Quality of Service (QoS) in Wireless Networks

Scott Dill
Department of Computer Science and Engineering
University of Texas at Arlington

Abstract – Mechanism for achieving quality of service in the Internet have been researched and are being implemented to support the new real-time and critical applications that are being used more and more in that environment. As users migrate to the wireless environment, they expect to be able to use the same applications at some level of assured quality. This wireless environment imposes new challenges on the implementation of QoS, but many techniques are available.

1. INTRODUCTION

When computing devices go mobile, user expectations may change from those they have for wired access, but that does not mean that they must accept poor service. This paper will first provide a very brief overview of what exactly is meant by the terms "quality of service", "wireless networks" and "mobile computing". Next the relationship between quality of service and mobile computing will be examined so as to illustrate the additional complexities encountered when combining the two. Finally a few examples of recent work in this field will be discussed.

It needs to be noted that this topic is incredibly broad and in order to maintain the proper scope of the paper, only limited aspects of the topic can be covered. As such, while other networks will be touched upon, the primary area of focus will be on wireless networks that allow access to the Internet and implications of this interaction. Also there is a tremendous amount of research going on in this area, so there is no attempt to be inclusive in any way.

Quality of Service

The basic design of the Internet Protocol is one that provides best-effort service. That is to say that provided that the applications and other higher layers operating on the Internet perform as specified, the underlying network layer will route all packets as fairly and quickly as its resources allow. This arrangement is fine for the original applications intended for the Internet (email, file transfer, remote terminal operation) and even for the "killer-app" which brought the Internet into everyday parlance – the World Wide Web. As demand for video, voice and real-time applications grew; this type of service has proven inadequate.

The classic use of QoS primarily refers data transfer characteristics. These characteristics can be grouped into three broad groups: timeliness (delay, jitter, response time), bandwidth (transfer rate, transaction rate) and reliability (packet loss, availability)[1, 18]. Specification of these characteristics allows applications to receive the required data at consistently fast enough speeds to meet their needs. This has been achieved in the wired Internet through Differentiated Services (DiffServ, DS), Integrated Services (IntServ, IS), (Reservation Protocol) RSVP and Multi-Packet Label Switching (MPLS)[4][18].

QoS can and should be defined more broadly, however to include all aspects of service that affect the user experience. These characteristics include, but are not necessarily limited to the following: criticality, perceived quality (picture detail, color accuracy, smoothness, synchronization, audio quality, etc), cost, and security level [1]. All of these aspects are related to the data transfer characteristics in that in order to achieve specific goals, certain levels of data transfer specifications must be maintained [1].

It can therefore, be seen that active management of QoS is required if it is to be maintained. QoS maintenance can be divided into two types: static maintenance – which focuses on those aspects, which do not change during the course of a session; and dynamic maintenance – which manages changes within the environment during a particular session [1].

Wireless Networks

In the literature, the terms "wireless" and "mobile" are often used interchangeably. Though this is technically incorrect, it generally suits the purposes of high to medium level discussion and the practice will be used in this paper. Wired mobile computing is more accurately referred to as "nomadic" [19, 11]. Fixed computers can be connected to the network via wireless means. The primary focus of this paper will be on mobile computers connected by wireless means, as this is the most interesting and challenging type.

2. QoS IN WIRELESS MOBILE COMPUTING

The wireless environment puts far greater strain on the ability to maintain specified QoS than is experienced in the wired world due to both static and dynamic constraints. The static constraints refer to the limitations of the available bandwidth and portability requirements. The dynamic limitations involve the changing nature of the wireless, portable environment. [1]

Constraints

Through Mobile IP, a solution for global mobility has been achieved; unfortunately, Mobile IP has no provision for quality of service requirements. In fact, the triangular routing and tunneling actually exasperates the problem. The triangular routing, or more specifically the indirect routing from the correspondent host through the home agent (HA) to the foreign agent (FA), actually increases network traffic by consuming a greater quantity of total network resources. Encapsulation associated with tunneling can also hide the flow identifiers that specify QoS, or flow identifiers in the IP header. [6]

Though the technology is improving, it is reasonable to assume that wireless technology will continue to have at least an order of a magnitude lower available bandwidth than wired connections. This makes guaranteeing and reserving bandwidth more expensive and less practical, particular for the high bandwidth required for streaming multi-media applications. Further complicating the situation is the fact that

just because the device has been able to garner the required bandwidth in the current cell of the wireless network, it does not mean that it will be able to keep it if conditions change. The available bandwidth can be reduced by environmental conditions, localized blind spots or simply by other users entering the cell with equal or higher priority. [1,7]

Movement creates another set of problems for maintaining QoS. One of the major obstacles comes from the handoff as the mobile device moves from one cell to another. This handoff can result in the loss of a certain number of packets as they are routed the mobile device's former location instead of its current. This problem can be avoided with protocols like Cellular IP [16], or other soft handoff techniques which roué packets to two access points during the transition phase. Also, work involving sophisticated prediction techniques [3] has proposed models to predict which cell a mobile user will be in and when so that its traffic can be routed there. This requires that base stations have both powerful prediction models and space to buffer the data for the host at all of the possible cells they may be appearing in soon as well as the current cell. It can also restrict access to the network due to the advanced reservations for resources made by the prediction model. These techniques have been demonstrated to be successful with low bit-rate data streams, like text and voice (8kbs), but with large data streams it is much more expensive and difficult. Another issue associated with handoff, is that even though the current cell is proving the required resources, there is no guarantee that the cell being migrated into will be able to. The new cell could be of a different standard (campus LAN to a public cellular system as you leave work) or could already be taxed by other users so that the new request cannot be serviced. In this case, either the connection has to be refused, or the QoS must be renegotiated. Some of these adaptive techniques will be discussed later in the paper. [1,2]

Portability has its own set of restrictions that put unique demands on QoS specifications and management. To take advantage of true mobile computing a device needs to be of appropriate size and weight to be comfortably used on the move. Currently this requires that they have simpler displays that have reduced resolution (pixel count and color) and simplified – less effective – input devices. This means that if a user uses certain services from different devices with varying capabilities (desktop vs. PDA) then varying levels of QoS must be provided for these services. Context aware computing aids in this and is discussed later in the paper. An even more important feature of portability is its power source. Today and into the foreseeable future portable energy storage devices (batteries) will occupy a large portion of any mobile device and restrict their capabilities. Since wireless communication, particularly transmission, requires a great deal of power, the techniques of providing and managing QoS must account for this by scaling to accept various power consumption curves. [1,7]

Cellular Telephony Issues

In the general case, much of the public access that we will use to achieve true mobile computing will be provided by what is now the cellular telephone network, or more accurately the growing group of cellular service providers. This is analogous to the case in the wired world were much of the last mile connection is provide trough home phone, ISDN or T1 lines owned by the various communications companies. A significant difference is that it will be built on an existing digital network that has its own

concept of QoS that does not necessarily map well with the IP version of QoS that will be needed for the type of applications and devices of interest in this paper.

Current cellular phone networks rely on call admission as one of their most important methods to achieving QoS [8]. Call admission control (CAC) generally gives priority to handoff calls in preference to new calls since it is considered worse to drop an ongoing call than to not be able to start a new one [8]. The static nature of this method does not scale well to handle the variable data rate connections that need to be made in a mobile computing environment [8]. Research is being done to merge the requirements of IP based QoS with traditional cellular QoS that allow effective coexistence [8,15].

Resource Reservation

As described in the section on the problems associated with movement in a wireless network, the handoff between base stations causes the greatest disruption in quality of service and in order to minimize this impact, the network must pre-allocate resources to the cell where the mobile host (MH) is going to be next. This is very difficult to accomplish, but there are several methods that have been proposed. [2]

One approach is to make no prediction as to where the mobile user is going to go next and simply try and reserve a certain amount of resources in all of the neighboring cells. If these reservations were "hard " reservations – actually lock up resources – there would be a great deal of waste. Instead they could be considered to be "passive" reservations in that the resources will not actually be tied up until the MH making the reservations actually begins to use them. This "Advanced Reservation Scheme" allows each base station to potentially reserve a certain amount of their bandwidth for incoming MHs while retaining another allotment for continuing local traffic. This scheme allows for efficient handoffs, but makes no attempt to predict the continuing path of the MH, so new negotiations for reservations must be made with adjacent cells each time a handoff is completed [2]. This disallows a continuous guarantee of QoS for many scenarios. [13]

Another, more complicated approach, is to actually attempt to predict where the MH is headed next. One method used to do this to simply take the MHs current position and velocity and use this to calculate its future position. This information can be gathered through cellular triangulation or through an onboard global positioning system (GPS) device [3]. Prediction can also be based on the past history of the MH. Depending on the level of service commitment intended to be provided, the future path can be predicted and advance reservations made through many potential steps in a potential path. It is also possible to create a probability based model in which reservations are only made at the cell most likely be next visited. It is important to note that the further along the path the prediction model allows reservations to be made, the higher the probability that the QoS specification can be maintained. [2]

Adaptation [1,9,7]

The environment generally envisioned for mobile computing is one in which there is a great variety of devices that vary in capabilities to process, display and transmit information. The nature of the wireless environment in one in which constant change is the only thing that really be counted on. The inconsistency is caused by both

environmental characteristics and the general scarcity of resources available when on the move. The ideal way to deal with this situation is to have all constituents of the network to adapt their schemes and approaches to suit the current state. There have been many proposals and efforts that require various constituents and layers to be aware of the needs and situations that require proactive changes to the computing environment. For example, a user happily watching a streaming video on a hand-held device may encounter a situation in which new users start communication sessions in the same cell, dramatically reducing the available bandwidth. Since new resources cannot be created and the original user cannot command exclusive use of the existing ones, some form of adaptation must occur if he is to continue using the application. Many approaches could be taken to adapt to this situation. Some examples are identified below:

- Buffer size can be large enough to continue the current playback rate while changing conditions are absorbed – assuming condition is transient and not persistent.
- The application can detect that it is receiving less data and signal the source to change to a lower resolution, lower frame rate or less colorful stream. If the situation is severe enough, the stream could be converted from video to audio only, or even a textual transcript.
- The server can offer the client, the opportunity to maintain its resource reservation through payment of higher rates – premium service.

The need for adaptation is not limited to bandwidth requirements and availability, but also needs to be considered for power and contextual situations.

Context awareness is also a key concept in adaptability because it can be an effective method of managing the static (or "large-grained" as they are called in [1]) constraints that are inherent to the wireless media – low bandwidth, scarce power availability, limited I/O resources. Context adaptability could be manifested in many ways: [1,14]

- Migration of data to nearby servers
- Utilization of nearby resources to conserve total network usage
- Specification of low fidelity media files due to remote location
- Pre-downloading of data that will be needed in destined location
- Stopping execution of distracting media when in inappropriate locations
- Selection of appropriate network interfaces based on available bandwidth, cost and security.

In the next sections, two different approaches to adaptation will be discussed.

3. QoS aware Applications [7]

In [7] an approach is studied in which only the applications take part in the adaptation process. What is particularly interesting about this approach is that is something that can be achieved by the developers of the end unit, or even just the operating system for it, and does not involve a system-wide implementation or standards

development. The basic idea is that when many applications share a scarce resource – bandwidth and power – there needs to be a way to gracefully scale back their demands for them as the supply decreases.

In order for this to be achieved, there are certain things that must be assumed. To be able to participate in this adaptation, an application must have the capability of operating in several discrete states that consume a range of resource relative to their utility. An example of this would be a media player that has several available resolutions, color levels and frame rates that it can choose based on its available share of power and or bandwidth. In this example, as well as most others in this paper, media applications are used as an illustration, but it cannot be inferred that these are the only applications that can impact adaptability. CPU intensive applications, like encryption, can be self-adjusted so that their level of detail - resolution - is changed based on available resources.

The fundamental problem is then for the device to decide how to downgrade (upgrade) application attributes to accommodate the changing resources while maintaining the maximum aggregate utility. [7] demonstrates that it is not necessary to perform a complete state space search in order to closely approximate this maximum utility value, but that what he calls a "stateful ratio" produces nearly optimal results. Mr. Geihs effectively describes the method used to calculate the stateful ratio as:

> *Compute the ratio of utility over resource demand for each level. If resources increase, select application level with biggest utility ratio, but only if the new level at time t is higher than the previous one at time t-1. Repeat until resource supply is met. Analogous for a decrease of resources. [7]*

The above technique was demonstrated to be effective for both static and dynamic utility / resource functions.

4. DYNAMIC QoS MANAGEMENT THROUGH AGENTS [9]

A proposal using agents to dynamically manage call admission, handoff and prediction in a mobile, wireless environment is discussed and reviewed. The summary of the protocol comes from [9]. I will first summarize the protocol used for MH migration prediction, then a protocol for dynamic bandwidth adjustment (adaptation) and will close up with a description of the final proposal to merge them into a single cohesive model. I will close with my personal comments on the proposal.

Proactive Bandwidth Reservation [9]

This portion of the proposal is a position-assisted reservation scheme like those described in section 2. The author calls it Predictive Mobility-Based Bandwidth Reservation (PMBBR) and it uses agents to predict where the MH will probabilistically go next and reserves resources in neighboring cells based on the calculated probability. It assumes that traffic in adjacent cells is not independent due to handoffs, because when an MA enters a cell, it not only consumes local resources, but places demands on

neighboring cells as well based on its probability of movement. PMBBR uses a *geolocation agent* in each MH to predict movement and calculate the total amount of bandwidth that needs to be reserved in neighboring cell j using the following equation:

$$R_j = D \sum_{i \in Nj} \sum_{k \in S'i} BW_k P_k$$

BW_k	bandwidth requirement of MH_k
P_k	probability of handoff in future
N_j	set of all cells neighboring j
S'_i	set of ongoing sessions currently in cell I and are predicted to handoff to j within interval
D	tunable constant

The implementation requires several types of agents:

- Message Agents (MA) – exchange information and management data among agents and managers.
- Message Managers (MM) – create and receive MAs.
- Geolocation Agents (GA) – contain algorithms for signal measurement and triangulation to estimate host's current position.
- Geolocation Managers (GM) – create and receive GAs.
- Bandwidth Reservation Agents (BRA) – bandwidth reservation algorithm sent to MHs.
- Bandwidth Reservation Managers (BRM) – create and send BRAs.

To request service, a MH creates and sends an MA to the access point within the cell. The access point creates a GA and BRA and sends them back to the MH. The MH uses the information from these two agents to calculate bandwidth requirements and migration probabilities. This information is periodically sent to the access point to keep the system current.

Reactive Bandwidth Reallocation [9]

This system does not perform any predictive bandwidth allocations, but instead evaluates the current resource allocation and the requirements of the new MH in order to determine if there is enough bandwidth for allocation and redistribute all of the MHs allotment if needed. The architecture is very similar to that of PMBBR except that there is no GA or GM since there is no mobility prediction and the BRA is only used to reserve bandwidth when the session is initialized and to reallocate when instructed to.

When a MH requests service it starts by sending a MA to the access point of the cell. The access point calculates the bandwidth reservation and sends a BRA back to the MH. When a MH moves into new cell, its sends a BRA to the new access point detailing its current reservation and its minimum acceptable level. If its minimum (or greater) level can be accommodated, the access point sends it a BRA with its new reservation. If

the minimum level cannot be given, then a reallocation is attempted. If this cannot be achieved, then the MH is dropped. Reallocations are performed in both directions so as to keep all of the active MHs within their acceptable QoS. Priority is given in the order of current MHs, handoffs, new sessions.

Model Combining Above Systems [9]

The model proposed assumes that there are two classes of traffic generated by MHs. Basically these correspond to real-time and best-effort traffic. Real-time traffic has a specified bandwidth requirement denoted as BW_1, but if it cannot obtain this amount, then it can continue at a single discrete lower amount denoted as BW'_1. Best-effort traffic has a desired bandwidth of BW_2, and a minimum requirement of BW'_2, but can operate at any value in between. The basic assumptions are that all current sessions have higher priority than any handoff sessions and real-time traffic has higher priority than best-effort traffic. If there is a bandwidth shortage, real-time traffic can borrow it from best-effort traffic, leaving at least BW'_2 for the best effort traffic. Reserved bandwidth does not only include active sessions within the cell, but also those reservations made by sessions in neighboring cells per the PMBBR protocol.

For new a new real-time session, the MH will first attempt to reserve BW_1, if it is available, it will start the new session. If it is not available, an attempt will be made to make a reservation at the lower BW'_1 level, if this to is not available, the connection is dropped. A best-effort session can only be initiated if there it can successfully reserve its ideal rate of BW_2. New best effort sessions are never started at sub-optimal QoS. The lower levels of acceptable QoS are used as a reserve to accommodate changing conditions and hand-offs.

When a real-time session requests a handoff into a new cell, a BRA is sent requesting BW_1. If this reservation can be accommodated, then the session is accepted with a return BRA, if it cannot be accommodated, an attempt to reserve the lower BW'_1 is made even if current best-effort traffic needs to be reallocated down to as low as its respective BW'_2 levels. As long as this reallocation is possible, the handoff is accepted. It is important to note that not all best-effort traffic is treated equally when being selected for downward reallocation. They are selected in order of highest probability of moving into a neighboring cell as determined by PMBBR. This keeps the reallocation overhead to a minimum.

Best-effort handoffs are handled slightly differently in that if it is at all possible to allocate the sessions at any level down to its BW'_2, then the handoff will be accepted. There will be no reconfiguration attempted unless it is absolutely necessary to accept the handoff. This is in order to reduce both processing and communication overhead.

Comments

This is a very simple model and approach that demonstrates an easy to understand example of managing QoS in a mobile / wireless environment. A weakness in the model is that it is heavily influenced by the cellular paradigm of call acceptance dominating QoS. In the expected future of ubiquitous computing, it will be necessary to have many idle connections with various and dynamic requirements. There also need to be

mechanisms in the protocols that allow the MH itself to make adjustments to better present the data using the available bandwidth.

5. Conclusions

Quality of Service in wireless networks is complicated by the fact that two strong technologies are being combined to form the service. QoS in the Internet environment has come to be dominated by diffServ implementations and focused on allowing real-time communication in the presence of and without blocking the dominant best-effort traffic that IP was designed for. The other technology, cellular telephony, has a strong background in QoS, but focuses of providing consistent service levels to all users instead of the variable requirements seen in the computing world.

Adding further complexity is the nature of wireless traffic and devices themselves. Several models have been demonstrated that show techniques that can help achieve desirable levels of QoS.

There is a tremendous amount of research in this field and while this paper covers the primary issues involved, it has only been able look at a very few proposed solutions to some of these issues.

REFERENCES

[1] Chalmers, Dan; Sloman, Morris "A Survey of Quality of Service in Mobile Computing Environments", *IEEE Online Communication Surveys* 2(2), 1999.

[2] Chan, Jonathon; Seneviratne, Aruna; Landfeldt, Bjorn; "The Challenges of Provisioning Real-Time Services in Wireless Internet", *Telecommunications Journal of Australia*, vol 50 no. 3 pages 37 – 48.

[3] Curran, Kevin; Parr, Gerard, "A Framework for the transmission of Streaming Media to Mobile Devices", *International Journal of Network Management*, 2002 vol 12, pages 41-59.

[4] Das, S.K.; Chatterjee, M.; Kakani, N.K. , "QoS provisioning in wireless multimedia networks" *Wireless Communications and Networking Conference*, 1999. WCNC. 1999 IEEE , 1999 Page(s): 1493 -1497 vol.3.

[5] Ei-Kadi, M.; Olariu, S.; Abdel-Wahab, H., "Rate-based borrowing scheme for QoS provisioning in multimedia wireless networks", *Parallel and Distributed Systems, IEEE Transactions on* , Volume: 13 Issue: 2 , Feb. 2002 Page(s): 156 –166.

[6] Garcia-Macias, J. Antonio et al, "Quality of Service and Mobility for the Wireless Internet", *1st Workshop on Wireless Mobile Internet – ACM*, July 2001 pages 34 – 42.

[7] Geihs, Kurt "Resource Management: Analysis of adaptation strategies for mobile QoS-aware applications", *Proceedings of the 5th ACM international workshop on Modeling analysis and simulation of wireless and mobile systems* September 2002 .

[8] Wu, Si; Wong, K.Y.M.; Li, Bo, "A dynamic call admission policy with precision QoS guarantee using stochastic control for mobile wireless networks", *Networking, IEEE/ACM Transactions on* , Volume: 10 Issue: 2 , April 2002 Page(s): 257 –271.

[9] Jian Ye; Papavassiliou, S.; Anastasi, G.; Puliafito, *A.*, "Strategies for dynamic management of the QoS of mobile users in wireless networks through software agents" ,*Computers and Communication IEEEs, 2002. Proceedings. ISCC 2002. Seventh International Symposium* on , 2002 Page(s): 369 -374

[10] Hawa, M.; Petr, D.W. "Quality of service scheduling in cable and broadband wireless access systems", *Quality of Service, 2002. Tenth IEEE International Workshop on* , 2002 Page(s): 247 –255.

[11] Kleinrock, Leonard , "Nomadicity anytime, anywhere in a disconnected world" , *Mobile Networks and Applications*, Volume 1 , Issue 4 (1996) Pages 351-357.

[12] Mahadevan , Indu; Sivalingam, Krishna M. "An architecture for QoS guarantees and routing in wireless/mobile networks", *Proceedings of the first ACM international workshop on Wireless mobile multimedia* October 1998.

[13] Mahadevan , Indu; Sivalingam , Krishna M. "Architecture and experimental results for quality of service in mobile networks using RSVP and CBQ", *Wireless Networks*, May 2000 Volume 6 Issue 3.

[14] Satyanarayanan, M., "Pervasive Computing: Vision and Challenges", *IEEE Personal Communications*, August 2001, Pgs. 10- 17.

[15] Tuoriniemi, A.; Eriksson, G.A.P.; Karlsson, N.; Mahkonen, A. "QoS concepts for ip-based wireless systems", *3G Mobile Communication Technologies, 2002. Third International Conference on* (Conf. Publ. No. 489) , 2002 Page(s): 229 –233.

[16] Valkó, András G. "Cellular IP: a new approach to Internet host mobility", *ACM SIGCOMM Computer Communication Review*, Volume 29 , Issue 1 (January 1999), pgs. 50-65.

[17] Guo, Yile; Antoniou, Zoe; Dixit, Sudhir, "IP Transport in 3G Radio Access Networks: an MPLS-based Approach", *IEEE Communications*, 2002 pages 11 – 17.

[18] Stallings, William *High-Speed Networks and Internets*, Prentice Hall 2001.

[19] Imielinski, Tomasz; Badrinath, B. R. , "Mobile wireless computing: challenges in data management", *Communications of the ACM* Volume 37 , Issue 10 (1994)Pages 18-28.