Dynamic SLA-based QoS Control for Third Generation Wireless Networks: The CADENUS Extension

Rajiv Chakravorty, Ian Pratt and Jon Crowcroft
{FirstName.LastName}@cl.cam.ac.uk
University of Cambridge, Cambridge CB3 0FD, U.K.

Maurizio D'Arienzo
maudarie@unina.it
University of Naples, Italy

Abstract— **With the evolution of QoS-capable third generation wireless networks, the wireless community has been increasingly looking for a framework that can provide an effective, network independent, end-to-end QoS control. In this paper, we first construct such a framework and then describe how dynamic SLA-based control can be used to achieve end-to-end QoS in a wired and wireless (UMTS) environment. The proposed framework, which is an extension to the IST CADENUS project, offers an effective wired-wireless QoS translation, an efficient QoS control and management, and a dynamic SLA policy-based QoS provisioning.**

Keywords—**QoS control, UMTS, SLA, Policy, CADENUS**

I. Introduction

The current Internet is *best-effort* that lacks any Quality of Service (QoS) support mechanism. This has lead the Internet Engineering Task Force (IETF) to seek for an appropriate QoS management solution, the outcome of which has resulted in two distinct Internet QoS paradigms – IntServ and DiffServ. Both paradigms typically attempt to address, in their own way, QoS management in the Internet. Meanwhile, the introduction of Universal Mobile Telecommunications System (UMTS) standard for third generation wireless networks is being widely looked upon as the Internet's future wireless extension. With Internet's ever expanding dimensions, constituting heterogeneous wired and wireless networks, the focus is gradually shifting towards a model for achieving an end-to-end QoS control rather than looking for simple QoS control for a specific network. Take for instance European Telecommunications Standard Institute (ETSI) UMTS Release 99 [4], where work based on similar lines to that of IETF – with a penchant to an "all-IP" architecture – targets on achieving an end-to-end QoS control rather than plain wireless QoS.

However, despite all the efforts made in this direction, the goal to realize an effective end-to-end QoS control scheme in a wired-wireless environment still remains to be seen. There are some fundamental challenges that has to be overcome in order to achieve this goal:

- **Network Level QoS Translation (Mapping) :–** Service differentiation based on a set of traffic classes, for both wireless as well as IP-core, needs an effective and reliable QoS translation (or QoS mapping) mechanism,
- **Dynamic QoS Management :–** This includes dynamic QoS monitoring and control. Proper monitoring of the network as per the user agreements, and control that can statically or even dynamically arbitrate or modify the QoS available from the network,
- **Control Management Infrastructure :–** Networks will have to ensure QoS control in their own way using their own sets of control mechanisms. These mechanisms will enable establishment, maintenance and termination of underlying network QoS. To achieve an interoperability strategy between these control mechanisms, in the quest for an end-to-end QoS control, would necessitate a framework that takes care of both – wired as well as wireless QoS.

In this paper, we address the aforementioned issues with a simple extension to the CADENUS (IST CADENUS Project, http://www.cadenus.org) framework (see, [1], [2]) that takes care of the end-to-end QoS control. The CADENUS framework defines an integrated architecture for creation, configuration and provisioning of end-user services for the networks that offer some sort of service differentiation (premium IP networks). It defines a logical architecture that is functionally partitioned into number of components (which we discuss later) for realization of such a framework. Our approach considers an extension to the existing CADENUS framework, and derives all the benefits of this framework to achieve a simple, yet effective, end-to-end QoS control for wired-wireless (UMTS) environment.

An important first step in this direction is to judiciously interwork QoS's of both wired as well as wireless (UMTS) network. ETSI specifically envisages different ways of interworking UMTS QoS with Internet QoS. The Third Generation Partnership project (3GPP) currently specify [4] the use of: (1) signalling (e.g. RSVP, LDP) along the flow path, (2) packet marking or labelling (e.g. DiffServ, MPLS), (3) interworking policy control with network resource mediators and, (4) using service level aggreements (SLAs) enforced by the network border routers. This paper will, however, limit the discussion to achieving such an end-to-end QoS control using a combined mix of dynamic SLA-based and policy control schemes.

The paper is structured as follows: Section II sets the context by reviewing QoS related issues pertaining to UMTS. Section III delineates the CADENUS framework and Section IV describes the proposed CADENUS extension for UMTS. Section V elaborates on the experimental test platform, while Section VI discuss related work. The last section concludes the paper.

II. UMTS Quality of Service Issues

A. UMTS Architecture

UMTS defines a system architecture that consists of a number of logical network elements, with each element having a defined functionality. Three main functional network elements are defined [3]: UMTS Terrestrial Radio Access Network (UTRAN) that handles all radio-related functionality, the circuit-switched (CS) and packet-switched (PS) domain along with the two fundamental support nodes - the serving GPRS support node (SGSN) and gateway GPRS support node (GGSN), and finally the core network (CN) in UMTS that has been adopted from the General Packet Radio Service (GPRS) architecture.

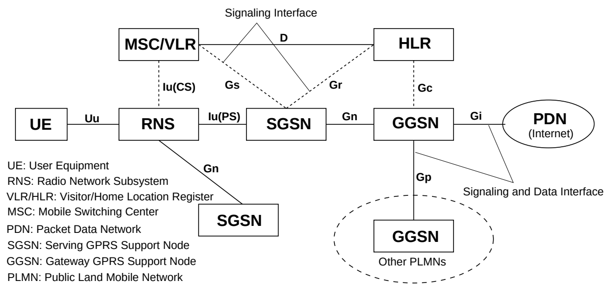

Fig. 1. UMTS Architecture

The core network in the PS domain connects UTRANs with the external networks through the GSNs. User equipment (UE) that interfaces the user and the radio-interface is also defined, which consists of the Mobile Equipment (ME) and the UMTS subscriber identity module (USIM). A packet-switched UMTS supports end-to-end IP services with guaranteed Quality of Service (QoS) as defined by the 3GPP [3], [4]. Thus an end-to-end service will involve the UE, UMTS access and the core network.

B. UMTS QoS Management

UMTS achieves QoS management using a layered architecture, with Bearer Services (BS) established between UMTS modules at different layers. Each bearer service deals with control signaling, user-plane transport as also QoS management. End-to-End QoS control is possible by inter-working TE/MT BS, the UMTS BS and External BS. To provide service differentiation, a UMTS network supports different bearer services that should correspond to a similar differentiation that can be applied in the IP-core network. For QoS control with an external IP network, 3GPP [4] dictates use of an IP BS Manager within the UMTS GGSN node (see figure 2). The IP BS Manager provides QoS control for an IP-core using DiffServ Edge functionality (or using a RSVP function).

The ETSI 3GPP Release 99 specifies [3] that GGSN *should* implement DiffServ Edge functionality, along with an IP Policy Enforcement Point (PEP) and optionally RSVP/IntServ function in its IP BS Manager. On the other hand, UE has a choice to support the same in its IP BS Manager, depending upon what capability it wants to have it's IP BS Manager be built-in with. While the UE may support all, or none of the functionality, it is typically the responsibility of the UMTS network to assure end-to-end QoS [3].

From 3GPP standardization point of view, the UE determines its QoS requirements using an application layer scheme (e.g. Session Description Protocol used with Session Initiation Protocol), which then maps those requirements to the Packet Data Protocol (PDP) context parameters or to the IP layer parameters (e.g. RSVP messages).

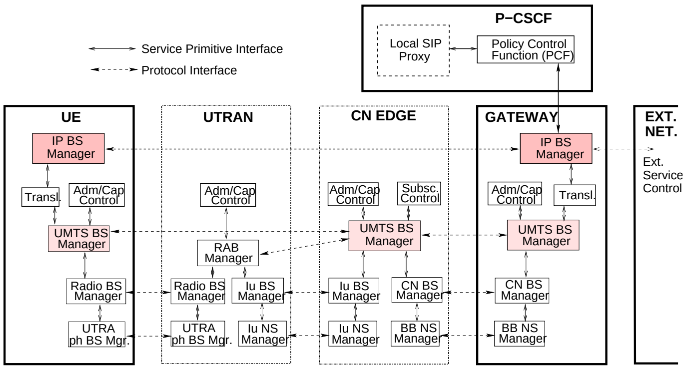

Fig. 2. UMTS QoS Management (control-plane)

When a GGSN gets to know about the QoS requirements (through PDP context or RSVP messages), it may use this IP level information to configure the DiffServ classifier functionality and provide inter-working between PDP context and backbone IP network. This information is made available using an *authorization token* located in the PDP context messages or from the RSVP messages directly. An authorization token consists of the policy and QoS inter-working functions: packer classifier, QoS information (SDP based), packet handling action and event generation information (for usage record purposes) .

If a UE supports RSVP, the GGSN can use RSVP instead of the PDP context to control QoS through the backbone IP network. In this case, RSVP session from a UE may be used to configure the DiffServ classifier functionality directly. However, if the GGSN does not support RSVP, the messages may transparently pass through the GGSN. In case of UE supporting only UMTS QoS mechanism, application QoS requirements can be signalled to the IP BS Manager at GGSN, by mapping the UMTS QoS parameters to the PDP context activation messages. The Admission/Capability function takes care of the resource allocation based on availability or some other policies and service invocation based on administrative issues. The subscription function authorizes usage of the service for a particular user, while translation function at the edge converts service primitives of the UMTS BS to external primitives of the external networks.

Further information describing a number of possible end-to-end QoS control-plane scenarios for UMTS is available in [3].

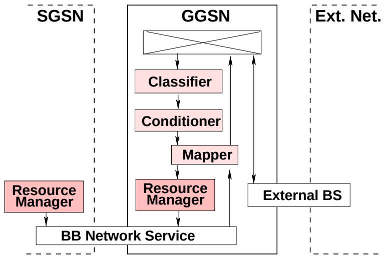

Fig. 3. UMTS QoS Management functions of GGSN (data-plane only)

The UMTS data-plane (figure 3) is responsible for traffic classification, mapping, conditioning and resource management. The classification function accommodates specific flow of packets, based on the marking on the packet headers or by some other means of classification, while the mapping function maps service classes in order to derive the intended QoS from the core networks. Policing function checks if the traffic profile is consistent to what was negotiated, and if it is not consistent, then the shaping function takes over to appropriately shape (mark, drop or delay) the packets for traffic compliance. Resource Manager schedule packets, and performs bandwidth and queue management functions. Unlike the downlink, the traffic conditioning for the uplink is performed directly by the MT.

C. UMTS QoS traffic classes

UMTS defines four QoS traffic classes: *conversational*, *streaming*, *interactive* and *background*. The conversational class provides strict delay guarantees, while background class offers no qualitative or quantitative guarantees. It can be at best referred to as a *best effort* class. When strict guarantees are necessary, for real-time voice and video applications, using conversational class seems to be a good choice. Slightly relaxed in terms of delays is the streaming class, to which all streaming applications can be mapped.

UMTS QoS Class	Conversational	Streaming	Interactive	Background
Maximum Bit Rate	√	√	√	√
Maximum SDU Size	√	√	√	√
Delivery Order	√	√	√	√
SDU format information	√	√		
SDU Error Ratio	√	√	√	√
Residual Bit Error Ratio	√	√	√	√
Delivery of Erroreous SDUs	√	√	√	√
Transfer Delay	√	√		
Gauranteed bit rate	√	√		
Traffic priority			√	
Allocation priority	√	√	√	√

TABLE I
ATTRIBUTES RELATED TO UMTS CLASSES

While interactive traffic follows a request-response pattern and can only justly provide qualitative guarantees, background class is similar to the best effort traffic consisting of bulk (e.g. ftp) and asynchronous traffic flows (e.g. e-mail) that will fall into this category.

A number of QoS attributes have been defined in support of the UMTS QoS classes. These attributes are highlighted in table I. The maximum bit rate should confirm a token bucket algorithm [4]. This rate corresponds to the maximum rate of the token bucket, and the bucket size indicates maximum service data unit (SDU) size. The maximum or the guaranteed bit rate cannot exceed 2Mbps, which is also the physical limit of this service. Finally, the transfer delay is the maximum delay for 95^{th} percentile of the delay distribution for all SDUs delivered [3].

III. THE CADENUS FRAMEWORK

The CADENUS framework investigates service and resource management aspects of networks, which an ISP or an Operator may consider for possible service differentiation. Specifically, the CADENUS project recognizes and emphasizes the importance of control management planes for effective end-to-end QoS control. It aims at the automation of a number of processes, collectively defined as a SLA-based service creation process for end-user services. The framework incorporates two key ideas – generalized mediation concept and the contract negotiation and translation feature for all the components.

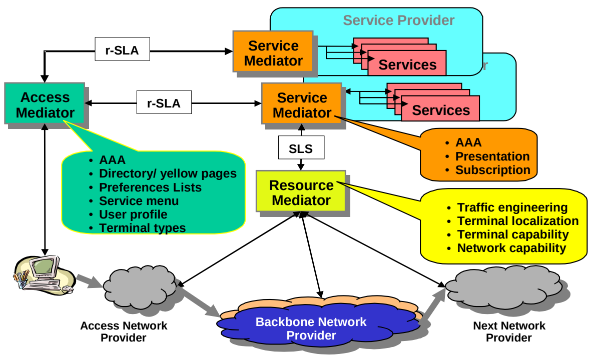

Fig. 4. CADENUS Architecture

A. CADENUS Components

To achieve the objectives sketched above, the framework proposes a CADENUS Mediation Component Architecture (CMCA) [2] that partitions system functionalities into 3 major components - Access Mediator, Service Mediator and a Resource Mediator (figure 4).

Access Mediator (AM) :– AM is the device into which users input their requests to the system. It adds value to the user by presenting a wider selection of services, ensuring lowest cost and offering a harmonized interface to the user. It serves to assist and ease the service selection process. It is also responsible in selecting the appropriate Service Mediator(s), according to the request made by the users. The other main functions include automation of r-SLAs (retail SLAs) process, *authorization-authentication-accounting (aaa)*, static or dynamic negotiation of r-SLAs and interaction with directory service for service specific information.

Service Mediator (SM) :– SM is the place where services are created and from where the impacts of the service re-configuration are communicated to the network resource management. It is responsible for finding the service, by requesting information from the appropriate Resource Mediator(s). In some cases, it may also build services from individual elements - the service itself. It maintains no direct contact with the end-users for SLAs, or for authentication or accounting. Instead, it is a unique entry point for presentation of the service and the subscriptions, the contract profile, and access to the service that was chosen.

Resource Mediator (RM) :– The Resource Mediator has the onus of selecting the appropriate network capabilities from multiplicity of network service providers and network technologies, given several available options. RM is the key functional entity in CADENUS architecture that has an end-to-end view of the complete QoS set-up. From an application view point, the RM only provides an "abstract-view" of the underlying network and differentiation that it can offer.

Notice that CADENUS offers a functional partition, where services are apportioned from resource control and management functions and also from the service creation process. It provides a clear separation of services from customers and service providers, as also between services and resources. It offers an open and flexible architecture, allowing the service providers to concentrate on the service and letting the network provider concentrate on the QoS control and on the scope of the corresponding Service Level Specification (SLS).

The combined role of these components is to efficiently manage user's access to the service, to present the portfolio of available services and to appropriately configure and manage the QoS-aware network elements available in the underlying network infrastructure.

B. CADENUS Service Definition

The CADENUS architecture offer services with QoS, on demand to an end user. The AM component allow users to select a service with a given or chosen QoS. This process invokes SLA and its corresponding SLSs involved in the transaction. While SLA is a human readable contract, it does not posses the technical sufficiency in achieving the desired end-to-end QoS control. Hence, SLAs are first converted into corresponding (one or more) SLSs, which can help fulfil the technical requirements imposed, for achieving the end-to-end QoS control.

In the CADENUS framework, service contract definition are of two types - (1) retail SLAs (r-SLA) as a unitary service between customer and the service provider and, (2) wholesale SLAs (w-SLA) for contracts between service providers. SLA is contracted (signed) by a customer, or more exactly, when the customers subscribes or modifies his service contract with his service provider. This modification step is off-line and not in real-time. However, from the perspective of wireless networks (UMTS), this presents a difficulty since QoS available in wireless networks may vary considerably during service usage (fading, handovers etc.). Hence, it should be possible to perform QoS (re)negotiation on the fly, necessitating use of dynamic SLAs within the CADENUS framework. Benefit of using dynamic SLAs as opposed to simple static ones is that decision about user's QoS needs can be done in real time. When a SLA is contracted by a customer, one or more SLSs will be created, which will essentially serve to configure or provision resources in the network. The scope of SLS includes: flow identification, traffic conformance, excess treatment, performance guarantees and reliability information. While the CADENUS framework specifies a clear functional separation for SM and RM, note that SM is still aware of the network topology as it has to generate SLSs for which it has to have some priori knowledge of the network.

In a section below, we show how with our proposed CADENUS extension for UMTS, SLAs can be dynamically provisioned in real-time for the wireless customers.

IV. PROPOSED CADENUS EXTENSION FOR UMTS

In this section, we present our proposed CADENUS extension for UMTS. We refer to this extension, in short as CUE, or CADENUS-UMTS Extension. As shown in the figure 5, the extension consists of two new, independent, functional components: the CUE-Service Mediator (CUE-SM) and CUE-Resource Mediator (CUE-RM). The extension lacks a third component from the original CADENUS framework - the Access Mediator. The need for an AM is not justified anymore for

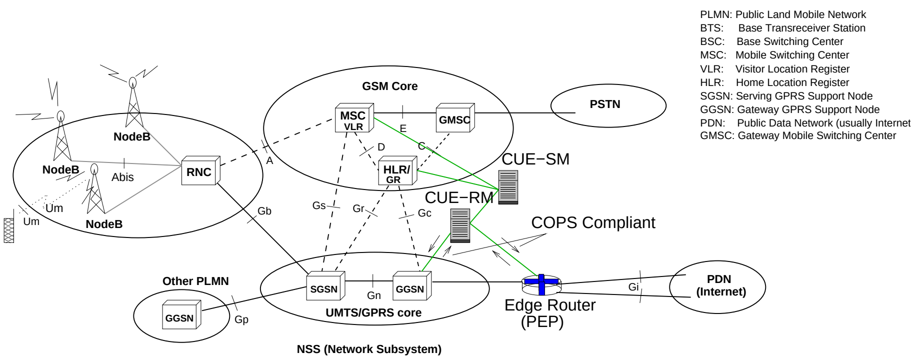

Fig. 5. Dynamic Policy-based CADENUS Extension for UMTS

the UMTS networks. This is because AM is used where users have to explicitly map their service requests to the system (as in case of premium-IP networks), whereas in UMTS such requests typically take the form of a PDP context activation/change request messages in real-time. This is also one reason why dynamic SLAs are more appropriate in the context of a UMTS framework. The SLA 'dynamic-ness' in UMTS is such that SLAs can change *on-the-fly* during a live session (for instance, during an on-the-fly codec change in a VoIP session). SLAs can be negotiated per session (during session initialization using PDP) or in the midst of a session typically through (re)negotiations (PDP context modify requests). However, the concept of static SLA for UMTS users is still applicable. This can correspond to the contract negotiated by the customer at start. Such contracts may provide SLAs that are typically *time-varying*, but are essentially static, yet application requirements may be change the SLA depending upon user demand. Once a contract is negotiated by a user, SLAs (or in particular retail-SLAs) are made available in the form of a residual user QoS profile located in the user's home UMTS network (mostly profiles will be located in HLR/LR of the UMTS network). Once a user profile is updated, it is converted into one or more SLSs by the CUE-SM, and communicated to the CUE-RMs. The CUE-RM uses the SLS to statically configure the GGSNs accordingly. Note that this step is indicative of a provisioning based resource allocation model rather than an outsourcing one. This is well expected since provisioning phase is necessary only immediately and just after the contract initialization.

Notice that there exists little functional distinction between CUE and IETF's policy-based framework [12]. In terms of functionality, a Service Mediator (SM) is a policy repository and policy translation system while a Resource Mediator (RM) can be compared to a Policy Decision Point (PDP). Contrast this with a 3GPP local policy-based scheme [4], where a P-CSCF (Proxy-Call Center Control Function) having built-in Policy Control Function (PCF) has functionality similar to that of a PDP. The PEP in this case is the GGSN. A PCF (here the PDP) can be located anywhere within the provider's network. Upon reception of a request from a user, the GGSN (PEP) retrieves the authorization token from a PDP context activation message and triggers the request to the PDP (P-CSCF) that performs QoS authorization (bandwidth, delay etc.) for the session. The interface used between GGSN and PCF is known as the `Go` interface [3], which allows service-based local policy and QoS inter-working on Common Open Policy Service (COPS) compliance. The initial authorization operation specified by 3GPP is a 'pull' operation (outsourcing based), but subsequent operations (modify) may be 'push' or 'pull' based.

In CUE, service-based policy and QoS inter-working is purely based on policy outsourcing (pull) model [9]. Also, a PEP as in CUE need not necessarily be located in the GGSN as specified by 3GPP; however, the basic functionality applicable to the data and control plane specific to the 3GPP standards would still remain. For instance, policy decisions may be stored by the COPS client in the Resource Mediator (CUE-RM) to achieve a degree of separation. Further, events in GGSN that demand change/modification (in a PDP change request message) will also trigger events to the CUE-RM, which will then decide upon the admission of this new session and respond appropriately. A pure 'outsourcing' based policy control model demands for a new COPS-client type that makes it possible to have dynamic policy-based resource allocation. We are currently finalizing the design of a new COPS client-type based on such an outsourcing model for the CUE.

A. Control Management in CUE

As per the 3GPP specifications, it is the responsibility of the UMTS network to ensure QoS from the core IP network even if a UE is not capable to do so. A UE may typically convey its application QoS requirements to the GGSN IP BS Manager using some appropriate signaling mechanism. This might be by using RSVP control messages, or using PDP context activation/modify request messages consisting of the authorization token.

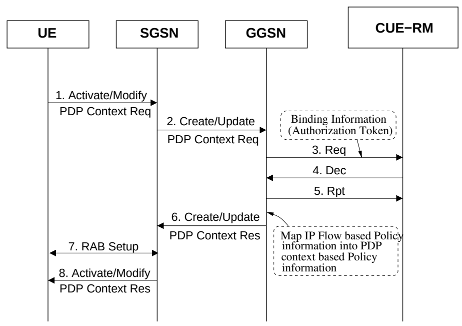

Fig. 6. Control Management in CUE with PDP context messages

Once the authorization token is available to GGSN, it can trigger an event (we discuss more on this in related work section) which is forwarded to the CUE-RM that decides based on some admission control scheme, whether to accept or reject the user application QoS requirements (see, figure 6). If the QoS requested by the application is below certain given threshold that was originally contracted by the user, then static SLAs take over. In this case, the QoS profile will be retrieved by the CUE-SM from user profile database and communicated to the CUE-RM. The final decision about admission is conveyed to the GGSN by CUE-RM, which then responds to the user in a PDP response message. This whole (re)negotiation process is based on the dynamic SLA-based policy outsourcing model, and is similar to the one used in [9].

From 3GPP view point, the control and data plane functionality is an integral part of the network elements. While the control-plane functionality has be tied to the GGSN, it may be a good idea to have part of the data plane functions co-located in an edge router. Specifically, since data traffic handling and scheduling can become computationally expensive, this part of the job can be located in a co-located edge-router (see figure 5). Further, the edge router can also take decisions on the traffic flows based on some sort of feedback available from the CUE-RM or alternately from GGSN.

B. CUE QoS Translation

An important goal in end-to-end QoS control over wired and wireless networks is to insure an effective translation of QoS traffic classes. Hence, mapping rules have to be defined so that functions can inter-operate efficiently. DiffServ framework offers Assured Forwarding (AF) [14] and Expedited Forwarding (EF) [13] as the *per-hop behaviour* (PHB) for a Differentiated Services (DS) compliant node. The EF PHB group is used when low loss, low latency, low jitter, and assured bandwidth for the end-to-end service is required. This group can offer deterministic service guarantees and a service likened to that of a virtual leased line. A single codepoint is defined for the EF class. The AF (read as $AFij$) class allow a DS domain to provide different levels of guarantees for forwarding IP packets. Currently, four classes ($N = 4, 1 \leq i \leq N$) with three levels of drop precedence in each class ($M = 3, 1 \leq j \leq M$) are defined for general use. An example usage in AF is that each class represents a higher level of service (e.g. platinum=1, gold=2, silver=3, bronze=4) with low, medium and high (j=1,2,3 respectively) drop precedence levels. Thus AF11 represents "best" while AF43 gives "worst" service levels.

UMTS QoS Class	Mapped DiffServ Class	Mapping Criteria
Conversational	EF only	Strict mapping for deterministic QoS guarantees, low latency and jitter
Streaming	AF11, AF12, AF21, AF22 etc.	low jitter
Interactive	AF13, AF23, AF22, AF32 etc.	relatively low latency and loss probability
Background	AF42, AF43 etc.	best effort

TABLE II

SAMPLE MAPPING OF UMTS QOS TRAFFIC CLASSES WITH DIFFSERV IP-CORE

For inter-working purposes, multiple QoS mapping levels can be defined. Typically, two different approaches can be envisaged:

- **One-to-one QoS Mapping:–** This is a strict one-to-one QoS mapping, which maps each UMTS QoS class to a corresponding DiffServ QoS class. Disadvantage here is that one-to-one mapping might not always be possible since networks may support different sets of QoS classes. A sample one-to-one mapping for DiffServ-UMTS QoS inter-working is given in [5]. In yet another example in [6], [7], mapping is done one-to-one. In this case, the traffic handling priority attributes of the UMTS interactive class map to the drop precedence of the AF class as they both share the same number of priority levels internally.
- **Many-to-one Mapping:–** This can map a number of DiffServ QoS traffic classes into a single UMTS QoS traffic class. When a DiffServ core defines many QoS traffic classes (using AF PHB let's say) as compared to limited QoS classes supported by UMTS, then a close set of DiffServ QoS traffic classes having almost similar QoS requirements can be merged into a single UMTS QoS class.

A proposal for a sample interworking of QoS traffic classes between DiffServ IP-core and UMTS QoS classes is shown in table II. Note that a number of QoS mapping levels can be defined depending upon how many QoS classes are supported by the UMTS and DiffServ-IP core. As opposed to a strict one-to-one mapping considered in all the previous studies [5], [6], [7], our proposal considers a *hybrid* QoS mapping scheme where strict one-to-one mapping is enforced only for some certain traffic class to support strict QoS translation when necessary (e.g. between EF-Conversational class for deterministic QoS guarantees), while one-to-many mapping may be used where QoS requirements are slightly relaxed. Thus our scheme is able to offer adequate flexibility and control on the QoS translation as desired by the network providers while not simultaneously trading-off any of the QoS requirements for the supported traffic classes.

V. EXPERIMENTAL TEST PLATFORM

Our experimental test bed consists of a GPRS test infrastructure from Vodafone. In this platform, shown in figure 7, the mobile users connect to the GPRS network via their GPRS phones. (Plans to upgrade the set-up to support UMTS 3G network are also being finalized by Vodafone). Currently, the platform lacks a proper service differentiation mechanism for the wireless network. However, this is soon expected to be available and since GPRS and UMTS 3G networks share the same QoS architecture, an experimental validation of our framework over GPRS should suffice even for UMTS.

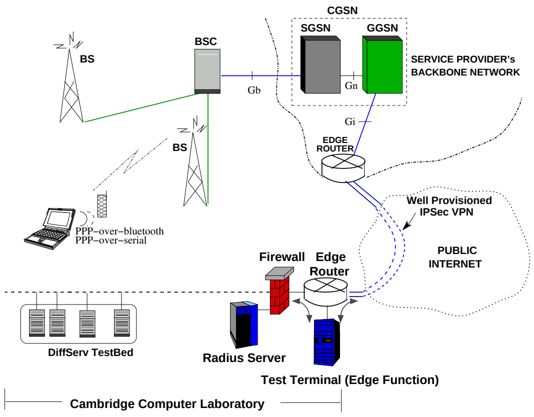

Fig. 7. CUE Experimental Platform

In the current Vodafone configuration, both the SGSN and GGSN nodes are co-located in a CGSN (Combined GPRS Support Node). Since we were unable to have direct access to the CGSN (for security reasons), we installed a test terminal near to the CGSN through a well provisioned IPSec VPN tunnel to route all traffic via the Computer Laboratory. The test terminal is located at the end of the tunnel, with the routing configured such that all packets flowing to and from the mobile host are passed to it for processing. This is fine since the terminal can be made to perform control as well data plane functions expected out of a GGSN – like traffic classification, mapping, scheduling etc. Integration of the CADENUS components - CUE-Service Mediators (CUE-SM), CUE-Resource Mediators (CUE-RM) on top of this infrastructure is the next step. Further, a DiffServ test-bed is also being set-up within the lab and appropriate inter-working as proposed between QoS traffic flows from both the networks are being envisaged. Future plans include performance evaluation of the experimental test set-up for different class flows under varying load, measuring QoS mapping effectiveness, etc..

VI. RELATED WORK

This section compares and contrasts our framework with some other proposals of significance. We first discuss the work of other alternative frameworks and then elaborate on issues related to dynamic SLA-based provisioning, as they are more relevant in the current context.

A. Other Contemporary Approaches

V. Marques *et al.* in [8] propose an "all-IP" architecture in the IST's MobyDick project. The MobyDick framework considers a DiffServ architecture along with the concept of admission control using QoS brokers [15]. However, the end-to-end QoS is implicit with no notion of end-to-end QoS at the IP layer; i.e. at the explicit per-flow level. Instead, the framework relies on proper network dimensioning and adequate network management to achieve a generic end-to-end QoS framework. Drawback with this approach is that network service can be arbitrated only during negotiation (at the application level using for example using SIP) and any changes to the network service is not possible. Thus dynamic service re-negotiation is not supported. Further, the framework assumes a rigid model where a UE (or MT) has to support DSCP (DiffServ Code Point) signalling. It also does not provide any timing constraints or explicit deterministic guarantees on QoS negotiated.

S. Maniatis *et al.* in [6] present QoS mapping for the traffic classes in both UMTS and core IP network. It introduces the *RCL Architecture* that largely trades end-to-end QoS based on static SLAs. The interworking between wired and wireless network is *direct*, which means that both the network can inter-operate with QoS management functions. However, unlike the MobyDick approach, it does not need a DSCP functionality with the MT. Deterministic QoS guarantees can be achieved since QoS mapping is direct with typically strict translation.

A. Yew *et al.* in [11] discuss a policy based framework for QoS in the Virtual Home Environment (VHE). In contrast to both of the schemes above, their framework make use of four separate logical entities, which they call - VHE provider, Access Network Provider, VHE Service provider, and the VHE user. In their servic provisioning model, they locate the PDP in the VHE provider as it is best suited to decide on the operational management of the user's home environment. A proxy with Content Capabilities/Profile Preferences (CC/PP) is also present in the access network, which exposes the user equipment capabilities to the PDP to allow for service adaptation. Typically, the access network takes help of the 'soft networks' solutions like Parlay APIs and Open Service Access (OSA) to aid QoS management using Dynamic SLAs, but in a more distributed manner. Using Parlay/OSA APIs on behalf of a user, the VHE provider can remotely configure a QoS capable agent at the egde router of the access network. Thus such an agent acts a PEP in a policy-based framework.

B. End-to-End Dynamic SLA-based provisioning Issues

The current DiffServ model lacks a standard definition for a signaling mechanism and an appropriate admission control framework. Hence, most proposals that consider resource negotiation in DiffServ networks are basically semi-static, meaning QoS can be negotiated only for the long-term (e.g. days), so also with the term *static* SLAs. IETF's RAP group has defined COPS protocol and its extension – COPS-PR (COPS provising model) that can be used to statically provision resources with a DiffServ network. A static resource provisioning model is simple; however, it can lead to under-utilization of network resources and may compound the adaptation process at times of high variability in the traffic demand [9].

To further evolve schemes that allow dynamic provisioning, S. Salsano *et al.* in [9] propose a new COPS client type, called COPS-DRA, that allows dynamic resource allocation and policy control for DiffServ based networks using COPS for signaling. The new client supports combination of both resource allocation models: outsourcing and provisioning. It uses signaling for resource admission control and QoS-aware call setups for SIP-based applications. It enables dynamic provisioning by exchanging resource allocation requests from the edge routers to a logically centralized bandwidth broker (BB), which it finally maps in a policy-based PDP-PEP relationship.

L-N. Hamer *et al.* in [10] consider COPS-PR for use in the UMTS framework. As mentioned earlier, a new interface called Go has been defined by 3GPP [4], which can be used within the provisioning framework of UMTS. This new interface allows service based policy control between GGSN (the PEP) and the Policy Control Function (the PDP). Their proposal considers a policy outsourcing scheme, but using COPS-PR, where a PEP informs the PDP (during initialization) what types of event it should trigger during policy outsourcing. Thus, when an event occurs the PEP triggers a COPS REQ to the PDP (here, UMTS PCF), which then retrieves session information and takes informed decision that is finally pushed to the PEP using standard COPS-PR semantics. Even if a COPS-PR model is typically considered a static one, L-N. Hamer *et al.* show that by suitably sharing the kind of events between a PEP and PDP using COPS-PR can help redefine the degree of 'dynamic-ness' of a resource provisioning model.

Our proposed CADENUS extension for UMTS, considers a 'outsourcing' model for resource provisioning, and hence, derives full benefits from a dynamic resource allocation framework.

VII. Conclusion and Future Work

Providing end-to-end QoS control in a wired-wireless environment necessitates effective QoS translation, proper control management and a dynamic SLA-based resource provisioning. We achieve all this in our CUE framework, which is a extension of the CADENUS architecture. To derive all the benefits of the CADENUS framework, the CUE architecture adds two new components: CUE-SM and CUE-RM that can be used to provision end-to-end QoS in a wired-wireless network. Further, the QoS arbitration used in the framework is dynamic, by using PDP context Activation/Modify messages, which can be changed in real-time in the midst of a live session.

We are currently working with the CUE experimental test platform for implementing the framework. We already have a working GPRS test-bed in collaboration with Vodafone, and are now tackling issues related to CUE component integration, DiffServ test-bed integration and implementation of the GGSN edge functions into our test terminal. Meanwhile, we are going to define a new-COPS client type, based on the dynamic SLA-based 'outsourcing' model for the CUE architecture. The integration of all the system components and a performance evaluation of the test platform will essentially serve as a validation of our proposed framework.

References

[1] M. D. Arienzo, M. Esposito, S. P. Romano, G. Ventre, "Dynamic SLA-based Management of Virtual Private Networks", In Proceedings of International Workshop on Digital Communications (IWDC 2001), Taormina, Italy, September 17-20, 2001, LNCS 2170, Springer 2001 source: http://www.cadenus.org/papers/

[2] G. Cortese et al., "CADENUS: Creation and Deployment of End-User Services in Premium IP Networks", IEEE Communications Magazine, pages 54-69, January 2003

[3] ETSI 3GPP, " End-to-End QoS Concept and Architecture (Release 5)", 3GPP TS 23.207 v5.4.0, 2002-06.

[4] ETSI 3GPP, "QoS Concept and Architecture (Release 1999)", 3GPP TS 23.107 v3.8.0, 2002-03.

[5] P. Stuckmann, "Quality of Service Management in GPRS-Based Radio Access Networks", Telecommunication Systems, 19:3,4, 515-546, Kluwer Academic Publishers, 2002

[6] S. Maniatis, E. G. Nikolouzou, and I. S. Venieris, "QoS Issues in the Converged 3G Wireless and Wired Network", IEEE Communications Magazine, pages 44-53, August 2002.

[7] S. Maniatis et al., "DiffServ-based Traffic Handling Mechanisms for the UMTS Core Network", In Proc. of IST Mobile and Wireless Telecommunications Summit 2002, June 2002, Thessaloniki, Greece

[8] V. Marques et al, "An Architecture Supporting End-to-End QoS with User Mobility for Systems Beyond 3rd Generation", In Proc. of IST Mobile and Wireless Telecommunications Summit 2002, June 2002, Thessaloniki, Greece

[9] S. Salsano et al., "QoS Control by Means of COPS to Support SIP-based Applications", IEEE Network, pages 27-33, March/April 2002.

[10] L-N. Hamer et al., "COPS-PR for outsourcing in UMTS: UMTS Go PIB" IETF Internet draft, draft-hamer-rap-cops-umts-go-00.txt

[11] A. Yew, A. Liotta and G. Pavlou, "Applying a Policy-based Framework to Manage Quality of Service Requirements in the Virtual Home Environment", In Proc. of the IEEE International Conference on Communications (ICC'02), New York, May 2002

[12] K. Chan et al., "COPS Usage for Policy Provisioning", IETF RFC 3084, March 2001

[13] V Jacobson et al., "An Expedited Forwarding PHB", IETF RFC 2598, June 1999

[14] J. Heinanen et al., "Assured Forwarding PHB Group" IETF RFC 2598, June 1999

[15] K. Nichols et al., "A Two-bit Differentiated Services Architecture for the Internet", IETF RFC 2638, July 1999