the WFQ fixed to $\phi_1 = 0.7$ for class 1 and $\phi_2 = 0.3$ for class 2. The result is shown in Fig. 3, where the network delay is viewed as a function of the total network load, λ_N.

We see that the behavior of BE class is essentially equal for the two queuing systems. This is expected since from the BE point of view the two queuing systems work as an FPQ with 2 service classes, the low-priority BE class and the high-priority class consisting of the original class 1 and class 2. The mutual ordering between class 1 and class 2 in the high-priority class is done with the corresponding queuing system.

Far more interesting is the result for class 1 and class 2. There we see that, in FPQ, the performance of class 2 is suppressed in favor of class 1, whereas in WFQ the resources are shared between the two service classes according to the resource allocation. Here, service class 1 is suppressed, compared with FPQ, to give service class 2 its fair share of the resources. With the resource allocation parameter, ϕ_i, we can adjust allocation and, in the limiting case, when $\phi_1 \to 1$ and $\phi_2 \to 0$, the WFQ will behave much like a FPQ.

Another interesting measure to look at is the throughput, which we define as the average number of packets per time slot delivered to their final destination. The throughput for the three service classes is shown in Fig. 4

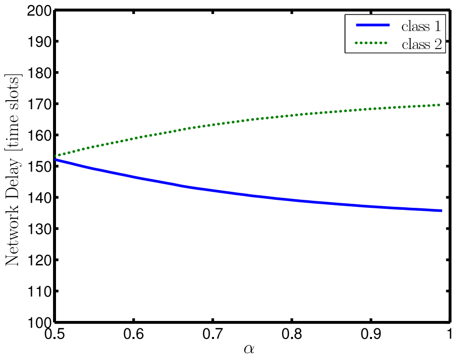
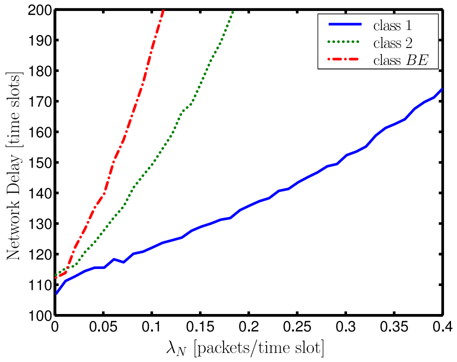
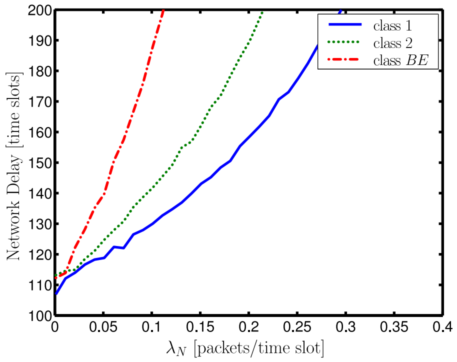


Figure 2: Network delay for a total network load of $\lambda_N = 0.1$ packets/time slot, in a traffic adaptive TDMA network with WFQ, as a function of the parameter α for two service classes with the resource allocation $\phi_1 = \alpha$ and $\phi_2 = 1 - \alpha$.

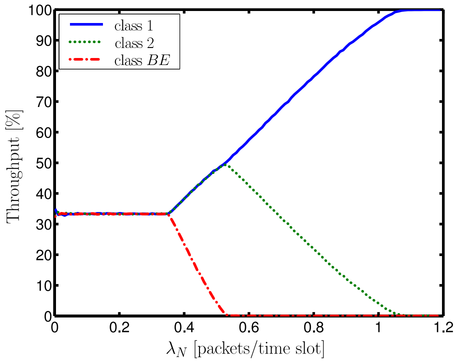


(a) FPQ

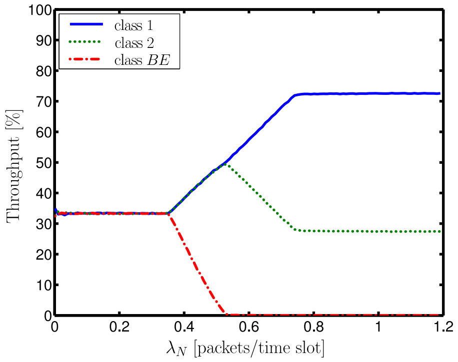


(b) WFQ with $\phi_1 = 0.7$ and $\phi_2 = 0.3$

Figure 3: Network delay in a traffic adaptive TDMA network with FPQ 3(a) and WFQ 3(b).



(a) FPQ



(b) WFQ with $\phi_1 = 0.7$ and $\phi_2 = 0.3$

Figure 4: Throughput for the three service classes as percent of total throughput in a traffic adaptive TDMA network with FPQ 4(a) and WFQ 4(b).

as a percentage of total throughput in the network.

There we more clearly see how the low-priority classes are suppressed in FPQ in favor of class 1. They are even suppressed to the extent that class 1 can take all the capacity in the network, whereas in WFQ the throughput for the two prioritized classes levels out on their specific resource allocation, which in this case is 70% for class 1 and 30% for class 2.

6. Conclusions

In this paper we have examined the possibility of providing a QoS mechanism in ad hoc networks by using priority queues in the MAC layer. More specifically we have studied the problem of providing QoS by the use of fixed priority queuing (FPQ) and weighted fair queuing (WFQ) in TDMA networks.

The evaluation of fixed priority queuing shows that it gives a very distinct delay differentiation, i.e. there is a very distinct difference in network delay between high-priority classes and classes with lower priority. The high-priority class can in fact dominate so much that no other traffic can pass through the network. This is because in fixed priority queuing, high-priority classes always take precedence over low-priority classes. Is this a desirable property? It certainly has its applications in a military context where, for example, *flash messages* should always take precedence over all other traffic. However, it might not be the best way to differentiate between traffic that has different priorities for technical reasons because in this case the priorities do not indicate the importance of the traffic and therefore it is no longer obvious that the prioritized traffic should always take precedence over other traffic.

Weighted fair queuing, on the other hand, provides the means to control how much of the resources that are dedicated to a specific service class. Consequently no service class can totally dominate the network. This might be better suited to give QoS for technical reasons.

We conclude that both of the evaluated queuing schemes have their advantages and disadvantages, and none of them alone is likely to be the answer to providing QoS. Instead, a combination of them could be used. For example; there could be a FPQ with three service classes on top; class 1 for *flash messages* and the like; class 2 for traffic that is prioritized for technical reasons; and class 3, a best effort class. Class 2 could then be divided into subclasses, and a WFQ could be used to determine the mutual ordering within that class.

6.1. Future work

In this work we have used a very simple Poisson model for the arriving traffic. A natural extension of our work would be to use a more realistic traffic model that models a *real-time* application such as a video conference or phone call. With such a model it would be interesting to study the sample probability distribution for the end-to-end packet delay for individual sessions, like a single phone call, and see how that is affected by different queuing schemes.

REFERENCES

[1] *Military Strategic Doctrine*. Swedish Armed Forces, 2002. M7740-774002.

[2] Abhay K. Parekh and Robert G. Gallager. A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks: The Single-Node Case. *IEEE/ACM Transactions on Networking*, 1(3):344–357, June 1993.

[3] Alan Demers, Srinivasan Keshav, and Scott Shenker. Analysis and Simulation of a Fair Queueing Algorithm. In *SIGCOMM*, pages 3–12. ACM, 1989.

[4] Steven Blake, David L. Black, Mark A. Carlson, Elwyn Davies, Zheng Wang, and Walter Weiss. RFC 2475: An architecture for differentiated services. ftp://ftp.rfc-editor.org/in-notes/rfc2475.txt, December 1998. Status: PROPOSED STANDARD.

[5] Börje Asp, Gunnar Eriksson, and Peter Holm. Detvag-90 – Final report. Scientific Report FOA-R–97-00566–SE, Swedish Defence Research Establishment (FOA), 1997.

[6] Constantinos Dovrolis and Parameswaran Ramanathan. A Case for Relative Differentiated Services and the Proportional Differentiation Model. *IEEE Network*, 13(5):26–34, 1999.

[7] João L. Sobrinho and A.S. Krishnakumar. Quality-of-Service in Ad Hoc Carrier Sense Multiple Access Wireless Networks. *IEEE Journal on Selected Areas of Communication*, 17(8):1353–1368, August 1999.

[8] Laura Marie Feeney, Bengt Ahlgren, and Assar Westerlund. Spontaneous Networking: An Application-Oriented Approach to Ad Hoc Networking. *IEEE Communications Magazine*, 39(6):176–181, June 2001.

[9] Leonard Kleinrock. *Queuing Systems, Volume II: Computer Applications*. Wiley-Interscience, 1976.

[10] Raphael Rom and Moshe Sidi. *Multiple Access Protocols Performance and Analysis*. Springer-Verlag, 1989.

[11] Ravindra K. Ahuja, Thomas L. Magnanti, and James B. Orlin. *Network Flows. Theory, algorithms, and applications.* Prentice Hall, 1993.