# Survey of Quality of Service in Mobile Computing Environments

Dan Chalmers, Morris Sloman

30th July 1998, revised 4th February 1999

Research Report 98/10

Department of Computing,

Imperial College, London, SW7 2BZ

Email: dc@doc.ic.ac.uk, mss@doc.ic.ac.uk

## 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 lap-top computers, palmtops and Personal Digital Assistants with integrated communication capabilities facilitates mobile computing. This paper is a survey of QoS concepts and techniques for mobile distributed computing environments. The QoS attributes typically specified and negotiated for general communication systems are described as well as some QoS models. A brief overview is given of some practical systems described in the literature. The design issues relating to both mobile and nomadic computing are explained and then the specific QoS issues related to mobile and nomadic systems are discussed. The conclusion summarises the important issues relating to supporting QoS for mobile systems.

### Keywords

Mobile systems, nomadic systems, quality of service, multimedia

# 1. Introduction

Reliable message transfer with error control and notification of non-delivery is common in many modern communication systems. However, it is only recently that much thought has been given to the ability to specify timeliness, and the perceived quality of the data arriving, particularly where more complex (multi-)media are being used. The underlying concepts of bandwidth, throughput, timeliness (including jitter), reliability, perceived quality and cost are the foundations of what is known as Quality of Service (QoS).

Many business are dependent on distributed, networked computing systems and are beginning to rely on high-speed communications for multimedia interactions and Web based services. The assumption that QoS will be provided and maintained, without some guarantee or notification of inability to deliver, is seriously flawed from a business and technological perspective. In many applications late information has ceased to have any value, and in “hard” real-time applications late information may be dangerous, or have financial repercussions. While many are prepared to tolerate slow computer interactions, others try to overcome problems by installing more processing power, communication capacity, or the latest version of software, but this may not solve the problem. However, in many cases over-dimensioning 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 requirements, and manage situations when the required QoS is not available. Formal notations, standards, and practical implementations, particularly in the field of networks, but also more recently in systems software, now exist for this, and are described below.

Another aspect of technology which is becoming more prevalent is that of mobile communications. An ever more mobile workforce, home working, and the computerisation of inherently mobile activities are driving a need for powerful and complex mobile computer systems and applications integrated with fixed systems. Mobile telephony and packaging computer technology for portability are both well established fields, with well known and understood problems, and ever improving solutions. The particular problems of highly variable connection quality (especially set-up time, throughput, and error rate); management of data location for most efficient access; the restrictions of battery life and screen size on portable systems; and cost of connection (e.g. through mobile or hotel telephones) all impact the ability to manage and deliver the required QoS in a mobile environment.

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 [DAVIES 96a, BLAIR 97b].

This report surveys the literature on QoS for mobile computing systems. While reference is made to managing QoS in general purpose networks and wireless communications systems, and other related topics such as location aware services, security and mobile software these topics are not the focus of this paper, as they each form a large and complex topic in themselves.

## 1.1 Outline of Contents

The literature available on QoS, mobile communications and mobile computing is vast and so it was not possible to provide an exhaustive literature survey. We have selected a subset of the available papers based on the following criteria. In order to better understand the problems and results of work in QoS and mobility, section 2 is a review of the literature which best describes generic concepts of QoS in order to understand how to specify and measure QoS in terms of its attributes and management techniques which can also be applied to mobile systems. We have also an emphasised QoS techniques for multimedia. Section 3 reviews the common problems of mobile computing and its dependency on mobile communications. As stated above, our survey does not cover the large literature on wireless telecommunications. Section 4 considers current work on provision of QoS in a mobile computing environment, examining where the convergence provides for reuse of known principles, or requires a different approach, and section 5 summarises our views on the most important issues and current state of QoS provision for mobile computing systems, and future research issues.

## 1.2 Abbreviations

| | |
|---|---|
| ATM | Asynchronous Transfer Mode |
| CCITT | International Telegraph and Telephone Consultative Committee |
| CDPD | Cellular Digital Packet Data |
| CITR | Canadian Institute for Telecommunications Research |
| CORBA | Common Object Request Broker Architecture |
| CoS | Class Of Service |
| DCE | Distributed Computing Environment |
| ETSI | European Telecommunications Standards Institute |
| FDDI | Fibre Digital Data Interface |
| GSM | Groupe Spécial Mobile now known as Global Systems Mobile |
| IETF | Internet Engineering Task Force |
| IMA | International Multimedia Association |
| ISO | International Standards Organisation |
| ITU | International Telecommunication Union |
| MTBF | Mean Time Between Failure |
| MTTR | Mean Time To Repair |
| MSS | Multimedia System Services |
| ORB | Object Request Broker |
| RSVP | Resource Reservation Protocol |
| PDA | Personal Digital Assistant |
| QoS | Quality Of Service |
| RM-ODP | Reference Model for Open Distributed Processing |
| TINA | Telecommunications Information Networking Architecture |
| UMTS | Universal Mobile Telecommunications Service |

# 2. Overview Of Quality Of Service

Many multimedia and real-time applications entail the concept of QoS, where there may be a scale of performance which is acceptable, and the boundary between success and failure of the system may be blurred or varying [BLAIR 97b]. For instance, a video may still be acceptable if presented with a lower frame rate, or reduced resolution, but be unacceptable for viewing if there are pauses or gaps in the film. The results of a search on data potentially returning many fields may still be acceptable if only the first few results are given initially, as long as this is done within a time limit, when appearance of efficiency is needed in front of customers, or where connection time waiting for results incurs cost.

Management of QoS includes various aspects, relating to the nature of perceived quality. This section provides an overview of these, and the treatment they have been given in the literature. The topics covered are: i) Definitions Fundamental to QoS management; ii) Techniques for the static management of QoS; iii) Techniques for the dynamic management of QoS; iv) QoS issues relating to multiple stream systems, common in multimedia applications; v) QoS issues relating to managing faults and availability; vi) A brief overview of the work described in the literature relating to formal specifications, standards and practical investigations where the previously described techniques are applied.

## 2.1 Definitions and Categories Of QoS

The ISO/IEC recommendation X.901-5, Open Distributed Processing Reference Model (RM-ODP) provides a de-jure definition of QoS. This subsection presents an overview of the most critical parts of what QoS is, and what it entails, based on [BLAIR 97b, HUTCHISON 94, STOREY 96], whose treatment is primarily concerned with multimedia.

### 2.1.1 QoS Characteristics

While systems are often defined in terms of their functionality, QoS defines non-functional 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 which we consider, and Table 2 summarises the main user-based parameters.

[HUTCHISON 97] describes perceived quality as user level QoS requirements, and then maps them to lower level QoS characteristics. [NAHRSTEDT 95a] describes a selection of quality characterisations in terms of QoS parameters and value ranges, for various data types.

**Table 1 Technology-Based QoS Characteristics**

| 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 | System level data rate | Bandwidth required or available, in bits or bytes per second. Basic mathematical models for concatenating throughput, delay, jitter and frame loss-rate are specified in [BOCHMANN 97, KNOCHE 97]. |
| | Application level data rate | Bandwidth required or available, in application specific units per second, e.g. video frame rate |
| | Transaction rate | Operations requested or capable of being processed per second |
| Reliability | Mean Time To Failure (MTTF) | Normal operation time between failures. See [STOREY 96] 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 which does not arrive as sent, e.g. network error rate |

*Cost* is a slightly different category to the others described, as it is not an intrinsic part of the visible results of most transactions. Cost is generally described in terms of a monetary value per interaction, or in terms of the time spent interacting. It is often the case that cost will be used to place upper and lower limits on other characteristics. For example, I am prepared to pay £5 per hour to watch this film, only if the quality of reproduction is at least half that of terrestrial television, but I will not pay more than £8 an hour, however good the quality of the results, and I want a refund if it stops half way through. Note that the quality of reproduction in this example will probably be decomposed to terms such as jitter, frames per second, resolution and colour depth. [BOCHMANN 97] includes treatment of this characteristic, but found that where multiple choices are possible within QoS requirements, the desired trade-off when incorporating cost may not be clear if requirements are not prioritised. [BOUCH 99] discusses pricing policies in relation to users, and observes that describing pricing in terms of application level characteristics is required. [KOISTINEN 98] discusses negotiation of a specification based on the perceived worth of the various alternatives offered to the user. It the model described the information source is not given a description of the client's worth descriptions for various parameters. However, it is not hard to imagine that one parameter being negotiated over is cost, or that the server could offer alternatives with various costs coupled with variations in other parameters corresponding to levels of perceived QoS from market research. Once user preferences are classified it would not be complex to offer standard packages based on the market's willingness to pay for a level of service, rather than the cost of providing that service, as already happens in many other industries.

**Table 2 User-Based QoS Characteristics**

| Category | Parameter | Description / Example |
|---|---|---|
| Criticality | Importance rating | Arbitrary scale of importance, may be applied to users, different flows in a multimedia stream, etc. |
| Perceived QoS | Picture detail | Pixel resolution |
| | Picture colour accuracy | Maps to colour 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 synchronisation | Video and audio stream synchronisation, 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 | Prevent 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 sent or received data and when this occurred. |
| | Authentication | Proof of identity of user or service provider to prevent masquerading., using public or secret encryption keys. |

The *Security* requirements, indicated in Table 2, may be specified as a QoS requirement in terms of discrete classes or levels as for other QoS parameters. We shall not examine security mechanisms here, as it forms a large and separate topic in itself. However, it is worth noting that security protocols increase the overheads in terms of extra messages and increased data, so the required security may also be subject to the trade-offs applied by QoS management [KOISTINEN 98]. See [PFLEEGER97, STALLINGS 98] as a starting point for the topic of security.

### 2.1.2 Class Of Service

A further important classification of QoS requirements, or more particularly the systems implementing the requirements, is the class of service provided. [BOCHMANN 97] subdivides classes of service (CoS) into five levels:

- Deterministic guarantee
- Statistical guarantee
- Target objectives
- Best effort
- No guarantee

These are not the only classes of service in common use. Many networking standards use CoS to describe levels of service for other parameters than the level of guarantee given to QoS specifications, as we are using. For instance, [HALSALL 92] describes the 5 CoS levels in the ISO/OSI reference model (ISO 7498) to describe transport layer protocol classes.

*Deterministic guarantees* will always be met or bettered, under all circumstances, while a *statistical guarantee* allows a percentage of time where the guarantee is not met. The last three levels provide no real guarantee, but offer varying levels of assistance in achieving the desired QoS. A system which takes account of *target objectives* will try to satisfy requirements, with some knowledge of their implications, which could then be used to determine scheduling priority. A *best effort* system, like the Internet, would provide the same QoS for all services i.e. with no real consideration of QoS factors. Some historic information about performance is then the only guide to the level of service to be expected, although there is a move to provide some QoS guarantees within the Internet – see section 2.5.4. *No guarantee* is a similar class to best effort, although it is unlikely any information about system performance is available with this class of service. There are often limits to the degree of guarantee available, for instance under multiple failure conditions, however both the above guaranteed classes become important under high load and overload conditions.

From the examples given here, and in the literature, it is apparent that all parts of a system must perform their jobs to a certain standard, to achieve a given overall QoS. This end-to-end specification of QoS has been a major practical hurdle, as while network technology e.g. ATM has well established QoS defining characteristics, often many parts of  the internet do not actually have the capability of supporting QoS specifications.

QoS management is defined by Blair as “the necessary supervision and control to ensure that the desired quality of service properties are attained and” (where applicable) “sustained ... QoS management applies both to continuous media interactions and to discrete interactions”, and as [BOCHMANN 97] proposes, it can reasonably be seen as a specialised area of distributed systems management. It is this management of systems for QoS which is described in the following sub-sections.

## 2.2 QoS Management

The various aspects of interaction and types of guarantees required, as described above, must then be synthesised into a specification of requirements, and relationships for trade-offs to enable the delivered QoS to be managed. We divide these first into static and dynamic functions: those which are applied at the initiation of an interaction, and those which are applied continuously or as needed during an interaction.

### 2.2.1 Static QoS Management Aspects

The static QoS management functions relating to properties or requirements which remain constant throughout some activity, are summarised in Table 3 and expanded below, drawing from [HUTCHISON 94, NAHRSTEDT 95a, BLAIR 97b].

**Table 2 Static QoS Management Functions**

| Function | Definition | Example Techniques |
|---|---|---|
| Specification | The definition of QoS requirements or capabilities. | Requirements at various levels of abstractions are described as combined parameter, value, allowed variation, and guarantee level descriptions. |
| 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 reached. The modification of requirements should consider the inter-relation of parameters and preferences of the user. |
| 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. |

*QoS specification* is the creation of a contract between producers and consumers of data, based on a specification of requirements. As described in 2.1 requirements may be at various levels of abstraction from user to low-level descriptions, and describe a range of interrelated characteristics. This may be the starting point for negotiation (see below). [McILHAGGA 98, BOUCH 99] discuss some psychological aspects of user level QoS specification, and the variation in user level specifications between groups of users with different levels of technical knowledge, temperaments and in different situations. Each element in an interaction may specify what it requires and what it is capable of delivering for each QoS parameter it manages. [FLORISSI 95, FRØLUND 98, LOYALL 98] are interesting examples of work on languages for the specification of QoS requirements, and behaviour in relation to actual QoS experienced.

*QoS negotiation* is the process of reaching an agreement between parties in an interaction, on the acceptable bounds on the QoS to be delivered, to form an agreed contract between all parties. This function must consider the specifications and dependencies of all parties, and may reject a contract, or submit a different proposal from that requested, if the original set of specifications cannot be honoured on an end-to-end basis. [FRY 97] uses a notation of user specified weights on parameters to enable some automatic trade-off to take place, and [KOISTINEN 98] describes the use of a ‘worth’ based mechanism, which combined with the use of constraint specification and weighted parameters allows a general mechanism for selection of a specification from a set of alternative possible provisions and a set of requirements. There are many algorithms described in the literature, which cannot all be covered here.

*Resource reservation* is a complementary function to admission control, where agreed requests are registered, and the required resources allocated from the available pool. This then assists in predicting and guaranteeing the performability of requests by tracking expected system usage. [DEGERMARK 95 &97, FERRARI 97] all provide a similar and elegant solution for resource reservation and admission control in a distributed environment (Degermark's using RSVP, Ferrari's in the Tenet Real-Time Protocol Suite 2), with particular emphasis on providing for advance reservation, and resource use with specified time limits or duration. The increase in the information available provides a significant improvement in the acceptance rate of requests for bandwidth due to better planning. This protocol also provides for more general immediate, unspecified duration channels alongside and dynamically sharing bandwidth with those with advance reservation, and specified duration. [DELGROSSI 93] describes the IP based protocols of ST-II and RSVP, which provide basic resource reservation in the context of multicasts.

These four static management functions all depend on being able to specify requirements and the current state of the system in an appropriate manner. This may be achieved with deterministic guarantees, for "hard" specifications, or using probabilistic or stochastic specifications. Deterministic requirements specify a precise value or range of values to be achieved for a given characteristic. The probabilistic and statistical methods require a value or range of values to be met for a given proportion of events or time. This is likely to be a more realistic specification than the deterministic specification, which often requires significant over reservation of resources to achieve.

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 it's 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, given people have thresholds of boredom, and finite reaction times. End-to-end QoS provision including description in terms of the user's perceptions is required, as it is the user that ultimately defines whether the result has the right quality level.

Each component involved in providing a service must be able to negotiate with the other components it interacts with. The contract negotiated with the end-consumers of the service should be based on the complete system's abilities (including both maximum and minimum capabilities per characteristic, and the interplay between characteristics), and each part of the system can then provide it's services at the same level as the others.

[BOCHMANN 97] proposes several supporting functions to aid in static QoS management:

- A database of multimedia documents may hold meta-data about versions of each, to aid in assessing resource requirements and contract proposals, for instance, recording the required throughput of two versions of a video stream, or historical delays between the user networks and systems holding various copies of a document.
- A profile manager to aid users in specifying their QoS requirements, and then storing them for future use. This could also be managed by a service provider, allowing certain defaults or ranges of QoS requirement to be available to different classes of user, or for updating tariffs when calculating acceptable cost.
- A network monitor (and similar systems monitors) could be provided, to allow monitoring of the system's state as an input to admission control, and dynamic functions (see below).

A QoS manager is then proposed to manage negotiation and adaptation, by decomposing tasks for negotiation with components, and selecting combinations of components, where there is a choice, which provide the service. This function then provides negotiation for users and applications, and transparency of the complexity of providing end-to-end specifications.

The provision of deterministic QoS requires that it be maintained in the presence of failures. We consider two types of failure:

i) *Transient data corruption or loss* is routinely catered for by networking systems i.e. loss of packets. There are two main options when addressing this problem: request a retransmission, or accept the loss. From a QoS point of view, retransmission has advantages where maintaining a data rate or perceived quality is concerned, however, the act of requesting and receiving retransmission may be time consuming, which also impacts the achieved QoS. In systems where a part of the system is known to be error prone, it may be that the QoS management functions have to negotiate QoS with an expectancy of a certain loss or retransmission rate.

ii) *Failure of part of the system* can be regarded as the situation where that part ceases to perform with the required QoS in a sustained manner e.g. in the case of network congestion. Two options are available: fail the tasks which depended on the failed part, or relocate the activities from the failed part elsewhere. The latter option requires redundancy of services, provision for commencement of activities at intermediate points, and possibly providing for partial renegotiation of QoS during the performance of a task. This may raise considerations of cost, both in equipment and computational and network resources.

[BILLOT 96] is an interesting example of allowing for availability to be part of the QoS negotiation specification.

### 2.2.2 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. Also, as [SREENAN 96] notes specifications are often inexact as resource usage and flow characteristics are not generally completely defined in advance. The dynamic management functions are summarised in Table 3 and expanded below, drawing from [HUTCHISON 94, NAHRSTEDT 95a, BLAIR 97bc, CAMPBELL 97b].

*Policing* is concerned with ensuring that all parties adhere to their part of the service contract. For instance, where a video frame rate of 25 frames per second is required, the provider must not consistently provide too few frames, or generate bursts which saturate some part of the system – see [BILLOT 96] for a description of input and output work-ahead limits to prevent impeding overall QoS provision to all users

*Renegotiation* is something of a last resort response to sustained failure to honour a QoS contract. [BOCHMANN 97] suggests taking the average QoS provided over an interval to avoid spurious renegotiation due to very transient fluctuations. A tuned probabilistic or stochastic model could incorporate current information about the causes of failure and resource characteristics. [ZHANG 97] describes client controlled renegotiation and examines network utilisation. They demonstrate graceful QoS degradation of deterministic guarantees, which they contend shows benefits over statistical guarantees. Renegotiation can be invoked by users, on deciding that they do not consider a given characteristic to be acceptable as specified, having experienced it. Renegotiation may also be invoked on long running tasks, where the characteristics of the underlying system vary with time.

**Table 3 Dynamic QoS Management Functions**

| Function | Definition | Example Techniques |
|---|---|---|
| Monitoring | Measuring QoS actually provided. | Monitor actual parameters in relation to specification, usually introspective. |
| Policing | Ensuring all parties adhere to QoS contract. | Monitor actual parameters in relation to contract, to ensure other parties are satisfying their part. |
| Maintenance | Modification of parameters by the system to maintain QoS. Applications are not required to modify behaviour. | The use of filters to buffer or smooth streams, in order to maintain stable delay, data rate and jitter [KNIGHTLY 97]. 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 | Renegotiation of a contract 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 functions. |
| 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. |

*Adaptation* by media scaling requires multiple versions of the source to be stored, and a system where swapping between versions during presentation, while maintaining position in the record, is required. Various scaling techniques are examined in [DELGROSSI 94, FRY 97]. In general, tasks such as changing resolution of an MPEG video stream in real time is computationally expensive. [SISALEM 97] describes the use of an algorithm for adapting multimedia flows in a multicast application in response to monitored network conditions. [LI 98] gives an example of a control-theoretical approach to adaptation. Many other algorithms and techniques exist in this field, each being designed with a particular application and/or network scenario in mind. We do not intend to discuss these at great length here, although we believe that the use of application level adaptation is a key feature in QoS support for mobile computing systems.

Most of the literature discusses maintaining a QoS contract under adverse conditions, or reducting a data stream to stay within limits. 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.

An important consideration in specifying these functions is the interplay of the granularity of their response, and the load they create on the system. Too little monitoring may cause out of specification performance for a period of time while these measurement and management functions in themselves will place a load on the systems they are monitoring. The overhead is often at least partly dependent on the granularity of the measurement, so that more frequent or accurate measurement causes greater overhead. For some parts of the system it is possible to piggy-back monitoring with the actual work resulting in a minimal overhead [FRANKEN 97].

### 2.3 QoS For Multiple Stream Systems

The effects of multiple streams have two main variations [BLAIR 97b]:

- Multicast and combinational receiving
- Synchronisation of multiple sources

Multicast is the transmission of the same information to multiple receivers. This is gaining support in networking protocols, but the most commonly found protocols use point to point connections, and multicast is often implemented using multiple physical messages abstracted as one logical multicast message. Fusion (combining streams) on the other hand, might occur in a multi-party video conference, where each participant sends video signals to each of the other participants. The application in this situation must be designed to manage multiple concurrent stream interfaces, both incoming and outgoing. It is obvious that this is the reverse process to multicast: if the arrows in Figure 1 represent data flow, they would be reversed for a combinational system.

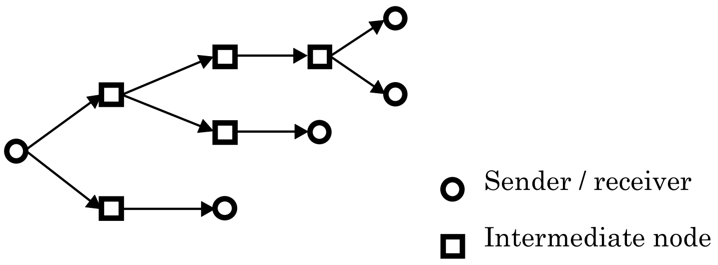

**Figure 1: A Multicast Graph, adapted from [BLAIR 97b]**

Synchronisation is required between separate audio and video streams so that words in audio are played in synchronisation with lip movement in video.

QoS may be specified in these situations, and they pose some slightly different problems compared to a point-to-point connection scenario.  In a multicast system, it may be that not all of the recipients, or the paths to the recipients, are capable of supporting the same QoS. In this case while recipients may be receiving the same logical data, each recipient may be receiving it at a different QoS level. This may be achieved by the use of media scaling filters at intermediate nodes, or by subdividing the recipients into groups, each with the capability to handle a different class of service. A system with combining streams must manage the cumulative effect of each stream through it’s resource allocation and/or monitoring and feedback functions.

A further effect of combining streams is that it may be possible to drop certain streams under adverse conditions, without  loss of comprehension by the user. It may be possible to assign some relative importance to individual streams within an application. For example, in a video conferencing system where audio, video and a “white-board” application are used, the video stream may reasonably be sacrificed to maintain quality in the other streams, without the system becoming unusable. [BOCHMANN 97] suggests this technique, but notes that a generic model for media replacement adaptation is not simple to define. He also suggests a further stage to this of converting speech to text data in certain restricted situations.

Synchronisation of streams is a special case of combining streams which places temporal QoS constraints between the two streams, rather than on the system, which is the case considered up to now. Generally this will involve representing each stream in a format where temporal information is stored with the data, allowing cross referencing between the streams, to assist in playback. QoS management functions can only assist in the provision of synchronised media by managing the QoS contracts of the multiple streams under the knowledge that they must be related. Synchronisation is a separate and complex topic in itself, and we shall not consider this further here.

## 2.4 Formal Reasoning About QoS

An object model for multimedia which supports QoS annotations, and a real time logic to aid this specification are described in [BLAIR 97b]. The real time logic, QL, has been designed to specify QoS annotations for timeliness, volume and reliability. The use of the logic allows analysis of properties of components with regard to specifications, and the combination of logic models for components, thus allowing a rigorous examination of systems, and their monitoring and policing functions. However, QL has limitations: it does not support probabilistic or stochastic specifications, which in turn inhibits it’s descriptiveness when faced with reliability concepts such as MTBF, rather than simpler reliability models such as lost frames. Several other logic models are mentioned which could be applied to QoS, or are in development.

An alternative representation is given by [FRANKEN 97], where Stochastic-Petri nets are used to describe models, with failure/repair rates, and delays. This model is used to build a moderately complex composition of components to simulate a video-phone, by defining generic component models. This technique also has limitations, particularly with regard to complex (e.g. Erlang) distributions of events, and modelling of jitter, however their models were compared to a monitored implementation, and provided very accurate predictions of throughput and delay for given frame rates.

## 2.5 Distributed Systems Models And Quality Of Service

Several standards, modelling constructs, and basic architectures have been defined to support development of distributed systems. Some of them address QoS issues. We summarise some of the work in this area below, for detailed information the reader is referred to the standards bodies themselves and [HUTCHISON 94, BLAIR 97b, AURRECOECHEA 98].

### 2.5.1 RM-ODP & TINA

The Reference Model for Open Distributed Processing (RM-ODP) is a framework for distributed systems standards, jointly produced by both ISO and ITU-T. It defines models to represent various viewpoints on a system in an object oriented manner, and transparencies to manage the complexities of distributed environments. Being a meta standard, it does not place any specific requirements on implementations, but is used extensively in TINA and by some authors, as a starting point for describing the design of a distributed system. There is specific work on QoS within the RM-ODP framework – see http://www.iso.ch:8000/RM-ODP/ and [BLAIR 97b, DAVIES 96b]. The Telecommunications Information Networking Architecture (TINA) is the result of work by several major telecommunications providers and manufacturers, and interested parties from the computing world. It has been designed specifically for implementation work in a telecommunications environment, but is based on the RM-ODP model. Specifically it provides a generic architecture for describing and controlling networks, and basic service and management implementation frameworks. These services could include multimedia and incorporate QoS management, although QoS has not been a specific emphasis of the TINA work.

### 2.5.2 CORBA

The Common Object Request Broker Architecture (CORBA) is an Object Management Group (OMG) object model, interface definition language, and architecture definition, which describes various standard services. There are a number of CORBA compliant distributed programming environments available. CORBA essentially supports remote object invocation but does not specify support for multimedia interactions or QoS management, although there is work within OMG on extensions for this. See http://www.omg.org/library/library.htm for details of work within the OMG. An example of using aspect oriented programming and CORBA as a basis for QoS enabled extensions to the OMG IDL is described in [BECKER 98].

### 2.5.3 IMA MSS

The International Multimedia Association (IMA) Multi-media System Services (MSS) is a CORBA based architecture specifically designed to support multi-media. As such it provides object classes for devices supporting various media types, connection types, QoS, and stream and synchronisation mechanisms. It’s support for QoS requirements are deterministic, providing only minimum and maximum values for delay, bandwidth and jitter on devices and “virtual connections”, but not on standard CORBA requests. [BLAIR 97b] notes, that MSS does not allow any extension to these QoS parameters. Three classes of service are defined: No guarantee, best effort, and guaranteed. Inter-object dependencies are not supported either, except through a grouping of objects, and there is no explicit framework for QoS management functions, although some static functions are implicitly present in virtual resource classes, and there are no suitable supporting functions for dynamic functions.

### 2.5.4 IETF

The Internet Engineering Task Force (IETF) are in the process of producing several standards for QoS management on IP to improve on best-effort service, including RSVP [RFC 2205], QoS Guarantees [RFC 2211, RFC 2212], int-serv [RFC 1633, RFC 2210, RFC 2215], and diff-serv [RFC 2474, RFC 2475]. The IETF work is still under development. Up-to-date information can be obtained from http://www.ietf.org/ and http://diffserv.lcs.mit.edu/ Further discussion and examples of work using these standards and techniques are to be found in [DELGROSSI 93, FLORISSI 99, TASSEL 97].

## 2.6 Examples Of Current Practical Work

We describe below some of the applications and supporting environments used to develop QoS management principles in a practical manner documented in the literature. These descriptions are not comprehensive, but are intended to show something of the diversity of work being undertaken, and illustrate the practical work being undertaken in relation to the issues and techniques being described above. [AURRECOECHEA 98] provides a fuller treatment of QoS Architectures, and many of the other referenced papers describe practical work.

### 2.6.1 QoS-A

[HUTCHISON 97] describes another ongoing research project at Lancaster University, UK, which provides for QoS in design and implementation. The Quality Of Service - Architecture (QoS-A) provides "a framework to specify and implement the required performance properties of continuous media applications". It provides a layered approach to QoS, abstracting the sub-systems (user, application, OS, device, network) considered in describing end-to-end QoS, thus allowing hiding of levels of complexity, while at the same time providing support for flow management, QoS management, and protocol support. Support for monitoring QoS parameters applicable to the level is given, and these are then intended to be mapped down to the parameters relevant at lower levels i.e. user-level concerns such as picture quality are mapped to bandwidth and delay concerns at lower levels.

The QoS-A framework was first defined in the early 1990s, and has used object techniques as specified in the ODP standard, and provided input to the ISO Quality of Service framework. There is an implementation based on this model, including "vertical issues" such as management and group support, and "horizontal issues" such as QoS at the various levels of abstraction (similar to a communications stack). Their work has encompassed operating system support (Chorus), networking including Internet and ATM, and work on filters of various types, and has been applied to various practical projects – LINK, ADAPT and BT-URI Management of Multi-service networks.

### 2.6.2 X-bind

The Comet group at the university of Columbia have developed xbind, which is a substantial (~30,000 lines of code) network centric QoS management system. It is now in use at various academic and industrial research facilities, and has been ported to several switch and computing platforms. It's focus is on supporting services within an ATM, IP and telecommunications networks, by providing standardised CORBA based QoS controlled APIs through a "broadband kernel" [LAZAR 97]. The kernel controls resource allocation, and provides a simplifying abstraction of the provision of these quality managed services. This work is particularly interesting due to it's grounding in telecommunications rather than pure data networks, and the consequent treatment of the topics of cost, and maximisation of resource utilisation. A comprehensive description is available at: http://www.ctr.columbia.edu/comet/xbind/.

### 2.6.3 Tenet Real Time Protocol Suite

The Tenet protocol suite is designed to provide guaranteed quality (deterministically and probabilistically) for real-time stream data in heterogeneous networks at the network and transport layers. [BANERJEA 96] describes the original protocol suite development on which started as early as 1987. A new version now exists building on these foundations. http://tenet.berkeley.edu/ provides links to full information about the current suite. [FERRARI 97] and [KNIGHTLY 97] also describe specific aspects of the work being undertaken, much of which focuses on the issues of delay, jitter and throughput management as these are of particular relevance to a low level investigation such as this.

### 2.6.4 Sumo

A description of a platform based on the Chorus environment designed for multimedia applications, called Sumo-CORE, and a CORBA based Object Request Broker (ORB) designed to support telecommunications, multimedia, and other real time applications, called Sumo-ORB is given in [BLAIR 97b]. The object modelling approach supports both static and dynamic QoS management functions, with appropriately expressive object bindings including QL as annotations to the object model. As the modelling language was developed particularly for multimedia applications, QoS is inherent in it's specifications, although it has limits as discussed for QL (above).

Sumo-CORE places particular emphasis on providing end-to-end bindings and support for QoS management in heterogeneous environments. In particular QoS management is provided for thread scheduling, resource reservation, particularly communications and buffer management, and monitoring, policing maintenance and renegotiation functions are applied to all bindings between components. It was developed at Lancaster University, UK.

Sumo-ORB expands upon the CORBA approach, while retaining compliance with existing CORBA implementations, by providing support for discrete (operational and signal), and stream interactions, reactive objects, a recursive binding model to allow complex compositions and abstractions to be built, and explicit QoS management on object bindings. The architecture is also designed to provide greater openness to internal services, and C++ and Esterel are currently supported languages for bindings. In was developed by CNET.

Sumo-ORB currently runs over an FDDI network, allowing deterministic guarantees on underlying network services, which, for many real world applications, is a rather optimistic generalisation, otherwise Sumo-CORE provides good basic QoS support, and SUMO-ORB defines only static functions.

### 2.6.5 QoS Broker

[NAHRSTEDT 95b] describes an approach using a QoS broker which mediates between the application and the OS and network, rather than providing explicit QoS support in the low level APIs. This approach provides a standard low level interface to all applications for managing QoS, while relying on underlying support for functions such as network resource reservation. The advantage of this organisation is that the application can provide it's own translation into system level QoS specifications, and negotiation of resources to be reserved at that level, while being shielded from the low level QoS management of network and OS resources. This paper describes the negotiation of acceptable QoS in an interesting way, which may find application in agent based systems, although the dynamic capabilities described are not as complete as some more recent work.

### 2.6.6 QuAL and QoSockets

[FLORISSI 95, FLORISSI 99] describe work undertaken at Columbia University (USA) in QoS specification and monitoring through a Quality Assurance Language (QuAL) and an extension to the sockets mechanism, QoSockets which, using QuAL, enable QoS specification, negotiation, and monitoring to provide management and statistics on delivered QoS. QuAL provides for network and application level QoS constraints to be specified. These may then be compiled into run-time QoS monitoring components, which can raise exceptions on QoS violations, and generate statistics on achieved performance and violations.

The sockets described are available on Solaris and Linux, and support a range of network protocols, including RSVP, ST-II and ATM. The API provided gives a single abstraction of the various network's QoS specification and negotiation facilities, and provides a best-effort negotiation of QoS given the available resources and protocols. The QoS management will select the most appropriate transport mechanism at runtime, from those supported on an end-to-end basis. Where QoS requirements are not strict lightweight protocols such as UDP and TCP may be selected over those with a greater overhead. The use of SNMP managers to monitor provided QoS is described, and it is noted that applications must still be designed to tolerate violations in QoS, as the underlying QoS support may be limited.

# 3. Overview of Mobile Computing Systems

In this section, we describe some of the most significant characteristics of mobile systems, drawing mainly from [KATZ 94, IMIELINSKI 94, DAVIES 96a]. The treatment of how these characteristics affects QoS is covered in the next section, although this section focuses on those attributes of mobile systems which are most relevant to QoS and distributed systems, as described in the previous section. The areas described are: i) The types of mobile application, mobile host and identification of two categories of mobility; ii) The effects of changing location and the management of location information; iii) The effects of mobility on link characteristics; iv) The impact of portability on hardware design, and the restrictions this then places on application design. The reader looking for a comprehensive set of references to work on mobile communications and computing may find Agrawal, Sreenan and Srivastava's bibliography and web resource pages [AGRAWAL 98] informative.

## 3.1 Categories of Mobility

There are many aspects to mobility, and the vital characteristics one considers important depend on the viewpoint being taken, however we suggest the following computing system oriented characterisation of mobile systems based on computing device, network support for mobility and application requirements, which draws some of it's underlying categorisation from [DAVIES 96a].

Firstly we look at the computing devices available for use. Mobile hosts may have very limited computing, storage and user interface facilities such as palmtop computers or personal digital assistants (PDAs), and devices integrated into cars etc. They are likely to be less powerful than typical desk-top workstations, and are not capable of running network services other than to support local users. More powerful laptops (also called notebook computers), could provide a remotely accessible service, such as a database. These are in essence scaled down versions of desktop workstations, providing similar or moderately limited computing, storage and user interface characteristics.

Next, we consider two categories of mobility supported by the available communications infrastructure. **Mobile** systems allow full, transparent mobility during use, by means of wireless communications methods. In **nomadic** systems, mobility is not transparent, requiring a new connection to be explicitly established by the user after relocation of the host [KATZ 94]. They are typically based on wired dial-up, or local area network communication facilities. For example, a nomadic user may carry a lap-top and connect to a network at various times from their home, office and various remote sites such as client's offices or hotels. Whilst travelling between these locations the lap-top is disconnected from the network. The user may then make large geographical movements between connections, and connections over equipment with widely ranging capabilities, but during connection will exhibit relatively static characteristics. In contrast a mobile user may be connecting to a network using their mobile phone whilst travelling in a train. During the course of a connection the radio reception experienced is likely to have varied widely, and the physical location of the device may be hundreds of miles from it's 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. In practice, most fully mobile communications systems, based on cellular telecommunications, are compositions of local communications systems (cells), with protocols for “hand-over” of communication between cells when mobility requires re-connection to a more appropriate cell. It is generally the case, with current technology, that this hand-over is not seamless, but occurs within time limits expected by the target application of the network, i.e. speech for mobile telephony. This problem is discussed further below. It can also be expected that mobile systems are also nomadic i.e. their position can be expected to change whilst disconnected as well as during connection.

## 3.2 Location Dependant Services

The property of mobility in itself affects the design of systems. We shall consider below the effects of the two types of mobility described above: nomadic systems, and those which are mobile during a connection.

Any form of remote access increases security risks but wireless based communication is particularly susceptible to undetected monitoring so mobility complicates some traditional security mechanisms. In addition there are legal and ethical issues raised in the monitoring a user’s location, but they will not be dealt with in paper, as they are both complex areas and application and jurisdiction dependant.

### 3.2.1 Relocation for Nomadic Users

As described in [KATZ 94, IMIELINSKI 94], a new style of computing is emerging, where mobile hosts will have intermittent connection to other hosts, to exchange data, and due to their mobility will often make connection from different locations, and using different types of host. In some cases this may be limited to a small set of locations, such as home and office. In other cases, the user may routinely connect from points never previously visited, possibly on a global range e.g. hotel rooms when travelling. This then presents the problem of making efficient provision of resources commonly required by the user.

In a fixed system, it would be normal to arrange each user’s network connection such that the cost of connection to their data (e-mail server, home file-system, databases etc.) is minimised by locating their data on server(s) local to their connection point. However a nomadic system may not have access to local servers, and anyway the user’s resources will not be on the local servers. If the nomadic user does have access to local servers, then it might be feasible to migrate resources from the home servers to the local ones.

Imielinski proposes that a mobile user should be registered with a “home location server” (c.f. “home location register” in cellular telephony) which provides a link to the required resources, and the connection between a user and their home location server is known throughout the system, as for fixed networks. In addition, it is proposed that a user may register as a visitor with local servers, which together with the home server then allow resources to be managed to facilitate access from the user’s current location. A similar scheme using proxy servers and Mobile IP [RFC 2002] for public access is suggested by [ZENEL 95].

This approach then introduces various system management trade-offs:

- Whether mobile hosts should always inform other hosts of their movements?
- Whether a search for a host should be global, or only at predefined locations?

- Whether the most efficient access method is by remote connection by the user to their home location, or by migrating or replicating their data to their current visiting location?
- What data filtering should be performed while at a given location?

Much work has been performed on distributed directory management, which is a fundamental requirement to locate users or their home location registers, and standards such as X.500 provide a suitable basis for this. Many of the issues affecting implementation of these possibilities are dependant on the use of the system. e.g. in some applications it is common that interactions are mainly between a remote user and resources which are at the user's "home" location. In general the costs associated with the types of access, and their effect on the location management policy may be best determined statistically. [ZENEL 95] extends the concept of filtering to include omission or delaying of data, and modification of protocols to suit the network characteristics.

Imielinski notes that in distributed systems, the cost of location management may become significant, to the point of overwhelming the actual data exchanged, as the size of the total system, and frequency of location-changes increases. The scalability (e.g. to Internet proportions) of a scheme is important for any general purpose implementation. This issue of scale, both in numbers of hosts and users, and geographical scale also affects the issues described above.

- The placement and mobility of resources may be affected – the cost of resource movement, and maintaining replicated or cached data rises with the frequency of location changes
- There is the trade-off between ease of locating resources and cost of maintaining the information as it changes.
- In general the less informed the sender is, the greater the cost of resolving the location of the destination.
- A suitable solution may be to maintain partial or approximate location information.
- Large-scale systems are more likely to have to cater for heterogeneity of both end-systems, and the environment experienced by users and their applications.

### 3.2.2 Relocation for Mobile Users

Connections with mobile hosts which move during a session are similar to the problems experienced within the cellular telephony field. We address the problems in terms of link quality arising from this mobility below. The primary problem of on-line relocation is the management of hand-over of connection between base stations as the host moves between communications cells. This can be considered an extension of the problems associated with nomadic systems, but with much more frequent connection re-establishment.

Hand-off is a standard feature and well understood for mobile telephony. The provision of wide ranging communications is achieved by splitting a geographic area into cells. The area covered by each base station is determined by geographical features such as whether it is a built-up area, limitations of the base-station's range, and the it's capacity in handling connections relative to the demand for connections in that area. Each cell's communications are managed by a base-station: a transmitter/receiver operating with the appropriate communication protocols to manage multiple connections from hosts within it's cell. Each cell's covered area will overlap somewhat with it's neighbours, and as a mobile hosts moves towards the edge of a base-station's coverage, that base-station arranges to hand the connection over to another base station which is better placed to handle the connection. The detail and practicalities of achieving this handover are beyond the scope of this paper. The current state of practical implementations is such that hand-over is generally achieved in a time which does not cause such a loss of information that the sense of spoken conversation is lost, however, where data intensive communications are concerned the time taken for hand-over may cause considerable loss. This is partly an effect of the comparative volume of the markets for mobile voice communications compared to that for mobile data communications, at this time. A final problem affecting hand-off, which may be resolved by network planning, is that of selecting a suitable base-station to which it can hand-over, which has sufficient spare capacity to support the connection [KATZ 94].

In order to support mobile users, it is necessary to move the connection to remote resources or to move the resources to local hosts while they are being used. However moving resources is not very common, particularly where they are providing stream based data. [BALACHANDRAN 97] describes an implementation where some computing facilities acting on data (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. [SRIVASTAVA 97] 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.

Mobile users result in a much higher rate of location change, and a reduction in the acceptable time taken to manage that change compared to nomadic users. In addition, a wireless cell network is likely to result in a much finer granularity of location change than nomadic use of fixed systems, again raising the complexity of managing location information.

### 3.2.3 Location Aware Services

Location aware services can be viewed as a generalisation of the techniques applied in migrating or selecting resources dependant on current location, as described above. Location awareness is a subset of context awareness which is crucial to effective mobile computing. [SCHILIT 94] provides a good introduction to this topic. To provide for effective provision of services, e.g. selection of servers, connection methods, protocols etc. the location of a system in the network is needed. This computer-centric approach can be broadened however: user level information services may be altered to provide the most appropriate information for the geographical or temporal location of the user. E.g. A telephone directory may provide the number for a person depending on where they were last seen by the system, a user requiring a street map may automatically be provided with one for the location of the system, detected using sources such as GPS, cell-phone cell identity, a diary detailing where the user expected to be, etc. Examples of work in this area are [LEONHARDT 98, YE 98].

## 3.3 Link Quality

The problem of link quality is particularly significant in mobile systems. Cellular telephones often experience poor quality sound and loss of connection when used in cars or trains. Variations in link quality may also apply to nomadic users. While link quality in mobile connections may improve with advances in technology [KAO 98], to some extent the problems are intrinsic in radio based technology, and must be worked with. The key considerations of bandwidth, range and cost are summarised for some popular communications technologies in Table 4.

**Table 4 Common Communication Systems**

| Comms. System | Typical Bandwidth | Range | Costs |
|---|---|---|---|
| Ethernet LAN | 10 – 100 Mb/s | Fixed, wired network. Usually for commercial locations. | Installation |
| Wireless LAN | 1 – 10 Mb/s | 100 – 500 m from base station | Installation |
| UMTS, GPRS | 144 Kb/s – 2 Mb/s | Aiming for Europe-wide coverage | Experimental |
| Infra-Red | 19.2 Kb/s – 1 Mb/s | Within room. | Installation |
| ISDN, ADSL etc. | 128 Kb/s – 1.5 Mb/s | Fixed, wired network. Usually for residential and small business locations. | A monthly fee, plus costs for usage, generally per second or minute.<br>Moderate Cost |
| Modem via Dial-up Telephone | 9.6 Kb/s – 56 Kb/s | Fixed, wired network<br>Available globally. | A monthly fee, and / or costs for usage, generally per second or minute.<br>Low Cost |
| DECT | 32 Kb/s | Cellular phone networks approaching national coverage. Some standards differences. | A monthly fee, plus costs for usage, generally per second or minute.<br>High Cost |
| CDPD | 19.2 kb/s | | |
| GSM | 9.6 Kb/s | | |

New wireless technologies with higher bandwidth may emerge in the near future, and this trend can be expected to continue, *however it is a reasonable assumption that wireless network technology will continue to provide throughput 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 throughput of a radio channel is generally related to the area of transmission/reception, and thus the degree of movement allowed while remaining connected, e.g. infra red typically has cells of room size, and a network rarely covers more than a building in continuous area, while GSM may cover several square km per cell, and provide continuous service on a continental basis – see Figure 2.

The use of highly data intensive applications is then problematic. For instance, speech quality audio with compression generally requires at least 8Kb/s, and 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.

There can be considerable variation in link quality experienced during a connection. Dial-up and fixed land lines have relatively stable QoS characteristics for a given connection, but radio-based connections may be affected by movement in relation to the environment, such as passing under a bridge or atmospheric conditions. The data rate provided by the Internet depends connection end-points, localised usage factors and global usage levels, which can fluctuate over time-periods of seconds. In the case of radio communications these fluctuations may be long, such as entering a building which blocks reception or short, such as those caused by driving under bridges, having lorries passing between the mobile and the base station, or during hand-over between cells, as described above. Not only do you get variation in duration of the problem, but the data rates available during a session can vary and sometimes drop to zero. Some protocols for wireless connection also affect the connection at a lower level, due to factors such as collisions caused by the multiplexing algorithm, causing the link jitter to vary considerably [SRIVASTAVA 97].

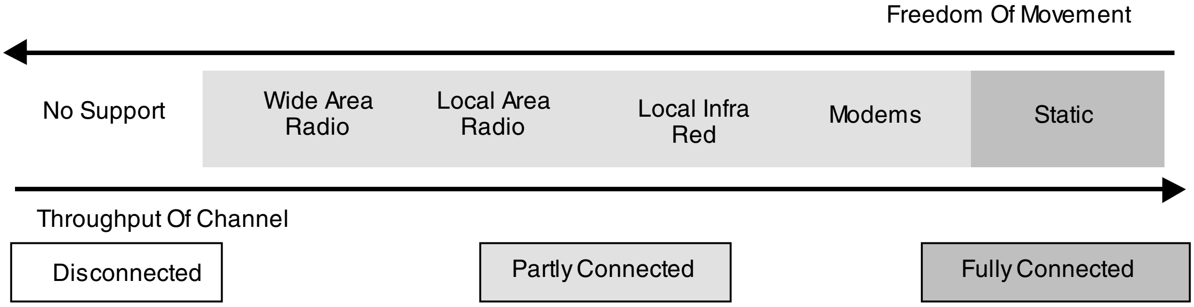

**Figure 2: The Relationship Between Mobility, Connection and Throughput, adapted from [DAVIES 96a]**

The use of highly data intensive applications is then problematic. For instance, speech-quality audio with compression generally requires at least 8Kb/s, and 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.

There can be considerable variation in link quality experienced during a connection. While some methods such as dial up connection are generally relatively stable on any given connection, the use of radio based connection may be affected by movement in relation to the environment such as passing under a bridge or atmospheric conditions, and the data rate provided by the Internet depends on the point of connection, the point connected to, localised usage factors, and the global usage levels, which can fluctuate over time-periods of seconds. In the case of radio communications these fluctuations may be long, such as entering a building which blocks reception or short, such as those caused by driving under bridges, having lorries passing between the mobile host and the base station, or during hand-over between cells, as described above. Not only do you get variation in duration of the problem, but the data rates available during a session can vary and sometimes drop to zero. Some protocols for wireless connection, such as CDMA also affect the connection at a lower level, due to factors such as collisions caused by the multiplexing algorithm, causing the link jitter to vary considerably [SRIVASTAVA 97].

On a more extended time-scale, as Katz suggests, users are likely to use the device most appropriate to their situation, so while one connection may be by a low bandwidth radio link, the following connection by the same user may be from an office workstation, with considerably greater network capacity.

While not always considered a quality consideration, the cost of mobile connections is often also much higher than for fixed systems.

A final effect of the link type is the delay involved in establishing connections. Where a connection is permanent, initial latency may still be low, even where data rate is also low. Where connections require establishing on-demand e.g. dial-up connection, the time to establish a connection at various protocol levels, and pass any authorisation mechanism may add tens of seconds to the connection time. In conjunction with the cost aspect, it is also possible that while virtual connections are maintained by hosts, the actual link may be dropped during periods of inactivity causing unplanned delays [DAVIES 96a]. This is a facility of ISDN modems to reduce costs.

### 3.4 Mobility Design Considerations

There are a number of limitations imposed by portability of the mobile computing device [DAVIES 96a, IMIELINSKI 94, KATZ 94]. The main limitation is in the physical size of mobile computer which results in the following restrictions:

- Restrictions imposed by the limitations of battery power
- Restrictions on the user interface due to size

Since 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 poser consumption in portable PCs through to Personal Digital Assistants (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 when in use, particularly for transmission, so network connection is commonly extremely intermittent, as described above.

The second point is that of user interfaces:mlarge 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. In PDAs screens are commonly very small, monochrome and low resolution, often tending towards text based, and have minimal miniature keyboards, pen based, voice or very simple cursor and selection, input 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.

Whilst 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.

## 3.5 Summary

The key requirements due to mobility can then be summarised as:

- Efficient location of both users and the resources they require. How to maintain sessions and continue routing information to mobile users as they move.
- Ability to manage large variations in link quality between and during connections, at various rates.
- The costs and technical difficulties involved in the various choices available to implement mobile connections.
- Ability to manage massive variation in capabilities, both in computing power and user interfaces, of mobile computers.

# 4. Quality of Service in Mobile Computing Systems

We shall now summarise the problems of mobility described above (link quality, movement of users and resources, restrictions of devices, security and cost) in direct relation to QoS, and then describe some of the ideas and solutions being developed specifically to manage QoS in a mobile environment, as reported in the literature.

## 4.1 The Impact of Mobility on QoS

Much of the mainstream work on QoS and multimedia concentrates on fixed networks and does not address the additional complications imposed by mobility. However mobility is an area of growing research in the last few years. Early on it was recognised that the special requirements of mobile systems, discussed above, would have great influence on their ability to support QoS, and that a likely solution was in providing the ability to adapt to the changes mobility brings about, rather than trying to provide hard guarantees of QoS [KATZ 94, SRIVASTAVA 97, BLAIR 97a, CAMPBELL 97a].

### 4.1.1 The Effects of Link Quality on QoS

The essential effect of mobile systems is that link quality becomes extremely variable, often in a random manner, although some parts of this effect can be predicted, or statistically modelled. The main QoS parameters – reliability of connection, bandwidth, latency and jitter are all affected by mobility, both during and between connections. In the first case the effect is generally due to environmental influences on the communications medium. The changes between connections also cover the more stable orders of change brought about by changes in the type of communications medium and type of end-system, where there will be periods of high quality connection (over fixed networks), and periods of poorer quality connection over dial-up or radio links. Much of the current literature focuses on provision within the areas of poorer quality connection, developing low bandwidth protocols. [KAO 98] describes one emerging high bandwidth wireless infrastructure, however, it is the *variation* in QoS which is the crucial factor and mobile communications is likely to remain limited in bandwidth in comparison to fixed networks. The distinction between nomadic and mobile systems is not hard-and-fast so QoS management of periods of disconnection between periods of high quality connections is also important for mobile systems.

QoS management can be applied to manage the effects of these changes. However, in contrast to the fixed network situation, the methods used to achieve delivery of quality may be significantly more drastic, and initiated more frequently. It is suggested that due to the nature of mobile communications it is not practical to guarantee QoS levels. Instead, an adaptive approach is needed, where QoS management specifies a range of acceptable results. The QoS management is responsible for co-operation with **QoS aware applications** to support adaptation, rather than insulating applications from variation in underlying QoS.

Further impacts which must be considered when designing solutions in mobile distributed computing are [IMIELENSKI 94, DAVIES 96a, SRIVASTAVA 97]:

- That connections will generally be intermittent, either as brief environmental or hand-over effects, or as longer-term disconnections due to end-system limitations.
- Increased latency in connections, and particularly in establishing connections.

- Connection cost.
- That connection bandwidth will be widely ranging.

These effects require then that algorithms employed must be capable of managing frequent entry and loss of nodes visible in the network, and that overhead should be minimised 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.

### 4.1.2 The Effects of Movement on QoS

The very act of movement has two direct results which impact QoS: matching resources to requirements in an efficient manner is complicated by the movement of the end-system requiring the resource; and secondly, in the fully mobile case, the migration of resource provision and associated QoS specifications is very difficult. These problems include those of managing dynamic replication e.g. of user's personal data; selecting generic resources for convenience, e.g. printers or mirrored data repositories; and directory management. However, once location information is being provided, it may become easier to manage QoS based on known information about historical network reliability and bandwidth in the local environment. This may be viewed as an extension to context aware computing. The problem of handover is similar to that in mobile telephony, with the addition of considerations regarding relocating processing and data in addition to connections. [LEVINE 97] describes a system for QoS driven resource estimation and reservation to support hand-off in wireless networks supporting mobile clients. His 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 can be sometimes 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 hand-off 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.

### 4.1.3 The Restrictions of Portable Devices on QoS

The effect of mobile devices on QoS is similar to that of link quality, in that it is characterised by extreme restriction, and widely varying heterogeneity, where different end-systems may be used. In general restricted computing power, restricted user interfaces, and intermittent availability due to the limitations of batteries and mobile communication systems characterise mobile devices. QoS management in a mobile environment should then 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 [Davies 96 a & b]. 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.

### 4.1.4 The Effects On Other Non-Functional Parameters

Radio communication used for mobile systems is more susceptible to monitoring than landlines and so increases security risks. Even nomadic systems will make use of less secure telephone and internet based communications than office systems using LANs. Some organisations may place restrictions on what data or serviced can be accessed remotely by authorised users. In many environments mobile systems thus require more sophisticated security techniques such as encryption, authentication and access control than is needed for office systems. However security is a sufficiently large topic that 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 costly to set up or connect to, 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.

## 4.2 Current Work On Management of QoS in Mobile Environments

### 4.2.1 Adaptivity and Measured Change

As stated above, one of the key concepts in applying QoS to mobile environments is that of change, and adaptation to change. These changes manifest themselves in various ways, we describe three classes below, although others such as [DELGROSSI 94] approach this issue with regard to transparent and non-transparent scaling of media:

- Large-grained change
- Fine-grained change
- Hideable change

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

*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 fluctuations can only reasonably be managed in conjunction with applications. Adaptation to this type of change has been a focus of interest recently [BLAIR 97a, DAVIES 96c, NOBLE 97, SRIVASTAVA 97, SREENAN 96, ZHANG 97]. It should be noted that depending on the application and the QoS management techniques employed the boundary between fine-grained, large-grained and hideable change may vary.

*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 filtering, and buffering techniques. Techniques applicable here include in-network filtering of packets by differentiating between those containing base and enhancement levels of information in multimedia streams, or by buffering to remove jitter by smoothing a variable (bit or frame) rate stream to a constant rate stream. However, while the filtering techniques are generally similar to those in fixed network systems [KNIGHTLY 97], new problems arise where these filters are deployed in the network. Filters may be deployed in the network to allow different scaling to be applied after splitting of streams towards destinations of different end-to-end capability in a multicast environment, without duplicating the stream over the whole path. Where the end-systems are mobile, network connections are formed and torn down during sessions as the end-system moves and makes connections with different base stations. It is possible then that where in-network filtering is taking place, filters must also be set up during a session, as well as moving connections. Further, in the general case, the new connection may not provide the same QoS as the previous one, and so the filtration applied may differ – this then being one cause of fine-grained change. To manage this then requires an extension of the traditional interactions for migrating connections between base stations. The selection and hand-over of control must take account of available and required QoS, and the capacity of the network to accommodate any required filters. Ideally much of this information would be present in, or could be added to the polling messages used already. Where the network cannot maintain the current level of service, base stations should initiate adaptation in conjunction with hand-off [BALACHANDRAN 97, CAMPBELL 97a].

Fine-grained change on the other hand, differs in that it is expected that changes in service provided will be either notified to or negotiated with the applications concerned. Typical causes of fine-grained change are:

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

It should be noted that these effects may be seen as failure, or transient loss, of parts of the system. The effects 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 [DAVIES 96c]. 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 (as previously described). However, the highly transient effects due to mobility must be considered. Geographical and hand-off effects may be seen as failures, but may only last seconds, so management systems which use a probabilistic or stochastic model of QoS requirement specification for given levels will be well placed 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 [NOBLE 97]. A user-level QoS parameter can be included to describe the interplay and trade-off between stable presentation and agile adaptation. [LU 96 & 97] suggests that where movement causes frequent fluctuations in service the maintenance of QoS at a steady level, thus providing seamless operation, is preferred by users, whilst users whose systems experience less frequent fluctuations would tend to prefer that the QoS provided is maximised, 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. Enhancements can be selected based on the difference between capacity used by basic-level services and the capacity available in a resource, priority or deadline schemes. It is common, in much of the literature, to concentrate on adaptation due to last-hop effects, as this fits the model of a wireless-enabled user terminal. 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 during the day due to telecoms providers' tariff structures.

### 4.2.2 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. This is in many ways reasonable, however, some resource allocation and admission control would seem prudent when resources are scarce, even if hard guarantees of resource provision are not practical. [LU 96 & 97] proposes that guarantees be made in admission control on lower bounds of requirements, whilst 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 hand-off, 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.

Resource management is covered below under the heading of context awareness, as awareness of available resources is fundamental to managing them in a heterogeneous system.

### 4.2.3 Context Awareness

A further aspect of resource management is that of large-grained adaptivity, and context awareness. [TURNER 98] 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 behaviour". 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 [ZENEL 95]. A fundamental paper on context awareness is [SCHILIT 94], 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. However, it should be noted that that Schilit's definitions also cover 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.

[DAVIES 96c] proposes that protocol management should analyse connections, and adapt to make best use of the available resources. [ZHAO 98] describes the use of Mobile IP [RFC 2002] 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.[BLAIR 97a] 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.

[GECSEI 97] describes a model of adaptivity within the available resources of a system at that time. 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. The mapping of a region of acceptable performance onto a region of the resource space is described, where that region in the resource space is then the adaptation space which may be moved within whilst maintaining the specified performance. Without adaptation the variation in the available resource space which may be accommodated whilst maintaining performance within the acceptance area is much smaller.

[LEVINE 97] describes a resource reservation system for mobile wireless networks, which has the capacity to make reservations in advance of hand-off using knowledge of the environment. For instance, base stations along a major road are likely to be used in sequence, and once one hand-off has taken place, reasonable prediction of subsequent cells to be moved into can be made with knowledge of the road layout. It is also suggested that were this to be integrated with in-car navigation systems, even more accurate predictions of resource usage could be made, particularly where many known paths may be taken, such as in urban areas. [LU 97] describes a similar system where each end of a flow is characterised as static or mobile, and advance reservations are made for mobile flows on the predicated next cell.

[TITMUSS 97] proposes a scheme for implementing services in a telephony environment, using agents, where an agent manages a resource, and a hierarchy of agents is formed for scalability and abstraction. Requests for service are passed to the agent best able to offer the service at an appropriate level of quality (both in results given and cost). The agent can choose the resources most relevant to the user, based on context e.g. geographic location or particular end-system. The use of agent hierarchies also allows tolerance of failures, and avoids the format compatibility and overhead problems associated with the WWW and

MIME e-mail approaches. An example application is given where documents are displayed in full on workstations, with any necessary conversion between formats, as text-only on PDAs, and a spoken summary given to telephone users, with additional consideration given to the ability and cost of the actions, including location information.

### 4.2.4 Use Of Standards

Many of the standards described earlier (e.g. RM-ODP, CORBA), and following from these many of the fixed approaches to QoS, are directed at maintaining transparency of platform, and hiding complexity from applications. [DAVIES 96a] contends that in an adaptive, mobile environment this approach is no longer relevant. However, some implementations e.g. [BALACHANDRAN 97, CAMPBELL 97a] are based on “standard” environments such as CORBA, or software components previously developed for fixed network QoS architectures [DAVIES 96c]. These systems provide adaptive connections using existing components, while retaining the benefits of known interfaces, and re-use of low level protocol implementations. [DAVIES 97] surveys mobile distributed systems platforms, including a variation on 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.

## 5. Conclusions

We summarise 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 in our opinion, 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 E.g. the same Java applications may be run on heterogeneous platforms, and use context awareness to adapt their operation [BLAIR 97a, KATZ 94, SCHILIT 94, TITMUSS 97, ZENEL 95].
- Context aware adaptation of protocols, with regard to overhead, and degree of synchrony depending on degree of connectivity [DAVIES 97, SCHILIT 94].
- Provision for the definition of adaptation paths from user-level QoS parameters, including trade-offs, using probabilistic or stochastic specification of parameters. Trade-off should take account of meta-data 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 [KATZ 94, FRY 97, NOBLE 97].
- 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, meta-data and experience in a context. Additionally the use of probabilistic and stochastic resource models to enable task allocation and resource reservation with fault tolerance [BILLOT 96, FERRARI 97, LEVINE 97].
- Filtering to include delay or rejection of data, as well as scaling. Selection of filters should be aided by meta-data, and available resources. Filters should act as "plug-in" modules on QoS aware components of the system [BALACHANDRAN 97, CAMPBELL 97a, KNIGHTLY 97, ZENEL 95].
- Control of in-service mobility, and migration of resources which is a mobile network oriented problem [CAMPBELL 97a, LEVINE 97].

Associated with these requirements are the following enabling functions, which are of general interest beyond QoS provision:

- Meta-data stored with resources such Web pages to enable accurate QoS aware decisions in advance of actual measured impact of actions.
- Context maps of resources, with resource models for QoS aware resource selection [SCHILIT 94, ZENEL 95].
- Performance monitoring as input to context models to permit adaptation by QoS management [KATZ 94, SCHILIT 94].
- Models of movement to enable intelligent caching, replication and migration of data for nomadic use [BLAIR 97a, DAVIES 97, LEVINE 97].
- QoS management incorporated into systems such that non-QoS-aware applications may continue to function.

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 interacting 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.

## 6. Acknowledgements

We gratefully acknowledge financial support from the EPSRC (Grant GR/L 06010) and British Telecom.

## 7. References

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

[AURRECOECHEA 98] Aurrecoechea, C., Campbell, A.T., Hauw, L.: A Survey of QoS Architectures *ACM Multimedia Systems Journal – Special Issue on QoS Architeure May1998,* Springer-Verlag,

[BALACHANDRAN 97] Balachandran, A., Campbell, A.T., Kounavis, M.E.: Active Filters: Delivering Scaled Media to Mobile Devices *Proceedings of the IEEE 7th International Workshop on Network and Operating Systems Support for Digital Audio and Video* p125-34, 1997

[BANERJEA 96] Banerjea, A., Ferrari, D., Mah, B.A., Moran, M., Verma, D. C., Zhang, H.: The Tenet Real-Time Protocol Suite: Design, Implementation, and Experiences *IEEE/ACM Transactions on Networking* vol4 no1 February 1996 p1-10

[BECKER 98] Becker, C., Geihs, K.: Quality of Service – Aspects of Distributed Programs *International Workshop on Aspect-Oriented Programming at ICSE'98, Kyoto/Japan, 1998*

[BILLOT 96] Billot, M., Issarny, V., Puaut, I., Banatre, M.: A proposal for Ensuring High Availability of Distributed Multimedia Applications *Proceedings of 15th Symposium on Reliable Distributed Systems* p220-7 (IEEE, 1996)

[BLAIR 97a] Blair, G.S., Davies, N., Friday, A., Wade, S.P.: Quality of service support in a mobile environment: an approach based on tuple spaces *Proceedings of the 5th IFIP International Workshop on Quality of Service 1997* from http://www.comp.lancs.ac.uk/computing/research/mpg/index.html

[BLAIR 97b] Blair, G., Stefani, J-B.: *Open Distributed Processing and Multimedia* (Addison-Wesley, 1997)

[BOCHMANN 97] Bochmann, G. v., Hafid, A.: Some principles for quality of service management *Distrib. Syst. Engng* vol4 p16-27 (IOP Publishing, 1997)

[BOUCH 99] Bouch, A., Sasse, M.A.: It ain't what you charge, it's the way that you do it: A user perspective of network QoS and pricing *to be presented at the IFIP/IEEE International Symposium on Integrated Network Management (*IM'99) Boston, 24-28 May 1999

[CAMPBELL 97a] Campbell, A.T.: Mobiware: QoS-aware middleware for mobile multimedia communications from http://comet.columbia.edu/~campbell/andrew/publications/publications.html

[CAMPBELL 97b] Campbell, A., Coulson, G.: A QoS adaptive multimedia transport system: design, implementation and experiences media *Distrib. Syst. Engng* vol 4 p48-58 (IOP Publishing, 1997)

[CERI 85] Ceri, S., Pelagatti, G.: *Distributed Databases Principles and Systems* (McGraw Hill, 1985)

[DAVIES 96a] Davies, N.: The impact of mobility on distributed systems platforms *Proceedings of the IFIP / IEEE International Conference on Distributed Platforms, Dresden, 1996* p18-25 (Chapman & Hall, 1996)

[DAVIES 96b] Davies, N., Blair, G.S., Cheverst, K., Friday, A.: Supporting collaborative applications in a heterogeneous mobile environment *Computer Communications Special Issue on Mobile Computing* p346-58 (Elsevier, 1996)

[DAVIES 96c] Davies, N., Friday, A., Blair, G.S., Cheverst, K.: 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)

[DAVIES 97] Davies, N., Wade, S.P., Friday, A., Blair, G.S.: Limbo: a tuple space based platform for adaptive mobile applications *Proceedings of the International Conference on Open Distributed Processing / Distributed Platforms 1997* from http://www.comp.lancs.ac.uk/computing/research/mpg/index.html

[DEGERMARK 95] Degermark, M., Köhler, T., Pink, S., Schelén, O.: Advance Reservations for Predictive Service *Proceedings of Network and Operating Systems Support for Digital Audio and Video 1995* (Springer-Verlag, 1995)

[DEGERMARK 97] Degermark, M., ., Köhler, T., Pink, S., Schelén, O.: Advance Reservations for Predictive Service in the Internet *Multimedia Systems* vol5 p177-186 (Springer-Verlag 1997)

[DELGROSSI 93] Delgrossi, L., Herrtwich, R.G., Vogt, C., Wolf, L.C.: Reservation Protocols for Internetworks: A Comparison of ST-II and RSVP *Proceedings of Network and Operating System Support for Digital Audio and Video, Sept 1993* p195-203 (Springer-Verlag, 1994)

[DELGROSSI 94] Delgrossi, L., Halstrick, C., Hehmann, D., Guido, R., Krone, O., Sandvoss, J., Vogt, C.: Media scaling in a multimedia communication system *Multimedia Systems* vol2 p172-180 (Springer-Verlag, 1994)

[FERRARI 97] Ferrari, D., Gupta, A., Giorgio, V.: Distributed advance reservation of real-time connections *Multimedia Systems* vol5 (1997) p187-98 (Springer-Verlag, 1997)

[FLORISSI 95] Florissi, P.G.S.: QuAL: Quality Assurance Language, PhD Thesis, Computer Science Department, Columbia University, New York, 1995.

[FLORISSI 99] Florissi, P.G.S., Yemini, Y., Florissi, D.: QoSockets: A New Extension to the Sockets API for End-to-End Application QoS Management, IEEE/IFIP Integrated Management Symposium (IM'99), Boston, May 1999.

[FRANKEN 97] Franken, L.J.N., Haverkort, B.R.H.M.: Quality of service management using generic modelling and monitoring techniques *Distrib. Syst. Engng* vol4 p28-37 (IOP Publishing, 1997)

[FRØLUND 98] Frølund, S., Koistinen, J.: QML: A Language for Quality of Service Specification *HP Research Report HPL-98-10* (Hewlett Packard, 1998)

[FRY 97] Fry, M., Seneviratne, A., Vogel, A., Witana, V.: QoS management in a World Wide Web environment which supports continuous media *Distrib. Syst. Engng* vol 4 p38-47 (IOP Publishing, 1997)

[GECSEI 97] Gecsei, J.: Adaptation in Distributed Multimedia Systems *IEEE Multimedia* April-June 1997 (IEEE, 1997)

[HALSALL 92] Halsall, F.: Data Communications, Computer Networks and Open Systems 3rd ed. (Addison-Wesley 1992)

[HUTCHISON 94] Hutchison, D., Coulson, G., Campbell, A., Blair, G.S.: Quality of Service Management in Distributed Systems *in Network and Distributed Systems Management* ed. Sloman, M. p273-302 (Addison-Wesley, 1994)

[HUTCHISON 97] Hutchison, D., Mauthe, A., Yeadon, N.: Quality of service architecture: Monitoring and control of multimedia communications *Electronics and Communications Engineering Journal* vol9 no3 p100-6 (IEE, 1997)

[IMIELINSKI 94] Imielinski, T., Badrinath, B.R.: Mobile Wireless Computing *Communications of the ACM* vol 37, no 10 p18-28 (ACM Press, 1994)

[KAO 98] Kao, S.: Speedy Wireless Networks *Byte International Supplement* vol 23 no 3 (March 1998) p40 IS15-17 (McGraw-Hill 1998)

[KATZ 94] Katz, R.H.: Adaptation and Mobility in Wireless Information Systems *IEEE Personal Communications* vol 1, no 1 p6-17 (IEEE, 1994)

[KNIGHTLY 97] Knightly, E.W., Rossaro, P.: On the effects of smoothing for deterministic QoS *Distrib. Syst. Engng* vol 4 p3-15 (IOP Publishing, 1997)

[KNOCHE 98] Knoche, H., de Meer, H.: Quantitative QoS-Mapping: A Unifying Approach *Proceedings of the 5th IFIP International Workshop on Quality of Service 1997* (Chapman & Hall 1997)

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

[LAZAR 97] Lazar, A.A.: Programming Telecommunications Networks *IEEE Network* September / October 1997 p8-18 (IEEE, 1997)

[LEONHARDT 98] Leonhardt, U.: Supporting Location-Awareness in Open Distributed Systems *PhD Thesis, Department of Computing, Imperial College, London* (Imperial College, 1998)

[LEVINE 97] Levine, D.A., Akyildiz, I.F., Nagshineh, M.: A Resource Estimation and Call Admission Algorithm for Wireless Multimedia Networks Using the Shadow Cluster Concept *IEEE/ACM Transactions on Networking* vol5, no1 p1-12 (IEEE/ACM, 1997)

[LI 98] Li, B., Nahrstedt, K.: A Control Theoretical Model for Quality of Service Adaptations *Proc. IEEE International Workshop on Quality of Service (IWQoS 98), May 1998, Napa, California* (IEEE, 1998)

[LOYALL 98] Loyall, J.P., Schantz, R., Zinky, J.A., Bakken, D.E.: Specifying and Measuring Quality of Service in Distributed Object Systems *Proc. ISORC '98, Kyoto, Japan* (IEEE, 1998)

[LU 96] Lu, S., Bharghavan, V.: Adaptive Resource Management Algorithms for Indoor Mobile Computing Environments *ACM SIGCOMM* p231-242 (ACM Press, 1996)

[LU 97] Lu, S., Lee, K-W., Bharghavan, V.: Adaptive Service in Mobile Computing Environments *Proceedings of the 5th IFIP International Workshop on Quality of Service 1997* (Chapman & Hall, 1997)

[McILHAGGA 98] McIlhagga, M., Light, A., Wakeman, I.: Towards a Design Methodology for Adaptive Applications *Proc. MOBICOM'98 Dallas, Texas, USA* p133-144 (ACM, 1998)

[NAHRSTEDT 95a] Nahrstedt, K., Steinmetz, R.: Resource Management in Networked Multimedia Systems *IEEE Computer* vol28 no5 May 1995 (IEEE, 1995)

[NAHRSTEDT 95b] Nahrstedt, K., Smith, J.M.: The QOS Broker *IEEE Multimedia* vol2 no1 Spring 1995 (IEEE, 1995)

[NOBLE 97] Noble, B.D., Satyanarayanan, M., Narayanan, D., Tilton, J.E., Flinn, J., Walker, K.R.: Agile Application-Aware Adaptation for Mobility *Proceedings of the 16th ACM Symposium on Operating Systems Principles* p276-87 (ACM Press, 1997)

[PFLEEGER 97] Pfleeger, C.P.: Security in Computing 2nd ed. (Prentice-Hall, 1997)

[RFC 2002] IETF Network Working Group: RFC 2002 IP Mobility Support ed. C. Perkins (IETF, 1996)

[RFC 1633] IETF Network Working Group: RFC 1633 Integrated Services in the Internet Architecture: an Overview ed. Braden, R., Clark, D., Shenker, S. (IETF, 1994)

[RFC 2205] IETF Network Working Group: RFC 2205 Resource Reservation Protocol (RSVP) - Version 1 Functional Specification ed. Braden, B, Zhang, L., Berson, S., Herzog, S., Jamin, S. (IETF, 1997)

[RFC 2210] IETF Network Working Group: RFC 2210 The Use of RSVP with IETF Integrated Services ed. Wroclawski, J. (IETF, 1997)

[RFC 2211] IETF Network Working Group: RFC 2211 Specification of the Controlled-Load Network Element Service ed. Wroclawski, J. (IETF, 1997)

[RFC 2212] IETF Network Working Group: RFC 2212 Specification of Guaranteed Quality of Service ed. Shenker, S., Partridge, C., Guerin, R. (IETF, 1997)

[RFC 2215] IETF Network Working Group: RFC 2215 General Characterisation Parameters for Integrated Service Network Elements ed. Shenker, S., Wroclawski, J. (IETF, 1997)

[RFC 2474] IETF Network Working Group: RFC 2474 Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers ed. Nichols, K., Blake, S., Baker, F., Black, D. (IETF, 1998)

[RFC 2475] IETF Network Working Group: RFC 2475 An Architecture for Differentiated Services ed. Blake, S., Black, D., Carlson, M. Davies, E., Wang, Z., Weiss, W. (IETF, 1998)[SCHILIT 94] Schilit, B.N., Adams, N., Want, R.: Context-Aware Computing Applications *Proceedings of the Workshop on Mobile Computing Systems and Applications, Santa Cruz, 1994* (IEEE, 1994)

[SISALEM 97] Sisalem, D.: End-to-End Quality of Service Control Using Adaptive Applications *Proceedings of the 5th IFIP International Workshop on Quality of Service 1997* (Chapman & Hall, 1997)

[SREENAN 96] Sreenan, C.J., Mishra, P.P.: Equus: A QoS Manager for Distributed Applications *Proceedings of the IFIP / IEEE International Conference on Distributed Platforms, Dresden, 1996* p496-509 (Chapman & Hall, 1996)

[SRIVASTAVA 97] Srivastava, M., Mishra, P.P.: On Quality of Service in Mobile Wireless Networks *Proceedings of the IEEE 7th International Workshop on Network and Operating Systems Support for Digital Audio and Video* p147-58 (IEEE, 1997)

[STALLINGS 98] Stallings, W.: Cryptography and Network Security 2nd ed. (Prentice-Hall, 1998)

[STOREY 96] Storey, N.: Safety-Critical Computer Systems (Addison-Wesley, 1996)

[TASSEL 97] Tassel, J., Briscoe, B., Smith, A.: An End to End Price-Based QoS Control Component Using Reflective Java *Proceedings 4th International COST 237 Workshop, Lisboa, Portugal, 1997* p18-32 (Springer 1997)

[TITMUSS 97] Titmuss, R., Crabtree, I.B., Winter, C.S.: Agents, Mobility and Multimedia Information *Software agents and soft computing. Towards enhancing machine intelligence. Concepts and applications* p146-59 (Springer-Verlag, 1997)

[TURNER 98] Turner, R.M.: Context-mediated behaviour for intelligent agents *International Journal of Human-Computer Studies* vol.48, p307-330 (Academic Press Ltd., 1998)

[YE 98] Ye, T., Jacobsen, H.A., Katz, R.: Mobile awareness in a wide area wireless network of info-stations *Proc. MOBICOM'98 Dallas, Texas, USA* p109-120 (ACM, 1998)

[ZENEL 95] Zenel, B., Duchamp, D.: Intelligent Communication Filtering for Limited Bandwidth Environments *Proceedings of the 5th Workshop on Hot Topics in Operating Systems, Rossario, May 1995* p28-34 (IEEE, 1995)

[ZHANG 97] Zhang, H., Knightly, E.W.: RED-VBR: a renegotiation-based approach to support delay-sensitive VBR video *Multimedia Systems* vol 5 p164-176 (Springer-Verlag, 1997)

[ZHAO 98] Zhao, X., Castelluccia, C., Baker, M.: Flexible Network Support for Mobility *Proc. MOBICOM'98 Dallas, Texas, USA* p145-156 (ACM, 1998).