GENERIC SYSTEM ARCHITECTURE FOR 4G MOBILE COMMUNICATIONS

Vangelis Gazis, Nikos Housos, Athanasia Alonistioti and *Lazaros Merakos*

Communication Networks Laboratory, University of Athens

***Abstract*— In the European Union, the debate about 4G mobile has spawned the vision of a system that enables an "Always Best Connected" (ABC) mode of communication for the citizen of the forthcoming information society. This now widely accepted vision sketches a heterogeneous communication landscape comprising different wireless access systems in a complementary manner, where the user, supported by his/her personal intelligent agent(s), enjoys untethered connectivity and ubiquitous access to applications over the most efficient combination of wireless systems available. In the present paper, we identify the major developments in the fourth generation mobile communication market, present the technical aspects of the fourth generation network architecture and analyze the implications of the "ABC" vision upon it in terms of functional requirements and overall service provision capabilities. In closing, we introduce a generic 4G system model, elaborate on its major functional entities and finally, identify its key enabling technologies and solution sets.**

***Index Terms*—4G, service provision, policy management.**

I. INTRODUCTION

Fourth (4G) generation mobile communication systems tend to mean different things to different people: for some it is merely a higher-capacity (e.g., 100 Mb/s) new radio interface, while for others it is an interworking of cellular and wireless LAN technologies that employs a variant of the Mobile IPv6 mobility management protocol (e.g., Hierarchical Mobile IPv6) for inter-system handoff and IETF AAA technologies for seamless roaming.

There is no doubt that 4G systems will provide higher data rates. Traffic demand estimates suggest that, to accommodate the foreseen amount of traffic in the 2010 – 2020 timeframe in an economically viable way, 4G mobile systems must achieve a manifold capacity increase compared to their predecessors. In the European Union, the debate about 4G systems has been taking place mostly within the context of its IST Framework Program activities. Out of this process has spawned the vision of a system that enables an "Always Best Connected" – or "ABC" for short – mode of communication [1]. This now widely accepted vision sketches a heterogeneous network infrastructure comprising different wireless access systems (e.g., GSM/GPRS, UMTS, DVB-T, HAPS, WLAN) in a complementary manner, where the user, supported by his/her personal intelligent agent(s), enjoys untethered connectivity and ubiquitous access to applications over the most efficient combination of available systems. Figure 1 provides an illustration of the future 4G mobile network architecture comprising ad-hoc, cellular, hot-spot, and satellite radio components.

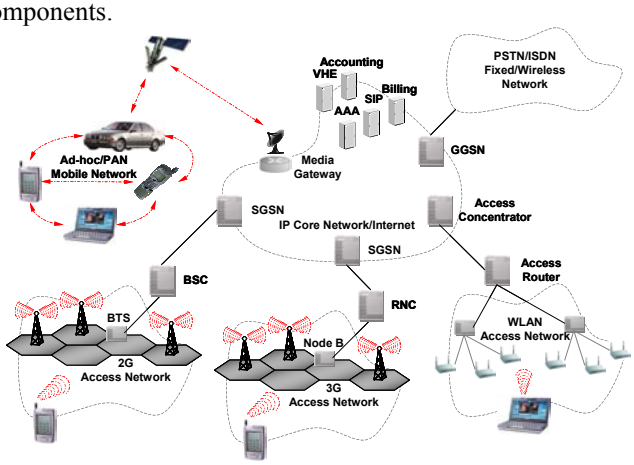

Figure 1. The generic 4G mobile network architecture.

Considering that a single system that optimally meets a wide range of use cases and satisfies diverse service requirements is likely to remain an engineering utopia, we understand that heterogeneous architectures that can exploit individual system capabilities' to optimally server the instant application and value-added service mix in a flexible manner are a plausible design approach. The goal of future mobile communication systems will be to incorporate and integrate different wireless access technologies and mobile network architectures in a complementary manner so as to achieve a seamless wireless access infrastructure. It is widely accepted that the individual (wireless and/or wireline) access networks will interface to core and/or backbone network elements over the IP protocol, the lingua franca of networking technology. Regardless of their particular technological blueprints (e.g., licensed or unlicensed frequency band, mobility management protocol, etc), these wireless access networks are expected to have the following in common:

1) A dynamic address assignment mechanism (e.g., DHCP, SLP, IPv6) that is capable of associating a short-lived or long-lived IP address to the respective wireless interface at the mobile terminal (e.g., Mobile IP COA association)
2) A transparent IP forwarding service that is accessible over the logical termination of the IP layer at the mobile terminal and one or more gateways (e.g., GGSN, Mobile

IP FA) at the wireless access network infrastructure. The IP forwarding service is set up by signaling procedures (e.g., PDP context signaling in the UMTS case) specific to the technical architecture of each wireless access network.

Technical reviews of existing wireless access standards [2] [3] and research contributions on the similarities of wireless network architectures [4] validate these assumptions.

II. SERVICE PROVISION IN THE 4G ERA

To reap the economical and developmental benefits of competition, namely diversified service offerings and rapid technological evolution, the mobile value chain must be open so as to foster and harbor the participation of multiple new players, e.g., value added service providers, content providers, application developers, etc). These players will cooperate with the incumbent mobile operators to contribute additional value to the mobile service provision process but will also compete for the lion's share of user revenue.

A. Analyzing the "ABC" vision

In the 4G mobile communication era, a plethora of disparate services and multimedia applications will have to be flexibly yet efficiently deployed over a heterogeneous multi-network environment, raising service management requirements [5]. Nonetheless, mobile users will expect seamless global roaming across these different wireless networks and ubiquitous access to personalized applications and rich content via a universal and user-friendly interface.

In studying the implications of the heralded "Always Best Connected" vision of 4G mobile systems, we identify the notion of utility, implicitly embedded in the "best" adjective. Utility is a fundamental concept in microeconomic theory that concerns a typically continuous function representation of the consumer's preference relation over a set of commodities [6].

B. User utility issues

Users engage communication-based applications to realize various subjective benefits. These applications depend on the timely and orderly provision of network bearer services to exchange application-specific signaling and to move various classes of user information (e.g., image, video, corporate data) between communicating application endpoints. Inasmuch as the network is unable to provide the required levels of service, application will become dysfunctional and any user-perceived benefits of these applications will remain elusive, thus leading to a degraded user experience.

Performance of communication-based applications depends on the accommodation of QoS requirements for their native signaling and the exchange of arbitrary user information. From a network viewpoint, these factors translate to traffic flows with different QoS requirements that will – in principle – levy different charges, thereby decreasing user satisfaction. Thus, ensuring an adequate performance for communication-based applications so as to maximize user satisfaction, translates to honoring the QoS requirements of their traffic flows while minimizing the overall charges incurred, i.e., solving the user's utility maximization problem. Providers of network bearer services face the dual problem, i.e., maximizing revenue and minimizing network resource usage whilst meeting QoS requirements for all serviced traffic flows.

However, having the network meet the QoS requirements of communication-based applications, does not – necessarily – maximize user utility. Considering QoS as a multi-dimensional space, user utility is a diminishing function of quantity along each of the individual dimensions of QoS (e.g., packet delay). For communication-based applications that can operate on multiple QoS levels and content resolutions, requesting the highest QoS level possible does not necessarily increase user utility. For, in general, the higher the QoS level chosen by the application, the more network resources must be allocated to support it and the more costly the use of the network will be.

Given the multitude and diversity in the product offerings of the value chain participants, the technological complexity of the overall heterogeneous system and the IT illiteracy of the major consumer segment, it is understandable that most users will be unable to engage and coordinate such service provision matters all by themselves so as to maximize their utility. Consequently, some kind of intelligent mediation as part of the mobile service provision process should be introduced to efficiently cater for the utility-related aspects.

We believe that such mediation is a task that cannot – and should not – be undertaken by any of the aforementioned roles in the mobile value chain (e.g., value-added service provider, mobile network operator). For each of them will find interest in biasing a solution to his/her own preference – and monetary benefit of course. Thereupon, we claim that a trusted user delegate (e.g., intelligent agent) should always provide for the mediation between the value chain participants in providing services and applications, as well as for an unbiased solution to the user's utility maximization problem. Fundamentally, that constitutes emergence of new role in the value chain; a role that will maintain the customer relationship and provide the user with a universal roaming and service access capability whilst accommodating personal preferences, regardless of the access network(s) and terminal equipment in use.

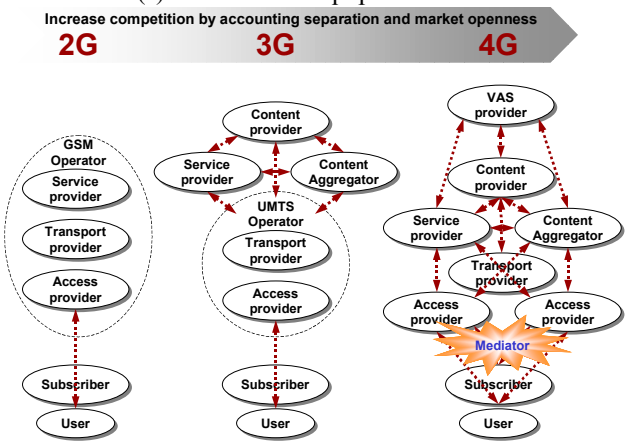

Figure 2. Evolution of the mobile value chain toward 4G.

Regardless of whether this new role will emerge through

fission from the incumbent mobile network operator role or through further evolution of existing MVNO approaches, one of its major tasks will be to provide billing services for its customer (i.e., the mobile user) by collecting related charging information from other players' equipment that is engaged in the mobile service provision process, correlating it and issuing a single itemized bill to the customer, thus fulfilling user requirements for one-stop billing. In addition, it will act as a clearinghouse, realizing accounting procedures that apportion revenue between the interested players according to bilateral or multilateral accounting agreements [7].

The intelligent mediation process could be part of a service provision platform [8] that mediates between independent application provider's and mobile network operator's domains to accomplish a flexible deployment model for value-added services and multimedia applications developed by the former over the network infrastructure managed by the latter. The alternative service management option is to impose several bilateral customer relationships between the user and all kinds of wireless access network operators he/she may contact when accessing any value-added service or application. However, that significantly complicates service provision by mandating the resolution of all technical issues (e.g., deployment details, activation preconditions, pricing structures, etc) on a per service/application provider-mobile network operator basis for each particular user – an approach that is clearly non-scalable.

For intelligent mediation to work, network and terminal domain functionality must be controllable by higher layer or third party entities besides the network and terminal entities engaging in protocol signaling related to the particular functionality. That is, *network control and management plane should be exposed* to higher layer entities (e.g., the "mediator" agent) and technologically agnostic interactions (e.g., IDL) that allow such higher layer entities to monitor and control network protocol signaling and thereby, overall network behavior, should be specified. Thus, rather than a concrete, all-encompassing architecture, a set of interworking approaches with standard technological solutions seems to be a more suitable approach for 4G. In the next section we outline the key technological capabilities that will characterize 4G mobile systems.

III. TECHNOLOGICAL SOLUTION SETS FOR 4G

A. Adaptable capability-aware service provision

To provide the mobile user with a consistent list of available applications that are supported by the mobile device he/she is currently employing, it is necessary to discover – and exploit – device capability information. Furthermore, given that wireless access networks differ significantly in terms of coverage area and supported bandwidth, mobile network capabilities (e.g. traffic load) should also be considered along with other important flavors of context information (e.g. user location and mobility patterns) so as to further refine the list of applicable services. In turn, that requires flexible representation formats (e.g., XML) and announcement procedures (e.g., CC/PP) like the ones adopted by the 3GPP MExE specification [9] as well as the employment of sophisticated user profiling mechanisms.

B. Transparent mobility and universal roaming capability

The variety of wireless access technologies that will coexist in the 4G mobile environment complicates the technical and regulatory aspects of roaming. Seamless user mobility across different wireless access technologies (e.g. WLAN, UMTS) with minimal or zero user intervention must be supported by efficient inter-system mobility management and handover procedures. To reduce signaling load, micro-mobility should be handled by the specific mobility management mechanisms of each wireless access network, while macro-mobility and roaming should built on cross-industry standard protocols and architectures, such as hierarchical Mobile IPv6 [10] and AAA [11]. Considering that handoff to a different system may entail different charges, it may be desirable to include QoS and pricing information as part of mobility management signaling.

C. Automated protocol configuration mechanisms

In the 4G era, the plethora of available – but disparate – applications combined with the existence of multiple wireless access systems and the need for a user profile-driven decision process, suggests that there may be multiple options capable of accommodating the same set of services. In example, a media stream can be transported to a mobile terminal by means of either a wireless LAN or a UMTS bearer service. However, the decision regarding which particular system to use depends on a number of factors, such as the respective cost of service, availability of network resources, radio link quality and user preferences. Continuing our previous example, we can imagine scenarios where end-to-end application signaling is routed via UMTS, because of its predictable performance and explicit QoS guarantees, while media streams are routed via a nearby WLAN to take advantage of its greater bandwidth capacity and lower cost. Since using different systems will result in accruing different charges, in 4G mobile environments service provision decisions must remain informed of users' pricing preferences.

D. Policy-based management and information models

Our previous point suggests that the system and protocol configuration procedures should be dynamic and automated to the highest degree possible but also open and flexible enough to facilitate higher-layer control over network bearer services. We find typical network management solutions (e.g., SNMP) as too narrow in scope, focusing on management of individual network elements within – rather than across – administrative domains. On the other hand, policy-based management [12] demarcates between enforcer entities and decision entities in the infrastructure, thereby allowing the realization of a flexible management architecture that spans across multiple administrative domains [13]. Furthermore, policy protocols support both outsourcing [14] and provisioning [15] modes of operation, making policy-based management an ideal approach for 4G mobile environments.

E. *Interoperable QoS management across different systems*

Currently, the prevalent QoS models for the IP protocol are Integrated Services (IntServ) [16] and Differentiated Services (DiffServ) [17]. Despite the differences in the scope of their design assumptions, control model and trade-off between accuracy and scalability, IntServ and DiffServ share a common subset of functionality, e.g., the traffic classification elements. It is possible to treat the common functional components of these QoS architectures as instances of a generic information model for network elements that provide QoS-aware treatment to IP packets [18]. That, in turn, would constitute part of a generic information model for an entire network infrastructure that provides an aggregate IP forwarding service. Combined with policy-based management, information models can provide a consistent view of network – and mobile terminal – functionality and facilitate its configuration and adaptation regardless of the particular technologies it is built upon. With a flexible open API (e.g. IDL) that exposes the functional features of the network infrastructure in a technologically opaque fashion and allows a (trusted) third party or application to control and coordinate the underlying network mechanisms, the realization of QoS management schemes distributed across multiple administrative domains becomes greatly simplified, especially in the case of heterogeneous infrastructures like 4G.

F. *Flexible pricing and billing mechanisms*

The clear demarcation between the network and the service domains that has been architected in existing 3rd generation systems suggests that any future-proof charging and billing architecture should portrait similar – if not greater – flexibility in its design. As a minimum requirement, network-related pricing models must be completely independent from service-related ones, with regard to formulation as well as application matters. In combination with policy-based management, information models support the interoperable specification of pricing models for specific domains (e.g., QoS-based model for the network domain) [19] and the configuration of affected components in the mobile network infrastructure, respectively.

G. *Application and mobile execution environment aspects*

To efficiently achieve mass scale deployment over millions of mobile terminals from different manufacturers and with disparate characteristics, application development must adopt the "write once, run anywhere" paradigm. Virtual machine that abstract differences in the operating system and the hardware platform promote a hassle-free application development, while interpreted languages lend themselves nicely to the restricted nature of mobile devices that may lack the resources required for a full compilation of a downloaded application. In addition, independent service providers will be relieved from the burden of developing, supporting and maintaining multiple versions of their applications for each possible client.

The execution environment at the mobile terminal should shield applications from mobility-induced events (e.g. change of IP address) in the underlying protocol stacks while enabling the realization of network-aware application behavior. Any API exposed to applications should refrain from using network related information fields (e.g. IP address, port numbers) in its class and method definitions. Information that may change due to network events (e.g., handover) should be handled internally via other libraries that are wrapped by the API that is visible to applications. Ideally, applications will use the transport API to instantiate so-called "flow" objects, allowing them to exchange information with their counterparts. These "flow" objects should be opaque with regard to the protocol details of the underlying connectivity service, thus shielding applications from undesirable network events (e.g. loss of transport socket connection). Such events should be handled by the execution environment that will send the appropriate notification to the application, allowing for the graceful termination of its active communication session, if necessary.

In addition, the mobile terminal must provide a local QoS Manager [9] to serve application requests for QoS treatment of their traffic flows. QoS signaling must proceed in a two-stage admission control; one at the mobile terminal, as it is a device limited in resources, and one at the intelligent mediation agent that will gather admission control decisions from the network agents of the wireless networks employed to carry the traffic flows of that particular application. Undoubtedly, there will be cases where the mobile terminal rather than the wireless interface is in scarcity of resources (e.g. drained power supply), so this approach can potentially minimize unnecessary signaling over the radio interface and preserve valuable bandwidth. Figure 3 below provides an illustration.

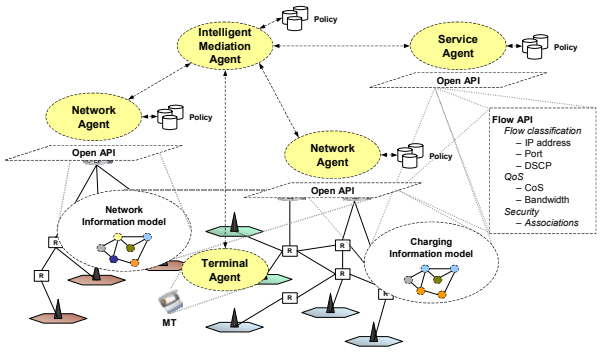

Figure 3. Intelligent mediation architecture for 4G.

In the above generic architecture, network agents advertise their bearer service offerings to the user's mediation agent. We consider bearer service offerings to be a combination of QoS level and pricing model supported and applied, respectively, by the wireless network represented by the network agent. We assume provision of wireless network bearer services under a wholesale model of peering agreements to adjacent network domains. Service agents refer to application functionality that may interact with mobile terminal agents and network agents for the purpose of QoS management and QoS adaptation. A service agent will inform the mobile terminal agent and the network agents of the QoS

requirements of their traffic flows and register an appropriate callback interface to support subsequent notifications and QoS adaptation for these traffic flows. Mobile network agents must abide by a common network information model so that an unambiguous definition of network bearer services is possible, e.g., when negotiating with other agents. Realizing that each agent represents the interests of a particular stakeholder in the mobile service provision (i.e., mobile network operator, application provider, mobile user), we assume that it operates under an individual policy. Notably, the resulting dynamics are that of an open market where different goods are advertised at widely known prices and selfish consumers may freely choose from a wide range of producers.

IV. Conclusions

Advances in mobile communication technologies have been rapid and their effects have frequently manifested themselves in ways and places far beyond the ones imagined by their inventors. Policy-based management and information model concepts, hierarchical Mobile IPv6 and AAA, flexible pricing and billing schemes, capability negotiation processes, and last but not least, open, technology-independent APIs are all important building blocks of 4G mobile systems. Properly combined, the aforementioned technologies can support a 4G system architecture that will be a far cry from its monolithic predecessors.

References

[1] J. Pereira, "Fourth generation – Beyond the hype, a new paradigm", IEE 3G Mobile Communication Technologies, March 28, 2001, London, United Kingdom.

[2] 3G TS 22.060 V5.0.0 (2001-10), "3GPP; Technical Specification Group Services and System Aspects; General Packet Radio Service (GPRS); Service Description, Stage 1", December 2001.

[3] ETSI TR 101 683 V1.1.1 (2000-02), "Broadband Radio Access Networks (BRAN); HIPERLAN Type 2; System Overview".

[4] M. Annoni, R. Hancock, T. Paila, E. Scarrone, R. Toenjes, L. Dell'Uomo, D. Wisely, "Radio access networks beyond 3G: A first comparison of architectures of 4 IST projects", Mobile Communications Summit 2001, 10 – 12 September 2001, Barcelona, Spain.

[5] N. Houssos, V. Gazis, S. Panagiotakis, S. Gessler, A. Schuelke, S. Quesnel, "Value added service management in 3G networks", Network Operations and Management Symposium (NOMS) 2001, 15 – 19 April 2002, Florence, Italy.

[6] A, Mas-Collel, M. D. Whinston and J. R. Green, "Microeconomic Theory", Oxford University Press, ISBN 0-19-510268-1.

[7] V. Gazis, N. Housos, A. Alonistioti, L. Merakos, "Evolving perspectives of 4G mobile communication systems", Personal Indoor Mobile Radio Communications (PIMRC) 2002, 15 – 18 September, Lisbon, Portugal.

[8] N. Alonistioti, V. Gazis, N. Houssos, S. Panagiotakis, "An Application platform for downloadable VASs provision to mobile users", IST Mobile Communication Summit 2000, 1 – 4 Octoberr 2000, Galway, Ireland.

[9] 3G TS 23.057 V4.4.0, "3GPP; Technical Specification Group Terminals; Mobile Station Application Execution Environment (MExE); Functional description, Stage 2", December 2001.

[10] INRIA Hierarchical Mobile IPv6 Technical Report.

[11] RFC 2904, "AAA Authorization Framework",

[12] "Primer on policy-based network management", http://www.hp.com/.

[13] M. L. Stevens et al., "Policy-based management for IP networks", Bell Labs Technical Journal, October – December 1999, pp. 75 – 94.

[14] RFC 2748, "The Common Open Policy Service (COPS) protocol".

[15] RFC 3084, "COPS usage for policy provisioning".

[16] RFC 1633, "Integrated Services in the Internet Architecture".

[17] RFC 2475, "An Architecture for Differentiated Services".

[18] Internet Draft, "Policy Framework QoS Information Model", work in progress.

[19] F. Hartanto, G. Carle, "Policy-based billing architecture for Internet Differentiated Services", IFIP Fifth International Conference on Broadband Communications, Hong-Kong, 10 – 12 November 1999.

Vangelis Gazis (gazis@di.uoa.gr) received his B.Sc. and M.Sc. (Communication Networks) from the Department of Informatics at the University of Athens, Greece in 1995, his and 1998, respectively, and his M.B.A. from the Athens University of Economics and Business in 2001. From 1995 until now, he has been with the research staff of the Communication Networks Laboratory (CNL) at the Department of Informatics in the field of mobile ad-hoc networks and cellular systems (MOBIVAS, ANWIRE). In parallel, he worked with a number of established companies in the IT sector as consultant. He is currently pursuing a Ph.D. in the Department of Informatics. His research interests include flexible and adaptable service provision, management of reconfigurable systems and services, quality of service, billing, and business model issues in 3G/4G mobile networks.

Nikos Housos (nhoussos@di.uoa.gr) obtained his B.Sc. degree in Informatics from the University of Athens, Greece in 1998 and his M.Sc. (with distinction) in Telematics (Communications & Software) from the department of Electronic and Electrical Engineering, University of Surrey, UK, in 1999. He is a staff member at the Communication Networks Laboratory of the University of Athens, working in the area of mobile service provision. He is involved in projects MOBIVAS (including workpackage leadership), ANWIRE and PoLoS of the European Union IST framework. He is also currently pursuing a Ph.D. at the Department of Informatics & Telecommunications, University of Athens. His current research interests relate to flexible value-added services provision in 3G/4G mobile communication networks and in particular to the design and implementation of service management and reconfiguration control platforms, intelligent service adaptation mechanisms, network reconfigurability and advanced business models. He has more than 15 publications in the above areas.

Nancy Alonistioti (nancy@di.uoa.gr) has a B.Sc. degree and a PhD degree in Informatics and Telecommunications (University of Athens). She had been working for 7 years at the Institute of Informatics and Telecommunications of NCSR "Demokritos". She has collaborated for one year as an expert at the Greek regulatory body and she is currently working as senior researcher and project manager in the Communication Networks Laboratory (CNL). She has participated in several national and European projects (RAINBOW, MOBIVAS, ANWIRE), undertaking technical management responsibilities. She specializes in mobile communications, reconfigurable systems and networks, adaptable service engineering, formal specification and testing of communication protocol and services, communications software engineering, with many publications in these areas. Her current research includes: reconfigurability and adaptability management, protocol/software download and open architectures and platforms.

Lazaros Merakos (merakos@di.uoa.gr) received his diploma in electrical and mechanical engineering from the National Technical University of Athens, Greece in 1978, and M.S. and Ph.D. degrees in electrical engineering from the State University of New York, Buffalo, in 1981 and 1984, respectively. From 1983 to 1986 he was on the faculty of electrical engineering and computer science at the University of Connecticut, Storrs. From 1986 to 1994 he was on the faculty of the Electrical and Computer Engineering Department at Northeastern University, Boston, Massachusetts. During the period 1993-1994 he served as director of the Communications and Digital Signal Processing Research Center at Northeastern University. During the summers of 1990 and 1991 he was a visiting scientist at the IBM T.J.Watson Research Center, Yorktown Heights, New York. In 1994 he joined the faculty of the University of Athens, Greece where he is presently a professor in the Department of Informatics and Telecommunications, and director of the Communication Networks Laboratory (UoA-CNL) and the Network Operations and Management Center. His research interests are in the design and performance analysis of broadband networks, and wireless/mobile communication systems and services. He has authored more than 140 papers in the above areas. Since 1995 he has led the research activities of UoA-CNL in the area of mobile communications, within the ACTS and IST framework programme.