QUALITY OF SERVICE IN MOBILE AD HOC NETWORKS, CARRYING MULTIMEDIA TRAFFIC

Benny M. Nyambo\textsuperscript{1}, Gerrit K. Janssens\textsuperscript{2} and Wim Lamotte\textsuperscript{3}

\textsuperscript{1}Computer Science Department, University of Zimbabwe, Harare
e-mail: nyambo@science.ua.ac.zw

\textsuperscript{2}Logistics Research Group, Hasselt University, Diepenbeek
e-mail: gerrit.janssens@uhasselt.be

\textsuperscript{3}Expertise Centre for Digital Multimedia (EDM), Hasselt University, Diepenbeek,
e-mail: wim.lamotte@uhasselt.be

Abstract: Mobile ad hoc networks (MANETs) have become an integral part of the ubiquitous computing and communication environment, providing new infrastructure for multimedia applications such as video phone, multimedia-on-demand, and others. In order to access multimedia information in a MANET, Quality of Service (QoS) needs to be considered, such as high success rate to access multimedia data, bounded end-to-end delay, security and others. Factors, like delay and jitter, bandwidth and throughput, are studied that affect quality of service in wireless mobile ad hoc networks. Various Quality of Service architectures on IEEE 802.11-based mobile ad hoc networks are discussed, concentrating on architectures that employ cross-layer interaction in the OSI protocol stack. Architectures discussed include IntServ, DiffServ, FQMM, CEQMM, INSIGNIA, SWAN and ASAP.

Key words: mobile ad hoc networks, quality of service, multimedia traffic.

1. INTRODUCTION

Wireless communication networking is one of the most significant technologies in the current century [1]. Whilst this is an exciting development, for many people, multimedia is the holy grail of networking technologies [2]. The former see big technical challenges in providing (interactive) video on demand to every home. The latter see equally immense profits in it. This justifies a great need for research on wireless networks which carry multimedia traffic.
Wireless networks can be classified into two distinct groups: with infrastructure or without infrastructure. Networks with infrastructure are composed of mobile nodes, base stations and access points. The base stations and the access points form the core of the network and mostly they are fixed. All routing information is stored in the core network and the host just needs to pass information to the access point and the necessary route is found. In wireless networks without infrastructure, only mobile nodes exist. Each node operates both as a host and a router. If a host receives information meant for another host, it finds the best route to the destination and forwards the information to the next host. The advantage of these networks is that they are easy and cheap to set-up. They find potential use in areas such as tactical communication disaster response, battlefield, remote areas, sensor networks and many other scenarios that may arise from time to time. Networks without infrastructure are also known as Ad hoc wireless networks. When coupled with mobility, they are called Mobile Ad hoc Networks (MANETs).

![Figure 1. A mobile ad hoc network (MANET) showing mobile nodes, wireless links and signal range for each node [8]](image)

In a MANET (see Figure 1), nodes within interference range share status information in a way that nodes are conscious of the presence of all their neighbours. MANETs are expected to become the future of wireless networks, because they are practical, versatile, easy to use and inexpensive to setup. Researchers project a world where the network instantly updates and reconfigures itself to keep people connected wherever they go.

On the other hand there is a great advance in multimedia transmission in networks. This has seen the emergence of Internet telephone or Voice over IP (VoIP), multimedia streaming and real-time Video over IP. Because of the increase in the use of multimedia in conjunction with wireless technologies, there is a demand for high speed wireless networks that are multimedia enhanced.
In future, there may also be a need to hook up MANET users to the Internet. In this way a MANET becomes part of a heterogeneous network. Although MANET users would require all real-time services that wired network users enjoy, there are still are a lot of challenges that need to be addressed. Real-time time traffic is more sensitive to network quality of service (QoS) as compared to best-effort traffic such as email and file transfer. But MANETs are very low on bandwidth and they are usually battery operated so they are energy sensitive.

2. QUALITY OF SERVICE

Quality of Service (QoS) describes the level of technical assurance provided by a network, while transporting a packet stream from a source node to a destination node. There are many definitions of Quality of Service available in literature. The International Telecommunications Union (ITU) in its 1995 recommendation, (ITU-T Rec. E.800) defines QoS as "the collective effect of service performance which determine the degree of satisfaction of a user of the service." They characterize quality of service by combining aspects of service support performance, service operability performance, service availability performance, service security performance and other factors specific to each service. The International Systems Organization in its ISO 8402 defines Quality of Service as "the totality of characteristics of an entity that bear on its ability to satisfy stated and implied needs." Similarly, ISO 9000 defines quality as the "degree to which a set of inherent characteristics fulfills requirements." The ISO 8402 definition seems better from the user's view. In any event, QoS is clearly a subset of overall quality.

In computer networks, the goal of QoS support is to achieve more predictable, reliable and deterministic communication behaviour, so that information carried by the network can be better preserved while optimally utilizing network resources. QoS can also be defined as the ability of a network element (e.g. an application, a host or a router) to provide some level of assurance for consistent network data delivery. QoS is based on an agreement or a guarantee by the network to provide a set of measurable pre-specified service attributes to the user in terms of available bandwidth, probability of packet loss (loss rate), throughput, network delay, delay variance (jitter), and security. Different applications require different QoS requirements, from the network. Real-time applications need packets by a certain time; otherwise the packets become worthless; non-real-time applications are concerned more on reliability instead. For multimedia traffic over the internet, the ultimate goal is to preserve both mission-critical data in the presence of multimedia voice and video and to preserve the voice and video quality in the presence of busy data traffic.

In its present form, the Internet does not offer any quality of service (QoS) guarantees to streaming media over the Internet [3]. Therefore there is a need for new models of transmitting data on the internet that guarantees quality of service to real-time multimedia traffic.
2.1. Quality-of-Service Metrics

QoS metrics are base parameters of quality for a network. QoS parameters include throughput or bandwidth, delay, jitter, packet loss or error rate, security, network availability, and battery life. The QoS can be defined in terms of one the parameters or set of parameters in varied proportions. Throughput or bit rate is the rate at which end-systems can exchange information.

Video and voice packets generally require large bandwidth. Otherwise bottlenecks will develop in the network links leading to packet losses. Packet loss refers to the percentage of packets that fail to reach their destinations for various reasons. A packet loss of 1% produces a jerky video, while loss of 2% will start to render video unusable, though audio can be acceptable.

There are two notions associated with rates at the interface between an end-system and a network. These are the access speed (bandwidth) and the bit rates. The access speed is the frequency at which bits may be sent or received over the interface between the end-system and the network. This frequency is always determined by the technology used by the network. In certain cases this frequency is determined by independent clocking signals and bits can only be sent or received when matching these signals.

The available bandwidth of a path is a concave metric that defines the width of the path. In practice it is a bottleneck which defines the bandwidth that a service can be allocated to.

\[ B_{\text{avail}} = \min \{B_x\} \]

where \( B_x \) is the bandwidth or the access rate at each node \( x \) in a given path from source \( i \) to a destination \( k \).

However, not all networks are capable of transporting data transmitted at the sustained access speed of the network interface. Several networks cannot accept data during certain periods because of internal congestion, lack of capacity, or because the user has subscribed to a bit rate lower than the access rate.

In MANETs several factors affect the overall throughput of any protocol operating in an ad hoc network. For example, node mobility may cause link failures, which will negatively impact routing and quality-of-service support. Network size, control overhead, and traffic intensity will have a considerable impact on network scalability. These factors along with inherent characteristics of ad hoc networks may result in unpredictable variations in the overall network behaviour.

When data is transferred over a communications medium, such as a MANET, the average transfer speed is often described as throughput. This measurement includes all protocol overhead information, such as packet headers and other data that is included in the transfer process. It also includes packets that are retransmitted because of network conflicts or errors. Goodput, on the other hand, only measures the throughput of actual data.
Certain networks cannot accept a sustained traffic at access speed of the network interface. Goodput can be calculated by dividing the size of a transmitted file by the time it takes to transfer the file. Since this calculation does not include the additional information that is transferred between systems, the Goodput measurement is always less than or equal to the throughput. For example, the maximum transmission unit MTU of an Ethernet connection is 1,500 bytes. Therefore, any file over 1,500 bytes must be split into multiple packets. Each packet includes header information (typically 40 bytes), which adds to the total amount of data that needs to be transferred. Therefore, the Goodput of an Ethernet connection is always slightly less than the throughput.

While Goodput is typically close to the throughput measurement, several factors can cause the Goodput to decrease. For example, network congestion may cause data collisions, which requires packets to be resent. Many protocols also require acknowledgment that packets have been received on the other end, which adds additional overhead to the transfer process. Whenever more overhead is added to a data transfer, it increases the difference between the throughput and the Goodput.

### 2.1.1. Network Delay

Network latency or delay refers to the total transit time for packets to arrive at the intended destination node. It is the time elapsing between the emission of the first bit of a data block by the transmitting end and its reception by the receiving end-system.

Store-and–forward packet networks, based on packet switches or routers, may have substantial transit delays, up to seconds for long-haul connections. The total end-to-end delay includes three components:

- The time necessary at the source to wait for the medium to be available or for the network to be ready to accept the block of information. This time is sometimes called access delay.
- The time necessary to actually transmit the sequence of bits of the blocks, one after the other, once the network is ready. This time is called the bit transmission delay. For a given block of size, this delay only depends on the access speed.
- The network transit delay which is a property of network machinery.

The end-to-end delay metric of a path is additive. It is the sum of the propagation delays of the path. It is also an indication of the length of the path. The propagation and queuing delays from a source of communication to the destination is additive. Suppose \( d_{ij} \) represents the delay for link \((i,j)\). The path \( p \) linking \( i \) to \( m \) nodes, \( p = (i,j,k,\ldots,l,m) \), has delay \( D \) given by equation 2.2.

\[
D = \sum_{x=i, y=j}^{x=l, y=m} d_{xy}
\]  

\[ (2.2) \]
In equation 2.2, $d_{xy}$ is the delay experienced in the link between nodes $x$ and $y$. This means that effort has to be made to reduce delay in all links is a path from a source to the intended destination.

In packet networks, packets are queued in a buffer for processing. Queuing delay depends on the packet scheduling algorithm, the buffer size, and other factors. A useful theorem on the relationship between the average queue length and the average queue delay is given by Little’s theorem as follows:

$$\eta_N = \lambda \eta_d$$

Where $\lambda$ is the arrival rate, $\eta_N = E\{N\}$ is the expected queue length and $\eta_d = E\{d\}$ is the expected delay [5].

In interactive applications of real time sound transmission, as well as in virtual reality, the overall one way delay needs to be short in order to give the user an impression of real time responses. A maximum value on the order of 0.1 to 0.5 seconds is required to accomplish this goal. Based on subject tests, the International Telecommunication Unit (ITU) G.114 specification recommends less than a 150 millisecond one-way end-to-end delay for high-quality real time traffic in telecommunication. (ITU G.114, 1996) The G.114 time limits are shown in Table 1 and Table 2 shows the effect of delay on human voice perception [4].
Table 1. End-to-end Audio Delay and the Human Ear

<table>
<thead>
<tr>
<th>Audio Delay (ms)</th>
<th>Effect of Delay on Human Voice Perception</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt; 600</td>
<td>Speech is unintelligible and incoherent</td>
</tr>
<tr>
<td>600</td>
<td>Speech is barely coherent</td>
</tr>
<tr>
<td>250</td>
<td>Speech is annoying but comprehensible</td>
</tr>
<tr>
<td>100</td>
<td>Imperceptible different between audio and real speech</td>
</tr>
<tr>
<td>50</td>
<td>Humans cannot distinguish between audio and real speech</td>
</tr>
</tbody>
</table>

Table 2. G.114 Limits for One-way Transmission Time

<table>
<thead>
<tr>
<th>One-way transmission time</th>
<th>Effect of Delay on Human Voice Perception</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 to 150 ms</td>
<td>Acceptable for most users</td>
</tr>
<tr>
<td>150 to 400 ms</td>
<td>Acceptable, but had impact</td>
</tr>
<tr>
<td>400 ms and above</td>
<td>Unacceptable</td>
</tr>
</tbody>
</table>

2.1.2. Delay Variation/Jitter

Jitter is the variation in latency for packets within a given data stream. In transmission technology, jitter refers to the variation of the delay generated by the transmission equipment. Jitter is generally caused by congestion in the Internet Protocol network. The congestion can occur either at the interfaces of a router which acts as a gateway device or in a provider or carrier network if the circuit has not been provisioned in a proper way.

The average jitter $q_j$ experienced can be expressed as:

$$q_j = \frac{1}{n-1} \sum_{j=2}^{n} (R_i - R_{i-1}) - (S_i - S_{i-1})$$  \hspace{1cm} (2.4)

where $S_i$ is the time the packet $i$ was sent from source and $R_i$ is the time at which packet $i$ arrives to its destination and $n$ is the number of hops [9].

![Figure 3. Packet jitter caused by the network](image)

The jitter present in packet networks complicates the decoding process in the receiver device because the decoder needs to have packets of data readily available at the accurately correct time instants. If the data is not available, the decoder is not able to produce smooth, continuous speech or a continuous video stream. Thus, in addition to adding to the delay, jitter leads to a timing problem for the receiver. Since the receiving decompression algorithm requires fixed spacing between the
packets, the typical solution is to implement a jitter buffer within the gateway, to make sure that packets are available when needed. The jitter buffer deliberately delays incoming packets in order to present them to the decompression algorithm at fixed spacing. The jitter buffer also fixes any out-of-order errors by looking at the sequence number in the RTP frames. The voice decompression engine receives packets directly on time; the individual packets are delayed further in transit, increasing the overall latency. Jitter causes either blocky, jerky or undesirable audio. Jitter for packets within a given stream should not exceed 20-50 milliseconds.

A lot of research on QoS has occurred, especially in wired networks [6], [7]. IntServ and DiffServ are two well-known, QoS models, designed for wired networks. Although much progress has been achieved on QoS for wire-based networks, a lot is still to be done when it comes to wireless networks. The unique characteristics like shared medium, mobility and the distributed multi-hop communication in wireless networks make it difficult to give a quality of service anticipated by the network user.

2.1.3. Human Perception to QoS

People are more sensitive to alterations of audio than visual signals. Our tolerance of transmission errors affecting audio streams is much lower than our tolerance of errors affecting motion video streams. If, in an application, audio and video are transmitted simultaneously, both streams might compete for resources. In such cases the audio stream must have priority over the video stream. A good example is audio-video conferencing in packet mode supported by personal computers of workstations. The bit rates required for multimedia traffic depends on the quality and standard of technology used. Table 3 shows the required bit rate for various audio standards [11].

<table>
<thead>
<tr>
<th>Quality</th>
<th>Technique/Standard</th>
<th>Bit Rate in Kbps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Telephone quality</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard</td>
<td>G.711 PCM</td>
<td>64</td>
</tr>
<tr>
<td>Standard</td>
<td>G.721ADCM</td>
<td>32</td>
</tr>
<tr>
<td>Improved</td>
<td>G.722 SH-ADCM</td>
<td>48, 56, 64</td>
</tr>
<tr>
<td>Lower</td>
<td>G728 LD-CELP</td>
<td>16</td>
</tr>
<tr>
<td>CD quality (stereo)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consumer CD audio</td>
<td>CD-DA</td>
<td>1411</td>
</tr>
<tr>
<td>Consumer CD audio</td>
<td>MPEG audio FFT</td>
<td>192</td>
</tr>
<tr>
<td>Improved (sound studio)</td>
<td>MPEG audio FFT</td>
<td>384</td>
</tr>
</tbody>
</table>

In the context of telecommunications, quality of service (QoS) definition borders on the degree of a user’s satisfaction with the service. The QoS is thought to be divided into, speech or voice and video communication quality, service
performance, and the necessary terminal equipment performance. The voice and video communication (or transmission) quality is more user-directed and, therefore, determines acceptability of the service from the user’s point of view. [10]

Although a lot of research has been devoted to mechanisms supporting the QoS in different types of networks, less has been done to support the unified, comparable assessment of the quality really achieved by the individual approaches. Many researchers constrain themselves to prove that a certain mechanism is capable of reducing the packet loss rate, packet delay or packet jitter considering those measures as sufficient to characterize the quality of the resulting multimedia transmission. However, the above mentioned parameters cannot be easily and uniquely transformed into a quality of the transmission. In fact such transformation could be different for every coding scheme, loss concealment scheme and delay/jitter handling as shown in Table 1.

Quality can be defined as the result of the judgement of a perceived constitution of an entity with regard to its desired constitution. For a perceiving person it is a characteristic of the identity of the entity. Applying this definition to multimedia, voice and video quality can be regarded as the result of a perception and assessment process, during which the assessing subject establishes a relationship between the perceived and the desired or expected multimedia signal. Multimedia quality can be defined as the result of the subject’s judgement on spoken language, which he/she perceives in a specific situation and judges instantaneously according to his/her experience, motivation, and expectation. Regarding voice communication systems, quality is the customer’s perception of a service or product, and multimedia quality measurement is a means of measuring customer experience of telecommunication services. The most accurate method of measuring multimedia quality would be to actually ask the callers during or after the call, for their opinion on the quality [8].

In practice, two broad classes of voice quality metrics exist: subjective and objective. Subjective measures are conducted by using a panel of people to assess the voice quality of live or recorded speech signals from the voice communication system/device under test for various adverse distortion conditions.

<table>
<thead>
<tr>
<th>Score</th>
<th>Quality of speech</th>
<th>Impairment</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Excellent</td>
<td>Imperceptible</td>
</tr>
<tr>
<td>4</td>
<td>Good</td>
<td>Perceptible, but not annoying</td>
</tr>
<tr>
<td>3</td>
<td>Fair</td>
<td>Slightly annoying</td>
</tr>
<tr>
<td>2</td>
<td>Poor</td>
<td>Annoying</td>
</tr>
<tr>
<td>1</td>
<td>Bad</td>
<td>Very annoying</td>
</tr>
</tbody>
</table>
The speech quality is expressed in terms of various forms of a mean opinion score (MOS), which is the average quality perceived by the members of the panel. Table 4 shows the mean opinion score scale against the user perception levels.

### 2.2. Quality of Service Issues in MANETs

While it is difficult to provide quality of service in wired networks, MANETs and wireless networks in general bring in more difficulties because of their characteristics. Traditional Internet QoS protocols cannot be easily migrated to the wireless environment due to the error-prone nature of wireless links and the high mobility of mobile devices. This is true for mobile ad hoc networks where every node moves arbitrarily, causing the multi-hop network topology to change randomly and at unpredictable times.

The main objective of QoS in MANETs is to achieve a more deterministic network behaviour so that information carried by the network can be better delivered and network resources are best utilized. This can be achieved by raising the priority of a traffic flow or limiting the priority of another flow. As multimedia requires high bandwidth, getting it to work over fixed connections is a hard task, but the requirement to work efficiently on MANETs is even harder. Besides the known interference problems faced by wireless networks, MANETs have their own characteristics that bring challenges in guaranteeing quality of service. They include the following:

- **Node mobility and non-infrastructure**: Node mobility is the basic cause of the dynamic network topologies in MANETs. The MAC layer allocation of bandwidth to each node changes dynamically according to mobility scenarios. The bandwidth is difficult to control due to the non-infrastructure feature. The roles of nodes as a host or router change together with node mobility and the dynamic topology. In MANETs there is no central infrastructure that can regulate the distribution of resources to nodes. The network is decentralized, where all network activity including discovering the topology and delivering messages are to be executed by the nodes themselves, that is, routing functionality will be incorporated into mobile nodes. The nodes are free to move about and organize themselves into a network, thus the network topology may change rapidly and unpredictably over time. The challenge is to design decentralized QoS schemes. The dynamic nature of MANETs causes the precise maintenance of network state information extremely difficult [21].

- **Limited Bandwidth and Network Size**: At first glance, this feature makes the scalability problem less likely to occur in MANETs. However, as fast radios and efficient low bandwidth compression technology develop rapidly, the emergence of high-speed and large-sized MANETs with plenty
of applications is foreseeable. At that time, MANETs will inevitably meet the scalability problem as in high-speed fixed networks today.

- **Time-Varying Feature**: In MANETs, the link capacity is time varying due to the physical environment of nodes, the mobility of nodes, and the dynamics of the network topology. This time-varying feature makes the service provision mechanisms in MANETs more difficult than in wired fixed networks. Take the signalling protocol for example. A signalling protocol generally comprises three phases: connection establishment, connection teardown, and connection maintenance. Literature predicts that a larger percentage of link capacity is allocated to control overhead in a network with smaller and time-varying aggregate network capacity. For MANETs with dynamic topology and link capacities, the overhead of connection maintenance usually outweigh the initial cost of establishing the connection [8].

- **Power Constraints** [21]: The processing capability of nodes is limited due to the limited battery power. This means there should be low processing overhead of nodes and thus, the control algorithms and QoS algorithms should use bandwidth and energy efficiently. QoS challenges due to limited capabilities of mobile nodes in terms of processing power, storage capacity, or energy. The limited capabilities challenge, influence, and shape the QoS design for instance by forcing a distributed approach, avoiding lookup tables, accommodating dormant devices, or adopting simpler lightweight algorithms.

- **Lack of central authority**: It concerns to maintain central information on flows, routes, or connections. QoS challenges due to Hidden and Exposed Terminal Problems. In a MAC layer with the traditional carrier sense multiple access (CSMA) protocol, multi-hop packet relaying introduces the “hidden terminal” and “exposed terminal” problems. The hidden terminal problem happens when signals of two nodes, say A and C, that are out of each other’s transmission ranges collide at a common receiver, say node B (see Figure 4a). The exposed terminal problem will result from a scenario where node B attempts to transmit data A while node C is transmitting to node D. In such a case, node B is exposed to the transmission range of node C and thus defers its transmission even though it would not interfere with the reception at node D (see Figure 4b).

All these challenges lead to serious concern in the provision of quality of service in ad-hoc networks. Some of these challenges influence greatly the issue of flow reservation in ad-hoc networks.
3. Quality of Service Models for MANETs

A QoS model specifies the architecture in which certain services are provided in the network. A QoS model for MANETs should first consider the challenges of MANETs, such as dynamic topology, low available bandwidth and time-varying link capacity. In addition, the potential commercial applications of MANETs require a seamless connection to the Internet. Thus, the QoS model for MANETs should also consider the existing QoS architectures in the Internet. In this section, existing QoS models for the Internet and MANETs: IntServ and DiffServ, INSIGNIA, SWAN and ASAP are discussed.
3.1. Integrated Services (IntServ)

The IntServ model [6] merges the advantages of datagram networks and circuit switched networks. It can provide a circuit-switched service in packet-switched networks. In circuit-switching, this path is decided upon and established before the data transmission starts. For the whole communication session, the route is dedicated and exclusive, and released only when the session terminates. In packet-switching there is no predetermined path and packets are sent towards the destination independent of each other. Each packet finds its own path to the destination making routing decisions at various nodes in the path.

The Resource Reservation Protocol (RSVP) was designed as the primary signalling protocol to setup and maintain the virtual connection. RSVP is also used to propagate the attributes of the data flow and to request resources along the path. Routers finally apply corresponding resource management schemes to support QoS specifications of the connection. Based on these mechanisms, IntServ provides quantitative QoS for every flow.

3.1.1. Resource Reservation Protocol (RSVP)

The (RSVP) is a classic two-pass protocol using out-of-band signalling. Figure 5 shows the classical operation of RSVP. The messages used are the Path message, which originates from the traffic sender and the reservation (Resv) message, which originates from the traffic receivers. The primary roles of the Path message are first to install reverse routing state in each router along the path, and second to provide receivers with information about the characteristics of the sender traffic and end-to-end path so that they can make appropriate reservation requests. Resv messages finally carry reservation requests to the routers along the distribution tree between receivers and senders. RSVP state is "soft-state", after a certain expire time, the state of the path and the reserved resource is released. Periodical issuing of Path or Resv messages are necessary to keep the reservation alive. Additional signalling information allows the soft state timeout to adapt to the refresh period. Furthermore, RSVP provides a routing triggered local repair [13] mechanism to overcome the need for a very fast refresh rate in order to react to route changes.

Figure 5. The operation of RSVP [13].
3.1.2. Disadvantages of IntServ/RSVP

The shortcomings of IntServ in MANET environments are in scalability and signalling. The amount of state information in IntServ increases proportionally with the number of flows since IntServ provides per flow granularity. Keeping flow state information will cost a huge storage and processing overhead for the mobile host whose storage and computing capacity are scarce. The scalability problem is less likely to occur in current MANETs considering the small number of flows, the limited size of the network and the bandwidth of the wireless links. However, as the quality of wireless technology increases rapidly, high speed and large size MANETs may be found in the future and the problem will manifest [11].

The signalling protocols have three phases: connection establishment, connection maintenance and connection teardown. Since MANETs have dynamic topologies, this approach is not reliable since routes may change quickly and the handshaking would not be fast enough. Due to its out-of-band approach, RSVP produces a significant signalling overhead. This means that RSVP signalling packets will contend for bandwidth with data packets and consume a substantial amount of bandwidth in MANETs. This may be of importance if the refresh rate is high because the message size is not negligible in RSVP. A high refresh rate might occur when no route-change notification service from the routing layer is available. This causes local repair to fail.

3.2. Differentiated Service (DiffServ)

DiffServ [7] has been designed to overcome the difficulty of implementing and deploying IntServ and RSVP in the computer network [13]. IntServ provides per-flow guarantees but Differentiated Services (DiffServ) maps multiple flows into a few service levels. DiffServ defines three types of nodes. An ingress node is a mobile node that sends data. Interior nodes are the nodes forwarding data for other nodes. An egress node is a destination node. At the boundary of the network, traffic entering a network is classified, conditioned and assigned to different behaviour aggregates by marking a special DS (Differentiated Services) field in the IP packet header which supersedes the TOS field in IPv4 and CLASS field in IPv6. Within the core of the network, interior nodes, packets are forwarded according to the per-hop behaviour (PHB) associated with the DSCP (Differentiated Service Code Point). An intermediate network node performs a PHB, which is a logical instantiation performing traffic forward behaviour. The forward behaviour normally follows the traffic resource allocation per link based on the priority defined in DSCP. The traffic resource is determined based on packet loss rate, propagation delay and jitter. This eliminates the need to keep any flow state information elsewhere in the network [11]. Figure 6 shows the DiffServ architecture with the Ingress and Egress nodes connected through a cloud.
3.2.1. Problems of DiffServ

The DiffServ approach also shows some drawbacks in MANETs. The first problem is on soft QoS guarantees in DiffServ. DiffServ uses a relative-priority scheme to map the quality of service requirements to a service level. This aggregation results in a more scalable but also in more approximate service to user flow.

The other drawback is on the SLA (Service Level Agreement). DiffServ is based on the concept of SLA’s which are contracts between customers and their Internet Service Providers (ISPs) that specify the forwarding service that each customer should receive. In a DiffServ domain it is important that sufficient resources are provisioned to support the SLA’s committed by the domain. Also, the boundary nodes must monitor the arriving traffic for each service class and they should perform traffic classification and conditioning to enforce the negotiated SLA’s.

If a customer acquires QoS parameters and pays for such parameters then some entity should exist to assure them. In a completely ad hoc topology where there is no concept of service provider and client and where only clients exist, it would be difficult to innovate QoS, since there is no obligation from someone to someone else what makes QoS almost infeasible.

The problem of an ambiguous core network. DiffServ has the benefit that traffic classification and conditioning only needs to be done at the boundary nodes. This makes quality of service provisioning easier in the core of the network. In MANETs, however, there is no clear definition of what is the core network because every node is a potential sender, receiver and router. This means that several separate flow states are maintained at intermediate nodes just like in IntServ [13].

3.3. Flexible Quality of Service Model for Mobile Ad Hoc Networks (FQMM)

Flexible Quality of Service Model for Mobile Ad Hoc Networks (FQMM) [8], is another QoS model which has been designed to combine the IntServ and the
DiffServ models in order to combine the strengths of both models whilst at the same
time trying to override the weaknesses and disadvantages stated above. Three kinds
of nodes are defined, exactly as in DiffServ. An ingress node is a mobile node that
sends data. Interior nodes are the nodes forwarding data for other nodes. An egress
node is a destination node.

Figure 7 shows a scenario where there are two connections: one is from M1 to
M6 and another from M8 to M2. The roles of the nodes change depending on what
part they are playing for a specific flow. Node M8 is an interior node for flow C1
and it is an Ingress node for flow C2.

The basic idea of FQMM is that it uses both the per-flow state property of
IntServ and the service differentiation of DiffServ. This is achieved by preserving
per-flow granularity for a small portion of traffic in the MANET, given that a large
amount of the traffic belongs to per aggregate of flows, that is, per-class
granularity. A traffic conditioner is placed at the ingress nodes where the traffic
originates. Components of the conditioner include traffic profile, meter, marker and
dropper. The traffic profile decides the policy of other components which change
the configuration according to the traffic profile.

It is responsible for re-marking or discarding packets according to the traffic
profile, which describes the temporal properties of the traffic stream such as
transmission rate and burst size.

3.3.1. Problems associated with FQMM

FQMM is a first and important attempt at proposing a QoS model for
MANETs. The first problem is the problem with scalability. FQMM aims to tackle
the scalability problem of IntServ. But without an explicit control on the number of
services with per-flow granularity, the problem still remains [14].
The other problem is that due to its DiffServ behaviour in ingress nodes, FQMM may not be able to satisfy hard QoS requirements. It could be difficult to code the PHB in the DS field if the PHB includes per-flow granularity, considering the DS field is at most 8 bits without extension [13].

Lastly, how to make a dynamically negotiated traffic profile is a well-known DiffServ problem and FQMM seems not to solve it.

3.4. Complete and Efficient Quality of Service (QoS) Model for MANETs (CEQMM)

Badis and Al Agha in [20] define CEQMM, a Complete and Efficient Quality of Service (QoS) Model for MANETs which combines the positive aspects of both IntServ and DiffServ. It uses a hybrid per-flow and per-class provisioning scheme. In such a scheme, QoS traffic of highest priority is given per-flow provisioning while other priority QoS classes are given per-class provisioning. To offer this scheme and to ensure that certain packets receive higher priority transmission than other packets, priority classifier, active queue management and packet scheduler are integrated. Figure 8 shows the CEQMM Architecture.
CEQMM applies the QOLSR protocol to support multiple-metric routing criteria and to respond quickly when changes in topology and/or QoS conditions are detected. Once a path is chosen for one QoS flow, CEQMM performs call admission control (CAC) at each intermediate node. For only QoS flows of highest priority, a node can proceed to soft, and later to hard bandwidth reservation on links during the CAC process. CEQMM implements congestion avoidance mechanisms to prevent a network from entering the congested state. However, in MANETs, network congestion can still occur frequently under mobility. In order to prevent performance degradation due to mobility-triggered congestion, CEQMM uses a new congestion control scheme. Preliminary simulation results show that CEQMM achieves better performance than the best-effort model [20].

One limitation of implementing CEQMM for MANETs is that in case of continuous node movement, the average delay is very long and as a result many packets are dropped, which makes it unfavourable for multimedia traffic [19].

3.5. INSIGNIA: In-Band Signalling Support for QoS in Mobile Ad Hoc Networks

INSIGNIA [15] is a signalling protocol designed explicitly for MANETs. It can be combined with a variety of routing protocols to come up with an effective QoS model. It supports fast flow reservation, restoration and adaptation algorithms that are specifically designed to deliver adaptive real-time service. INSIGNIA implements an in-band signalling approach by encapsulating some control signals in the IP option of every data packet, which is now called the INSIGNIA option as shown in Figure 9.

![Figure 9. The INSIGNIA IP option in a packet [15]](image)

Flow state information is kept in every node in a particular path. This is done in such a way that the flow state information is periodically refreshed by the received signalling information. This is called soft-state reservation. When a source node wants to establish a reservation to a destination node, it sets the reservation (RES) mode bit in the INSIGNIA IP option service mode of a data packet and sends the packet towards the destination. The bandwidth request field allows a source to specify its maximum (MAX) and minimum (MIN) bandwidth requirements. On reception of a RES intermediate routing nodes execute admission control to accept or deny the request. When a node accepts a request, resources are
committed and subsequent packets are scheduled accordingly, otherwise the reservation is denied and packets are treated as best effort (BE) mode packets.

In the case where a RES packet is received and no resources have been allocated, the admission controller attempts to make a new reservation. This is a reactive local repair mechanism and commonly occurs when flows are rerouted during the lifetime of an ongoing session due to host mobility. When a node receives a request packet with the bandwidth indicator bit set to MAX indicates that all nodes before this node have sufficient resources to support the maximum bandwidth requested. If the bandwidth indicator is set to MIN it implies that at least one of the intermediate nodes is a bottleneck node and the maximum bandwidth requirement may not be met. As a result "partial reservations" may exist between source and bottleneck node, these resources remain reserved until explicitly released.

When a reservation is received at the destination node, INSIGNIA checks the reservation establishment status. QoS reporting message can be sent by destination nodes to inform source nodes of the ongoing status of flows. These messages do not have to travel on the reverse path towards a node.

The report commands can be either scale-down or scale-up commands. A scale-down command requests a source either to send with the rate specified as MINIMUM instead of MAXIMUM or to send its packets as best effort instead of MINIMUM depending on the current sending rate of the source node. This will
have the effect of clearing any partial reservation. A scale up requests a source node to initiate a reservation for some MINIMUM or MAXIMUM rate, depending on the actual flow state. Figure 10 shows some of the operations associated with reservation and adaptation of flows in INSIGNIA.

3.5.1. Disadvantages of INSIGNIA in MANETs

The most obvious drawback of INSIGNIA is its scalability problem due to the flow state information which is kept within the nodes of a certain path. INSIGNIA’s bandwidth usage is not efficient. The extra reservation on the path from the sending node to the bottleneck is a waste of bandwidth until an explicit release message is sent. Although this waste does not last long, topology changing of a MANET will make this reservation waste propagate frequently. Furthermore releasing partial reservations using QoS reports enforces source nodes either to set the bandwidth indicator of the INSIGNIA option field to MINIMUM or to send the packets as best effort depending on the actual flow state. In both cases the opportunity to scale up is lost.

INSIGNIA does not provide any mechanism to dynamically change the frequency by which control signals are inserted into the data packets. This imposes a major processing overhead on the network. Only two bandwidth levels to be used are offered, MINIMUM and MAXIMUM. A more fine-grained approach would be needed in order to satisfy application requirements and to fully exploit the resources available.

INSIGNIA differentiates traffic into best effort (BE) traffic and Quality of Service traffic which is split into base or enhanced quality of service (BE/EQ) depending on the payload of the network. Multimedia traffic however, comes in different types varying from online games, internet telephone, video conferencing, video streaming and many others. These different kinds of traffic have to be treated with different priorities by the network just like it is done in DiffServ.

In INSIGNIA, if the available bandwidth is just enough to only meet the minimum bandwidth requirement needs of the base QOS, enhanced QOS packets are degraded to best effort packets at bottleneck nodes by changing the service mode for EQ packets from RES to BE. When a node encounters degraded packets, it releases bandwidth that would have been allocated to enhanced QOS packets. In this way unused resources are released in intermediate nodes, it does not give a guarantee that the flow received will be at a useful quality of service level since it is below the minimum required level. There is a need to keep alive only those flows that have a guaranteed quality of service level.

3.6. Stateless Wireless Ad-Hoc Networks (SWAN)

A Stateless Wireless Ad-Hoc Network (SWAN) [16] is a stateless network model that has been specifically designed to provide service differentiation in wireless ad-hoc networks employing a best-effort distributed wireless MAC. It
distinguishes between two traffic classes: real-time UDP traffic and best-effort UDP and TCP traffic.

A classifier differentiates between real-time traffic and best-effort traffic. Leaky-bucket traffic shaper delays best-effort packets at a rate previously calculated, applying an AIMD (Additive Increase Multiplicative Decrease) rate control algorithm. Every node measures the per-hop MAC delays locally and this information is used as feedback to the rate controller. Rate control restricts the bandwidth for best-effort traffic so that real-time applications can use the required bandwidth. On the other hand the bandwidth not used by real-time applications can be efficiently used by best-effort traffic. The total best-effort and real-time traffic transported over a local shared channel is limited below a certain ‘threshold rate’ to avoid excessive delays.

Moreover, SWAN uses sender-based admission control for real-time UDP traffic. The rate measurements from aggregated real-time traffic at each node are employed as feedback. This mechanism sends an end-to-end request/response probe to estimate the local bandwidth availability and then determines whether a new real-time session should be admitted or not. The source node is responsible for sending a probing request packet toward the destination node. This request is a UDP packet containing a “bottleneck bandwidth” field. All intermediate nodes between the source and destination must process this packet, check their bandwidth availability and update the bottleneck bandwidth field in the case that their own bandwidth is less than the current value in the field.

The available bandwidth can be calculated as the difference between an admission threshold and the current rate of real-time traffic. The admission threshold is set below the maximum available resources to enable that real-time and best-effort traffic are able to share the channel efficiently. Finally, the destination node receives the packet and returns a probing response packet with a copy of the bottleneck bandwidth found along the path back to the source. When the source receives the probing response it compares the end-to-end bandwidth availability and the bandwidth requirement and decides whether to admit a real-time flow accordingly. If the flow is admitted the real-time packets are marked as RT (real-time) packets and they bypass the shaper mechanism at the intermediate nodes and are thus not regulated.

The traffic load conditions and network topology change dynamically so that real-time sessions might not be able to maintain the bandwidth and delay bound requirements and they must be rejected or readmitted. For this reason it is said that SWAN offers ‘soft QoS’. Intermediate nodes do not keep any per-flow information and thus avoid complex signalling and state control mechanisms making the system more simple and scalable [16].

3.6.1. Disadvantages of SWAN

It is unclear how the amount of bandwidth available for RT traffic should be chosen in a sensible way. Choosing a large value results in a poor performance of
RT flows and starvation of BE flows, and choosing a low value results in the denial of RT flows for which the available resource would have sufficed.

There would also be no flexibility to tolerate channel dynamics. The total rate of aggregated RT traffic may be dynamic due to node changes in traffic patterns and node mobility. Due to node mobility, for example, intermediate nodes may need to maintain RT traffic in excess of resources set-aside for RT traffic. An intermediate router making this observation sets the explicit congestion notification flag in RT packets’ headers.

SWAN fails to fully utilise the DiffServ field which is used only for two classes of traffic. It would be more useful if full advantage had been taken of differentiating the traffic into various classes that exist in practice.

Thus, though SWAN can be a candidate QoS model, it cannot be a complete QoS solution for a highly dynamic network like a MANET. It can be concluded that SWAN tries to maintain delay and bandwidth requirements of RT traffic by admission control of UDP traffic and rate control of TCP and UDP traffic.

3.7. Adaptive Reservation and Pre-Allocation Protocol (ASAP)

The Adaptive Reservation and Pre-Allocation Protocol (ASAP) [17] provides adaptive QoS support to real-time applications in infrastructure-based wireless IP networks. The purpose of this analysis is to extend the ASAP framework which can be used in mobile ad hoc networks.

In the ASAP architecture, a reservation concept, soft/hard reservation is introduced for efficient resource allocation. Soft reservation can be considered as the claim of a traffic flow for a certain bandwidth to be used in the future. Hard reservation enables a traffic flow to exclusively reserve some bandwidth.

The actual reservation mechanism is two pass based. When a new real-time flow is about to start, a soft reservation request is sent first. If there are sufficient resources available, the requested bandwidth will be soft reserved for that flow.

![Figure 11(a). ASAP reservation – soft reservation [18].](image-url)
After a soft reservation is established, the end node sends a hard reservation message requesting the same amount of bandwidth. This hard reservation removes all the traffic occupying the corresponding soft reserved bandwidth. So after a hard reservation, the QoS traffic can immediately start running with its necessary QoS support. Introducing these two kinds of reservations is to achieve good performance in QoS monitoring.

![Diagram showing the ASAP reservation process](image)

**Figure 11(b). ASAP reservation – hard reservation [18].**

Every node within the network stores information for each real-time flow having a reservation on that specific node. The per-flow information stored comprises a flowID uniquely identifying the flow and the actual soft and hard reservation for the flow. The set of all tuples stored within a node is called the QoS table. Table updates are triggered upon receiving signalling messages [18].

<table>
<thead>
<tr>
<th>Flow Label</th>
<th>SrcAddress</th>
<th>SoftResv</th>
<th>HardResv</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Host1</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>1</td>
<td>Host1</td>
<td>50</td>
<td>100</td>
</tr>
<tr>
<td>0</td>
<td>Host2</td>
<td>150</td>
<td>0</td>
</tr>
</tbody>
</table>

**Table 5: QoS Table for ASAP**

QoS monitoring packets periodically investigate the QoS situation on every node within a certain path. Hard reservation messages are sent whenever the end-to-end QoS changes. The monitoring interval can be changed dynamically. For example, more frequent monitoring is needed if the network is unstable, in order to adapt to bandwidth fluctuations. If the network is stable, processing overhead can be saved by keeping the monitoring rate low. ASAP also provides efficient in-band signalling for resource reservation, management, adaptation and releasing. The signalling is designed to produce minimum possible overhead and to provide maximum flexibility.
3.7.1. Problems of ASAP in MANETs

Studi [18] explained the problems associated with the ASAP quality of a service framework. These problems include the flow restoration problem, the reverse path problem, lost hard-reservation messages and incomplete differentiation.

The first problem in ASAP is the flow restoration problem. Assume a QoS path has been established from source to destination node and a maximum quality of service is provided along this path. If at a certain time one node moves out of the others’ transmission range breaking the path, routing then finds a new path for the flow. Because no reservation is established in the new path the flow is transmitted using best effort. This state is kept until the next SR message detects the missing reservation and triggers the source node to send a hard reservation message, which finally repairs the reservation on the new path.

The second problem is the reverse path problem. In ASAP, a hard reservation message is supposed to follow the reverse path that is previously established during soft reservation. This might be hard to achieve for several reasons. First, routes may change quickly in MANETs. A path established during soft reservation may be outdated while hard reservation is going on. Second, routes do not have to be symmetric. Although physically two nodes can reach each other in one hop distance, it does not mean routing also behaves like this. This could result in a big latency for hard reservation messages. The other problem related to reverse paths occurs when wireless links are not symmetric. Even if a node A can reach B in one hop distance, it is not sure that node B is able to reach A as well. As a consequence there may be no way for a hard reservation to pass through.

Another problem that is encountered when implementing ASAP refers to lost hard-reservation messages. If a hard reservation message during adaptation gets lost after sending, no subsequent soft reservation message will trigger any hard reservation if the path condition (bandwidth allocations on the nodes) stays the same because the adaptation process already did update its bandwidth allocation value. This state is kept until the end-to-end bandwidth for the flow changes somehow, i.e. until a soft reservation message arrives at destination having an actual bandwidth value that is different from the one stored by the adaptation process. If no node is moving and bandwidth is not fluctuating either, this may take a while. So a concept is needed to overcome this shortcoming. Hard reservation messages must be triggered until the reservation is actually done.

The differentiation problem of INSIGNIA and SWAN still appears in ASAP. ASAP differentiates traffic into Quality of Service and Best-effort traffic only. There is a need for better and more elaborate way of differentiating multimedia traffic according to some kind of priorities depending on the bandwidth needs and maximum allowable delay for each kind of traffic.
4. CONCLUSION

In this paper we defined quality of service, identified the factors affecting quality of service in mobile ad hoc networks. We described the quality of service metrics and factors that affect them in mobile ad hoc networks. We looked at quality of service models already in existence for mobile ad hoc networks. They include: IntServ, DiffServ, FQMM, INSIGNIA SWAN and ASAP. We identified that the IntServ and DiffServ are designed for wired networks so they do not fit for mobile ad hoc networks. While FQMM is a good model which combines the strengths of IntServ and DiffServ, it also carries most of their disadvantages with it which makes it not quite suitable a model for mobile ad hoc networks. INSIGNIA is a well designed signalling approach for MANETs but it exhibits some inherent problems. These drawbacks of INSIGNIA are its scalability problem due to the flow state information, which is kept within the nodes of a certain path and inefficient bandwidth usage. The bandwidth management of SWAN, though very good, is not very good for MANETs since it offers not a complete QoS solution for a highly dynamic network like a MANET. Although ASAP makes use of in-band signalling and fast adaptation but the protocol still fails to meet some MANET specific demands. Several problems of ASAP in a mobile ad hoc environment include flow restoration problem, reverse path problem and lost hard-reservation messages.

From all QoS models discussed in this paper, it is very clear that when designing a QoS framework, there is need for a classification of traffic into real time and non-real time classes. This is because real-time traffic has an immediate deadline that needs to be met otherwise the traffic becomes less useful. Delayed audio makes conversation very difficult and irritating. On the other hand delayed video may be viewed although with great difficulty. So the network should be able to treat these types of traffic with greater preference than general internet data. Therefore there is a need for a model that can classify traffic that takes into account different types of traffic that make up multimedia traffic and their varying bandwidth requirements. It does not make sense to give all real-time traffic the same priority since they come with different bandwidth, throughput and delay needs and therefore they must be treated differently according to their needs. The authors proposed a model [22] that is able to offer good bandwidth management based on an intelligent adaptation method that recognizes the priorities of the traffic. This proposal is made possible by an excellent bandwidth estimation method that makes the base of the management system. The main functionalities of this model in Mobile Ad Hoc Networks (MANETs) are resource estimation, admission control, resource reservation and bandwidth adaptation. These functionalities should be handled in a way that avoids the waste of resources and interference with other on-going communications.
REFERENCES


Information about the authors:

Benny M. Nyambo received BSc. Hons. Physics in 1999 and MSc Applied Physics in 2002 form University of Zimbabwe. He is currently studying towards a PhD in Computer Science at the University of Hasselt in Belgium. He has 8 years of lecturer in Computer Science. His research interests include performance modeling of communication networks, embedded systems.

Gerrit K. Janssens received degrees of M.Sc. in Engineering with Economy from the University of Antwerp, Belgium, M.Sc. in Computer Science from the University of Ghent, Belgium, and Ph.D. from the Free University of Brussels, Belgium. After some years of...
work at General Motors Continental, Antwerp, he joined the University of Antwerp until the year 2000. Currently he is Professor of Operations Management and Logistics at Hasselt University. His main research interests include the development and application of operations research models in production and distribution logistics.

Wim Lamotte obtained his PhD in Computer Science in 1998 at Hasselt University, Belgium, where he is now a full-time professor teaching courses in networking and multimedia. His current research interests include networked virtual environments, computer networks and networked multimedia. He leads the Multimedia and Communication Technology research group of the Expertise Centre for Digital Media (EDM), specializing in Networked Virtual Environments. He is member of IEEE and ACM.

Manuscript received on 07 May 2014