Multimedia Tools and Applications 7, 67–82 (1998)

QOS-aware Middleware for Mobile Multimedia Communications

ANDREW T. CAMPBELL campbell@ctr.columbia.edu
The COMET Group, Center for Telecommunications Research, Columbia University, Room 801 Schapiro Research Building, 530 W, 120th St., New York, NY 10027-6699
http://comet.columbia.edu/~campbell, http://comet.columbia.edu/wireless

Abstract. Next generation wireless communications system will be required to support the seamless delivery of voice, video and data with high quality. Delivering hard *Quality of Service* (*QOS*) assurances in the wireless domain is complex due to large-scale mobility requirements, limited radio resources and fluctuating network conditions. To address this challenge we are developing *mobiware*, a QOS-aware middleware platform that contains the complexity of supporting multimedia applications operating over wireless and mobile networks. Mobiware is a highly programmable software platform based on the latest distributed systems technology (viz. CORBA and Java). It is designed to operate between the application and radio-link layers of next generation wireless and mobile systems. Mobiware provides value-added QOS support by allowing mobile multimedia applications to operate transparently during handoff and periods of persistent QOS fluctuation.

Keywords: middleware, mobile communications, adaptive algorithms, active transport, QOS

1. Introduction

Recent years have witnessed a tremendous growth in the use of wireless communications in business, consumer and military applications. The number of wireless services and subscribers has expanded with systems for mobile analog and digital cellular telephony, radio paging, and cordless telephony becoming widespread. Next generation wireless networks such as *wireless ATM* (*WATM*) will provide enhanced communication services such as high resolution digital video and full multimedia communications.

The main challenge in a combined wireline/wireless ATM networks derives from *complexity*. Complexity is present in various forms in mobile multimedia communications. First, the combination of multi-rate multimedia connections with mobility proves difficult to achieve in practice. A connection with certain capacity reserved at a particular cell may have to be re-routed when the mobile device changes its location. The new path to the desired location may not have the original required capacity. Therefore, re-negotiation of resources allocated to the connection is needed. At the same time, the flow (e.g., audio or video) should be transported and presented ‘seamlessly’ to the destination device with a smooth change of perceptual quality. This motivates the need for QOS with mobility.

Next, audio and video flows are characterized by the production, transmission and consumption of single media streams (both unicast or multicast) with associated QOS. For multicast flows, individual receivers (both wired and wireless) may have differing capability to consume flows [20]. This could be due to either fluctuating network resources with mobility or imposed by individual applications. Bridging this heterogeneity gap [9] in mobile multicast environments [6] while simultaneously meeting the individual mobile devices' QOS requirements is an area of research that remains to be resolved.

Third, radio channel's varying QOS characteristics and device mobility, fundamentally impact our ability to deliver hard QOS guarantees in the WATM environment. QOS controlled mobility and a QOS adaptive transport system share a common link in that they must be able to respond or "adapt" to the changes in the delivered quality due to QOS-fluctuating wireless channel or application-level mobility, respectively.

In this paper, we propose a unique solution to the overall problem of complexity that is based on a methodology of networking programming based on a QOS-aware middleware platform called *mobiware*. We extend an open control methodology that has been previously developed at Columbia [14] to control ATM networks to the mobile ATM domain. The basic tenet of this methodology is to separate the network hardware from software.

The structure of this paper is as follows. Section 2 presents an overview of mobiware and its architecture which provides a framework for network programming and adaptation of flows. Following this we describe a set of adaptive and active algorithms which lay at the heart of mobiware's adaptation strategy. We describe a QOS controlled handoff algorithm in Section 3, an adaptive and active transport algorithm in Section 4 and an adaptive network service in Section 5. Finally, in Section 6 we provide some concluding remarks.

2. Mobiware: Programmable mobile networking

Mobiware is a software middleware platform that runs seamlessly on mobile devices, base stations and mobile-capable ATM switches. The platform is built on distributed system and Java technology and incorporates new architecture and novel adaptive algorithms to support QOS controlled mobility. The goal of the mobiware adaptive algorithms is to transport scalable flows, reduce handoff dropping and improve wireless resource utilization. We use the term "controlled QOS" in this paper to distinguish it from hard QOS guarantees offered by fixed ATM networks. Implicit in the term is the notion that flows can be represented and transported as multilayer scalable flows at mobile devices. Adaptive algorithms help scale flows during handoff based on the available bandwidth and an application-specific *flow adaptation policy* [7]. This policy characterizes each audio and video flow as having a minimum QOS layer and a number of enhancements.

Mobiware provides a highly programmable platform for ease in the service creation, monitoring and adaptation of multimedia flows. The concepts of programmability and adaptability are fundamental when addressing the complexity of supporting mobile multimedia applications over QOS-varying mobile networks. By adaptability we mean as mobiles roam mobiware's adaptive algorithms conspire to scale flows to match available bandwidth at the bottleneck node, e.g., the base station. By programmability, we mean that mobiware's APIs are 'high-level' enough to allow adaptive algorithms to be implemented using distributed systems technology.

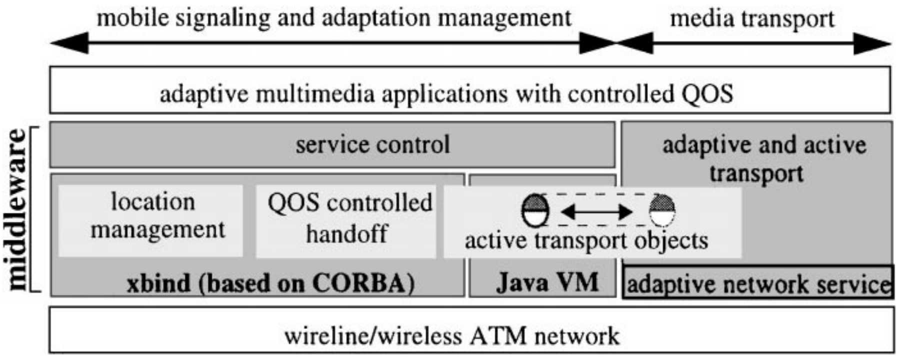


Figure 1. Mobiware architecture.

2.1. Architecture

Mobiware promotes the separation between mobile signaling and adaptation management on the one hand and media transport on the other. As illustrated in figure 1 mobiware utilizes xbind (based on CORBA) and Java for signaling and adaptation management during handoff or periods of persistent QOS fluctuation. The Java Virtual Machine executes on mobile devices, base stations and mobile-capable ATM switches and supports the dynamic execution of *active transport objects* (*ATOs*). These transport objects constitute an ‘active’ component of the mobiware transport system which can dispatch ATOs to strategic points in the network or end-systems to provide value-added QOS support. The concept of ATOs is derived from work on protocol boosters [11].

The realization of end-to-end QOS control and the exploitation of scalable flows is achieved in mobiware through; (1) resource binding between mobile devices, base stations and ATM switches; and (2) provision of a set of QOS-aware adaptive algorithms. These algorithms operate in unison under the control of mobiware:

- *QOS controlled handoff*, provides signaling for handoff which exploits the use of: (1) soft-state and hard-state to represent flows; (2) aggregation of flows to/from mobile devices; and (3) routing and QOS renegotiation anchor points to limit the impact of small-scale mobility on the wider fixed network;
- *adaptive network service*, provides hard QOS guarantees to base layers (BL) and soft QOS guarantees to enhancement layers (viz. E1 and E2) of multimedia flows based on the availability of resources in the wireless environment; and
- *adaptive and active transport*, supports the transfer of multilayer flows through the provision of a QOS-based API and a set ATOs (e.g., media scaling [9]) and *static transport objects* (*STOs*), e.g., playout control. STOs are statically configured and execute at mobile and fixed devices only. In contrast, ATOs are dynamically dispatched to the mobile devices, base stations or ATM switches to support valued-added QOS at strategic nodes.

Media scaling plays an important role in processing flows on-the-fly to match available resources at a bottleneck node, e.g., the air-interface. Media scaling exploits the

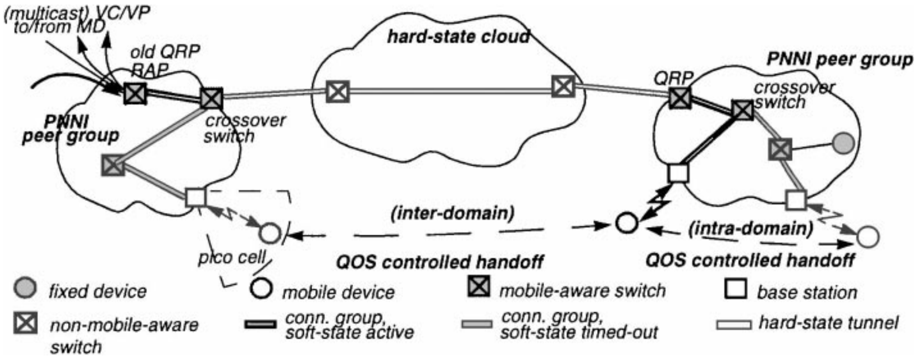


Figure 2. QOS controlled handoff.

intrinsic scalable properties of multi-resolution audio and video flows. In addition, adaptive algorithms take into account the knowledge of the user's flow adaptation policy to actively *filter* flows at critical nodes. This is achieved using an adaptive and active transport system to dispatch media scaling agents called *mobile filters* at critical nodes in the network or end-systems. Mobile filters are one of a class of ATOs which best utilize the available bandwidth to seamlessly deliver media with smooth change in the perceptual quality during handoff. Mobiware updates a location management algorithm during handoff. The mobiware location management uses a logical-name-to-logical-name mapping to express the current position of each mobile device.

The mobiware platform models the wireless portion of the ATM network as being divided into pico-cells each served by a base station connected to a wired ATM network as illustrated in figure 2. Base stations are cell relays which translate the ATM cell headers from radio ATM format to that used by standard ATM. Each base station supports signaling, QOS control and adaptation of flows based on semantics of an adaptive network service. The existing wired ATM network provides connectivity between base stations. We organize the wireless network into domains. A domain consists of a set of base stations which are under the management of mobile-aware ATM switches. A domain corresponds to a logical partition of the wireline network and the physical location of the base stations in hierarchies for scalable routing (in ATM Forum PNNI routing, these domains are peer groups).

3. QOS controlled handoff

The goal of the QOS controlled handoff algorithm is to dynamically re-route a set of connections associated with a mobile device from one base station to another without significantly interrupting flows in progress. The design of QOS controlled handoff is driven by two conflicting design goals: (1) support mobility; and (2) minimize the impact that small scale mobility has on the wireline portion of the network during handoff. To achieve these two design goals we introduce:

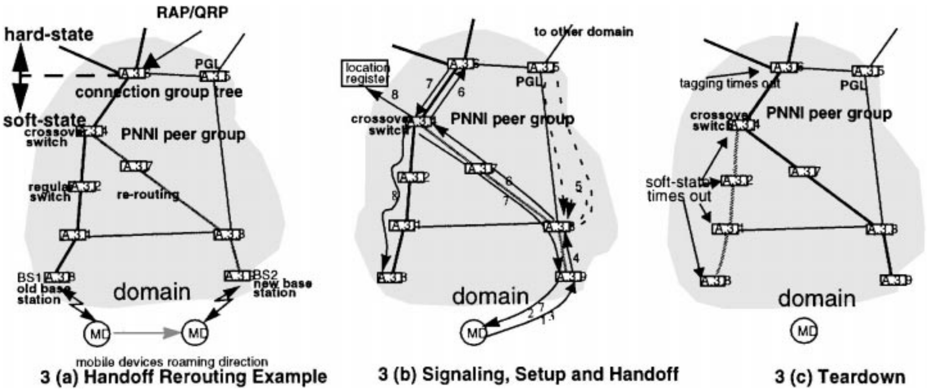
- *mobile soft-state*, models the dynamics of mobility through the continuous re-routing and QOS re-negotiation of flows as a mobile device roams. Soft-state is established between a per mobile *QOS re-negotiation anchor point* (*QRP*) and the mobile device. Hard-state is used between the QRP and the fixed network portion of the network;

- *connection group* (*CG*), provides a common routing representation for all virtual paths and virtual circuits destined to/from the same mobile. Connection groups decouple handoff re-routing from resource allocation at a per mobile *routing anchor point* (*RAP*). The RAP allows collective control and mobility management of all flows associated with a mobile device during handoff. Having a single reference point to manage all connections greatly simplifies handoff; and
- *logical anchor points*, provide an interface between the hard-state and soft-state portions of flows. The RAP and QRP are logical anchor points which localize the periodic re-routing and re-negotiation and during handoff processing, respectively.

Mobiware allows collective control and management of all connections associated with a mobile device using a single connection group identifier (CGI), which uniquely represents a single reference point to manage all connections. Connection groups are setup using multicast connection management operations, e.g., *addBranch* to a connection group tree. Removing a branch of a connection group tree after handoff is managed automatically through the semantics of the soft-state operations used in the mobile environment.

Mobiware handoff achieve this through the interaction of distributed QOS handoff algorithms with a set of mobiware *virtual resource objects* which model physical hardware devices and QOS as CORBA objects [19]. Mobiware models base stations [6] as a set of virtual resource objects: *virtualBasestation*, which is used to represent and manipulate the GCI/VP/VC routing table; *virtualWirelessLink*, which is used to represent and allocate QOS to a flow based on the concept of a *scheduler region* [14, 16]; and *virtualQOSFilter*, which is used to scale media at the base-to-mobile link.

In the following sections we will step through the QOS controlled handoff. See figure 3 for illustration of the handoff phases.



3 (a) Handoff Rerouting Example

3 (b) Signaling, Setup and Handoff

3 (c) Teardown

Figure 3. QOS controlled handoff walk-through.

3.1. *Signaling phase*

The first phase of QOS controlled handoff (see figure 3(b)) determines whether a new base station can provide a stronger signal at the desired level of QOS. A QOS monitoring algorithm resident at mobile devices monitors beacon messages which are periodically broadcast by all neighboring base station. In addition to indicating the strength of neighboring base station signals beacon messages indicate the residual capacity currently available at base stations. By periodically monitoring beacons from all neighboring base stations the mobile device is able to determine link qualities and occupancy of adjacent base stations and use this QOS state information as a basis to initiate handoff after a suitable dwell time.

The next phase of handoff is the establishment of a new signaling channel between the mobile device and the new base station. A mobile device issues a (1) signalRequest to the new base station over a dedicated meta signaling channel using the base station address found in the beacon message. This results in the creation of the signaling channel when the base station responds with a (2) signalResponse. The signaling channel carries all signaling and QOS management messages between the mobile and wireline network.

3.2. *Setup phase*

Once the signaling channel has been successfully created the mobile device initiates a forward handoff to the new base station. It does so by issuing a (3) *reservation message* (for details on the *res* message see Section 5) which includes connection group route state information and desired QOS required. The handoff management algorithm located at the new base station uses this state information to establish a new branch (between the crossover switch and new base station) to the existing connection group tree with the desired QOS. Mobile devices express desired QOS in terms of the semantics of the adaptive service and connection groups. Connection group QOS requirements are specified in terms of connection group base layer requirements and enhancement layer requirements, respectively. Admission control located at the new base station first determines whether sufficient resources are available to support the requested handoff.

Reservation messages can be updated by the distributed handoff algorithm as they are forwarded from the new base stations toward the rendezvous switch based on the availability of resources at the traversed nodes. The semantics of the adaptive service provide hard guarantees to the base layers and admit enhancements layers based on the availability of residual resources; that is, the resources remaining once all base layers have been guaranteed. The new base station only drops the handoff (*handoffDrop*) if insufficient residual capacity is available to meet the group connection base layer resource requirements. Assuming that sufficient resources are available to meet the minimum QOS requirements, the *res* message is routed toward the crossover and rendezvous switch reserving resources on route. Connection group routing information is used in combination with PNNI routing to determine (5) the shortest path between the new base station and the existing connection group tree which meets the desired QOS reflected in the *res* message. The PNNI peer group leader is interrogated (5) should the mobile device roam outside the current peer group domain of the old base station. This may suggest that the existing RAP would provide a sub-optimal routing point and a new RAP along with a new QRP is required. Generally, the new QRP is located close to the mobile device, and over a longer timescale a RAP re-routing algorithm determines when and where to move the RAP.

In the case where roaming is within the current peer group domain, the new base station issues a (6) *res* message to the crossover point. Each intermediate switch on route between the base station, crossover point and QRP provides admission testing and resource reservation. The QRP terminates QOS re-negotiation. The existing connection group traverses the crossover switch. Generally, the crossover switch processes the *res* message and forwards it to the rendezvous switch which also processes it and then responds by sending an (7) *adaptation message* (for details on the *adapt* message see Section 5) to the mobile via crossover and the new base station. The adapt message is used to commit switch and base station resources at the downstream nodes for the new connection group branch. Once the base station receives the adapt message it commits resources and broadcasts the adapt message to all mobile devices in the cell.

3.3. Handoff control

The adapt message serves two purposes. First, it confirms the level of QOS provided by the new base station and wireline portion of the network to the new mobile device as it enters a new cell. Second, it informs existing mobile devices of any resource changes which may have occurred during handoff to accommodate the new mobile device. The semantics of the adaptive service state that base layer resource reservation requests receive precedence over requests for any higher layer qualities. To this extent the quality delivered above the base layer to existing mobiles may be altered to allow a new mobile device to enter a new cell.

The handoff algorithm interacts with the media scaling agents (which interact with mobile filters—see Section 4.2) at the base station and QRP during connection group setup. The level of media scaling is dependent on the current utilization of the wireless link, application specific desired QOS and the semantics of the adaptive service. Media scaling may result in the “scaling-down” of mobile delivered quality when a mobile device enters a pico-cell and “scaling-up” when they leave. Mobile devices upload Java-based mobile filters (see Section 4.2 on active transport objects) to the base station or rendezvous switches in the case where the requested QOS could not be supported. This results in enhancement layers being dropped or filtered at specific points on a connection group tree (e.g., at the new base station or QRP). QOS filters support the delivery of different combinations of layers to particular mobile devices based on the available resources. Network-based mobile filters are essential to support multicast QOS to heterogeneous receivers.

Handoff registers new mobile devices with the domain location management which in turn allocates a new proxy ATM address as mobile devices roam into cells within a new peer group domain. A (8) *locationUpdate* message is used to register the mobile devices at the home location in this case and update the cache register at the old base station (8). QOS controlled handoff is based the notion of soft-handoff for roaming mobile receivers and hard-handoff for roaming mobile senders. In the case of soft-handoff, the mobile device simultaneously interacts with the old and new base stations. Once the QRP responds with an adapt message to the mobile device it begins to forward the new flow to the mobile. This results in duplicate cells (for old and new flow) arriving at the mobile device via the old and new base stations. The rendezvous switch uses cell tagging to preserve ATM cell level sequence integrity at the mobile device during handoff.

After tagging has commenced, mobile devices determine suitable synchronization points between old and new flows and initiate flow switching from the old to the new flow. After flow switching the old flow is rendered redundant. Old flows continue to arrive at the mobile devices as long as the route between the old base station and QRP is active; that is, old flows are switched through to the mobile while the mobile soft-state is still installed and has not timed-out. Mobile devices do not, however, have to process old flows.

3.4. Teardown

After flow switching the new base station refrains from sending any further periodic *res* messages to the old base station and closes the signaling channel to the old base station. Once the mobile soft-state timer expires the old branch of the connection group tree between the QRP to the old base station is timed out and removed; i.e., after the mobile soft-state timer expires resources are deallocated and switching tables flushed accordingly. This is defined as teardown.

Media scaling is once again invoked at the old base station to determine if deallocated resources can be utilized by any existing mobile devices at the old base station. Mobile devices located at the old and new base station periodically probe the base station and network for more resources using the *res* and *adapt* message pairs. During handoff, mobile devices resident at the old base station issue reservation messages toward the QRP and receive an adapt message which indicates any extra resources made available after a mobile has left the current cell. The new adapt message reflects any new resource availability and adjusts any QOS filters via interaction with media scaling agents at the old base/rendezvous switch. We describe the condition whereby resources freed up during handoff are distributed to remaining mobile devices at the old base station as scaling-up.

4. Adaptive and active transport system

A fundamental aspect of our work is the development of an adaptive and active transport that incorporates a QOS-based API and a full range of transport algorithms to support the delivery of continuous media over mobile networks. The mobiware transport operates in two modes:

- *adaptive mode*, which provides a set of STOs (viz. playout control, flow control, flow scheduling and shaping, flow monitor and adaptation manager) that best assists multimedia applications when adapting to minor QOS fluctuations as a consequence of cell/packet loss and delay variation; and
- *active mode*, which provides a set of ATOs (viz. mobile filters [2], mobile error control [12] modules and mobile snoop [3] modules) that can be dynamically dispatched to mobile devices, base stations or mobile-capable ATM switches to provide value-added QOS during conditions of persistent QOS fluctuation that may emerge during handoff.

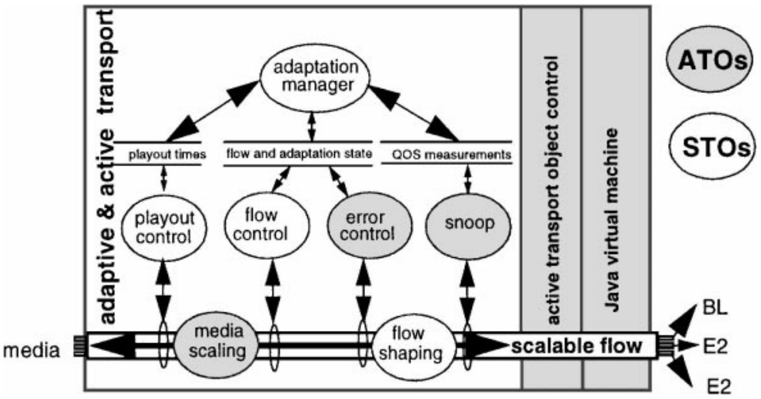


Figure 4. Adaptive and active transport system.

In the active mode, a local adaptation manager monitor the loss available bandwidth characteristics of flows and interact with ATO control to select, dispatch, bootstrap, configure and tune the appropriate ATO to the requesting target node.

4.1. Transport API

The transport API formalizes the end-to-end QOS requirements of the adaptive multimedia application (client or server) and the potential degree of media scaling acceptable to the application [7] (based on flow adaptation policy) which bounds its perceptual range. As illustrated in figures 1 and 4, applications interact directly with the service control API to establish, control and maintain the requested service. The service control algorithm governs the point at which a QOS controlled handoff is initiated.

The adaptive transport API assumes a client-server model where servers interact with service control to create *QOS groups* specifying their *QOS profile* (i.e., QOS requirements: traffic class, delay and bandwidth) for each multi-resolution of the flow and *flow adaptation policy* (i.e., the type of coding and prioritizing of the various resolutions used, and flow-spec for each flow) of the source media. The traffic class and delay bounds are common for each resolution of the scalable flow. The user can prioritize connections so that during handoff certain connections receive preferential treatment over others in light of reduced bandwidth (e.g., drop the video connection before the audio). The bandwidth for each resolution is specified in a flow-spec [7] by the clients and servers. Clients join QOS groups, inspect the QOS profile of the source and then select the appropriate resolutions by matching their capability to consume source media. For full details on the adaptive algorithms and API see [7].

4.2. Active transport objects

Adaptation managers operate on all mobile devices, base stations and mobile-capable ATM switches and continuously monitor the performance of the channel and each flow.

Adaptation managers and QOS controlled handoff algorithms interact with local ATO control to request the remote loading of ATOs to support QOS during periods of service degradation.

Currently mobiware supports three types of ATOs:

- *mobile filters*, which are used during periods of limited bandwidth to either drop layers of a flow (we call these types of mobile filters media selectors) or process audio and video in the compressed domain to meet an available bandwidth [2, 10, 20]. In general, mobile filters are dispatched once during handoff and are tuned based on the available bandwidth;
- *mobile error control modules*, which provide hybrid ARQ and FEC mechanisms for improved reliability of audio, video and data over the air-interface during periods of excessive cell/frame loss [12]; and
- *mobile snooping modules*, which help increase the performance of flows (e.g., TCP data flows) by snooping [3, 12] end-to-end protocol messages as a means to trigger the local value-added error control mechanism over the air-interface.

ATO are application specific and interact with differing algorithms to provide value-added support. For example, mobile filters are driven by an available bandwidth indication and interact explicitly with the adaptive network service algorithm. In addition, "*ATO state*" can be transparently moved during handoff. For example, a mobile filter executing on a base station can be automatically moved and reconfigured at the new base station. Movement of ATO state like this is dependent of the specific ATO and the operating conditions existing in the new cell. ATO state can propagate along with the connection state from the old to the new node over the wireline portion of the network. In this case the ATO state can take advantage of the wireline ATM interconnect to achieve a fast handoff. ATOs are capable of flexible autonomous actions in QOS fluctuating environments. Operationally, ATOs can be dispatched, configured and executed at any ATO-capable networked node.

4.3. Active transport object management

ATO management consists of a distributed algorithm which manages the installation of new ATOs anywhere in the network. ATO management uses a client-server approach to select, dispatch, bootstrap, configure and tune new ATOs between an ATO server and the target node.

As illustrated in figure 5 ATO management is divided into three operational modes:

- *ATO control* is a distributed signaling algorithm which comprises mobiware ATO control objects. These objects are permanently resident at the base stations, mobile capable ATM switches and mobile devices. ATO control objects support a set of methods to *select* (1), *dispatch* (2) and *configure* (4) mobile filters;
- *ATO instantiation* fetches remote Java bytecode classes (which are representations of ATOs) and *bootstraps* (3) them into Java VM environment based at base stations or mobile capable ATM switches. Once an ATO has been loaded and booted into a switch

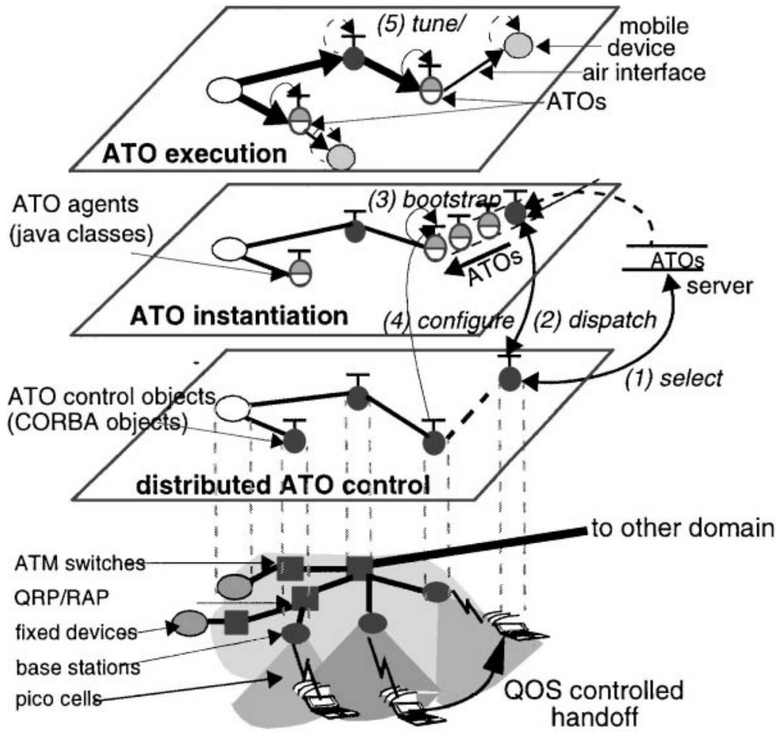


Figure 5. Active transport object management.

or base station the local ATO control object initiates a configure (4) operation to complete the instantiation phase; and

- *ATO execution and tuning*, at this point ATO can act autonomously to provide value-added QOS support to the transport of continuous media or data. Some ATOs interact with existing algorithms during the execution phase and can be periodically tuned (5).

As illustrated in figure 4, the adaptation manager periodically monitors the QOS delivered at the channel. In the case where the platform needs value-added software. Mobile filters are capable of flexible autonomous actions in QOS fluctuating environments. Operationally, mobile filters can be dispatched, configured and executed at the base stations or ATM switches. Mobile filters can be periodically tuned via a filter interface to match the available resources at a particular bottleneck node (e.g., base station or mobile capable ATM switches). In addition to being QOS adaptive, mobile filters automatically propagate during handoff. For example, in the case of a handoff between two WATM radio ports existing mobile filters (called *mobile filter state* in mobiware) propagate along with the connection state from the old to the new base station over the wireline portion of the network. In this case, the mobile filter state can take advantage of the wireline ATM interconnect to achieve a fast handoff.

4.4. Mobile filter ATOs

Currently we have implemented a media selector ATO in Java that drops either E1 (i.e., P pictures) and E2 (i.e., B pictures) frames of an MPEG-1 stream based on the available resources. Media selectors ATOs do not process the media unlike other computationally intensive mobile filters e.g., dynamic rate shaping mobile filters [10]. One of the key performance issues related to mobile filter technology is the time taken to dispatch, bootstrap and configure new agent over wireless and wireline interfaces. Another important performance concern relates to the performance penalty paid by flows as they are processed at ATM switches and base stations. The amount of delay introduced by such operations as flows traverse ATO is dependent on the computational complexity of the ATO, the additional overhead of the Java VM and the cost of derailing ATM cells for filtering. For some initial performance results relating mobile filter ATO see [2, 12].

5. Adaptive network service

The adaptive service is based on previous work on adaptive service for wireline ATM networks first proposed in [7]. This service model provided “hard” guarantees to the base layer (BL) of a multilayer flow and “fair share” guarantees to each of the enhancement layers (e.g., E1 and E2) supported by the service. To achieve this, the BL undergoes a full end-to-end admission control test [7]. In contrast, enhancement layers were admitted without any such test but competed for residual bandwidth with all other adaptive flows which traversed a particular switch along a specific route.

Similar schemes have subsequently been adopted by a number of practitioners in the wireless communications field [4, 5]. However, one drawback of these adaptive service schemes is that a QOS fluctuation at a particular switch (e.g., due to a new call being admitted) in a wireline/wireless environment can potentially impact other flows traversing the link/port. This results in a chain-reaction as distributed adaptive resource management algorithms resolve the new change in state. Unbounded adaptive service will typically occur in highly QOS fluctuating environments such as in mobile wireless systems. Therefore, there is a need to bound this chain-reaction while at the same time offering an adaptive service to mobile terminals to reduce the probability of handoffs being dropped as new mobiles roam into a bottleneck pico-cell.

This is the motivation behind our adaptive service. The adaptive service primarily operates over the wireless segments of end-to-end flows; that is, between the QRP, base station and the mobile devices as illustrated in figure 2. Therefore, QOS adaptation is limited to the pico-cellular area and does not directly impact the wired portion of the network. Mobile devices compete for wireless resources by interacting with an adaptive service algorithm at the base station. This limits adaptation to where it is more likely to occur: at the wireless link between the base station and QRP. Enhancement layers are rate controlled based on explicit feedback mechanisms about the current state of the ongoing flow and the availability of residual bandwidth at a base station.

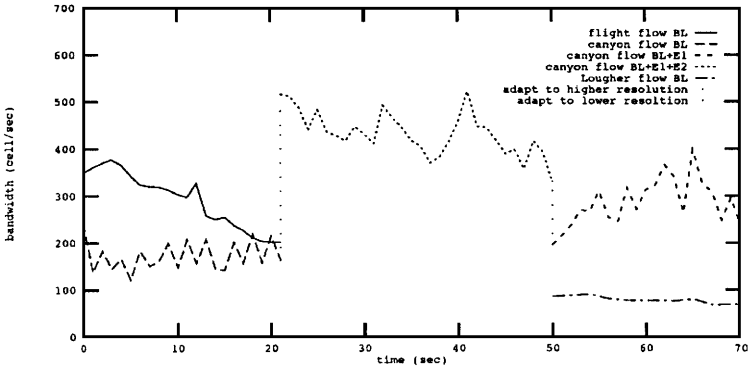


Figure 6. Scalable video flows.

A number of objectives motivate the design of a new wireless adaptive service for mobile QOS fluctuating networks. The first objective is to admit as many base-layers as possible across the wireless link. As more base-layers are admitted the guaranteed capacity region grows to meet the hard guarantees for all base signals. In contrast, the residual capacity region shrinks as enhancement layers compete for diminishing residual bandwidth resources. Our second objective is to fairly share this wireless residual capacity among competing enhancement layers based on an algorithm called *weighted fairshare* [7]. In addition, another associated objective is to limit the impact of the wireless adaptive service on the wired network. Our fourth objective is to adapt flows both discretely and continuously based on an adaptation mode supplied in the user-supplied adaptation policy. In the discrete mode, no residual bandwidth is allocated by the wireless adaptive service algorithm unless a complete enhancement can be accommodated. In contrast, in continuous mode any portion of the residual capacity can be made available and be utilized by the adaptive flow [6].

Figure 6 illustrates the operation of the adaptive service over a wireline ATM link. The bandwidth of the link which the adaptive algorithm shares between three different video flows (akin to mobile devices) was setup to be similar in bandwidth capability to a low bandwidth wireless ATM link (e.g., admission control is preset to 600 ATM cells/s to be shared by all competing flows). The scenario shows the consumption of three video clips beginning at time zero. The *canyon* and *flights* video flows start at time zero and only have their BL supported; i.e., the minimum quality is provided. No residual capacity is available to support higher qualities because of the limitation of the available resource (i.e., 600 cells/s). At 20 s into the trace, the *flights* video flow terminates freeing up resources for remaining flows (i.e., remaining mobile devices). This is akin to a mobile device roaming out of a pico-cell. At this point the adaptive service attempts to switch into the higher resolutions for remaining flows. The trace shows that the *canyon* video flow receives the best quality (i.e., BL + E1 + E2) after the *flights* video terminates. This situation remains stable until another video flow starts up at 50 s into the trace which is akin to a mobile device roaming into the cell and competing for wireless link resources. Resources are thus allocated to meet the BL QOS requirements of the new mobile device. The *canyon* video flow is adapted down to the BL + E1 quality at 50 s into the scenario as a result of admitting the new flow (akin to a new mobile). While the scenario is taken from wireline ATM experimentation into adaptive services [7] the results indicate what the service would look like to mobile devices as they roamed between cells.

5.1. Mobile soft-state

Rendezvous switches serve as points above which re-routing and QOS re-negotiation only occur when a mobile device roams between two adjacent routing domains. Inter-domain roaming operates at a much lower frequency than intra-domain roaming. Occasional inter-domain roaming requires re-routing and QOS re-negotiation in the wireline network. We contain the frequency of re-routing and QOS re-negotiation caused by small-scale mobility from impacting the wider wireline network. This is contained at the QRP. Above the QRP only infrequent re-routing and QOS re-negotiation occur. Below the QRP (between the mobile devices and QRPs) frequent re-routing and QOS may be observed.

Based on our understanding of the dynamics of small-scale (intra-domain roaming) and large-scale (inter-domain roaming) mobility we model end-to-end flow through a combination of “hard-state” (above the QRP in the wireline network) and “soft-state” (below the QRP in the wireless network). We argue that soft-state is suited to support the dynamics of mobility and QOS adaptation found in wireless and mobile networking.

A significant contribution to our approach to handoff is the use of soft-state [21] to support the dynamics of mobility in the wireless and wireline network. Soft-state is used between the mobile devices and QRP because it best suits the dynamic nature of QOS adaptation and device mobility. Mobile devices periodically send reservation messages (*res*) toward the QRP in the wireline network. These reservation messages carry the mobile devices desired QOS requirement and are interpreted by the base station and fixed network infrastructure. Resources are allocated to the mobile device over a particular wireless/wireline route for the duration of the mobile soft-state timeout. If during that time the infrastructure receives another *res* message it refreshes the state held in the base station and switches between the mobile device and the QRP. As mobile devices roam between adjacent cells the periodic *res* messages are used by the mobile devices to establish a new state (i.e., a route with QOS attached) at the new base station and intermediate switches. In addition, as the mobile device roams into a new cell the old connection group soft-state times out and is deallocated automatically. The periodic *res* message is used in combination with an adapt message to continually probe the base station and network for better QOS.

6. Conclusion

This paper has introduced a QOS-aware middleware platform for mobile multimedia communications. Mobiware has been specifically designed to address the complexity of proving QOS support for adaptive multimedia applications over wireless and mobile networks. Mobiware includes two key attributes which contains complexity. These are programmability and adaptability. In this paper, we presented the key adaptive algorithms that govern the available strategies for adaptability in mobiware. These include a QOS controlled handoff scheme which promotes the use of soft-state, connection groups and logical anchor points for fast and seamless handoff. We have also described a new transport systems which utilizes adaptive and active transport objects to provide value-added QOS support in the end-systems and network. The final mobiware adaptive strategy is based of a novel adaptive network service which has been designed to provide hard QOS guarantees for minimum flow quality and best effort delivery for higher quality.

The mobiware testbed consists of 4 ATM switches (viz. ATML Virata, Fore ASX100/ ASX200s, NEC Model 5, Scorpio Stinger) and 4 base stations. The base stations are multihomed 200 MHz Pentium with 25 Mbps wireline access to the wireline ATM network and 2 Mbps WaveLan air-interfaces to a number of mobile devices based on Pentium PCs and notebooks. The PCs run Linux, Windows/NT and xbind (based on CORBA). An early version of mobiware runs on PCs, base stations and the ASX100 ATM switch.

Finally, we have implemented a beta version of the handoff protocol which support soft-state, connection groups and logical anchor points [12]. In addition, we have completed the implementation of active transport object management and support mobile filters for media selection during handoff [8]. Currently we are implementing other adaptive and active transport objects and plan to investigate the integration of mobiware with wireless ATM radios. The results from this stage of our research will be submitted as contributions to the new IEEE standardization initiative on “Programmable Network Interfaces” [13].

Performance results of our experimental testbed can be found in [22].

References

1. C. Aurrecoechea, A.T. Campbell, and L. Hauw, “A survey of QOS architectures,” Multimedia Systems Journal, Special Issue on QOS Architecture, 1996 (to appear).
2. A. Balachandran and A.T. Campbell, “Mobile filters: Delivering scaled media to mobile devices,” Technical Report, Center for Telecommunications Research, Columbia University, Oct. 1996.
3. S. Balakrishnan, E. Amir Seshan, and R.H. Katz, “Improving TCP/IP performance over wireless networks,” 1st International Mobile Computing and Networking (MOBICOM’95), Berkeley, Nov. 1995.
4. V. Bharghavan, “Adaptive resource management algorithms for mobile computing environment,” Proc. OPENSIG Workshop, New York, April 1996.
5. K. Brown and S. Singh, “A network architecture for mobile computing,” INFOCOM’96, San Francisco, March 1996.
6. A.T. Campbell, “Towards end-to-end programmability for QOS controlled mobility in ATM networks and their wireless extensions,” Proc. 3rd International Workshop on Mobile Multimedia Communications (MoMuC-3), Princeton, Sept. 1996, and Wireless ATM Workshop, Espoo, Finland, Sept. 1996 (invited presentation).
7. A.T. Campbell and G. Coulson, “Implementation and evaluation of the QOS-A transport system,” 5th IFIP International Workshop on Protocols for High Speed Networks, Sophia Antipolis, France, Oct. 1996.
8. A. Campbell, R.-F. Liao, and Y. Shobatake, “Using soft-state for handoff in wireless ATM networks,” The Sixth WINLAB Workshop on Third Generation Wireless Information Networks, March 1997.
9. Delgrossi et al., “Media scaling in a multimedia communications system,” ACM Multimedia Systems Journal, Vol. 2, No. 4, 1994.
10. A. Eleftheriadis and D. Anastassiou, “Meeting arbitrary QOS constraints using dynamic rate shaping of coded digital video,” in Proc. 5th International Workshop on Network and Operating System Support for Digital Audio and Video, Durham, New Hampshire, April 1995, pp. 95–106.
11. D. Feldmeier, “Protocol booster,” COMET Group Seminar, Feb. 1996.
12. http://comet.columbia.edu/wireless
13. IEEE Standardization Initiative on “Programmable Network Interfaces.”

14. A.A. Lazar, S. Bhonsle, and K.S. Lim, “A binding architecture for multimedia networks,” Journal of Parallel and Distributed Computing, Vol. 30, No. 2, pp. 204–216, 1995.
15. D.G. Messerschmitt, J.M. Reason, and A.Y. Lao, “Asynchronous video coding wireless transport,” Workshop on Mobile Computing Systems and Applications, Santa Cruz, Dec. 1994.
16. M. Nagshineh and A. Acampora, “QOS provisioning in micro-cellular networks supporting multimedia traffic,” INFOCOM’95, Boston, April 1995.
17. J. Porter, A. Hopper, D. Gilmurray, O. Mason, J. Naylon, and A. Jones, “The ORL radio ATM system, architecture and implementation,” Technical Report, ORL Ltd., Cambridge, UK, Jan. 1996.
18. D. Raychaudhuri (NEC USA), L. Dellaverson (Motorola), M. Umehira (NTT Wireless Systems), J. Mikkonen (Nokia Mobile Phones), T. Phipps (Symbionics), J. Porter (Olivetti Research), C. Lind (Telia Research) and H. Suzuki (NEC C&C Research), “Scope and work plan for proposed wireless ATM working group,” ATM Forum Technical Committee, ATM Forum/96-0530/PLEN, April 1996.
19. The Common Object Request Broker: Architecture and Specification, Revision 1.2, published by the Object Management Group (OMG) and X/Open, Dec. 1993.
20. N. Yeadon, F. Garcia, A. Campbell, and D. Hutchison, “QOS adaptation and flow filtering in ATM networks,” in Proc. 2nd International Workshop on Advanced Teleservices and High-Speed Communication Architectures, Heidelberg, Germany, Sept. 1994.
21. L. Zhang et al., “Resource reservation protocol (RSVP)—version I functional specification,” Working Draft, draft-ietf-rsvp-spec-07.ps, 1995.
22. Angin et al., “A programmable mobile network: design, implementation and evaluation,” IEEE Personal Communication, 1998.

Andrew T. Campbell is an assistant professor in the Department of Electrical Engineering and member of the COMET Group at the Center for Telecommunications Research, Columbia University, New York. He is currently leading a research effort in Wireless Media Systems focusing on the development of QOS programmable middleware for mobile multimedia networks that comprise ad-hoc, broadband and next generation Internet technologies. Before joining academy Dr. Campbell spent 10 years in the industry focusing on the design and implementation of network operating systems and communication protocols for packet-switched local area and tactical wireless networks.