IEEE
COMMUNICATIONS
SURVEYS

A Survey of Quality of Service in Mobile Computing Environments

Dan Chalmers and Morris Sloman
Imperial College, London

Abstract

The specification and management of quality of service (QoS) is important in networks and distributed computing systems, particularly to support multimedia applications. The advent of portable laptop computers, palmtops, and personal digital assistants with integrated communication capabilities facilitates mobile computing. This article is a survey of QoS concepts and techniques for mobile distributed computing environments. The requirements of current and future mobile computing are examined and the services required to support mobility are discussed. Generic concepts of QoS specification and management are overviewed followed by an analysis of the QoS work specific to mobile computing environments.

The availability of lightweight, portable computers and wireless communications has made mobile computing applications practical. An ever more mobile workforce, home working, and the computerization of inherently mobile activities are driving a need for powerful and complex mobile computer systems and applications integrated with fixed systems. Mobile cellular telephony is widely available and computers are being integrated with these telephones to form mobile computing devices. Many businesses are dependent on distributed, networked computing systems and are beginning to rely on high-speed communications for multimedia interactions and Web-based services. Users are now requiring access to these services while travelling. In addition, new multimedia applications are emerging for Web-enabled telephones and mobile computers with integrated communications.

Multimedia applications require more sophisticated management of those system components, which affect the quality of service (QoS) delivered to the user, than for simpler voice or data-only systems. The underlying concepts of bandwidth, throughput, timeliness (including jitter), reliability, perceived quality and cost are the foundations of what is known as QoS. However, portable computers introduce particular problems of highly variable communication quality; management of data location for efficient access; restrictions of battery life and screen size; and cost of connection, which all impact the ability to manage and deliver the required QoS in a mobile environment.

The assumption that QoS will be provided and maintained, without some guarantee or notification of inability to deliver, is seriously flawed from business and technological perspectives. In many applications late information has ceased to have any value, and in "hard" real-time applications may be dangerous, or have financial repercussions. Many are prepared to tolerate slow computer interactions, or try to overcome problems by installing more processing power, or communication capacity. In many cases, overdimensioning of the system is not economic and mission-critical real-time applications cannot simply trust this uncertain approach.

Much progress has been made on providing the ability to manage QoS. Formal notations, standards, and practical implementations, particularly in the field of networks, but also more recently in systems software now exist for this. While the underlying technologies of QoS and mobile systems are well understood, the combination of the two problems has only recently started to be addressed [1, 2]. This article surveys the literature on QoS for mobile computing systems, rather than for wireless telecommunications.

The following section gives an overview of current and future mobile computing and some of the services needed to support mobility. This is followed by a summary of the generic concepts of QoS in order to understand how to specify and measure QoS in terms of its attributes and the management techniques which can also be applied to mobile systems. The QoS issues and current state of QoS provision for mobile computing environments are then discussed, with an examination of where the known QoS principles can be reused, or where mobility requires a different approach. The conclusion summarizes our views on the most important research issues in this field.

Mobile Computing Systems

In this section we examine current and future mobile computing applications and then discuss the services needed to support nomadic and mobile systems. The reader looking for a comprehensive set of references on mobile communications and computing may find [3] informative.

Mobile Computing Applications

There are many potential applications of mobile computing which will become important in the future, as the power of portable computing devices increases and the cost of wireless communications decreases.

Portable computing devices are commonly used for access to electronic mail, sending faxes, accessing the Web or remote databases, and using cellular telephones or local networks when users are travelling. Palmtop computers and personal digital assistants are being integrated with cellular telephones as part of the increasing convergence between telecommunications and computing. This type of very lightweight device could be used as an electronic "newspaper" capable of delivering selective news that is personalized according to individual user preferences and is more up-to-date than a newspaper. This could include financial information on a user's stock portfolio [4]. Web-based news delivery is already available, but currently only really used from desktop workstations. Web-based access to multimedia entertainment, videos, music, and games is also likely to increase in the future but this would require a considerable decrease in wireless communication costs to be practical for mobile computing. However, Web-enabled cellular phones and in-car communications are beginning to emerge and are considered a future growth area.

There are a range of location-aware applications in which the computing device is able to determine the users physical location, e.g., using Global Positioning Systems (GPS) which can be used to display current position on maps, receive traffic and weather information, and act as a car-navigation aid [5, 6]. This could be enhanced to provide local information such as nearest, hospitals, hotels or restaurants for travellers. Similar applications based on hand-held devices can be used as guides within museums, art-galleries or towns. These determine current position and provide information on exhibits or buildings near the user and also provide access to additional information such as a painter or architect's biography. Location aware services are discussed in the section on "Mobility Supprot Services" to follow.

The utility services are potential users of mobile computing services. Emergency services such as fire and ambulance can access plans of buildings or details of hazardous chemicals from remote databases. The use of experimental mobile computers by engineers working on power distribution systems is described in [7, 8]. Portable computers with cellular telephones were used to obtain maps indicating the current state of the power distribution system in the area in which they were working. The engineers were also able to communicate with each other and with their control center to coordinate activities and safely resolve switching requirements. There are many other applications where field workers would benefit from use of mobile computing devices for access to detailed 3-D drawings or plans, e.g., aircraft maintenance engineers, architects, or construction workers [9].

Access to educational material from digital libraries extends the concept of home or remote learning to education and training while on the move in trains or buses.

The common factor in the above future applications is they are likely to be based on multimedia interactions and not just textual data or voice.

Mobility Support Services

Mobile systems can be categorized depending on whether they use fixed or radio communication services.

Nomadic systems are typically based on wired dialup, or local area network communication facilities. Mobility is not transparent, requiring a new connection to be explicitly established when the user moves to a new location [4, 9]. For example, a nomadic user may carry a laptop and connect to a network at various times from their home, office, and various remote sites such as clients' offices or hotels. While travelling between these locations the laptop is disconnected from the network. The user may then make large geographical movements between connections, and connections over equipment with widely ranging capabilities, but will exhibit relatively static characteristics during a connection. Nomadic users may also use different computers to connect into their normal working environment. For example, many people have both an office and home computer or may connect using a local computer at a remote site of their organization. This then presents the problem of making efficient provision of resources commonly required by the user.

Mobile systems use wireless technology for transparent communications while travelling in a train, car, plane or even while walking. During the course of a connection the radio reception is likely to vary considerably, and the physical location of the device may be hundreds of miles from its starting point. While mobile telephony is generally implemented in terms of a series of discrete connections to base stations providing "cells" of coverage, a sophisticated system may be implemented such that discontinuous connection is abstracted or hidden from the user or application.

There is no absolute distinction between nomadic and mobile systems, e.g., a wireless-based mobile system may move from one site to another while disconnected so it is both nomadic and mobile. Many of the services required to support nomadic and mobile systems are very similar (as described in the next section), so we use the term mobile to cover both.

Mobile computers move, so the fundamental problem is for the system to track the current location of the mobile device in order to be able to communicate with it. Current network addressing mechanisms contain implicit location information — a subnet component maps onto a network at a particular place or within a particular organization. Telephone numbers also contain implicit location information — country code, area code, and local exchange. In the mobile telephone service, the phone number identifies a home location register (HLR) which knows the current location of the mobile device. When the telephone roams to a different service provider's coverage, it registers with a visiting location register which informs the HLR of the mobile phone's current location. Similar techniques have been proposed for mobile computing [4, 10, 11]. An alternative is to use a global directory, such as X500 which maintains the current location information and is accessed using a name rather than an address.

Location-aware services and applications require information on a user's geographical location in order to display a position on a map or provide local information as described in the section on "Mobile Computing Applications." This then requires a generalized location service, accessible by applications, to track the current position of the user. This can be accomplished using GPS, cellular telephone base stations, active badges, or determining which fixed computer is being used [6, 12]. Although the cellular telephone service providers know a subscriber's current location, they do not permit access to this information by applications running on a computer connected via a telephone, for legal or commercial reasons.

Context-aware applications require knowledge not only of location but the user's context which includes characteristics of the particular computing device being used (e.g., PDA or notebook) and information about the users current environment (e.g., who else is in the vicinity, or whether the user is in quiet or noisy environment). The application then adapts the presentation of information or quality of service provided to the user's current context [13]. This is discussed further in the section on "Context Awareness."

Mobile computing devices may also need access to local

Category	Parameter	Description/Example
Timeliness	Delay	Time taken for a message to be transmitted
	Response time	Round-trip time from request transmission to reply receipt
	Jitter	Variation in delay or response time
Bandwidth	Systems level data rate	Bandwidth required or available, in bits or bytes per second
	Application level data rate	Bandwidth required or available, in application specific units per second, e.g., video frame rate
	Transaction rate	Number of operations requested or processed per second
Reliability	Mean time to failure (MTTF)	Normal operation time between failures. See [18] For a further treatment of reliability issues
	Mean time to repair (MTTR)	Down time from failure to restarting normal operation
	Mean time between failures (MTBF)	MTBF = MTTF + MTTR
	Percentage of time available	MTTF/MTTF + MTTR
	Loss or corruption rate	Proportion of total data that does not arrive as sent, e.g., network error rate

■ **Table 1.** *Technology-based QoS characteristics.*

servers supporting electronic mail, printing, file service or databases. This could imply the need to migrate resources from the user's home servers to local ones, rather than just maintaining network connections to the home servers, in order to provide the required QoS or to reduce communication cost. [14] describes an implementation where some computing facilities acting on data (e.g., filters) are relocated in a cellular system, and discusses the problems of negotiating resource allocation as a result of location changes. Other systems, e.g., [15] simply treat the base stations as connection points into a fixed network, and do not engage in the complexities of reconfiguring in such limited time scales.

Category	Parameter	Description/Example
Critically	Importance rating (priority)	Arbitrary scale of importance, may be applied to users, different flows in a multimedia stream, etc.
Perceived QoS	Picture detail	Pixel resolution
	Picture color accuracy	Maps to color information per pixel
	Video rate	Maps to frame rate
	Video smoothness	Maps to frame rate jitter
	Audio quality	Audio sampling rate and number of bits
	Video/audio synchronization	Video and audio stream synchronization, e.g., for lip-sync
Cost	Per-use cost	Cost to establish a connection, or gain access to a resource
	Per-unit cost	Cost per unit time or per unit of data, e.g., connection time charges and per query charges
Security	Confidentiality	Preventing access to information, usually by encryption but also requires access control mechanisms
	Integrity	Proof that data sent was not modified in transit, usually by means of an encrypted digest
	Non-repudiation of sending or delivery	Signatures to prove who and when data was sent or received
	Authentication	Proof of identity of user or service provider to prevent masquerading, using public or secret encryption keys

■ **Table 2.** *User-based QoS characteristics.*

Overview of QoS

Management of QoS includes various aspects, relating to the nature of perceived quality. This section provides an overview of QoS concepts and both static and dynamic techniques for managing QoS. A more comprehensive overview of architectures supporting QoS is given in [16].

Definitions And Categories Of QoS

This overview of what is QoS and how to specify it is based on [2, 17, 18] whose treatment is primarily concerned with multimedia.

QoS Characteristics — QoS defines nonfunctional characteristics of a system, affecting the perceived quality of the results. In multimedia this might include picture quality, or speed of response, as opposed to the fact that a picture was produced, or a response to stimuli occurred. Table 1 shows the main technology-based QoS parameters. Basic mathematical models for concatenating throughput, delay, jitter, and frame loss rate are specified in [19, 20].

Table 2 summarizes the main user-based parameters. User level QoS requirements, are described as perceived quality and then mapped to lower level QoS characteristics in [21]. Reference [22] describes a selection of quality characterizations in terms of QoS parameters and value ranges, for various data types.

QoS Management

QoS management is defined in [2] as the necessary supervision and control to ensure that the desired quality of service properties are attained and sustained which applies both to continuous media interactions and to discrete interactions. It can be considered a specialized area of distributed systems management. The various aspects of

Function	Definition	Example Techniques
Specification	The definition of QoS requirements or capabilities.	Requirements at various levels of abstraction are described as combined parameter, value, allowed variation, and guarantee level descriptions. [23, 24, 25, 26] are interesting examples of work on specification of QoS requirements, and behavior in relation to actual QoS experienced.
Negotiation	The process of reaching an agreed specification between all parties.	A comparison of specifications in admission control with modification of requirements on Failure, and resource reservation when an agreement is reach. The modification of requirements should consider the inter-relation of parameters and preferences of the user. [27, 28].
Admission control	The comparison of required QoS and capability to meet requirements.	The available resources may be estimated with the aid of resource reservation information, and performance models.
Resource reservation	The allocation of resources to connections, streams etc.	A time-sliced model of capacity reserved is common, e.g., [29, 30, 31].

Table 3. *Static QoS management functions.*

interaction and types of guarantees required, as described above, must be synthesized into a specification of requirements, and relationships for trade-offs to enable the delivered QoS to be managed. We divide these first into static functions, applied at the initiation of an interaction, and dynamic functions, applied as needed during an interaction.

Static QoS Management Aspects — The static QoS management functions relating to properties or requirements which remain constant throughout some activity, are summarized in Table 3, drawing from [2, 17, 22].

In determining requirements, and agreeing to contracts it is important that the end-to-end nature of the requirements are considered. For instance, a video server may be able to computationally service a frame rate which neither its disk interface or all parts of the network passing the data to the recipients can sustain. In some situations it is necessary to consider human users as part of an end-to-end system, treating them as active participants, rather than passive receivers of information. For instance, people have thresholds of boredom, and finite reaction times. A specification of the user's perceptions is thus required, as it is the user that ultimately defines whether the result has the right quality level.

Dynamic QoS Management Aspects — The dynamic aspects of QoS management respond to change within the environment, allowing a contract to be fulfilled on an ongoing basis. Contract specifications are often inexact as resource usage and flow characteristics are not generally completely defined in advance [32]. The dynamic management functions are summarized in Table 4. These issues are expanded on in [2, 17, 22, 33].

Function	Definition	Example Techniques
Monitoring	Measuring QoS actually provided.	Monitor actual parameters in relation to specification, usually introspective. Frequency of monitoring affects monitoring traffic load but reducing frequency may result in out of specification performance for a period of time. See [34] for discussion on Piggy-back monitoring with other traffic.
Policing	Ensuring all parties adhere to Qos contract.	Monitor actual parameters in relation to contract, to ensure other parties are satisfying their part [35].
Maintenance	Modification of parameters by the system to maintain QoS. Applications are not required to modify behavior.	The use of filters to buffer or smooth streams, in order to maintain stable delay, data rate and jitter [36]. QoS aware routing to maintain network characteristics. Scaling media, e.g. by modifying levels of Detail provided within a stream.
Renegotiation	The renegotiation of a contract.	This is required when the maintenance functions cannot achieve the parameters specified in the contract, usually as a result of major changes or failures in the system. Usually invoked by exceptions raised by the monitoring, policing and maintenance function. [19, 37].
Adaptation	The applications adapts to changes in the QoS of the system, possibly after renegotiation.	Application dependent adaptation may be needed after renegotiation or if the Qos management functions fail to maintain the specified QoS. Often achieved by media scaling [27, 38, 39].
Synchronization	Combining two or more streams with temporal QoS constraints between them, e.g., synchronization of speech and video streams.	This involves representing each stream in a format where temporal information is stored with the data, allowing cross referencing between the streams.

Table 4. *Dynamic QoS management functions.*

Communications systems	Typical bandwidth	Range	Costs
Ethernet LAN	10–1000 Mb/s	Fixed, wired network.	Infrastructure & interfaces
Wireless LAN	1–10 Mb/s	100–500 m from base station.	Infrastructure & interfaces
Infra-Red	19.2 kb/s–1 Mb/s	Within room.	Infrastructure & interfaces
Satellite systems	Up to 2Mb/s in the Immediate future	World-wide in the future	Probably a monthly fee, plus cost for usage. Expected to be high.
Modem via dial-up telephone	9.6–128 kb/s	Fixed, wired network available globally. Usually for residential and small business locations.	A monthly fee, and/or costs fir usage. Low cost.
DECT CDPD GSM	32 kb/s 19.2 kb/s 9.6 kb/s	Cellular phone networks approaching national coverage. Some standard differences.	A monthly fee, and/or costs for usage. High cost.

Table 5. *Common communication systems.*

Most of the literature discusses maintaining a QoS contract under adverse conditions, often by reducing data volume. However, it should also be noted that QoS functions may be applied to increase data transfer rates when the system improves its ability to provide a service, i.e., a quality ordered sequence of alternatives due to media scaling or renegotiation may be traversed in the directions of both improvement and degradation when QoS passes given thresholds.

QoS in Mobile Computing Systems

We shall now summarise the problems of mobility in direct relation to QoS, and then describe some of the ideas and solutions being developed specifically to manage QoS in a mobile environment.

The Impact of Mobility on QoS

One of the key differences between mobile and fixed system is that the former have to be able to adapt to the changes in QoS resulting from mobility, rather than trying to provide hard guarantees of QoS [9, 15, 40, 41].

The Effects of Link Type On QoS — Nomadic systems may connect via a local area network and then reconnect via a modem or wireless link at a later time. A wireless link is obviously needed to support mobile computing. The key considerations of bandwidth, range, and cost are summarized for some popular communication technologies in Table 5.

Although new wireless technologies with higher bandwidth are emerging [42], it is a reasonable assumption that wireless network technology will continue to provide bandwidth at least an order of magnitude lower than that of fixed networks for some time, and continue to have characteristics which are more susceptible to environmental variations than wired connections. The area of coverage, and thus the degree of movement allowed while remaining connected, is related to bandwidth. Wireless LAN can cover a cell of about 500–1000 m diameter while Global Systems Mobile (GSM) may cover several square km per cell but provide country-wide and international coverage through widely deployed base stations. Satellite systems may provide similar global coverage but at very high cost in the medium term future.

The use of multimedia applications requiring high data throughput is problematic for mobile systems. Whereas speech-quality audio with compression requires only 8 Kb/s, even low-fidelity video tends towards Mb/s data rates. In addition, it is not desirable to simply limit the capabilities of systems to the lowest common denominator. It is better to try to manage the variations in data rates of the connection due to mobility and if possible, make applications adapt to these variations. Hence, the static QoS management functions must support a greater range of baseline capabilities to support mobile use.

The Effects Of Movement On QoS — One of the main problems of movement is due to handover as the mobile device moves from a cell covered by one base station to an adjacent cell of a different base-station during a connection. This handover time may result in a short loss of communication which may not be noticeable for voice interaction but can result in loss of data for other applications. Another problem is that of selecting a suitable base-station to which it can handover, which has sufficient spare capacity to support the connection [9]. For mobile computing, the base station may have to provide local processing, storage or other services as well as communication. [43] describes a system for QoS driven resource estimation and reservation to support handover. Their approach is based on a connection casting a “shadow” of advance requirement on neighbouring cells, where the shadow is stronger in the direction of movement. This cansometimes be established by including geographical knowledge of likely paths of movement. A stronger shadow represents a greater likelihood of the resource being required. The rate of handover may also be measured, suggesting reservation of more than one cell in advance (the cell currently occupied then casts a longer shadow of advance reservation). This gathering of information in conjunction with knowledge of the environment then allows confident predictions of future requirements to be made, enabling higher resource usage as fewer resources in the network are reserved unnecessarily. Another form of context aware resource reservation is described in [44], where each end of a flow is characterized as static or mobile, and advance reservations are made for mobile flows on the predicated next cell.

However, these techniques cannot completely hide all mobile link effects. Mobile wireless networks have blind spots under bridges, behind buildings or hills, where the signal may be very weak resulting in temporary quality reduction or connection loss when the mobile device is in a moving car or train. Variations in link quality can also be caused by atmospheric conditions such as rain or lightning. These effects require more sophisticated dynamic QoS management than fixed systems.

It is thus the variation in QoS which is the crucial difference between mobile systems and communications based on wired networks. This implies the need for adaptive QoS management which specifies a range of acceptable QoS levels, rather than trying to guarantee specific values. The QoS management is also responsible for cooperation with QoS aware applications to support adaptation, rather than insulating applications from variation in underlying QoS. The effects of mobility on QoS require then that algorithms employed must be capable of managing frequent loss and reappearance of mobile device in the network, and that overhead should be minimized during periods of low connectivity. This is in contrast to traditional distributed applications, where reasonably stable presence and consistently high network quality are often assumed.

The Restrictions Of Portable Devices On QoS — There are a number of limitations imposed by portability of the mobile computing device [1, 4, 9]. The main limitation is in the physical size of mobile computers, as discussed below. Mobile systems typically are designed with the limitations of batteries in mind, even where a mains power alternative is possible. Current battery technology still requires considerable space and weight for modest power reserves, and is not expected to become significantly more compact in the near future. This then places limits on the design due to the need to provide low power consumption as a primary design goal: low power processors, displays and peripherals, and the practice of having systems powered down or "sleeping" when not in active use are common measures to reduce power consumption in portable PCs and PDAs. Low power consumption components are generally a level of processing power below their higher consumption desktop counterparts, thus limiting the complexity of tasks performed. The practice of intermittent activity may appear as frequent failures in some situations. Similarly, mobile communications technology requires significant power, particularly for transmission, so network connection must be intermittent.

The second point is that of user interfaces: large screens, full-size keyboards, and sophisticated and easy to use pointer systems are commonplace in a desktop environment. These facilitate information-rich, complex user interfaces, with precise user control. In portable computers, screen size is reduced, keyboards are generally more cramped, and pointer devices less sophisticated. PDAs have small, low-resolution screens which are often more suited to text than graphics and may only be monochrome. They have minimal miniature keyboards, and pen-based, voice, or simple cursor input and selection devices. These limitations in input and display technology require a significantly different approach to user interface design.

In environments where users may use a variety of systems in different situations, the interface to applications may then be heterogeneous, and be required to scale with available devices, in a similar manner to the network connection's scaling depending on the medium used. Ideally there should be a consistent user interface for particular applications across a range of computing devices but this is not always easy to achieve.

While the limitation in battery size and power are expected to remain, I/O device technology is becoming more sophisticated: headset technology developed for virtual reality, and traditional display technology's resolution and colour representation in thin packages are areas of much development. Advances in computing power are enabling handwriting and speech based input technologies, although traditional keyboard input, and information display are unlikely to become significantly different or more advanced, due to the limitations of eyesight and dexterity of users.

QoS management in a mobile environment must allow for scaling of delivered information, and also simpler user interfaces when connecting using a general mix of portable devices and higher-power non-portable devices [1, 6]. Again the field of context aware computing provides groundwork in this area, where rather than treating the geographical context (as for mobility), one can treat the selection of end-system as giving a resource context.

The Effects On Other Non-Functional Parameters — Any form of remote access increases security risks but wireless based communication is particularly susceptible to undetected monitoring so mobility complicates traditional security mechanisms. Even nomadic systems will make use of less secure telephone and internet based communications than office systems using LANs. Some organizations may place restrictions on what data or services can be accessed remotely, or require more sophisticated security than is needed for office systems. In addition, there are legal and ethical issues raized in the monitoring of users' locations. However these topics are complex, and application- and jurisdiction-dependent, so full consideration is not possible here.

Cost is another parameter which may be affected by the use of mobile communications. However, while wireless connections are frequently more expensive, the basic principles of QoS management in relation to cost are the same as for fixed systems. The only major additional complexity is created by the possibility of a larger range of connection, and thus cost, options, and the possibility of performing accounting in multiple currencies.

CURRENT WORK ON MANAGEMENT OF QOS IN MOBILE ENVIRONMENTS

Management Adaptivity — As stated in the section "The Effects of Movement on QoS," one of the key concepts in managing QoS for mobile environments is adaptation to changes in QoS. In the following we discuss three classes of change which have to be catered for, although others approach this issue with regard to transparent and non-transparent scaling of media [38].

Large-grained change is characterized as changes due to types of endsystem, or network connection in use. Typically these will vary infrequently, often only between sessions, and thus are managed largely at the initialization of interaction with applications, possibly by means of context awareness.

Hideable changes are those minor fluctuations, some of which may be peculiar to mobile systems, which are small enough in degree and duration to be managed by traditional media-aware buffering and filtering techniques. Buffering can be used to remove jitter by smoothing a variable (bit or frame) rate stream to a constant rate stream. Filtering of packets may differentiate between those containing base and enhancement levels of information in multimedia streams, e.g., moving from color to black and white images and are similar to those in fixed network systems [35]. However, as mobile systems move, connections with different base stations have to be set up and connections to remote servers re-routed via the new base stations. This requires moving or installing filters for these connection. A new connection may not provide the same QoS as the previous one, and so the required filter technique may differ. To manage this requires an extension of the traditional interactions for migrating connections between base stations. The selection and handover of control must take account of available QoS, required QoS, and the capacity of the network to accommodate any required filters. Where the network cannot maintain the current level of service, base stations should initiate adaptation in conjunction with handover [14, 41].

Fine-grained change are those changes which are often transient, but significant enough in range of variation and duration to be outside the range of effects which can be hidden by traditional QoS management methods. These include:

- Movement between base stations in wireless networks.
- Environmental effects in wireless networks.
- Other flows starting and stopping in part of the system thus affecting resources available.
- Changes in available power causing power management functions to be initiated, or degradation in functions such as radio transmission.

These types of change must either be notified to or negotiated with the applications concerned, as they require cooperation between QoS management and the application for adaptation [7, 15, 32, 37, 40, 45, 46]. These effects may be seen as transient failure or loss, of parts of the system and can be similar to QoS degradation which can occur due to overload in fixed networks such as the Internet. Some notion of time outs on quiet connections is one simple way of differentiating between failure and absence of data or connection fade, without imposing costly polling protocols on low bandwidth connections [7]. A more advanced approach may be to absorb acceptable transient losses by probabilistic or statistical QoS specifications, which will also cause downward adaptation towards failure under sustained degradation. However, speedy reaction to degradation is important, as lossy protocols manifest themselves as severe jitter, or performance not meeting specifications. This may be achieved with techniques already developed for path adaptation, media scaling and selection, fault tolerance, and monitoring. Geographical and handover effects may be seen as failures lasting one or two seconds, so management systems should use a model of QoS requirements that allows them to absorb the more transient changes, and thus reduce unnecessary adaptations. Typical adaptations are likely to involve large steps in quality presented to users, as storage or media scaling to many levels for data intensive streams is generally expensive. Very frequent changes in presented quality may be more intrusive than small losses, or continued lower quality presentation. However, it is also important that QoS management should be able to react quickly to change when appropriate — agile response to fluctuations in QoS is considered in [45]. A user-level QoS parameter can be included to describe the trade-off between stable presentation and agile adaptation. [44, 47] suggests that where movement causes frequent fluctuations in service, the maintenance of QoS at a steady low level to provide seamless operation, is preferred by users. However, users whose systems experience less frequent fluctuations would tend to prefer that the QoS provided is maximized, at the expense of occasional disruption. This then may lead to a sliding scale of agility as a function of rate of variations causing adaptation. Another technique which is applicable in this scenario is to guarantee (as far as is possible) to provide a service at a basic level, and give best-effort management to enhancements.

It is common, in much of the literature, to concentrate on adaptation due to last-hop effects, as this fits the model of a mobile device with wireless link. In many situations it is a reasonable assumption that the wireless connection will determine the overall QoS. However, an end-to-end QoS management philosophy is still required, particularly for multicast systems, and those using the Internet for some part of their connection.

The impact of cost on patterns of desired adaptivity also becomes more pronounced in mobile systems, where connections typically have a charge per unit time or per unit data. Adaptation paths related to QoS management should be able to describe how much the user is willing to pay for a certain level of presentation quality or timeliness. The heterogeneity inherent in systems which may provide network access through more than one media will also be a factor here, as certain types of connection will cost more than others, and cost of connection may vary due to telecoms provider tariff structures.

Resource Management And Reservation — Some researchers contend that resource reservation is not relevant in mobile systems, as the available bandwidth in connections is too highly variable for a reservation to be meaningful. However, some resource allocation and admission control would seem prudent when resources are scarce, even if hard guarantees of resource provision are not practical. [44, 47] proposes that guarantees be made in admission control on lower bounds of requirements, while providing best-effort service beyond this. This is achieved by making advance reservation of minimum levels of resources in the next predicted cell to ensure availability and smooth handoff, and maintaining a portion of resources to handle unforeseen events. The issue of resource reservation is given some consideration by those working on base-stations and wired parts of mobile infrastructures, as these high bandwidth components must be shared by many users, so the traditional resource management approach still applies.

[48] describes a model of adaptivity within currently available resources. Adaptation is divided into levels of description based on the user, the application and the system - recognising that change may be required by the user or the system, and take place in the application or the system. A region of acceptable performance is mapped onto a region of the resource space in which adaptation can take place.

Resource management is related to context awareness, discussed below, as awareness of available resources is fundamental to managing them in a heterogeneous system.

Context Awareness — A further aspect of resource management is that of large-grained adaptivity, and context awareness. [49] defines situation as “the entire set of circumstances surrounding an agent, including the agent’s own internal state” and from this context as “the elements of the situation that should impact behavior.” Context aware adaptation could include migrating data between systems as a result of mobility; changing a user interface to reflect location dependent information of interest; selecting a local printer or power-conscious scheduling of actions in portable environments. The QoS experienced is also dependant on awareness of context, and appropriate adaptation to that context [11]. A fundamental paper on context awareness is [13], which emphasises that context depends on more than location, i.e., proximity to other users and resources or environmental conditions such as lighting, noise or social situations. In consideration of QoS presentation, the issues of network connectivity, communications cost and bandwidth, and location are obvious factors, affecting data for interactions as well as how end-systems are used and user’s preferences. For instance, network bandwidth may be available to provide spoken messages on a PDA with audio capability, but in many situations text display would still be the most appropriate delivery mechanism — speech may not be intelligible on a noisy factory floor, and secrecy may be needed in meetings with customers. “Quality” can thus cover all non-functional characteristics of data affecting any aspect of perceived quality.

[7] proposes that protocol management should analyse connections, and adapt to make best use of the available resources. [50] describes the use of Mobile IP [10] to provide

location transparency for mobile hosts, and the selection between interfaces to provide the most suitable communications interface and protocol for the situation and QoS requirements. The selection between alternative network interfaces then becomes a first level of context and QoS aware resource management. [40] describes an approach based on tuple spaces, which allow time and space decoupled modelling of connections, which supports fault tolerance, mobility, heterogeneity and change in a natural manner. The use of agents acting over tuple spaces provides the various aspects of management required, such as admission control, resource reservation, security, etc. This approach then allows tuple spaces to manage the context based variation in services received, and also smaller changes by the use of filter agents. An architecture for exporting environment awareness to mobile computing applications, based on the use of events to indicate changes, is described in [51].

Use Of Standards — The International Standards Organization (ISO) and International Telecommunications Union (ITU) have a joint working group defining a reference model for Open Distributed Processing (RM-ODP). They are working on a framework for specifying QoS and its components in an ODP system, but there is no specific consideration of mobility [52]. The Object Management Group have developed the Common Object Request Broker Architecture (CORBA) specification with vendors providing CORBA compliant platforms for implementing distributed systems [53]. QoS support is to be included in the CORBA 3.0 specification scheduled for release in mid 1999. The work within the Internet Engineering Task Force (IETF) has concentrated on mechanisms to support QoS management within the internet [54, 55, 56, 57, 58]. Some of these can be adapted to manage QoS for mobile systems.

The ODP and CORBA approaches are directed at maintaining transparency of platform, and hiding complexity from applications with respect to fixed computing devices. However [1] contends that in an adaptive, mobile environment this approach is no longer relevant. Some implementations e.g., [14, 41] are based on CORBA, or software components previously developed for fixed network QoS architectures [7]. These systems provide adaptive connections using existing components, while retaining the benefits of known interfaces, and re-use of low level protocol implementations. [59] surveys mobile distributed systems platforms, including a variation on the Open Group's Distributed Computing Environment (DCE), called mobile DCE. All the platforms examined (apart from Lancaster's tuple space based platform) use remote procedure call (RPC) based interaction semantics, with relaxed synchrony requirements. However, his conclusions are that the essentially synchronous nature of these protocols are unsuitable for use under degrading network QoS, due to periods of disconnection, which is his reason for suggesting asynchronous communication via tuple spaces.

CONCLUSIONS

We summarize the critical issues in managing QoS in a mobile environment, and the most interesting work relating to these issues. We consider the following to be important topics, both in existing work described in the literature, and for future development:

- The provision of context awareness, and adaptability to large-grained system dynamics, including end-system heterogeneity, and network heterogeneity. Context information must be accessible to applications to enable adaptation of QoS by user interfaces [9, 11, 13, 40],
- Context derived maps of resources, with resource models for QoS aware resource selection [11, 13]. Performance monitoring as input to these models to permit adaptation by QoS management [9, 13]. This enables context aware adaptation of protocols, with regard to overhead, and degree of synchrony depending on degree of connectivity [13, 59].
- Provision for the definition of adaptation paths from user-level QoS parameters, including trade-offs, using variation tolerant specification of parameters. Trade-off should take account of metadata relating to objects involved in requests, priority and deadline information, and available filters. QoS specification may include stability/agility and adaptation/underlying QoS effect hiding trade-off controls [9, 27, 45].
- Reservation without guarantees to increase confidence in the system's ability to perform tasks as required, particularly during periods of stability in the underlying QoS of the system. Reservation to include concepts of priority, deadlines, duration, and volume of data derived from user or application specification, metadata, and experience in a context. Additionally the use of probabilistic and stochastic resource models to enable task allocation and resource reservation with fault tolerance [31, 35, 43].
- Filtering to include delay or rejection of data, as well as scaling. Selection of filters should be aided by metadata, and awareness of available resources. Filters should act as "plug-in" modules on QoS-aware components of the system [11, 14, 35, 41].
- Control of in-service mobility, and migration of resources which is a mobile-network-oriented problem [41, 43]. Models of physical and network location and movement patterns to enable intelligent caching, replication, and migration of data for nomadic use [40, 43, 59].

In summary, much progress has already been made in providing QoS in various mobile and fixed environments. We believe that the techniques developed for QoS provision in specific environments should be brought together in a generic and flexible QoS management system so that the most appropriate methods can be deployed. Key factors to achieve this in a heterogeneous environment are the ability to define perceived QoS at the user interface level; how to relate this to underlying QoS supported within the underlying system, and how QoS-aware applications can adapt. Rather than isolating mobile systems as a special case, infrastructure and applications should be able to adapt to their environment, whatever that might be.

ACKNOWLEDGEMENTS

We gratefully acknowledge financial support from the EPSRC (Grant GR/L 06010) and the useful suggestions from the referees.

REFERENCES

[1] N. Davies, The impact of mobility on distributed systems platforms Proceedings of the IFIP/IEEE Int'l Conf. on Distributed Platforms, Dresden, Chapman & Hall, 1996, pp. 18–25.

[2] G. Blair, J-B.Stefani, Open Distributed Processing and Multimedia, Addison-Wesley, 1997.

[3] P. Agrawal, C. J. Sreenan, and M. Srivastava, Bibliography and Web Resources for ICMCS '98 Tutorial on Mobile Computing & Multimedia, http://gawain.janet.ucla.edu/tutorials/icmcs98

[4] T. Imielinski and B. R. Badrinath, *Mobile Wireless Computing Commun. of the ACM*, vol. 37, no. 10, pp. 18–28, 1994.

[5] S. Shekar and D. Lin, Genesis and Advanced Traveller Information Systems (ATIS): Killer Applications for Mobile Computing, *MOBIDATA: An Interactive J. of Mobile Computing*, vol. 1, no. 1, Nov. 1994, available

at http://www.cs.rutgers.edu/~badri/journal

[6] U. Leonhardt, Supporting Location-Awareness in Open Distributed Systems Ph.D. Thesis, Dept. of Computing, Imperial College, London, 1998.

[7] N. Davies *et al.* Distributed Systems Support For Adaptive Mobile Applications ACM Mobile Networks and Applications, Special Issue on Mobile Computing — System Services, vol. 1, no. 4, ACM Press, 1996.

[8] N. Davies *et al.* Supporting collaborative applications in a heterogeneous mobile environment Computer Communications Special Issue on Mobile Computing, Elsevier, 1996, pp. 346–58.

[9] R. H. Katz, Adaptation and Mobility in Wireless Information Systems, *IEEE Personal Commun.,* 1st Qtr 1994, vol. 1, no. 1 pp. 6–17.

[10] IETF Network Working Group: RFC 2002 IP Mobility Support, C. Perkins, Ed., 1996.

[11] B. Zenel and D. Duchamp, Intelligent Communication Filtering for Limited Bandwidth Environments, *Proc. 5th Wksp on Hot Topics in Operating Systems,* Rossario, May 1995, pp. 28–34, 1995.

[12] T. Ye, H. A. Jacobsen, R. Katz, Mobile awareness in a wide area wireless network of info-stations, *Proc. MOBICOM'98,* Dallas, Texas, pp. 109–120, 1998.

[13] B. N. Schilit, N. Adams, R. Want, Context-Aware Computing Application, *Proc. Wksp on Mobile Computing Systems and Applications,* Santa Cruz, 1994, 1994.

[14] A. Balachandran, A. T. Campbell, M. E. Kounavis, Active Filters: Delivering Scaled Media to Mobile Devices, *Proc. IEEE 7th Int'l Wksp on Network and Operating Sys. Support for Digital Audio and Video,* pp. 125–34, 1997.

[15] Srivastava, M., Mishra, P.P., On QoS in Mobile Wireless Networks, *Proc. IEEE 7th Int'l Wksp on Network and Operating Systems Support for Digital Audio and Video,* pp. 147–58, 1997.

[16] C. Aurrecoechea, A. T. Campbell, L. Hauw, A Survey of QoS Architectures, *ACM Multimedia Sys. J. — Special Issue on QoS Architecture*, May 1998.

[17] D. Hutchison *et al.*, QoS Management in Distributed Systems in Network and Distributed Systems Management, M. Sloman, Ed., pp. 273–302, Addison-Wesley, 1994.

[18] N. Storey, *Safety-Critical Computer Systems,* Addison-Wesley, 1996.

[19] Bochmann, G. v., Hafid, A., Some principles for quality of service management, *Distrib. Syst. Engng,* vol. 4, pp. 16–27, IOP Publishing, 1997.

[20] H. Knoche and H. de Meer, Quantitative QoS-Mapping: A Unifying Approach, *Proc. 5th IFIP Int'l Wksp on QoS 1997,* 1997.

[21] D. Hutchison, A. Mauthe, and N. Yeadon, Quality of service architecture: Monitoring and Control of Multimedia Communications, *Electronics and Commun. Engineering J.,* vol. 9, no. 3, pp. 100–6, 1997.

[22] K. Nahrstedt and R. Steinmetz, Resource Management in Networked Multimedia Systems, *IEEE Computer,* vol. 28, no. 5, May 1995, 1995.

[23] P. G. S. Florissi, QuAL: Quality Assurance Language, PhD Thesis, Computer Science Dept., Columbia University, New York, 1995.

[24] S. Frølund and J. Koistinen, QML: A Language for QoS Specification HP Research Report HPL-98-10, Hewlett Packard, 1998.

[25] J. P. Loyall *et al.*, Specifying and Measuring QoS in Distributed Object Systems, *Proc. ISORC '98,* Kyoto, Japan, 1998.

[26] C. Becker, K. Geihs, QoS — Aspects of Distributed Programs Int'l Wksp on Aspect-Oriented Programming at ICSE'98, Kyoto/Japan, 1998

[27] M. Fry *et al.*, QoS management in a World Wide Web environment which supports continuous media, *Distrib. Syst. Engineering,* vol. 4, pp. 38–47, 1997.

[28] J. Koistinen and A. Seetharaman, Worth-Based Multi-Category Quality-of-Service Negotiation in Distributed Object Infrastructures HP Research Report HPL-98-51, Hewlett Packard, 1998.

[29] L. Delgrossi *et al.*, Reservation Protocols for Internetworks: A Comparison of ST-II and RSVP, *Proc. of Network and Operating System Support for Digital Audio and Video,* Sept. 1993, pp. 195–203.

[30] M. Degermark, *et al.*, Advance Reservations for Predictive Service, *Proc. of Network and Operating Systems Support for Digital Audio and Video,* 1995.

[31] D. Ferrari, A. Gupta, and V. Giorgio, Distributed Advance Reservation of Real-Time Connections, *Multimedia Systems,* vol. 5, 1997, pp. 187–98.

[32] C. J. Sreenan and P. P. Mishra, Equus: A QoS Manager for Distributed Applications, *Proc. IFIP/IEEE Int'l Conf. on Distributed Platforms,* Dresden, 1996, pp. 496–509.

[33] A. Campbell, and G. Coulson, A QoS Adaptive Multimedia Transport System: Design, Implementation and Experiences, *Media Distrib. Syst. Engineering,* vol. 4, 1997, pp. 48–58.

[34] L. J. N. Franken and B. R. H. M. Haverkort, Quality of Service Management Using Generic Modelling and Monitoring Techniques, *Distrib. Syst. Engineering,* vol. 4, 1997, pp. 28–37.

[35] M. Billot *et al.,* A Proposal for Ensuring High Availability of Distributed Multimedia Applications, *Proc. of 15th Symp. on Reliable Distributed Systems,* 1996, pp. 220–27.

[36] Knightly, E.W., Rossaro, P., On the effects of smoothing for deterministic QoS Distrib. Syst. Engng vol. 4 pp. 3–15 (IOP Publishing, 1997. [37] Zhang, H., Knightly, E.W., RED-VBR: a renegotiation-based approach to support delay-sensitive VBR video Multimedia Systems vol. 5 pp. 164–176 (Springer-Verlag, 1997.

[38] Delgrossi, L., Halstrick, C., Hehmann, D., Guido, R., Krone, O., Sandvoss, J., Vogt, C., Media scaling in a multimedia communication system Multimedia Systems vol2 pp. 172–180 (Springer-Verlag, 1994.

[39] B. Li, and K. Nahrstedt, A Control Theoretical Model for QoS Adaptations, *Proc. IEEE Int'l Wksp on QoS (IWQoS '98)*, Napa, CA, May 1998.

[40] G. S. Blair, *et al.* Quality of service support in a mobile environment: an approach based on tuple spaces, *Proc. 5th IFIP Int'l Wksp on QoS,* 1997. from http://www.comp.lancs.ac.uk/computing/research/mpg/index.html

[41] A. T. Campbell, Mobiware: QoS-Aware Middleware for Mobile Multimedia Communications, http://comet.columbia.edu/~campbell/andrew/publications/publications.html

[42] S. Kao, Speedy Wireless Networks, *Byte Int'l Supplement,* vol. 23, no. 3, March 1998, pp. 415–17 (McGraw-Hill 1998.

[43] D. A. Levine, I. F. Akyildiz, and M. Nagshineh, A Resource Estimation and Call Admission Algorithm for Wireless Multimedia Networks Using the Shadow Cluster Concept, *IEEE/ACM Trans. on Networking,* vol. 5, no. 1, 1997, pp. 1–12.

[44] Lu, S., Lee, K-W., Bharghavan, V., Adaptive Service in Mobile Computing Environments, *Proc. 5th IFIP Int'l Wksp on QoS 1997,* Chapman & Hall, 1997.

[45] B. D. Noble *et al.,* Agile Application-Aware Adaptation for Mobility, *Proc. 16th ACM Symp. on Operating Systems Principles,* pp. 276–87, 1997.

[46] M. McIlhagga, A. Light, I. Wakeman, Towards a Design Methodology for Adaptive Applications, *Proc. MOBICOM '98,* Dallas, Texas, pp. 133–144, 1998.

[47] S. Lu, V. Bharghavan, Adaptive Resource Management Algorithms for Indoor Mobile Computing Environments, *ACM SIGCOMM,* pp. 231–242, 1996.

[48] Gecsei, J., Adaptation in Distributed Multimedia Systems, *IEEE Multimedia,* April-June 1997.

[49] Turner, R.M., Context-Mediated Behaviour for Intelligent Agents, *Int'l J. of Human-Computer Studies,* 1998, vol. 48, pp. 307–330.

[50] X. Zhao, C. Castelluccia, M. Baker, Flexible Network Support for Mobility, *Proc. MOBICOM '98,* Dallas, Texas, 1998, pp. 145–156.

[51] Welling G, Badrinath B., An Architecture for Exporting Environment Awareness to Mobile Computing Applications, *IEEE Trans. on Software Engineering,* vol. 24, no. 5. May 1998, pp. 391–400.

[52] ISO/IEC JTC1/SC33 1593, Open Distributed Processing — Reference Model — QoS (First CD), June 1998.

[53] OMG The Common Object Request Broker: Architecture and Specification Revision 2.2, Feb. 1998.

[54] IETF Network Working Group: RFC 2205 Resource Reservation Protocol (RSVP) - Version 1 Functional Specification, B. Braden *et al.*, Eds., IETF, 1997.

[55] IETF Network Working Group: RFC 2212 Specification of Guaranteed QoS, S. Shenker, C. Partridge, R. Guerin, IETF, 1997.

[56] IETF Network Working Group: RFC 1633 Integrated Services in the Internet Architecture: An Overview, R. Braden, D. Clark, S. Shenker, Eds., IETF, 1994.

[57] IETF Network Working Group: RFC 2474 Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers, K. Nichols *et al.*, Eds., IETF, 1998.

[58] IETF Network Working Group: RFC 2475 An Architecture for Differentiated Services, S. Blake *et al.*, Eds., IETF, 1998.

[59] N. Davies *et al.*, Limbo: A Tuple Space Based Platform for Adaptive Mobile Applications, *Proc. Int'l Conf. on Open Distributed Processing/Distributed Platforms,* 1997, http://www.comp.lancs.ac.uk/computing/research/mpg/index.html

[60] R. Titmuss *et al.*, Mobility and Multimedia Information Software Agents and Soft Computing, Towards Enhancing Machine Intelligence: Concepts and Applications, Springer-Verlag, 1997, pp.146–59 .

BIOGRAPHIES

DAN CHALMERS (dc@doc.ic.ac.uk) (http://www-dse.doc.ic.ac.uk/~dc) obtained his B.Sc. in software engineering from University of Manchester Institute of Science and Technology and an M.Sc. in advanced computing from Imperial College. He worked as a Systems Engineer for Ericsson Ltd., Burgess Hill, U.K. before doing his M.Sc. He is currently a research associate in the Dept. of Computing, Imperial College and studying part-time for his Ph.D. His research interests include architectural description languages and dynamic reconfiguration of components and services in response to changes in QoS or context, particularly to support mobile multimedia applications.

MORRIS SLOMAN (mss@doc.ic.ac.uk) (http://www-dse.doc.ic.ac.uk/~mss) obtained his B.Sc. in electronic engineering from University of Cape Town, South Africa and a Ph.D. in computing from University of Essex, U.K. He has been in the Dept. of Computing, Imperial College, since 1976. He has managed many research projects funded by the UK Engineering and Physical Science Research Council (EPSRC), European Union and various industries on management, security and design of distributed systems, multimedia systems, and mobility. He is editor of a reference book, *Management of Network and Distributed Systems (*Addison Wesley), co-editor of *IEE/IOP/BCS Distributed Systems Engineering J.,* and a member of the editorial board of the *J. of Network and Systems Management.* He is program co-chair of the First IEEE Enterprise Distributed Object Computing (EDOC) workshop and is a member of the EDOC steering committee. He program co-chair of the 1999 IEEE/IFIP Integrated Management Symposium (IM '99). He is chair of the EPSRC Multimedia and Network Applications Funding Programme.