

Current Approaches towards Improved Quality-of-Service Provision in Mobile Ad-hoc Networks

March 2003

Michael Gerharz, Christian Vogt, Christian de Waal
Computer Science Department IV • Communications Systems
Rheinische Friedrich Wilhelm University at Bonn, Germany
{gerharz,vogt,dewaal}@cs.uni-bonn.de

1 Introduction

With mobile computers and wireless networking hardware becoming widely and increasingly cheaply available, a large body of effort has over the past years gone into incorporating mobility support into the existing, prevalently wired Internet. Today's wireless networks, however, are either restricted to some sort of immobile infrastructure (e.g., UMTS), or to direct, single-hop connections (e.g., IEEE 802.11 DCF), or to both (e.g., DECT). In an infrastructure-oriented network, all communication is switched over a fixed, wired backbone. Stations may freely roam within the network's extent, yet the network itself cannot be relocated. Building the backbone calls for substantial investments in terms of both money and time. A single-hop wireless network has no routing capability, and two stations cannot exchange data unless they are within each other's radio range. This type of network seldom grows beyond the scale of a local area network.

Mobile Ad-hoc Networks (Manets) are multi-hop wireless networks without an explicit backbone, lifting the restriction and the expenditures of an unmovable infrastructure. In Manets, all stations combine the functionality of clients and routers. The network topology is continuously tracked, and routing paths are constantly updated, as stations freely roam about. Manets can be set up quickly and without effort. They are predestined for spontaneous employment at conventions, military engagements, or during disaster-recovery operations. Research on Manets is still emerging, and while the basic routing problem has received much attention, many challenges are yet to be approached. This includes Quality-of-Service issues: Due to the fundamental difference between wireless multi-hop networks and wired networks (perhaps with a wireless "last hop"), existing approaches for wired networks cannot be conferred without adaptation.

2 Overview

Communication sessions differ in complexity and demands. Classic Internet applications like Web browsing, email, or file transfer are adaptive to fluctuations in bandwidth availability and datagram-delivery delay. These applications manage to satisfy user expectations even when multiple communication sessions compete for the shared network resources. In contrast, video conferencing or VoIP exemplify a much more demanding type of application. These *real-time applications* fail to deliver appropriate quality unless bandwidth availability and datagram-delivery speed are satisfactory without disruption.

In the internet world, the IntServ [braden1994] and DiffServ [blake1998] concepts have been developed to employ QoS in wired networks. As networks are considered static in this context, these concepts basically specify in which manner resources are reserved.

An obvious difference in Manets is that routes between source and destination fulfilling certain QoS characteristics have to be found in the first place. The first category of publications deals with this

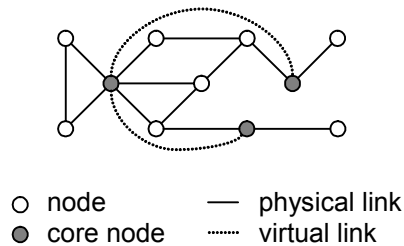


Figure 1: Core in CEDAR

problem, e.g. CEDAR [sivakumar1999] (cf. section 3), MMWN [ramanathan1998] (cf. section 4), and ticket based probing [chen1999] (cf. section 5).

The other category of publications presents mechanisms that enable QoS support independent of the routing protocol, loosely related to IntServ and DiffServ: INSIGNIA [lee2001a] (cf. section 6), SWAN [ahn2002] (cf. section 7), FQMM [xiao2000] (cf. section 8).

These protocols are described in detail in the following sections.

A fundamentally different aspect of Quality-of-Service support in mobile wireless ad hoc networks is that the characteristics of the routing should meet the demands of realtime applications to the largest extent possible.

One approach for improving the quality of communication sessions is to enhance the quality of the route selection process. Associativity Based Routing (ABR) [toh1997] and Signal Stability Adaptive Routing (SSA) [dube1997] both try to establish stable routes, i.e. routes that have a high probability of being available for a long period of time. [gerharz2002] provides a detailed analysis of the implications of different mobility models on the stability of links.

A somewhat complementary approach to provide more robust connections is taken by Preemptive Routing [goff2001] and Routelifetime Assessment Based Routing (RABR) [agarwal2000] which try to detect a link break before it actually happens in order to issue a new route discovery before the old route breaks.

Another approach to achieve robust connections is multipath routing, where several alternative paths are established at once. Upon a link break, this approach has the advantage to have a fallback route at hand. The Ad Hoc On-demand Multipath Distance Vector protocol (AOMDV) [marina2001] and the Split Multipath Routing protocol (SMR) [lee2001b] represent this approach.

Since the relation of these approaches to QoS in the classical sense is rather vague and a detailed description would require a high amount of space, they are not discussed any further.

Another research area within the context of QoS in Manets is the development of wireless MAC layers with QoS features and multi-hop capability. Again, going into details would exceed the focus of this report. However, it can be noted that providing QoS in a wireless multi-hop environment is a very complex task, and many proposals make assumptions that limit the usage scenarios, e.g. a global synchronisation of the devices. Some implications of this difficulty on level 2 are discussed at the end of this report.

3 Core-Extraction Distributed Ad-hoc Routing

A common feature of most ad hoc routing protocols is that they distribute routing functionality across all network participants. These protocols spend substantial bandwidth on control messages in order to coordinate the large number of routers. In proactive routing protocols, stations constantly notify each other about link-state changes. Amongst the reactive approaches, flooding is an integral but expensive part of destination discoveries.

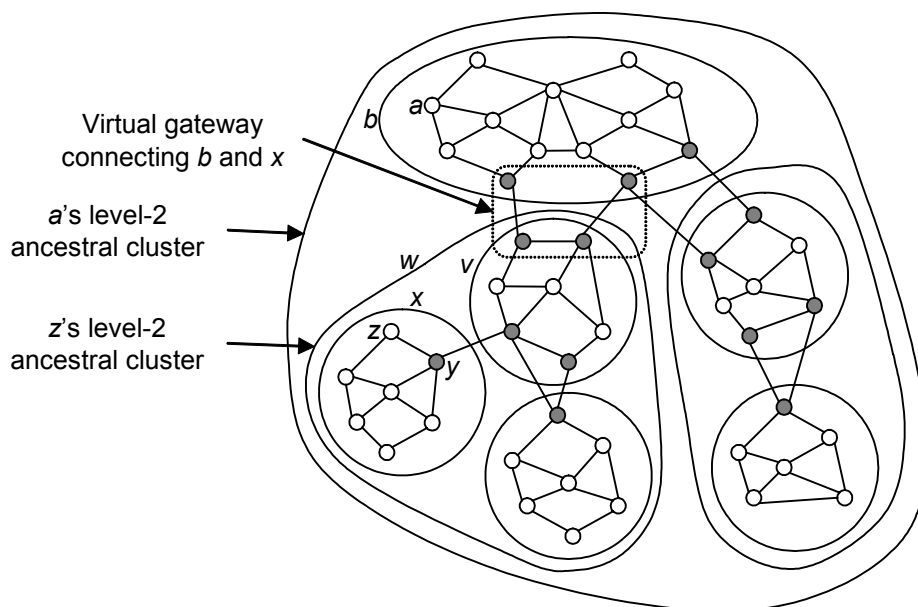


Figure 2: MMWN routing architecture

Sivakumar et al. propose the *Core-Extraction Distributed Ad-hoc Routing (CEDAR)* protocol, which attempts to reduce the coordination overhead [sivakumar1999]. CEDAR selects a subset of stations that together form a dominating set of the entire network. The dominating set is called the *network core*, and the stations within are referred to as *core routers* (cf. figure 1). All routing functionality is limited to the core routers. The core routers exchange notification messages about stable, high-bandwidth links, based on which they can compute routing paths that satisfy the applications' QoS demands. While most routing protocols use unacknowledged indirect transmissions to distribute control messages across the network, CEDAR transmits directly and, therefore, reliably. CEDAR uses a modified MAC-level frame format to identify redundant message transmissions as early as during the sender-receiver handshake that precedes all direct transmissions. Aborting a needless transmission during the handshake makes it unnecessary to transmit the message payload.

Each station monitors the available bandwidth on the links to its neighbors. When a new link comes up, an old link goes down, or the bandwidth along an existing link changes significantly, the adjacent stations take care that this news is distributed across the network core. It depends on the news' importance how far it is being relayed. Positive LSUs – i.e., those about new links or bandwidth increases – are artificially delayed at each core router, whereas negative LSUs – i.e., those about link failures and bandwidth reductions – propagate without delay. This strategy allows intercepting a positive LSU describing an unstable link when that LSU is being caught up by a negative LSU describing the same link.

Despite the promising new concepts of CEDAR, a number of issues ought to be considered. First, although direct transmissions within the network core are certainly more reliable than indirect ones, it remains to be seen whether CEDAR really reduces routing overhead. After all, direct transmissions go with an initial handshake and a concluding acknowledgement. Indirect transmissions avoid this overhead, which may exceed the bandwidth consumption of a few redundant transmissions. Another issue with CEDAR is the required modification of MAC-layer control messages. Wireless-networking equipment has over the past years become a commodity product, and functional modifications are expected to be hard to realize. Finally, the propagation delay of positive LSUs is a parameter that needs to be wisely determined. Its optimal value depends on other parameters like network size and traffic patterns. Choosing it offhand may have profound consequences.

4 MMWN: Multimedia support for Mobile Wireless Networks

Whereas CEDAR establishes a virtual backbone to reduce the information overhead, MMWN [ramanathan1998] clusters the network and abstracts over cluster characteristics.

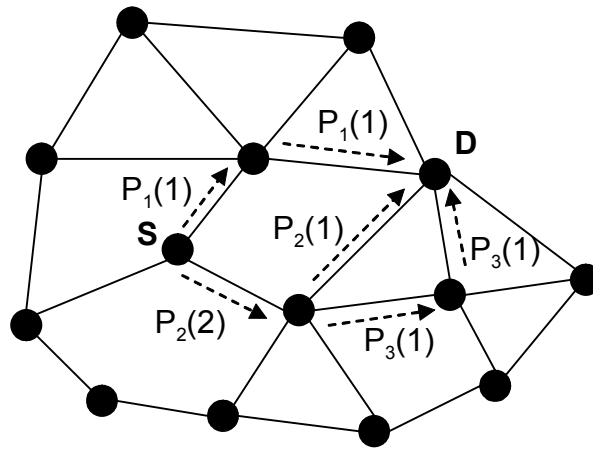


Figure 3: Ticket based path discovery

This clustering is done in a recursive manner. Each station is a level-0 cluster, and on highest level, the whole network is a single cluster. A level- i cluster containing a level- $(i-1)$ cluster is called *parent* cluster of this level- $(i-1)$ cluster. Two clusters are called *siblings*, if they share the same parent cluster. If there are one or more direct links between two siblings, the *virtual gateway* between these clusters is defined as the links between these two clusters together with the border nodes within the clusters adjacent to these links. These relations are illustrated in figure 2.

Each cluster contains a designated QoS manager that creates abstracted link state information for the cluster and distributes it to the cluster members. Therefore, ultimately, every station receives link state information for all clusters it participates in. This enables a station to reach another station based solely upon this knowledge, if it knows the destination's cluster memberships; as packets traverse other lower level clusters than the source station participates in, the stations in these clusters have the detailed knowledge necessary to route the packet to the destination. Furthermore, the link state information allows the stations compute routes fulfilling certain QoS constraints, despite their lack of detailed knowledge on the network.

As an example, in figure 2, station z were to address a packet to a , then it would compute the route

$$w.x.z \rightarrow w.x.y \rightarrow w.x \rightarrow w.v \rightarrow w \rightarrow b \rightarrow b.a$$

While clustering is a promising approach to abstract and reduce information, the MMWN architecture seems suitable only for rather stationary networks, because it itself would otherwise introduce a great complexity. A clustering has to be found and maintained, and with every topology change, the previous cluster structure might become outdated. If the structure remains the same after a topology change, it is already required that the new QoS characteristics are calculated and spread throughout the network. If the cluster structure becomes unusable, additional effort is needed to create a new structure which may include electing new QoS managers. Furthermore, also some stations' addresses will change, resulting in a necessity for a mobility management mechanism.

5 Ticket based probing

The ticket based probing presented in [chen1999] tries to provide QoS-constrained paths (e.g. maximum delay or minimum bandwidth paths) with a two staged approach. It uses proactive routing to provide stations with rough knowledge about the state of the network. Upon a connection request, a QoS-constrained path is discovered by a "directed" reactive route discovery based on this imprecise state information.

Three instances of a proactive distance vector protocol provide stations with hints on the shortest path length, the maximum available bandwidth on a single path, and the delay to all other stations. In order to keep the overhead within low bounds, these are operated with a low frequency of information exchange between the stations.

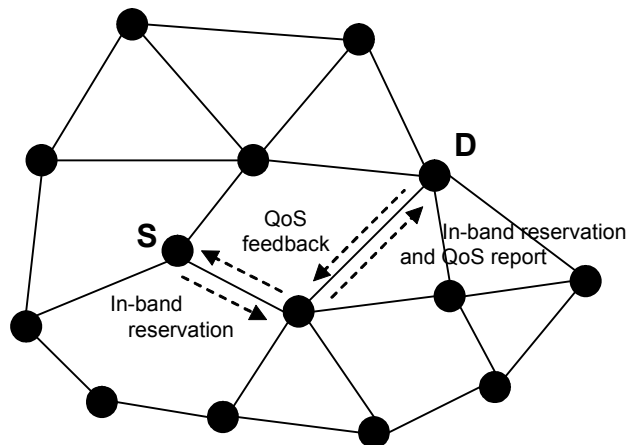


Figure 4: INSIGNIA soft-state reservation

The actual path discovery process utilises the state information to initiate route request messages, so-called *probes*, along promising paths. Probes may be split at each station and propagated across different links. Each probe carries a certain number of *tickets* which defines how often the probe may be split. Probes that are sent along more promising paths are assigned more tickets. This way, route requests are likely to reach the destination over paths with the desired characteristics, if such a path exists, without having to flood the entire network.

Figure 3 illustrates an example of a route discovery with ticket based probing. Assume a source node S tries to establish a QoS-constrained connection to a destination D . Based on its initial knowledge of the network it issues two probes P_1 , carrying 1 ticket, and P_2 , carrying 2 tickets. P_1 may not be further split up and is propagated directly on the most promising path towards D . P_2 seemed more promising to S and thus has been equipped with 2 tickets. Consequently, it is split into two probes P_2 and P_3 by an intermediate node, each carrying just 1 ticket. Upon reception of all three probes, D evaluates the probes and issues a response to S along the reverse route.

In order to protect certain flows against unpredictable topology changes, the protocol allows for multiple levels of redundancy for *fault-tolerance*. Depending on the importance of a flow, reservations may be made on multiple paths at once and packets may be even copied and transmitted along several of these paths simultaneously.

The protocol produces much overhead. Although the authors suggest to run the proactive part of the protocol at a very low frequency, still the overhead of three instances of a proactive routing protocol has to be taken into account. Furthermore, the more imprecise the knowledge of the network state gets, the more tickets are needed to find suitable routes. However, probes are sent as unicast frames, producing higher overhead than broadcast frames.

6 INSIGNIA: In-band Signalling for QoS in Ad-Hoc mobile networks

In contrast to the approaches presented in the previous sections, INSIGNIA [lee2001a] is a simple signalling mechanism which can be combined with a variety of routing protocols.

The INSIGNIA protocol provides per-flow QoS by piggybacking soft-state reservations onto data packets in an IP option field. If soft-state reservations are not renewed with a certain time interval, they are withdrawn. Each INSIGNIA IP option is self-contained in the sense that it carries all the information necessary to setup or maintain a reservation. If an intermediate node on the source-destination path is not able to provide the requested bandwidth it nevertheless forwards the packet, but modifies the QoS header with an indication to its lacking resources. The destination may then provide feedback on the QoS provision of the route to the source node which in turn may decide to abandon the session or to adapt the flow to the available resources. These procedures are illustrated in figure 4.

The idea behind this approach is to avoid explicit signalling and hard state reservations in order to deal with the dynamics of mobile ad hoc networks in a flexible way.

7 SWAN: Stateless Wireless Ad-hoc Networks

SWAN [ahn2002] classifies data packets into high-priority real-time traffic and normal, low-priority traffic. A *classifier* decides which type an incoming datagram belongs to. If the datagram carries real-time data, it enjoys precedence and is processed as soon as possible. Otherwise, the datagram is considered low-priority and must normally wait until all real-time datagrams are dispatched. A *shaper* limits the relay of low-priority traffic to reduce the contention between neighbouring stations.

To keep the delay of realtime traffic in sensible bounds on one hand, and to avoid starvation of bulk traffic on the other hand, the bandwidth available for realtime traffic is bounded. To determine the amount of bandwidth yet available for realtime traffic, stations monitor packets sent by neighbouring stations to find out how much realtime traffic is currently present on the shared medium.

SWAN cooperates with almost any routing protocol. When a source station wishes to establish a realtime session to another station, it probes the path to the destination to identify the bandwidth available for realtime traffic. If the available resources are sufficient, the source station launches the new communication session. Otherwise, it refuses the new communication session, or adapts it to the resources available. Note that the originating station is fully in charge of admitting or denying its own data flows. Intermediate stations along the routing path solely attach information about their local workload onto the originating station's probe. They do not contribute to evaluating this data.

Network-topology changes may cause quality degradations to ongoing communication sessions, because the upper bound on realtime traffic is exceeded due to re-routing of previously admitted flows. An intermediate router making this observation sets the explicit congestion notification (ECN) flag in realtime packets' headers. When a datagram with ECN arrives at its destination, the destination notifies the data flow's originator about the congestion, who is then responsible for dropping or adapting the communication session.

What remains unclear is how the amount of bandwidth available for realtime traffic should be chosen in a sensible way: Choosing this value too high results in a poor performance of realtime flows and starvation of bulk flows, and choosing it too low results in the denial of realtime flows for which the available resource would have sufficed.

8 FQMM: Flexible QoS Model for Manets

FQMM [xiao2000] is a different approach to employ DiffServ in ad hoc networks. In contrast to SWAN, it allows for differentiation of more than just two service classes; actually, a service class may consist of a single flow.

With FQMM, every station plays the role of an ingress node for the flows that it originates: It classifies and meters its own packets, and marks them accordingly. The source and intermediate stations perform traffic shaping according to those marks. The FQMM authors suggest the use of a token bucket metering algorithm to mark packets as in-profile and out-of-profile. In case of network congestion, out-of-profile packets are discarded with a higher probability than in-profile packets. In [xiao2001], the same authors suggest and compare a priority buffer and a priority scheduling scheme for this purpose.

A downside of this approach is that the source stations have to take great care in regulating their traffic, since the rate of in-profile traffic must be processable in all network regions, including bottleneck areas where traffic from different sources accumulates. However, FQMM (which is described in close relation to DiffServ by its authors) actually lacks the counterpart to DiffServ's

service level agreements, and it remains an open question how the source stations should determine the dynamic parameter for their token bucket metering.

9 Discussion

In the previous sections, the most representative work concerning QoS in Manets has been introduced. Sections 3 through 5 have presented methods to find QoS constrained routes. Each of these methods has an own strategy to reduce routing overhead, yet it is not clear whether any of these strategies achieves this goal. Sections 6 through 8 have presented methods that are more or less decoupled from the routing strategy. Rather, they specify how flows or traffic classes should be handled to achieve a certain degree of QoS.

A common assumption in many of the presented ideas is that devices are able to reserve bandwidth on certain links. However, it is not possible for a station to make bandwidth reservations or even to gain knowledge about the available bandwidth with wireless hardware that is currently available or will be available in the near future. To be concise, this problem persists as long as a single network interface is used to communicate on several links.

This is an important general issue of QoS in Manets that is not adequately considered in the literature yet. Although the technical aspects of wireless networks are quite different from wired networks, researchers have adopted many assumptions from wired networks. This is especially the case with the routing approaches presented in sections 3 through 5. In contrast, the SWAN mechanism presented in section 7 tries to take the special characteristics of wireless links into account, but even there, the bandwidth which is reserved for realtime flows has to be predefined, which is not a simple task, as already noted.

Our conclusion is that while some effort has been spent to solve the routing challenge in Manets, QoS approaches developed so far are rather theoretical and lack practicability for Manets to be deployed in a not too distant future. Practical solutions might be even more basic than e.g. SWAN and should be as layer-2 independent as possible.

References

- agarwal2000 Sulabh Agarwal, Ashish Ahija, Jatinder Pal Singh and Rajeev Shorey: "Route-lifetime Assessment Based Routing (RABR) Protocol for Mobile Ad-hoc Networks." In *Proceedings of the IEEE International Conference on Communications 2000 (ICC'00)*, pp. 1697-1701.
- blake1998 S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss: "RFC 2475: An Architecture for Differentiated Services." December 1998.
- braden1994 R. Braden, D. Clark, and S. Shenker: "RFC 1633: Integrated Services in the Internet Architecture." June 1994.
- ahn2002 Gahng-Seop Ahn, Andrew T. Campbell, Andras Veres, and Li-Hsiang Sun: "Supporting Service Differentiation for Real-Time and Best-Effort Traffic in Stateless Wireless Ad Hoc Networks (SWAN)." In *IEEE Transactions on Mobile Computing*, vol. 1, no. 3, July-September 2002, pp. 192-207.
- chen1999 Shigang Chen and Klara Nahrstedt, "Distributed Quality-of-Service in Ad Hoc Networks." In *IEEE Journal on Selected Areas in Communication*, vol. 17, no. 8, August 1999.
- dube1997 Rohit Dube, Cynthia D. Rais, Kuang-Yeh Wang and Satish K. Tripathi: "Signal Stability based Adaptive Routing (SSA) for Ad-Hoc Mobile Networks." In *IEEE Personal Communications*, February 1997.
- gerharz2002 Michael Gerharz, Christian de Waal, Matthias Frank and Peter Martini: "Link Stability in Mobile Wireless Ad Hoc Networks." In *Proceedings of the 27th Annual IEEE Conference on Local Computer Networks (LCN'02)*, November 2002, pp. 30-39.
- goff2001 Tom Goff, Nael B. Abu-Ghazaleh, Dhananjay S. Phatak and Ridvan Kahvecioglu: "Preemptive Routing in Ad Hoc Networks." In *ACM Seventh Annual International Conference on Mobile Computing and Networking (MOBICOM'01)*, July 2001, pp. 43-52.
- lee2001a Seoung-Bum Lee, Gahng-Seop Ahn and Andrew T. Campbell: "Improving UDP and TCP Performance in Mobile Ad Hoc Networks with INSIGNIA." In *IEEE Communications Magazine*, June 2001, pp. 156-165.
- lee2001b Sung-Ju Lee and Mario Gerla: "Split Multipath Routing with Maximally Disjoint Paths in Ad hoc Networks." In *Proceedings of the IEEE International Conference on Communications, Helsinki, Finland, June 2001*, pp. 3201-3205.
- marina2001 Mahesh K. Marina and Samir R. Das: "On-demand Multipath Distance Vector Routing in Ad Hoc Networks." In *Proceedings of the IEEE International Conference on Network Protocols, Mission Inn, Riverside, California, USA, November 2001*, pp. 14-23.
- pearlman2000 Marc R. Pearlman, Zygmunt J. Haas, Peter Sholander, and Siamak S. Tabrizi: "On the Impact of Alternate Path Routing for Load Balancing in Mobile Ad Hoc Networks." In *Proceedings of the IEEE/ACM International Symposium on Mobile Ad Hoc Networking and Computing, Boston, Massachusetts, USA, August 2000*, pp. 3-10.
- perkins2002 Charles E. Perkins, Elizabeth M. Belding-Royer, and Samir R. Das: "Ad hoc On-Demand Distance Vector (AODV) Routing." Internet Draft version 12 <draft-ietf-manet-aodv-12.txt>, IETF Manet Working Group, November 2002.
- perkins2000 Samir R. Das, Charles E. Perkins, Elizabeth M. Royer: "Performance Comparison of Two On-demand Routing Protocols for Ad Hoc Networks." In *Proceedings of the IEEE Conference on Computer Communications, Tel Aviv, Israel, March 2000*, pp. 3-12.
- ramanathan1998 Ram Ramanathan and Martha Streenstrup: "Hierarchically-organized, multihop wireless networks for quality of service support." In *ACM Mobile Networks and Applications, volume 3, no. 1, June 1998*, pp. 101-119

- sivakumar1999 Raghupathy Sivakumar, Prasun Sinha, and Vaduvur Bharghavan: "CEDAR: A Core-Extraction Distributed Ad hoc Routing Algorithm." In *IEEE Journal on Selected Areas in Communications, Special Issue on Ad Hoc Networks*, volume 17, no. 8, August 1999, pp.1454-1465.
- toh1997 Chai-Keong Toh: "Associativity Based Routing For Ad Hoc Mobile Networks." In *Wireless Personal Communications Journal, Special Issue on Mobile Networking and Computing Systems*, vol. 4, no. 2, March 1997, pp. 103-139.
- xiao2000 Hannan Xiao, Winston K.G. Seah, Anthony Lo, and Kee Chaing Chua: "A Flexible Quality of Service Model for Mobile Ad-Hoc Networks." In *Proceedings of the IEEE Vehicular Technology Conference, Tokyo, Japan, May 2000*, pp. 445-449.
- xiao2001 Hannan Xiao, Winston K. G. Seah, Anthony Lo, and Kee Chaing Chua: "On Service Prioritization in Mobile Ad-hoc Networks." In *Proceedings of the IEEE International Conference on Communications 2001 (ICC'01)*, June 2001, pp. 1900-1904.