♦ IP-Driven Access-Independent Resource Management in Converged Access Networks

Markus Bauer, Armin Dekorsy, Peter Schefczik, and Michael Soellner

A key requirement on future networks is to provide the user with seamless broadband access for triple-play services via any available access technology with high end-to-end quality of service. In this paper we propose a novel resource management concept on the network layer that utilizes service- and user-specific cross-layer metrics to control the resources in an all-Internet Protocol (IP) access network optimally via IP mechanisms such as routing and traffic engineering. We interpret any link, wireless or wireline, as an IP-hop in an all-IP network where major lower layer information is included in such metrics. This approach turns out to be access technology–independent, leading to a unified resource management concept in a "converged" wireline/wireless environment. With this proposal, we leverage flat IP-based cellular architectures like the Alcatel-Lucent Base Station Router (BSR) technologies and show how wireless and wireline access nodes can be integrated into future converged all-IP networks. The paper is based on results from ScaleNet, a European research initiative that is dedicated to advancing a vision of a scalable and flexible next-generation access network that provides an IP-optimized integration of heterogeneous access systems for future broadband multimedia services. © 2007 Alcatel-Lucent.

Introduction

Future communication systems have to support user needs more comprehensively, and they need to be designed in a more user-centric manner irrespective of the specific wireline or wireless access technologies involved. With upcoming fourth-generation (4G) systems, we expect two main characteristics: (1) wireless broadband access with peak data rates of 50 Mbps to 100 Mbps for nomadic, quasi-stationary to low-mobility users and (2) a horizontal integration of different heterogeneous wireless systems with comparable scales (e.g., cellular, wireless metropolitan area network (WMAN), hot-spot) via a common service backbone network with IP Multimedia System (IMS) functionality. With the anticipation of the upcoming trends toward a prevalence of content rich entertainment such as "mobile TV" or "multiplayer online gaming anywhere and anytime," it is obvious that the trend toward higher wireless data rates is still unbroken. In the home, the interconnection of multimedia equipment such as Internet Protocol television (IPTV), high-definition television (HDTV), and multichannel audio and multimedia recorders will be facilitated by auto-configured gigabit wireless local area networks (LANs) that are connected via high-speed broadband access to an IP-based network infrastructure. Obviously, fiber to the home

Bell Labs Technical Journal 12(2), 37–62 (2007) © 2007 Alcatel-Lucent. Published by Wiley Periodicals, Inc. Published online in Wiley InterScience (www.interscience.wiley.com). • DOI: 10.1002/bltj.20235

WILEY InterScience® DISCOVER SOMETHING GREAT

Panel 1. Abbreviations, Acronyms, and Terms

3G—Third generation
3GET—3G evolved technologies
3GPP—3rd Generation Partnership Project
3GPP-LTE—3GPP long-term evolution
4G—Fourth generation
AD—Access domain
AF—Application function
AN—Ambient network
ATM—Asynchronous transfer mode
BAN—Body area network
BGCF—Breakout gateway control function
BGF—Border gateway function
BPSK—Binary phase shift keying
BSR—Base station router
CQI—Channel quality indicator
CSCF—Call server control function
DSL—Digital subscriber line
ETSI—European Telecommunications Standards Institute
FMC—Fix-mobile convergence
FTP—File Transfer Protocol
FMC—Fixed mobile convergence
GLL—Generic link layer
HARQ—Hybrid automatic retransmission request
HDTV—High-definition television
HO—Handover
HSDPA—High-speed downlink packet access
IMS—IP Multimedia Subsystem
IP—Internet Protocol
IPRM—IP-based resource management
IPTV—IP television
ITU—International Telecommunication Union
ITU-T—ITU, Telecommunication Standardization Sector
L2T—Layer 2 termination function
LAN—Local area network
LPDP—Local policy decision point
LTE—Long-term evolution
MAC—Medium access control
MIMO—Multiple input, multiple output
MIP—Mobile IP
MM—Mobility management
NACF—Network attachment control functions
NAPT—Network address port translation
NASS—Network attachment subsystem
NAT—Network address translation
NE—Network element
NGN—Next-generation network
OA&M—Operations, administration, and maintenance
OASIS—Organization for the Advancement of Structured Information Standards
OSA—Open service architecture
OSPF—Open shortest path first
PAN—Personal area network
PD-FE—Policy decision functional entity
PDP—Policy decision point
PE-FE—Policy enforcement functional entity
PEP—Policy enforcement point
PES—PSTN/ISDN emulation subsystem
PIN—Policy ignorant node
PIP—Policy information point
PR—Policy repository
PU—Periodic update
QoS—Quality of service
QPSK—Quadrature phase shift keying
RACF—Resource and admission control functions
RACS—Resource and admission control subsystem
RC—Resource control
RCEF—Resource control enforcement function
RLC—Radio link control
RM—Resource management
RRC—Radio resource control
SAE/LTE—System architecture evolution/long-term evolution
SCF—Service control functions
SIP—Session Initiation Protocol
SN—ScaleNet
SPDF—Service-based policy decision function
TD—Timeout delay
TISPAN—Telecommunication and Internet Converged Services and Protocols for Advanced Networking
TRC-FE—Transport resource control functional entity
TRE-FE—Transport resource enforcement functional entity
TV—Television
UE—User equipment
UMTS—Universal Mobile Telecommunications System
UR—Unified router
VDSL—Very-high-speed digital subscriber line
VNW—Virtual network
VoIP—Voice over IP
W-CDMA—Wideband code division multiple access
WiMAX—Worldwide Interoperability for Microwave Access
WLAN—Wireless local area network
WMAN—Wireless metropolitan area network
WR—Wireless access router

or at least enhanced very-high-speed digital subscriber line (VDSL) connections would be a viable solution. However, we do not necessarily presume a 100 percent availability of wired broadband access in any region at any home, because of the related immense installation costs. To overcome this issue, powerful fixed wireless broadband solutions that can be interconnected to the home environment as well as to high-speed cellular infrastructures might be applied. Since highly efficient high-speed wireless technologies for mobile users with peak data rates of 100 Mbps and beyond are limited in their range to about a few hundred meters, backhaul networks for these cellular infrastructures are faced with the same availability problems as the distribution of broadband wireline networks based on VDSL with comparable ranges.

Recent discussions in standardization [4], industrial trends [19, 29, 31], and ongoing research on the evolution of cellular wireless communication show a strong trend toward applying IP-based principles and striving for "flatter" network architectures [6, 8, 10, 22, 23, 28, 36].

Hence, as a further step beyond these trends, we envision an actual convergence by means of a full combination and integration of wireless systems with different scales, e.g., multihop wireless mesh backhaul, cellular hot spots, home personal area networks (PANs), and body area networks (BANs), and wireline broadband access systems. From that, we anticipate that the concept of a fully converged access network integrating the control functions and transport capabilities for both fixed access and mobile access appears to be an attractive future evolution step.

Such a converged access network based upon a common IP-based infrastructure provides ubiquitous broadband access for fixed, nomadic, and mobile users and could help to overcome the digital divide between urban and rural regions by a flexible combination of wireline and wireless backhaul and access. The challenge to design truly converged access networks is addressed in European government-funded research programs like the German project 3G Evolved Technologies (3GET) [1] and especially in the ScaleNet project [35] launched in 2005. This topic is also expected to play a key role within the setup of the European 7th Framework Research Projects (FP7), for example, especially in the "eMobility" Technology Platform [14].

To meet the demands for converged networks, resource management mechanisms coordinating several different access technologies play a key role for providing services with high end-to-end quality of service (QoS). Current approaches, such as that in the European FP6 Ambient Networks [5, 17] project, propose a generic link layer (GLL) concept by tightly integrating multiradio resource mechanisms into wireless access network resource management concepts at the radio link control (RLC) or medium access control (MAC) layer. The drawback of such a tight coupling is less flexibility toward an extension with new access systems, a high computational effort on the RLC or MAC layer, and a cost-intensive and hierarchical network infrastructure.

As opposed to this approach, we propose in this paper a novel resource management concept that leverages existing IP principles and hence enables an IP-optimized integration of different access technologies. We call this concept IP-based resource management (IPRM). Integrating wireless access systems is based on the view that every link—wireless as well as wireline—has to be interpreted as an IP link in a common all-IP network. We then employ IP-based mechanisms such as routing or traffic engineering and appropriate protocols for controlling the resources in such an all-IP access network. Moreover, with IPRM the control of resources is performed on the network layer rather than on the RLC or MAC layer.

In such an all-IP access network, each wireline or wireless network link is characterized by its transmission characteristics or link state information such as available bandwidth (throughput), error performance, and transmission delay. Thus, an open and challenging task for controlling resources with IP-based mechanisms is to evaluate reliable metrics that capture main characteristics of the different access systems, including the first, possibly wireless, link to the end user, over a reasonable period. Hence, we abstract from each access system and exploit lower layer information on the network layer representing a cross-layer approach. The distributed downloading

concept that is also introduced by Bell Labs is an example of exploiting lower layer information on the transport layer while for IPRM, the cross-layer metrics combine the link, MAC, and RLC layers.

The paper is organized as follows: We first introduce the concept of IP-based resource management and describe its main features as well as the underlying network architecture. After a discussion of the benefits of IPRM, we discuss technological challenges and proposed solutions, which mainly address the issues of network architecture and scalability, handling of mobile users by means of routing protocols and algorithms, and the dynamics of the wireless link metrics. We then present an approach to the IPRM concept within the context of an ongoing next-generation network (NGN) standardization framework in ITU-T. After that, we show how user mobility can be handled by applying IP routing principles based on generic link metrics, and we discuss first performance simulation results. The problem of time-variant link metrics, as experienced in wireless systems, is addressed next. We derive a method for the prediction of "higher-layer" link metrics based on "cross-layer" considerations that are governed by the underlying wireless packet transmission. The theoretical models presented are derived for high-speed download packet access (HSDPA) transmission in Universal Mobile Telecommunications System (UMTS) networks but are similarly applicable to next-generation wireless broadband access networks. The concluding section provides an outlook on future work.

IP-Based Resource Management

Service convergence as a type of fixed-mobile convergence (FMC) has recently emerged as a consideration in designing future communication systems in a user-centric manner. To provide service convergence, research and system design activities have focused on the integration of different wireless and fixed broadband access systems via a common IP-based service platform (like IMS, IPTV, or any Web-enabled platform) in the core backbone network. **Figure 1** provides a schematic comparison of the convergence and resource control architecture as realized in today's systems and in our view of next-generation "converged" access.

In Figure 1a, we show a common service platform in the operator core network together with a common resource control functionality that is connected to different system-specific access platforms. This architecture ensures service convergence as it is usually seen today. The common service platform and the access-specific platforms cooperate with the underlying transport networks on the management of resources. The different access networks (e.g., UMTS, WiMAX, and DSL) are seen as separate transport networks. As Figure 1a indicates, in this approach to convergence, the diverse transport networks still perform their own access technology dependent dedicated resource management (RM) and mobility management (MM) functions based on specific protocol architectures tied to the underlying access technologies. As an example, a UMTS transport network is built upon a specific UMTS hierarchical architecture [23] where specific protocols and concepts manage the UMTS resources and mobility. This ensures high QoS delivery within the UMTS network. However, any interaction (e.g., seamless mobility or intersystem handover) between the UMTS transport network and any other access network can become unnecessarily complex, because system-specific interfaces and procedures are required on either side and must be specifically designed, standardized, and implemented for the involved transport networks. Moreover, additional exchange of information with the core network entities may be required, for instance, in case of nomadic mobility.

This can be observed in current and ongoing 3rd Generation Partnership Project (3GPP*) standardization activities to provide seamless mobility between a legacy UMTS transport network and a future system architecture evolution (SAE)/long-term evolution (LTE) network [4]. Obviously, any interaction between different networks via interfaces strongly impacts both future and currently deployed networks. As a consequence, the approach of service convergence across heterogeneous access technologies as depicted in Figure 1a turns out to be quite involved, strongly impacts standardization efforts, and, finally, leads to unnecessarily high network costs and efforts because of duplication of system-specific functions like mobility

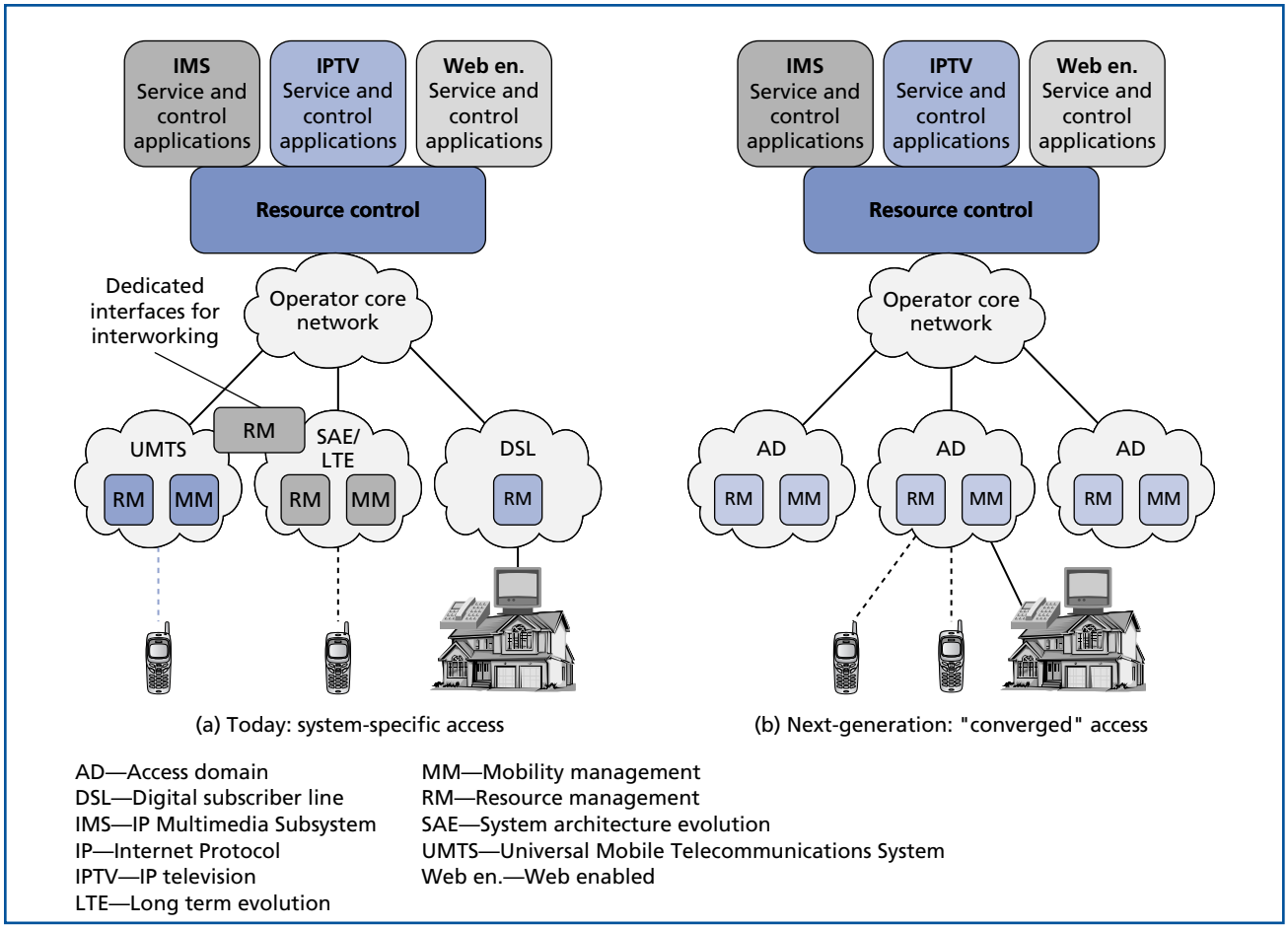


Figure 1.
Comparison of resource control architectures.

management and resource management rather than using common functionalities.

Taking the idea of convergence a step beyond IMS to its natural conclusion, we envision convergence by means of an integration of the different transport networks. Thus, we do not aim for service convergence only, but also for network convergence: that is, leveraging the benefits of a common "converged" infrastructure for transport and control planes.

Figure 1b shows the converged resource control architecture, as we introduce it in this paper. Compared with Figure 1a, the distinction between the different access systems is pushed further out to the edge of the access networks, thus making use of a common transport plane with common IP-based functionalities. Hence, the approach discussed herein preserves the service platform, such as IMS, and core resource control (RC) functionality—succinctly, we still facilitate service convergence—but we additionally enable transport convergence by an actual integration of the different transport networks. In the following discussions, we denote this type of convergence as access network convergence. With full-blown access network convergence we no longer differentiate between the various transport networks. Instead, we assume that access domains (ADs) of one type and resources within an access domain are controlled by an access-independent and unified resource and mobility management. Figure 1b shows that we apply the same type of resource and mobility management architecture within every access domain. Because of the convergence of fixed, mobile, and wireless

systems, we use IP mechanisms as a unifying method for controlling the resources and the mobility in each access domain. In particular, the main idea is to apply IP-based protocols, routing concepts, addressing techniques, and traffic engineering algorithms to manage an access domain. Up to this point, we have referred to this approach as IP-based resource management (IPRM). In the following sections, we will explain the key characteristics of IPRM and elaborate the related opportunities for next-generation converged access networks.

To explain IP-based resource management, let us start by defining any link in an access domain as an IP hop. Consequently, highly error-prone and dynamic wireless links are also IP hops. We further assume that within each access domain, IP routers are connected via wireline technologies while the last hop to end user terminals is either wireless or wireline, depending on the access technique applied. Thus, we obtain an all-IP access network topology encompassing wireline as well as wireless IP links.

As indicated, we use IP mechanisms such as routing protocols, routing algorithms, and traffic engineering to control the resources for this all-IP access network. Up to this point, the network layer performs resource and mobility management. This is in contrast to resource management on the radio resource control (RRC) or MAC layer, as proposed in [30] and references therein. **Figure 2** shows a schematic view of the embedding of IPRM. As shown, IPRM is a control plane technique that exploits information in order to control packet forwarding on the user plane.

The information that IPRM exploits is provided by access-independent metrics. These metrics capture the main characteristics of lower layer access technologies. Typical examples are throughput, delay, and transmission quality, all representing QoS parameters. Since IPRM operates on the network level, the metrics merely gather performance measures based on layer 1 techniques (e.g., modulation, forward-error detection, and multiaccess schemes) and layer 2 methods, such as hybrid automatic repeat request (HARQ) and scheduling. Thus, by introducing IPRM, wireline and wireless systems only differ in their layer 1 and layer 2 transmission schemes. The higher layer protocols—the access network architecture, and even the policy framework—are unique and independent of whether the user communicates via wireless or wireline transmissions. Compared to simple approaches in wireline networks that just use hop counts as distance metrics to compare and optimize routes, wireless link metrics typically vary over time because of instabilities such as fading or interference of wireless transmission, and/or terminal movement causing variable radio link path losses. Moreover, wireless links between end user terminals and certain access routers may completely break down, whereas others are newly established. So the underlying potential network topology is changing continually. Given these link metrics, IPRM routing protocols are used to propagate these user- and service-specific metrics throughout the access domain. Finally, IPRM traffic engineering algorithms exploit these metrics and calculate appropriate service-dependent routes for efficient packet forwarding in the access domain.

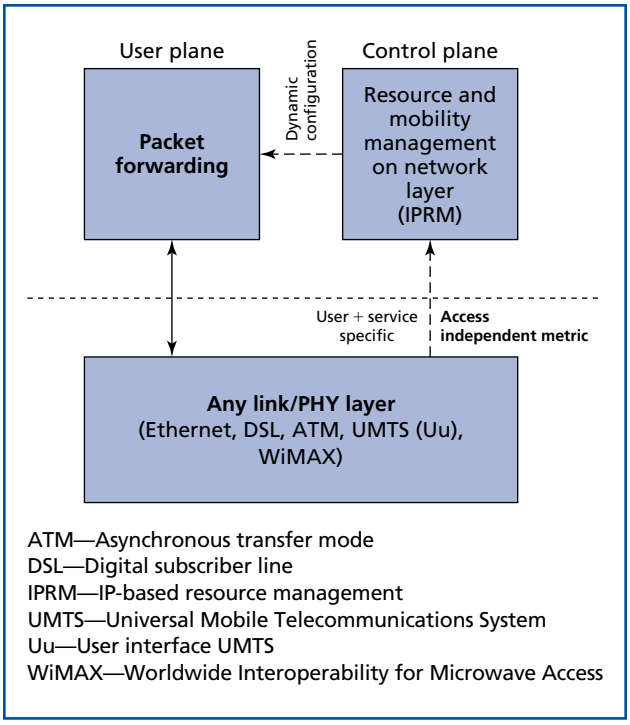


Figure 2.
Schematic view.

As opposed to the promising opportunities of the unified IPRM approach, there are still various challenges to overcome when introducing IPRM as an IP-driven, access-independent resource management architecture. First, IPRM is associated with an all-IP flat

architecture and a policy framework. Both have to be standards compliant and we have to ensure a smooth integration of IPRM into standard architecture and policy frameworks. We will describe later how to integrate IPRM into the broader NGN framework of ITU-T [26] as an example for related NGN standardization activities by the European Telecommunication Standards Institute (ETSI) Telecommunications and Internet Converged Services and Protocols for Advanced Networks (TISPAN) committee or 3GPP-LTE.

Next, the design of appropriate protocols to deal efficiently with highly dynamic wireless links proves to be a major research challenge. We started our investigations by considering protocols as used today in wireline or ad hoc networks. First results on this topic are given in the section, "IPRM and Mobile Routing."

The third decisive challenge is the evaluation of reliable metrics. Our approach is described in the "Estimating Wireless Link Metrics" section, where we focus on the example of deriving the throughput of a packet-based transmission with HARQ protocol and proportional fair scheduling from lower layer characteristics. Results are given for the specific case of UMTS-HSDPA.

Network Evolution

IPRM supports access network convergence on the network layer, as seen in Figure 2, as it allows the use of generic IP-driven networking and routing functions for all underlying access technologies. Access systems may involve different layer 1 and possibly layer 2 technologies, even in a heterogeneous manner. Strictly speaking, in the future, access systems will differ only in lower layer techniques while today's access systems (like UMTS, WIMAX, and DSL) often are specifically designed to include higher layer protocols. For cellular radio systems, this seems feasible since flat architectures like the Alcatel-Lucent BSR concept have been proved to be viable [6].

As a further step in network evolution, we assume that for network routers, interface or line cards will become available supporting a specific radio or air interface technology on a single board including layer 1 and layer 2 functionalities. This is under consideration for certain radio technologies like WiMAX and even seems feasible in the near future for conventional cellular technologies based on 3G and beyond, as achievement in other Bell Labs projects show [18]. This wireless development will be comparable with today's wired line cards, supporting fiber or copper transmission technologies. The interface of such a board will deliver performance measures of the connected air link and exchange these metrics with the network layer and forwarding module. Using IP-based networking and routing technologies, just by simply plugging these cards into the network node, the network can be extended by new radio interface technologies. IPRM ensures seamless interworking to already available access technologies without the need of introducing new standardized and specific interfaces. Moreover, by means of IPRM, seamless support of different wireless link technologies is achieved without any impact on the access techniques already deployed.

Concurrent Access, Load Balancing, and Handover

As parts of IPRM, traffic engineering algorithms are used to calculate the specific path through the network for each service given a network utilization function and service constraints (QoS requirements). Moreover, predefined policies can be taken into account to calculate the optimal routes and distribution of traffic flows. We assume a multiple commodity network flow problem where each network node can send data to many destinations and receive data from many resources. Each commodity is assigned a metric, and multiple metrics are associated with a single network link. This generic approach facilitates concurrent access where, as an example, a Voice over Internet Protocol (VoIP) call is transmitted via UMTS as a File Transfer Protocol (FTP) download is simultaneously forwarded via the wireless local area network (WLAN) to the end user terminal. Furthermore, packets belonging to the same service can be delivered via different network paths by splitting a stream across multiple paths. With IPRM, we are able to evaluate both the percentage split and the splitting point at the router and thus achieve load balancing across the network and the available wireless paths in particular.

Finally, since we employ traffic engineering methods, mobility management or handover execution is just part of the traffic engineering algorithms, as the triggered search for optimal routes includes the

switch-over from a low-quality to a high-quality link at a certain point in time. This allows for seamless mobility without any further interworking via specifically designed interfaces between access systems. However, these decisions have to be based on instantaneous measurements and observed metrics, and they are obviously time-critical. Later in this paper, we will show how to adapt standard protocols to overcome possible performance and inconsistency issues due to terminal mobility.

Distributed MIMO Technologies

If we assume that we are able to manage multipath routing in an access domain, then another major wireless technology comes into play: multiantenna multiple input multiple output (MIMO) technology. With this technology, several encoded data streams are transmitted simultaneously via different antennas, most of which are no longer collocated at a site, and can be separated or decoded by spatial signal processing algorithms at the mobile terminal [20, 37]. As a result, we obtain a higher throughput, a more robust transmission, and/or we can reduce transmission delay. At present, there is an increasing research interest at the physical layer on distributed and cooperative MIMO techniques, as seen in [13] and [32] and references therein. While past and ongoing research on multiantenna technologies has mostly focused on the wireless link layers between an access point such as a base station and a mobile station using collocated antenna sites, in the future, a network level technology like IPRM will be required to provide time-critical control of data streams if multiple antennas are distributed over network nodes.

Moving Networks

In the IPRM context, mobile terminal nodes as well as access nodes can perform the same IPRM functionalities such as routing and traffic engineering. In general, they will be able to host multiple network interfaces that may use the same or different link technologies. In addition, both node types use the same IP network and subnetwork addressing scheme. Hence, from this point of view, a mobile terminal can serve not only as an end system, but also as a subnetwork representing a "moving network."

Potential of IPRM

From the introduction, we expect that IPRM as an IP-driven resource management concept has a powerful potential to be realized in the next-generation network context. This is based on the following benefits:

- Open and flat architecture that allows simple and scalable extension by any access technology, whether wireline or wireless, and
  - No impact on architecture,
  - No impact on the resource management concept,
  - Simple exploitation of new metrics associated with new access technology, and
  - Easy incorporation of policies, such as those defined by network operators for their resource usage.
- Opportunity for a unifying technology-independent resource management platform that also enables the convergence of heterogeneous access networks including IP-based mobility support.
- Common user- and service-specific metrics, which enable flexible network routing by multiple paths or flows, diversity, and/or multiplexing.
- Simultaneous multiservice access via different wireless link technologies (e.g., voice via UMTS and FTP via 802.x).

In the following sections, we will discuss details of proposed solutions that address some issues still under study to ensure and improve the acceptance of the IPRM concept. These include the embedding of IPRM into ongoing ITU-T NGN activities, the impact of user mobility on routing performance, the prediction of dynamic wireless link metrics, and the scalability issues raised by propagation of the variable link metric.

Integration of IPRM in NGN Frameworks

We next describe a general architecture classification that is characteristic of most of the current next-generation network concepts under study. Moreover, we describe how IPRM fits into and augments these concepts. We also give an example how the IPRM concepts could be combined with the ITU-T NGN framework and describe the resulting access network infrastructure.

Resource Management Architecture Classification

In [38], the author provides an overview and description of several standard network architectures with respect to resource management and gives a general classification of resource control architectures. All those architectures are considered as having a service plane and a decoupled transport plane. To deliver an assured service, designers must solve the coordination problem between those two planes. The resource control (RC) platform is the functional entity between those two planes that solves the coordination problem. This leaves the service plane with its application functions (AFs) on top of the RC doing the signaling, for example, based on Session Initiation Protocol (SIP). The RC also provides instructions to the network elements in the transport plane, so that certain resources can be configured, reserved, and enforced for the service-related user packets. As an example, the latter can be realized by the policy enforcement point (PEP) inside the network elements (NEs) of the transport plane. **Figure 3** provides an overview. The classification described in [38] also includes other policy architectures such as the as ITU-T resource admission and control function (RACF) and ETSI TISPAN resource admission control subsystem (RACS) [15, 16].

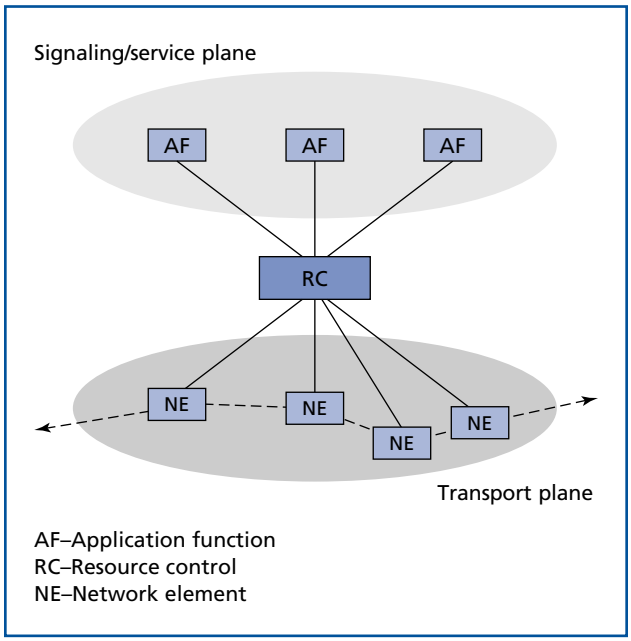


Figure 3.
General resource control architecture classification.

Though the policy-based approach is widely adopted in principle, it is still under discussion in various standards organizations. Beneficially, the policy-based methodology shows potential to be used in conjunction with IPRM. On the one hand, it can cooperate with IPRM in various areas including QoS handling, mobility, dynamic interworking between heterogeneous networks, routing, resource management, admission control, and traffic engineering. On the other hand, the policy-based method allows further concepts to be implemented, such as rules for user preferences and profiles, operations administration and maintenance (OA&M), security, content and service provisioning, charging, and accounting [3].

Network Architecture Overview

As described, IPRM is a key factor to introduce the concept of network convergence. This is done in an integrative approach though IPRM does not demand the introduction of a new network architecture concept. Instead, it establishes its concept of network convergence in a seamless manner within the already defined and ongoing NGN activities. Such smooth integration into NGN is ensured by adopting and utilizing the NGN's general resource control architecture depicted in Figure 3. Consistently with the general ideas of this architecture, IPRM supports the decoupling of the service and transport plane just as it builds upon on the idea of the resource control QoS mediator between the service and the transport plane.

In that framework, IPRM is realized by deploying its dedicated resource and mobility control mechanisms within the transport plane, fully transparent to the service plane and QoS mediator. From the NGN perspective IPRM is one dedicated implementation of the transport plane control functions in the access network, and we show how IPRM can be integrated with the current ITU-T NGN [26] architecture.

In **Figure 4** the general network setup of an access network that supports IPRM in the transport plane is depicted. The access network architecture is flat and IP-based. The architecture is composed of a number of IPRM-enabled network nodes, which are the unified router, the unified transport node, and the aggregation router. Despite different naming, we introduce the concept of a unified IPRM-enabled access

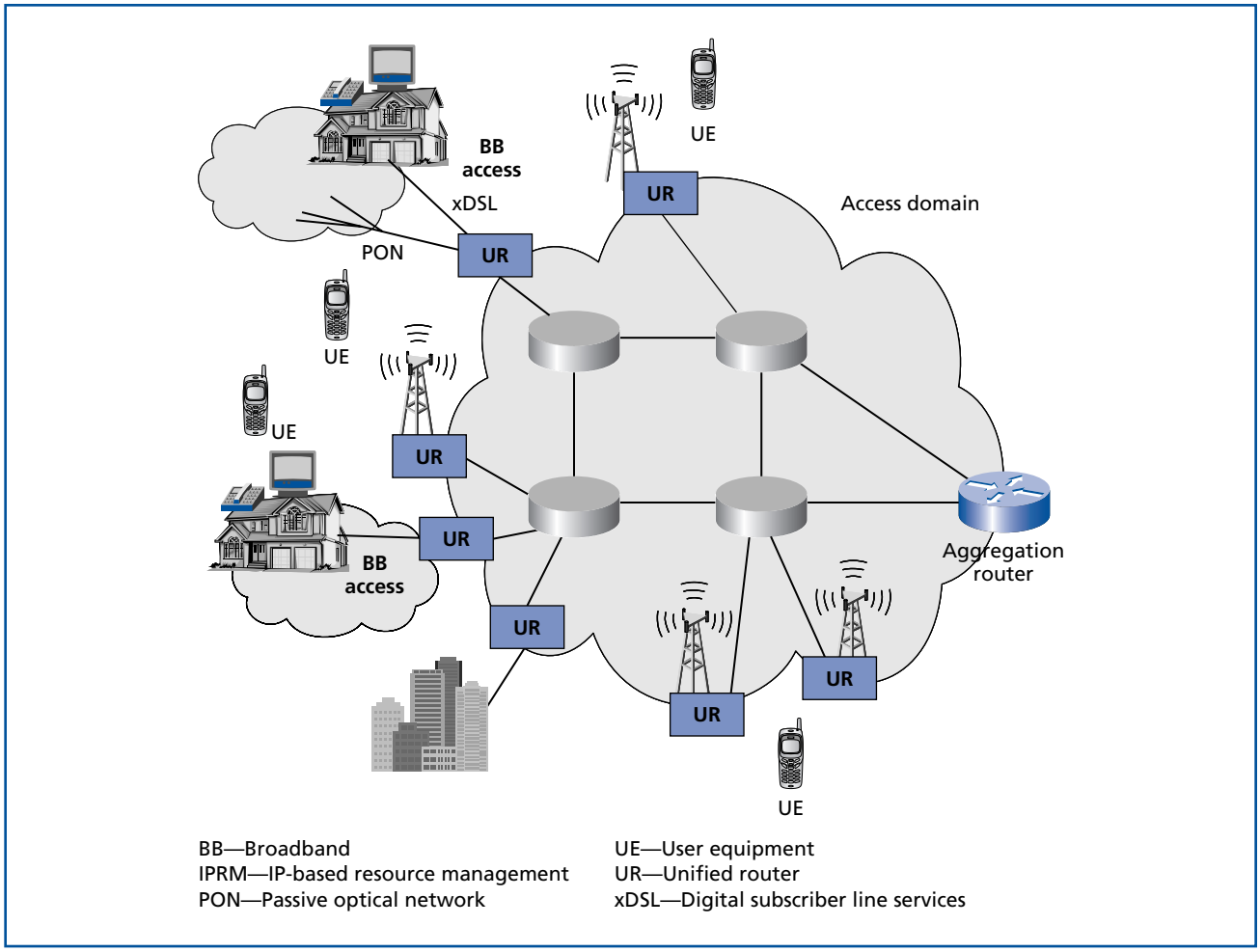


Figure 4.
IPRM enabled access network architecture.

network node common to each of these nodes. The general idea of this concept is that each of these nodes provides the same unified IPRM functionality (i.e., software modules). In such a way, we deploy a unified scheme that is responsible for access domain–wide resource and mobility management. Note that the underlying transport technology is fully transparent to the network nodes' resource and mobility management. All transport technology dependent functionality is handled on access modules that can be attached to any IPRM-enabled unified network node. Access module–specific resource management is executed exclusively on the link layer. Metrics are exchanged between the unified network element and the access modules being used by IPRM for access-independent resource and mobility management.

A unified network element has the capability to operate a number of access modules. Note that a unified network element is capable of supporting both wireline and wireless access modules (e.g., Ethernet, ATM, WiMAX, or UMTS). Any access module connected to a unified network element provides an "adaptation" function. This function maps a link technology dependent "link state metric" into a link technology independent "QoS metric." IP routing protocols and routing algorithms exploit these QoS metrics to perform link layer independent resource management as described. Via the metrics, any access module transport technology specific interface functionality is hidden from the IPRM in the unified network element. As a consequence, wireline and wireless network modules are treated equally by IPRM and

therefore the wireless link becomes "just another IP hop."

Mobility Management

Mobility management is performed in an access network–independent way in conjunction with IPRM. We differentiate three levels of mobility management:

1. Session handling (SIP-based, e.g., IMS),
2. Macromobility management based on MIPv6, and
3. Micromobility handling based on IPRM.

That means we propose to use Mobile IPv6 mechanisms between access networks, while within an access network, IPRM is used. Thereby the access networks can be totally different in nature, for instance, wireless or wireline, and can be based on different technologies. It should be noted that IPRM-based mobility management only takes place at the edge of the network and uses metrics that measure the access technology involved. Thus we expect that link layer technology can synchronize a handover via the IP-driven access domain if needed. All three levels of mobility management are independent of underlying link layer technology.

Macromobility between such administrative IP domains can be based on a centralized "mobility anchor" scheme as depicted in **Figure 5**. If a moving network is simultaneously connected to two "administrative

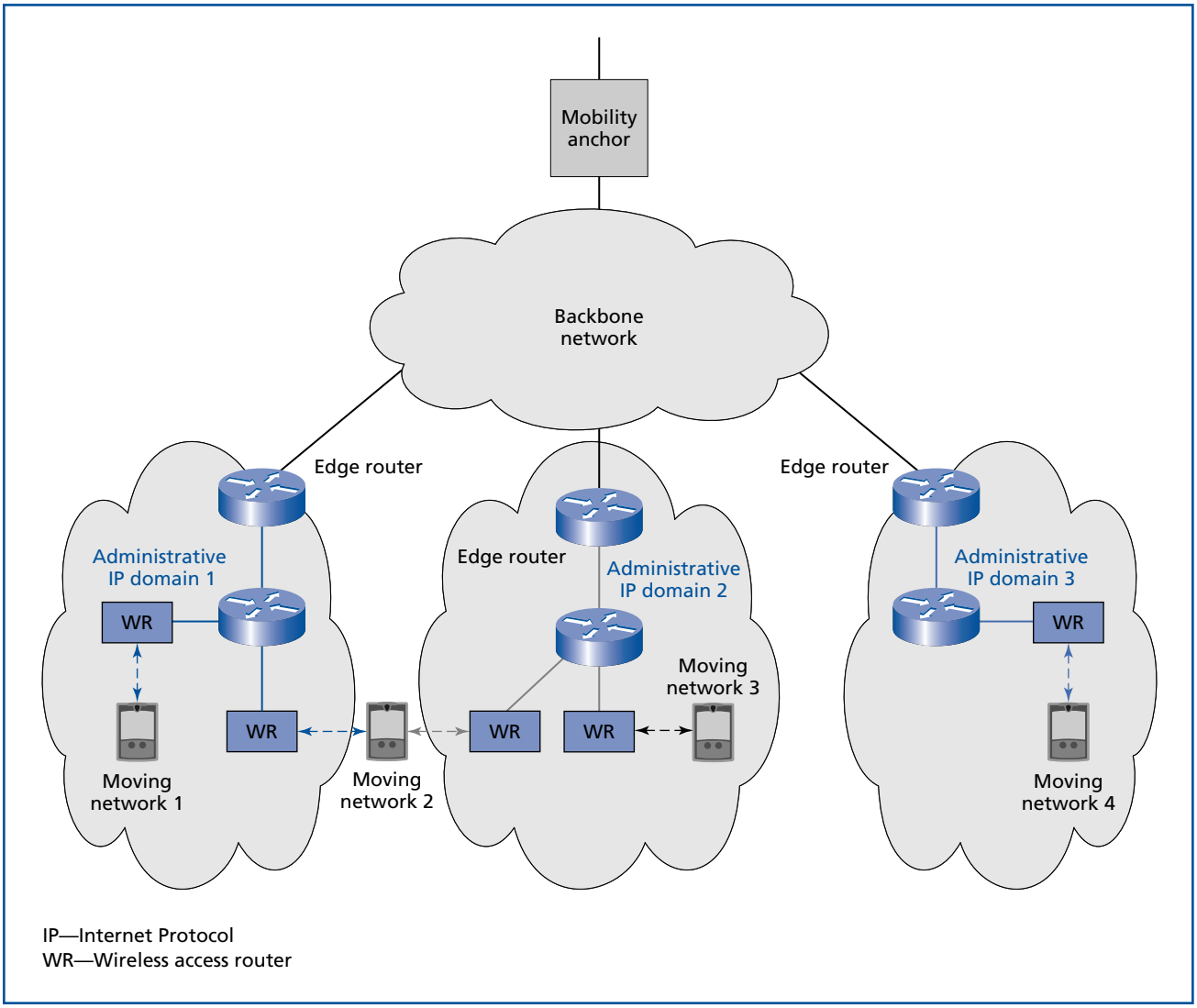


Figure 5.
Clustering of access domains.

IP domains," as shown for moving network 2 in Figure 5, an appropriate split of traffic in the core network is required. Most likely this split can be achieved in the mobility anchor. "Multihoming in Mobile IPv6" [27] and "distributed downloading" are the first candidate concepts to be investigated in more detail.

Clustering Concept

In this section, we discuss the scalability properties of the IPRM-based concept. For the IPRM concept, user- and service-specific metrics have been defined to capture the main link characteristics dynamically. As routing algorithms jointly optimize the routes in the wireline and wireless parts of the network based on these metrics, the link metrics information must be distributed through the IP network. In addition, in contrast to today's traffic engineering approaches, a significant amount of metrics data has to be distributed throughout the network, as the IPRM introduces metrics per user and per service on each link. Furthermore, as we use the metrics to capture the dynamics of the wireless links, the corresponding metrics must be updated to provide up-to-date information to the routing and traffic engineering algorithms. Finally, the IPRM operates within a network that dynamically changes its topology at the edges of the network, since the mobile terminal is considered as a general node within the all-IP network. Topology changes are caused by frequent setup and removal of links between wireless access routers (WRs) and "moving networks." To reflect this dynamic change of the network topology, link metrics and topology information of dynamically upcoming and vanishing links must be distributed whenever such topology changes occur.

In summary, IPRM requires distributing a significant amount of metrics and topology information throughout the entire IP network due to:

- User- and service-specific metrics per link,
- Wireless link metrics reflecting varying link performance, and the
- Dynamic changes of network topology due to wireless link setups and releases.

Routing protocols are used to distribute this information. As routing and traffic engineering algorithms always need up-to-date metrics details, the amount of metrics information transferred between the involved routers within the converged IP network is significantly higher than that of pure wireline networks.

To overcome the scalability issues introduced through our routing protocol-based IPRM approach, we introduced the concept of clustering the entire access network into a number of disjoint administrative domains, as shown in Figure 5. An edge router is used as an aggregation and gateway router that interconnects the operators' core network with a single administrative domain. Each of these domains performs its own and dedicated IP-based resource management. Therefore, we limit the required frequent updates and the processing of metrics and topology changes to the network nodes that are part of one administrative domain. The edge router hides all IPRM-related dynamics from the core network and the remainder of the access network. Consequently, there is no impact on any network node beyond the scope of the administrative domain, and no impact on the remaining part of the access network. This ensures the required scalability of the envisioned routing protocol-based IPRM approach.

IPRM and the ITU-T Resource and Admission Control Function Concept

In this section we show how the IPRM-based concept fits into the next-generation networks concept of the ITU-T. The general principles and a reference model for next-generation networks are laid down in [25]. Therein the key cornerstone of NGN characteristics is the separation of the services stratum from the transport stratum, allowing them to be offered separately and to evolve independently. For both the NGN service stratum and the NGN transport stratum, the general architectural concepts of data or user plane, control plane, and management plane can be logically identified.

ITU-T's NGN initiative conceived a generic resource and admission control functional (RACF) architecture in [26]. The functional architecture and its functional entities and relevant reference points are depicted in **Figure 6**.

The RACF is designed to offer real time, application-driven, and policy-based transport resource management for a wide range of services and for a variety

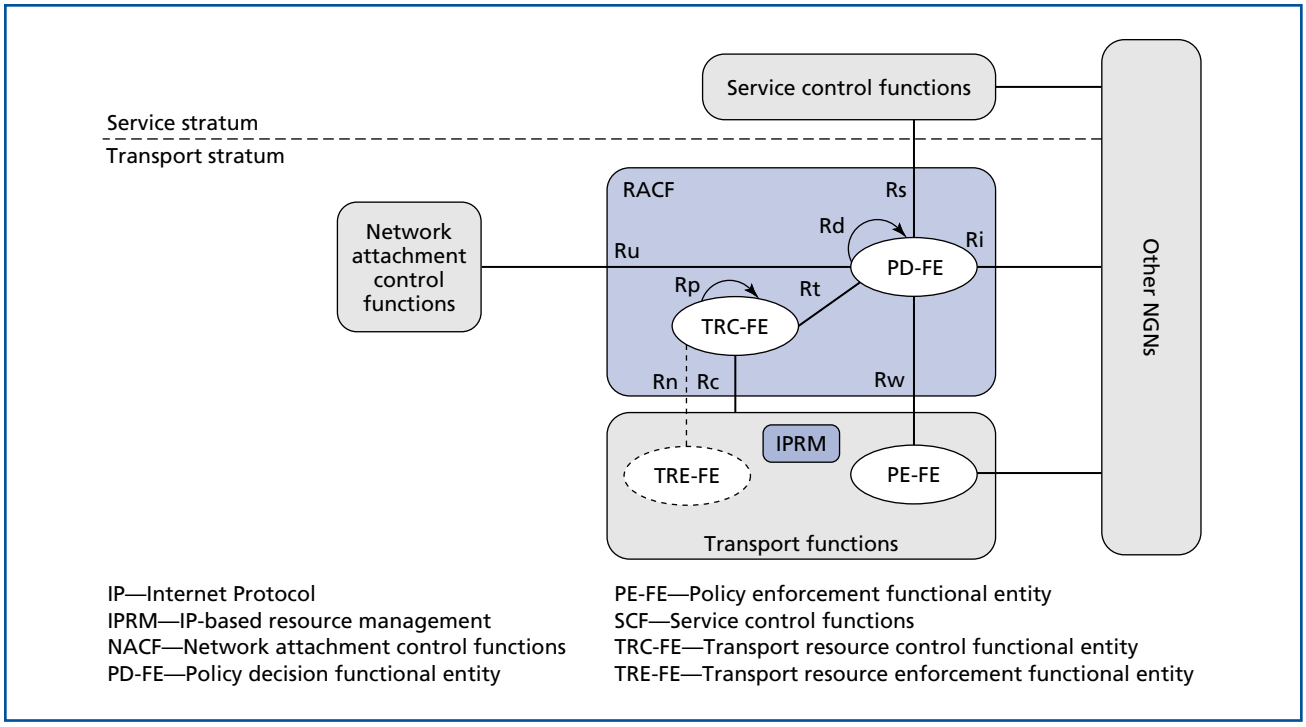


Figure 6.
ITU-T RACF functional architecture.

of transport technologies both fixed and mobile. This involves a variety of access and core transport technologies and supports end-to-end quality of service (QoS), gate control, network address translation (NAT), network address port translation (NAPT), firewall control, bandwidth reservation and allocation, packet filtering; traffic classification, marking, policing, and priority handling. By this means, the RACF is not service-specific and services can make use of RACF whether the IP Multimedia Subsystem is involved or not. The RACF interfaces with the service control functions (SCFs), the network attachment control functions (NACFs), and the transport functions. The RACF itself contains the following policy entities:

- Policy decision functional entity (PD-FE),
- Transport resource control functional entity (TRC-FE),
- Transport resource enforcement functional entity (TRE-FE), and
- Policy enforcement functional entity (PE-FE).

At the request of the SCF, the PD-FE performs the final policy decision, which is based on the information from the SCF and the resource information gathered by the TRC-FE. The TRC-FE thereby supports the monitoring of domainwide resources, allowing the RACF to work within a core network. The TRE-FE in the transport stratum enforces the transport resource policy rules instructed by the TRC-FE. The PE-FE enforces the network policy rules instructed by the PD-FE on a per-subscriber and per-IP flow basis.

Furthermore, the RACF QoS resource control supports both a push and a pull mode for exchanging QoS resource request information among the RACF, the service control functions, and the transport functions.

Concerning the general classification given in [38] the RACF constitutes the resource control that sits in between the service control functions (service plane) and the functional entities of the transport function (transport plane). As Figure 1 is totally inline with that general classification, this lets us integrate the RACF architecture with the IPRM system architecture. The policy-based concept thereby can be exploited, thus enabling coordinated control, monitoring, and management of heterogeneous networks, their services, and their resources. Those are all

complicated tasks today. The same is true for the deployment, installation, and configuration of network applications and user services. Policy-based networking is one technique seen to remedy the situation for the complex communication networks in the future. This includes policies for handling QoS, security, mobility, dynamic interworking between heterogeneous networks, and content and service provisioning. Other applications for policies are policy-based routing and policy-based resource management (e.g., IPRM), policy-based traffic engineering, policy rules for user preferences, charging and accounting (e.g., "always-best connection"), admission control, service requirements, user profile, terminal status, resource availability, and auto- and self-configuration.

IPRM and Mobile Routing

One of the key features of IPRM is the mobility and handover handling by routing protocols. We first give an overview of the basic IPRM handover concept. In addition, we show in detail how a distance-vector routing protocol could be used for handover management and we describe how this protocol's most important parameters affect handover performance. On the basis of these investigations, we present solutions around how to modify existing routing protocols to adapt them better to the "mobile routing" functionality.

Basic IPRM Handover Concept

To be able to handle handover execution via routing protocols, we define a schematic architecture of an IPRM-enabled terminal as shown in **Figure 7**. The main IPRM function within the terminal is the capability to monitor the state and the metrics of the underlying external links, to run the IPRM protocols and algorithms, and to route packets between the internal interfaces and one or more external interfaces. External interfaces connect the terminals' internal network interfaces with different IP subnets of the access network via any physical access technology provided by the terminals' external interfaces. Though this is a generic architecture for any terminal with multiple interfaces, obviously it becomes more relevant when applied to mobile terminals. Figure 7 shows two external interfaces, which connect the terminal to two IP-subnets, NW1 and NW2, located in the access network. Note that the physical transport technology used on those external interfaces can be of any wireline or wireless technology. The IPRM routing functionality interconnects the external IP subnetworks NW1 and NW2 with an internal virtual network VNW via the terminals' internal network interfaces.

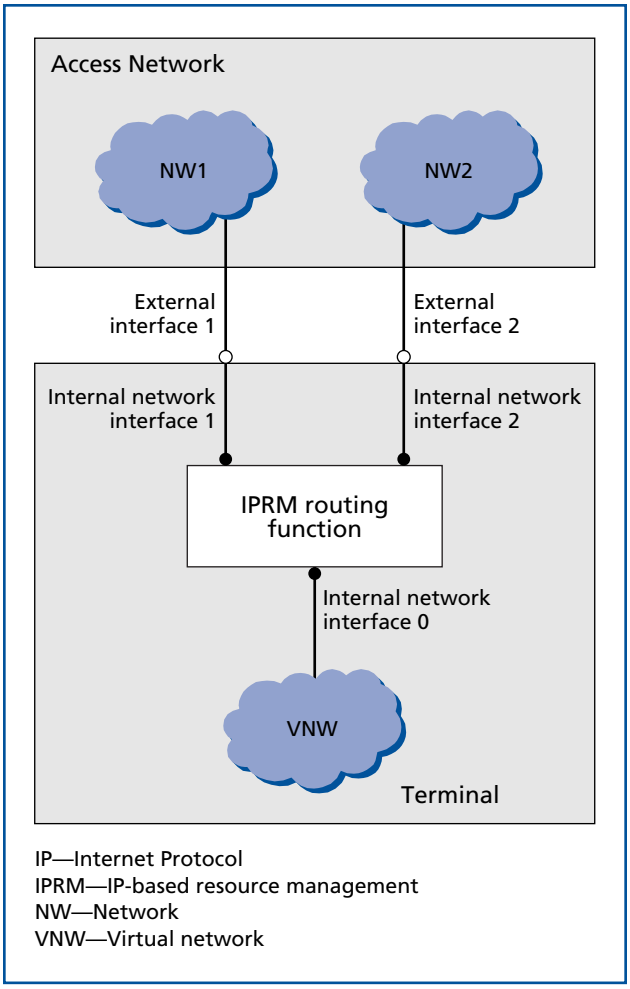


Figure 7.
Schematic terminal architecture.

For external communication, any application on the terminal will use an IP address from the address space of the virtual network. This ensures that the terminal's application IP address is preserved even if interfaces (e.g., interface 1 or interface 2) change the network they are connected to and accordingly change their IP addresses.

A unified router access module that supports one dedicated access technology (e.g., UMTS or WiMAX) makes up one IP-subnetwork to which IPRM-enabled mobile terminals can connect. When performing a hard handover from one unified router to another, the mobile will remove its interface from the IP-subnetwork of the old unified router and free the corresponding IP address. Subsequently, it connects its interface to the IP-subnetwork of the access module of the new unified router to which it moves. Accordingly, the new unified router assigns an IP address from the access module's address space to the mobile terminal's interface. Routing protocols detect the new point of attachment for the mobile terminal's interface, the IPRM traffic engineering algorithm recalculates the best routes in the network, and the routing protocol distributes this information within the network and adjusts the forwarding plane of each IPRM-enabled unified network element accordingly.

From the IPRM routing protocol perspective, each handover results in a network topology change as the mobile terminal's virtual network's point of attachment is moved from an old to a new unified router. Further, as handover execution becomes a common procedure in the converged networks of the future, the routing protocols we intend to use have to deal with a constantly changing network topology.

Mobile Routing

Classical fixed network routing protocols are designed and optimized to maintain routes in a more or less stable network topology. They are tuned to deal with moderate network dynamics like the rare event of a link breakdown or the addition of a new router to the network topology. They are not designed to maintain seamless connectivity within a constantly changing network topology. This takes us to the question whether routing protocols could deal with network dynamics we expect from our IPRM mobility management scheme. Handling mobility with routing protocols is a very ambitious task. Nevertheless, there are several peculiarities that help us to overcome the challenges we face. First of all, there is the fact that the network topology changes imposed due to hard handover execution always occur at the very edge of the network topology. It is very likely that such changes at the edge of the network reflect only in limited impact on routes at the edge of the network and do not reflect in route changes within the whole network. Furthermore, only one network (i.e., the mobile terminal's virtual network) is moved during handover execution. As no other terminal uses this network for routing its traffic, there is no impact on the traffic handling of other terminals.

Second, IPRM is used to manage mobility in the access part of a future converged next-generation network. Such an aggregation network is expected to be built up from a tree-shaped network topology. Clearly, network topology changes at the leaves of such a tree-shaped network can be handled quite efficiently with routing protocols. One can expect comparable performance to that of a hierarchical MobileIP mobility management approach. Nevertheless, the IPRM-based scheme can exploit and benefit from all advantages described previously in the section, IPRM.

Third, with the introduction of the access domain (AD), we already defined a network structuring element that helps to overcome the issues imposed by the constantly changing network topology. With the AD, we can limit the network size that the IPRM routing protocols have to keep in sync, as described in the clustering concept. In addition, IPRM routing protocol design can benefit from ad hoc routing protocol concepts that are currently the focus of intense research. These protocols are specifically designed to deal with constantly changing network topologies, and we expect many of the concepts developed for ad hoc routing also to be applicable for the IPRM concept.

Despite the perfect fit of ad hoc routing protocol concepts for IPRM, at the first glance, we expect a solution that comprises ideas from both fixed network and ad hoc routing protocols. The reason is that ad hoc routing protocols are designed and optimized for a constantly changing wireless network topology, unlike in the network we address with IPRM. Such an access network is based on a largely fixed topology with dynamic network "topology changes" occurring only at the very edge.

Routing Protocol Managed Handover: An Example

In this section we describe in detail how routing protocols could be used to perform handover

management. We do these investigations on the basis of a simple distance-vector routing protocol implementing the Distributed Bellman-Ford algorithm [7]. The most important parameters affecting protocol performance are the following:

- *Periodic Update Interval* (PU). Full routing tables are exchanged between adjacent routers every PU seconds.
- *Timeout Delay* (TD). When a PU for a dedicated route has not been received for TD seconds, the expectation is that the route is no longer available and it is marked as invalid in the IP forwarding table.

Adjacent routers exchange two kinds of messages to keep their routing protocol databases synchronized. First, there are periodic update messages sent out every PU seconds. With a periodic update, a router forwards its full database to each directly connected/adjacent router. These messages are used for keeping track of the network topology in order to detect topology changes. Second, there are triggered update messages, which a router sends out whenever it detects that the network topology has changed, and therefore that the router's database has changed. This kind of message is used to distribute network topology changes throughout the whole network quickly.

A simple network setup as depicted in **Figure 8** was chosen for initial investigations on the achievable performance of a routing protocol managed handover.

In this setup, the user equipment, that is, the wireless terminal designated as UE, is moving between two unified routers, UR1 and UR2. During handover execution, the mobile terminal's interface is disconnected from the old unified router's IP subnetwork and connected to the new router's IP subnetwork. This is done by freeing the terminal's IP address and tearing down layer 2/1 connectivity to the old UR. Subsequently, layer 2/1 connectivity is built up to the new UR and an IP address from the new UR's address space is assigned to the terminal's interface. For our initial investigations, we rely on the unchanged routing protocol behavior to detect the tear-down and set-up of layer 2/1 connectivity to the old and the new unified router so as to adjust

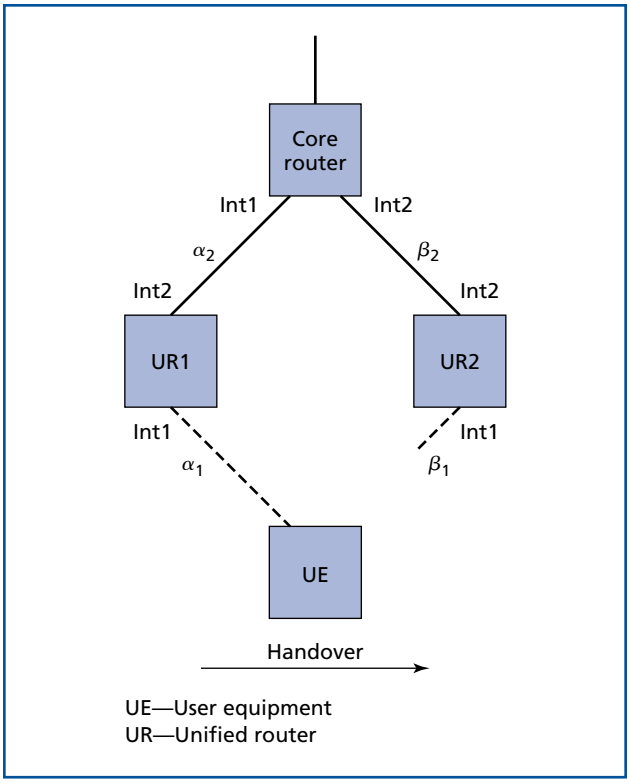


Figure 8.
Network setup.

the forwarding plane to the new network topology provoked by the handover. To provide a key indicator for achievable handover performance, we analyzed the mean handover time a terminal experiences during handover execution. Mean handover time is defined as the time from tear-down of layer 2/1 connectivity to the old UR until the IP forwarding plane has adapted to the topology change and the terminal can receive downlink data again.

For the investigations, we assume an ideal network with infinite link capacity and no processing delay for routing protocol messages on the various network elements involved as we are interested in handover performance solely invoked as a result of routing protocol behavior.

Figure 9 shows the timing diagram that illustrates the activities in the network during the handover.

The top lines depict the way the layer 2/1 connectivity of the mobile terminal's wireless interface goes down to the old and goes up to the new UR.

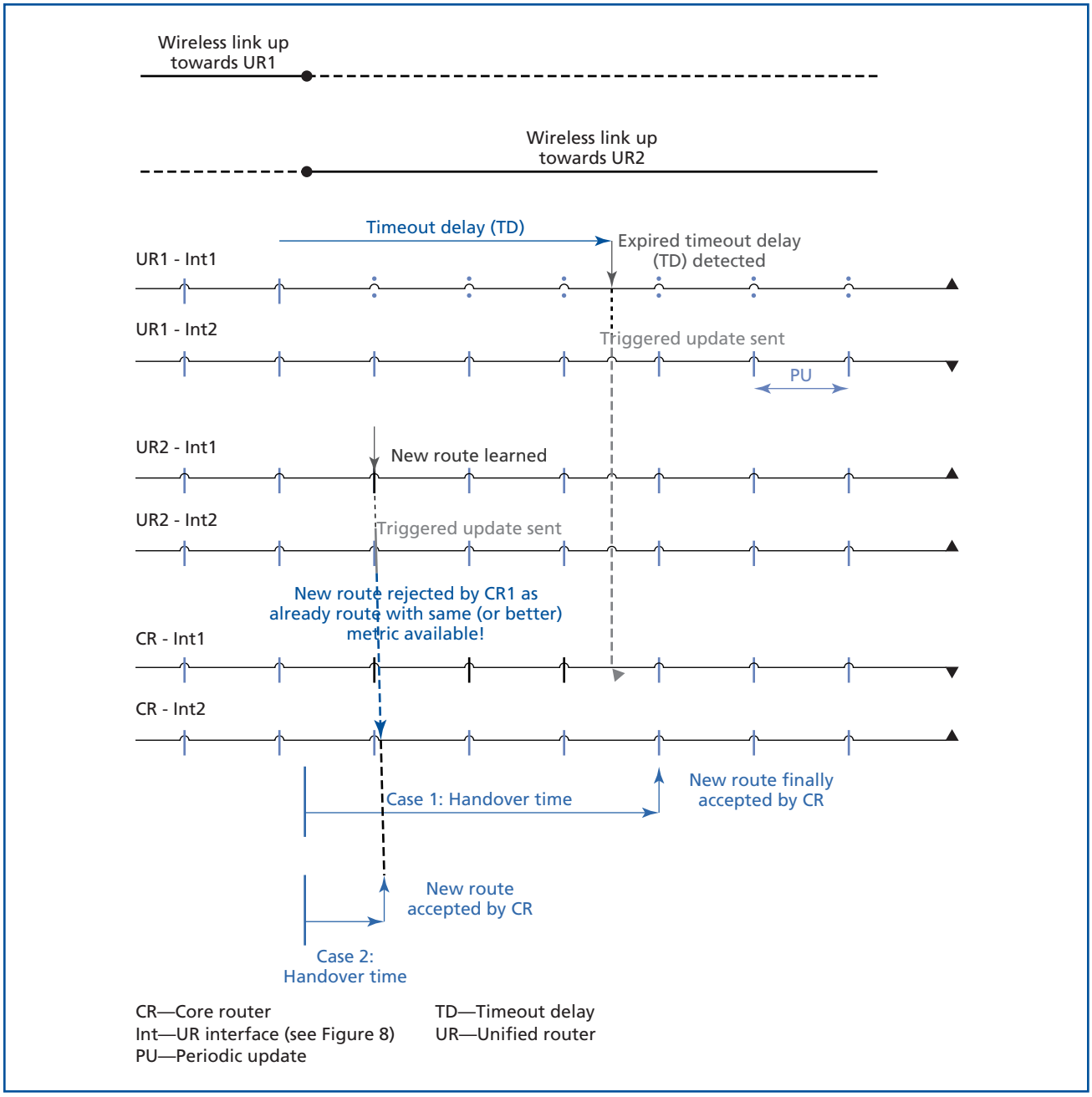


Figure 9.
Timing diagram.

Beneath that, the timing diagram depicts the activities happening on the various interfaces of UR1, UR2, and the core router during handover execution. The timelines are normalized to the occurrence of periodic updates on the various interfaces. At the start of handover execution, the mobile terminal tears down its connectivity to UR1. As a consequence, UR1 no longer receives the terminal's periodic updates on interface 1, and UR1 therefore does not reset the TD timer for the terminal's VNW entry in its routing protocol database.

The mobile device has already established its connectivity to UR2 when it sends out the next PU. Therefore, UR2 receives this PU on its wireless

interface. Note that since the mobile terminal's VNW has changed its point of attachment, cleaning up the terminal's routing database is critical. All entries but those regarding the terminal's own interfaces must be removed. This is necessary, for otherwise the mobile terminal would distribute with this PU the network topology that was valid before handover execution. If this were not done, the terminal would relay the information that it can reach, such as the UR1 wireless interface. But this is no longer true, as the link to UR1 is already shut down. From this PU, UR2 learns that it can reach the mobile terminal's virtual network via the terminal's interface 1 as indicated by "new route learned" in Figure 9. This new information is stored in the routing protocol database of UR2 and the IP forwarding table is adjusted accordingly on UR2. As UR2 has just learned that the network topology has changed, a triggered update is sent out on interface 2 to the core router. From this triggered update, the core router learns about the VNW connected to UR2 via the mobile device. The core router checks this information against its routing protocol database and detects that there is already an entry for the terminal's VNW that was valid when the mobile was connected to UR1 before executing the handover. The core router still assumes it can reach the VNW via UR1, since the TD timer on UR1 has not yet expired, and therefore UR1 has up to now not detected that it has lost connectivity to the mobile terminal via interface 1. Consequently, it has not updated the core router.

Two cases have to be distinguished. They differ in the value of metrics the core router holds for the old route to the VNW: $\alpha_1 + \alpha_2$ in its database and the metrics it receives for the new route to the VNW via the triggered update from UR2 $\beta_1 + \beta_2$.

In the first case, the new route has a higher or at least the same metrics as the terminal's old route. In Figure 8, this is the case if the mobile moves from UR1 to UR2 with metrics $\alpha_1 + \alpha_2 \leq \beta_1 + \beta_2$. As the routing protocol algorithm performs least cost routing, that is, the route with the lowest cost is chosen, the core router ignores the new route advertised by UR2 as it assumes the old route to be the better one. Even if UR2 constantly reports the new route via regular updates every PU seconds to the core router, this information is ignored. This changes when the TD timer on UR1 expires. At that time, the UR assumes that the mobile device and the associated VNW are no longer available. UR1 updates its routing protocol database and the IP forwarding table accordingly, and a triggered update is sent out via interface 2. The core router receives this update and consequently removes the VNW from its database and IP forwarding table. Note that the new route to the mobile is still not known by the core router. Only with the next regular update sent by UR2 does the core router add the VNW to its database and set up the IP forwarding table, as this is now a new route that has to be learned. At this point, the mobile handover is finished solely by routing protocol mechanisms. If we assume that the start of any handover is evenly distributed within a PU timeslot, the mean handover (HO) time can be calculated as follows:

$$meanHOtime = \left[Round\left[\frac{TD}{PU} + \frac{1}{2} \right] PU - \frac{1}{2} PU \right]$$

In the second case, that is, the metrics of the new route to the VNW, the core router receives, via UR2s, a triggered update that has a lower value than the old route in its database, that is, $\alpha_1 + \alpha_2 > \beta_1 + \beta_2$. Therefore the core router at once updates its database and the corresponding IP forwarding table since it has received a route to the VNW with "better" metrics. If we again assume an evenly distributed start of any handover within a PU timeslot, the mean HO time can be calculated as follows:

$$meanHOtime = \left[\frac{1}{2} PU \right]$$

With the assumptions we made, the mean HO time for the investigated link-state Bellman-Ford routing Protocol can be described by the following term,

$$meanHOtime = \theta \left[\frac{1}{2} PU \right] + (1 - \theta) \left[Round\left[\frac{TD}{PU} + \frac{1}{2} \right] PU - \frac{1}{2} PU \right] \quad (1)$$

with θ the relative frequency of the probability of $\alpha_1 + \alpha_2 > \beta_1 + \beta_2$.

On the basis of the example of a distance-vector routing protocol, we have shown that in principle it is possible to perform a handover via routing protocols based on our IPRM mobility management concept. We have also identified how important protocol parameters (i.e., PU and TD) affect achievable handover performance, and, from the investigations, we can also derive fundamental issues regarding performance and scalability. The major issues we face can be summarized as follows:

- Handover detection by routing protocol, and
- Database handling on IPRM enabled network elements.

In the following, we give a more detailed explanation on these two issues and set the direction for further studies.

Handover detection by routing protocol. Routing protocols deploy their dedicated layer 3 mechanisms to detect available routing peers. They rely on the exchange of routing protocol messages to detect routing peers and to build up associations with these routing peers. Examples include the PU messages of distance-vector routing protocols, or the exchange of "hello messages" used by the OSPF link state routing protocol to build up an association with peer entities. Such a mechanism always entails a trade-off between performance and control plane overhead. Consequently if we want to exploit such routing protocol mechanisms for our handover management we have to deal with this trade-off. This issue can easily be seen in the distance-vector routing protocol example. Equation 1 indicates that handover performance can be improved by decreasing the TD and PU interval. Unfortunately, this reflects a rampant increase of routing protocol control plane traffic in the network. As this overhead clearly is not tolerable for PU intervals that give sufficient handover performance, we can conclude that we cannot rely on such mechanisms for effective handover management.

A well-known concept to overcome this issue is the exploitation of lower layer information. We can use this information on the IPRM-enabled network elements that are involved in the handover to adjust the routing protocol databases to the movement of a terminal quickly. There is dedicated information related to the handover available on the terminal and on the involved URs. In the following, we give an example of how such lower layer information could be exploited in the preceding distance-vector routing protocol. We used layer 2/1 information (i.e., the information when the terminal's wireless link goes down and up during handover execution) to break up the handover performance dependency from PU. This can be achieved by exploiting available layer 2/1 information on UR1 and on the mobile terminal during handover execution.

First, UR1 should send out a triggered update to the core router when its connection to the mobile terminal is going down at start of handover. This triggered update communicates that the terminal's VNW is no longer reachable, and consequently the core router immediately removes the network from its database and IP forwarding table. Second, the mobile terminal should send out a triggered update as soon as layer 2/1 connectivity to UR2 is established. This ensures that the core router will quickly learn about the new attachment of the terminal's VNW. Note that the core router will now accept this initial triggered update, as the terminal's old route was already removed via the triggered update initiated by UR1. Network simulations on these enhancements indicate achievable handover performance sufficient to effect seamless mobility.

Database handling on IPRM-enabled network elements. The second issue that needs to be addressed is the amount of database state information and the associated routing protocol overhead each IPRM-enabled network element must handle. If we use a flat routing protocol approach throughout an AD, each terminal's VNW has to be maintained in the routing protocol's database on each IPRM-enabled network element. Moreover, each protocol-relevant change in a terminal's protocol state (e.g., a network topology change due to handover or metrics variation on the access link) must be distributed throughout the whole AD. Consequently we have to deal with two issues, (1) how to keep the size of the routing protocol database on the network elements in

reasonable size, and (2) how to keep the databases efficiently synchronized throughout the AD.

We approach these issues with a routing concept that was designed for a special class of ad hoc networks called reconfigurable wireless networks [33]. This concept is a hybrid framework for routing introducing two routing zones for any terminal. One routing zone utilizes a proactive routing protocol, while the other utilizes a reactive routing protocol. Proactive protocols constantly keep the routes within a routing zone up to date, so that when packets must be forwarded, the required routes are already established. In contrast, reactive protocols calculate and establish routes only when packets for a dedicated destination must be forwarded.

We can apply this scheme to the IPRM resource management in the following way: There is a proactive terminal-specific routing domain that is composed of the network elements located in the neighborhood of the terminal. Within this domain we use a proactive routing concept that ensures that the terminal can fully exploit the benefits of the IPRM resource management scheme. Beyond this proactive routing zone we use a reactive routing protocol. This ensures that we have limited knowledge of the full dynamics of the mobile terminal's routing protocol state only to the IPRM-enabled network elements within the proactive routing domain. Within the reactive routing domain, only network elements actively involved in communication with that mobile terminal have to maintain the terminal's network state in their routing protocol databases. This ensures that we do not need to maintain the terminal's state information throughout the whole domain. Instead, we limit this knowledge to a few network elements within the routing domain that are those most likely needed to serve the mobile terminal. In this way, we also limit protocol traffic overhead, since terminal mobility is now limited to the proactive routing domain.

Estimating Wireless Link Metrics

In this section we present one example of metrics that can be exploited by the described IPRM concept in order to route traffic in a flat all-IP network. In this paper, we focus on the highly error-prone wireless link and are interested in the long-term throughput the wireless link can provide. The wireless access system performs hybrid automatic repeat request (HARQ) processing [39] on the MAC layer and performs adaptive modulation and coding on the physical layer. Both techniques are commonly used to obtain robust and high data rate transmissions, as in HSDPA [2, 23]. For this evaluation, we follow an information theoretic approach based on two major concepts: (1) the renewal-reward theory and (2) the cut-off rate of digital transmission. Out of our scope are practical impacts such as parameter estimation variances or parsimonious channel state information, which are open tasks for further work. The paper recalls the theory described in [9]. Further information on the proposed metrics is available in [11] and in [12], where the evaluation approach of [9] has been extended to MIMO-HARQ transmission.

As introduced earlier, we use the renewal theory [39] to evaluate the throughput of a wireless transmission if HARQ processing is applied. The analysis is valid under the following four assumptions:

1. Infinite buffer occupancy for each user. Thus, as soon as a packet is successfully delivered at the receiver, the next packet is encoded and transmitted in the next available transmission interval.
2. Error- and delay-free feedback channels.
3. Each user accesses the channel with an average probability $\bar{p}$. This means that on each transmission interval, each user transmits a message with probability $\bar{p}$. This approach enables modeling of scheduling strategies.
4. Each user runs its own HARQ protocol independently of the other users likely to transmit.

Let us define the (average) throughput per user as

$$\eta = \lim_{l \to \infty} \frac{R(l)}{l} \text{ [b/s/Hz]} \tag{2}$$

with $R(l)$ as the sum of the number of bits per second per Hertz (mutual information) successfully decoded up to time slot l. Note that the halt of packet transmission is a recurrent random event associated with the random variable R. By applying the renewal-reward theorem [39] we obtain for

$$\eta = \frac{E\{R\}}{E\{T\}} \tag{3}$$

where R is the random reward with $R = R$ b/s/Hz if transmission stops as a result of successful decoding. If the transmission of the packet fails, we have $R = 0$ b/s/Hz. The random variable T describes the random time between two consecutive occurrences of the recurrent event and is called interrenewal time.

Furthermore, we assume a maximal number of transmissions per packet denoted as M. Transmission of a current packet stops if (1) the packet is successfully decoded at the m-th transmission with $m \leq M$ or (2) no successful decoding occurs up to the maximal number of transmissions with $m = M$. In the following, we denote $p(m)$ as the probability of unsuccessful decoding of a packet in the m-th transmission.

In [9] the authors provide a closed-form expression of equation 3 by computing $E\{R\}$ and $E\{T\}$. Because of size restrictions, we skip the detailed evaluation and simply recall the result in [9]. With [9], the per user throughput can be expressed by

$$\eta = \overline{p}\ \frac{R(1 - p(M))}{\sum_{m=0}^{M-1} p(m)} \tag{4}$$

with $p(0) = 1$.

For the ALOHA protocol $p(m) = p(1)^m$, which leads to the expression $\eta = R(1 - p(1))$ if the user always accesses the channel ($\overline{p} = 1$). Equation 4 is indicative of a cross-layer throughput metric. On the one hand, the physical layer is captured by the probability of unsuccessful decoding $p(m)$. On the other hand, parameter $\overline{p}$ represents the medium access control layer since scheduling a number of users all likely to access the channel is applied in the MAC. In the following, we provide examples of how to evaluate the probability of unsuccessful decoding $p(m)$ and the channel access probability $\overline{p}$.

Let us start with the probability of unsuccessful decoding that strongly depends on the retransmission protocol applied. Two main protocols—or a mixture of both—are commonly used in mobile communication systems: incremental redundancy and chase combining [23]. The latter approach is also known as *retransmission diversity* [9]. Herein, for illustration purposes, we focus on incremental redundancy only. Further information on chase combining can be found in [9] and [23]. For the assumptions made previously, it has been shown in [9] and [11] that the probability of unsuccessful decoding can be expressed by the outage probability

$$p(m) = P\{R^m_{0|\{H\}} < R\} \tag{5}$$

where $R^m_{0|\{H\}}$ is the conditional cut-off rate [34] for any number $m \leq M$. The condition is on the channel fading coefficient. Parameter R denotes the transmitted rate (b/s/Hz) given by the modulation alphabet chosen, such as binary phase shift keying (BPSK), or quadrature phase shift keying (QPSK). We further assume that channel fading remains constant during the transmission of a transmission block, time slot, or time transmission interval, a valid assumption in practice. Then, we have for incremental redundancy

$$R^m_{0|\{H\}} = \sum_{i=1}^{m} R_{0|H_i} \tag{6}$$

where $R_{0|H_i}$ represents the cut-off rate for a multiple-input multiple-output (MIMO) channel with H_i as channel coefficient matrix [20, 34, 37]. In the case of a single scalar transmission, the conditional cut-off rate per time slot i assuming channel coefficient h_i can be expressed by [21]

$$R_{0|h_i} = -\log 2\left[\frac{1}{q}\sum\nolimits_{j=0}^{q-1}\exp(-\gamma_i(h_i)\sin^2(\pi j/q))\right] \tag{7}$$

where q denotes the modulation index, for example, $q = 4$ for QPSK, and $\gamma_i(h_i)$ is the signal-to-interference noise ratio subject to the i-th transmission block. By inserting equation 7 in equation 6 we can evaluate the probability of unsuccessful decoding given by equation 5. For many practical systems it is valid to assume channel quality information at the transmitter. As an example, in HSDPA, the channel quality indicator (CQI) is fed back to the base station [2, 23]. On the basis of this channel information it is possible to estimate the signal-to-interference-and-noise ratio the transmission suffers. Consequently, by CQI estimation it is finally possible to calculate the cut-off rate (equation 7) and the outage probability (equation 5).

To accomplish the evaluation of the average throughput given by equation 4, let us focus on the average channel access probability $\overline{p}$. Obviously,

the average access probability of a user strongly depends on both the scheduling policy applied and the service required by the user. Assume given traffic models with diverse service mixes, it is possible to evaluate the channel access probabilities for the scheduling policy designed. As an example, for the often applied proportional fair scheduling rule [24], all users have the same amount of time allocated (in a long-term sense), if all transmissions are subject to the same fading conditions. Hence, channel access probability is reciprocally proportional to the number of users likely to have access to the channel.

HSDPA Results

In order to confirm the suitability of the metrics evaluated, we compared results for the proposed metrics (equation 4) with results obtained via a computationally intensive HSDPA simulation platform. **Figure 10** shows the long-term throughput of a specific user k versus the geometry [2] to which user k is subject. The parameter setting was as follows: 10 users required access to HSDPA, a maximum of four users were scheduled simultaneously, five codes are assigned to user k for all slots i user k was scheduled, user k applied a maximal $M = 4$ transmissions, and one active HARQ process per user. Moreover, the data rate R of user k was chosen such that the block-error-rate in the first transmission was below 10 percent.

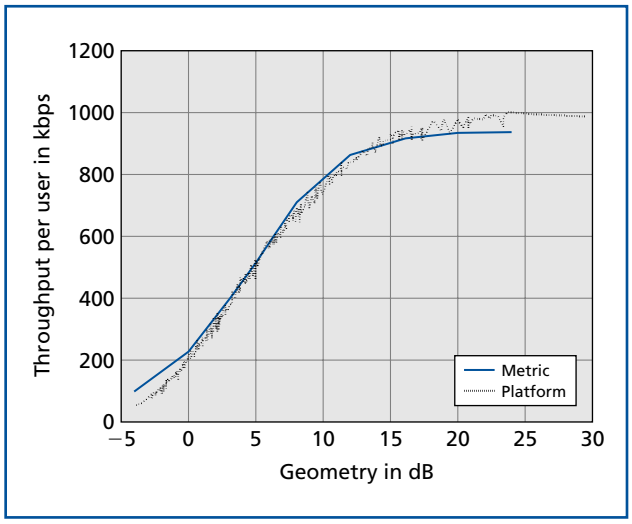


Figure 10.
Comparison: metrics via high-speed downlink packet access simulation platform.

The service assumed was an FTP download with 10 megabyte file size ensuring full buffer occupancy for all users. Finally, we applied a pedestrian A channel [2] with orthogonality factor $F = 0.089$.

Figure 10 depicts quite a good match of the results indicating the suitability of our metric. The minor mismatch in the lower geometry region can be explained by coding performance loss of turbo encoding compared with the cut-off rate approach we applied for the metric. For high geometries, a long thinking time of approximately 30 seconds results in an average of about 9.6 users for the simulation platform while 10 users were assumed for the metric.

Summary

In this paper, we presented an introduction to a novel concept for access-independent IP-based resource management that provides the opportunity to unify control and management of both wireless and wireline access next-generation networks and thus saves effort needed for duplicate standardization of similar functions. This new concept, which can be applied in flat and open architectures, and even in cellular environments, is currently being realized in Alcatel-Lucent's Base Station Router concept.

First proposals to solve key challenges of the IPRM approach such as scalability of the architecture, performance of mobile routing algorithms, and the stability of wireless link metrics have been presented. However, further research is needed to improve performance and ensure acceptance of the concept.

These further studies will be continued by Bell Labs in Germany under the ScaleNet European research initiative, which is dedicated to advancing the vision of a scalable and flexible next-generation access network that provides an IP-optimized integration of heterogeneous access systems for future broadband multimedia services. The ScaleNet project funded by the German government is a cross-industrial research project that will practically demonstrate its major achievements to the public in 2008.

Acknowledgment

The authors would like to thank the German Federal Ministry of Education and Research (BMBF) for funding a significant part of this work within the project ScaleNet under grant FKZ 01BU536.

The responsibility for this publication is solely with the authors.

***Trademarks**

3GPP is a trademark of the European Telecommunications Standards Institute (ETSI).

References

[1] 3GET, <http://www.dlr.de/pt_it/kt/foerderbereiche/mobile_breitband_kommunikationssysteme/3get/>.

[2] 3rd Generation Partnership Project, "High Speed Downlink Packet Access: Physical Layer Aspects (Release 5)," 3GPP TS 25.858, 2002, <http://www.3gpp.org/ftp/Specs/html-info/25858.htm>.

[3] 3rd Generation Partnership Project, "Policy and Charging Control Architecture (Release 7)," 3GPP TS 23.203, v0.4.0, Feb. 2006, <http://www.3gpp.org/ftp/Specs/html-info/23203.htm>.

[4] 3rd Generation Partnership Project, "3GPP System Architecture Evolution: Report on Technical Options and Conclusions (Release 7)," 3GPP TR 23.882, v1.2.3, June 2006, <http://www.3gpp.org/ftp/Specs/html-info/23882.htm>.

[5] Ambient Networks, <http://www.ambient-networks.org>.

[6] M. Bauer, P. Bosch, N. Khrais, L. G. Samuel, and P. Schefczik, "The UMTS Base Station Router," Bell Labs Tech. J., 11:4 (2007), 93–112.

[7] D. Bertsekas and R. Gallager, Data Networks, Prentice-Hall, Englewood Cliffs, NJ, 1987.

[8] P. Bosch, S. Mullender, and S. Samuel, "Mobility in UMTS Packet IP Networks," Proc. IEEE Wireless Commun. and Networking Conf. (WCNC '06) (Las Vegas, NV, Apr. 2006).

[9] G. Caire and D. Tuninetti, "The Throughput of Hybrid-ARQ Protocols for the Gaussian Collision Channel," IEEE Trans. Inform. Theory, 47:5 (2001), 1971–1988.

[10] G. Daley, E. Wu, A. Sekercioglu, and S. Narayanan, "Packet Tunneling for Route Optimization in MN-to-MN Communications," IETF Internet Draft, Mar. 2005.

[11] A. Dekorsy, "A Cutoff Rate Based Cross-Layer Metric for MIMO-HARQ Transmission," Proc. IEEE 16th Internat. Symposium on Personal, Indoor and Mobile Radio Commun. (PIMRC '05) (Berlin, Ger., 2005), vol. 4, pp. 2166–2170.

[12] A. Dekorsy and M. Bauer, "A Cross-Layer Metric for IP Based Resource Management in Next Generation Networks," Proc. Internat. Wireless Summit (IWS '05) (Aalborg, Den., 2005).

[13] M. Dohler, A. Gkelias, and H. Aghvami, "A Resource Allocation Strategy for Distributed MIMO Multi-Hop Communication Systems," IEEE Commun. Lett., 8:2 (2004), 99–101.

[14] EMobility, <http://www.emobility.eu.org/>.

[15] European Telecommunications Standards Institute, "Telecommunications and Internet Converged Services and Protocols for Advanced Networking (TISPAN), NGN Functional Architecture Release 1," ETSI ES 282 001, v1.1.1, Aug. 2005, <http://www.etsi.org>.

[16] European Telecommunications Standards Institute, "Telecommunications and Internet Converged Services and Protocols for Advanced Networking (TISPAN), Resource and Admission Control Sub-System (RACS), Functional Architecture," ETSI ES 282 003, v1.1.1, Final draft, Mar. 2006, <http://www.etsi.org>.

[17] European Union 6th Framework Program (FP6), Ambient Networks project, <http://www.ambient-networks.org/project.html>.

[18] G. Fischer, "Next-Generation Base Station Radio Frequency Architecture," Bell Labs Tech. J., 12:2 (2007), 3–18.

[19] Flarion Technologies, "The Benefits of a Packet-Switched, All-IP Mobile Broadband Network," White Paper, Feb. 2004.

[20] G. J. Foschini, "Layered Space-Time Architecture for Wireless Communication in a Fading Environment When Using Multi-Element Antennas," Bell Labs Tech. J., 1:2 (1996), 41–59.

[21] B. Friedrichs, Kanalcodierung, Springer, Berlin, Ger., 1996.

[22] Y. Guo, Z. Antoniou, and S. Dixit, "IP Transport in 3G Radio Access Networks: An MPLS-Based Approach," Proc. IEEE Wireless Commun. and Networking Conf. (WCNC '02) (Orlando, FL, 2002), vol. 1, pp. 11–17.

[23] H. Holma and A. Toskala, HSDPA/HSUPA for UMTS: High Speed Radio Access for Mobile Communications, John Wiley, Chichester, Eng., 2006.

[24] J. M. Holtzman, "CDMA Forward Link Waterfilling Power Control," Proc. Veh. Technol. Conf. (VTC '00) (Tokyo, Japan, 2000), vol. 3, pp. 1663–1667.

[25] International Telecommunication Union, Telecommunication Standardization Sector, "General Principles and General Reference Model for Next Generation Networks," ITU-T Rec. Y.2011, Oct. 2004, <http://www.itu.int>.

[26] International Telecommunication Union, Telecommunication Standardization Sector, "Resource and Admission Control Functions in Next Generation Networks," ITU-T Draft Rec. Y.2111 (formerly Y.RACF), v9.6, July 2006, <http://www.itu.int>.
[27] Internet Engineering Task Force, "Multihoming and Mobile Nodes in IPv6 (monami6)," IETF Working Group, <http://www1.ietf.org/mail-archive/web/ietf-announce/current/msg01581.html>.
[28] J. Kempf and P. Yegani, "OpenRAN: A New Architecture for Mobile Wireless Internet Radio Access Networks," IEEE Commun. Mag., 40:5 (2002), 118–123.
[29] Lucent Technologies Inc, "Lucent Base Station Router (BSR) for UMTS Networks," 2006, Brochure Doc. No. BSR.10.2006, <http://www.alcatel-lucent.com>.
[30] J. Lundsjoe, R. Aguero, E. Alexandri, F. Berggren, C. Cedervall, K. Dimou, J. Gebert, R. Jennen, L. Jorguseski, R. Karimi, F. Meago, H. Tang, and R. Veronesi, "A Multi-Radio Access Architecture for Ambient Networking," Proc. 14th Information Society Technologies Mobile and Wireless Commun. Summit (IST '05) (Dresden, Ger., 2005).
[31] Motorola, Inc., "Motorola AXPT: Reducing the Cost and Complexity of Delivering Indoor Mobile Broadband Using HSDPA Access Points," Concept Paper (2006), <http://www.motorola.com/networkoperators/pdfs/axpt_solutions_paper.pdf>.
[32] O. Oyman and A. J. Paulraj, "Design and Analysis of Linear Distributed MIMO Relaying Algorithms," IEE Proc. Commun., 153:4 (2006), 565–572.
[33] C. E. Perkins, "Ad Hoc Networking," Addison-Wesley, Upper Saddle River, NJ, 2001.
[34] D. Rainish, "Diversity Transform for Fading Channels," IEEE Trans. Commun., 44:12 (1996), 1653–1661.
[35] ScaleNet, <http://scalenet.mwrl.net/bin/view/ScaleNetPublic/>.
[36] M. Siebert, B. Xu, M. Grigat, E. Weis, N. Bayer, D. Sivchenko, T. Banniza, K. Wunstel, S. Wahl, R. Sigle, R. Keller, A. Dekorsy, M. Bauer, M. Soellner, J. Eichinger, C. Fan, F. Pittmann, R. Kuhne, M. Schlager, I. Baumgart, R. Bless, and S. Stefanov, "ScaleNet—Converged Networks of the Future," IT—Inform. Technol., 48:5 (2006), 253–263.
[37] E. Telatar, "Capacity of Multi-Antenna Gaussian Channels," European Trans. Telecommun., 10:6 (1999), 585–595.
[38] I. Widjaja, "Next Generation Packet Network Architectures with Decoupled Service Plane and Transport Plane," Proc. 3rd Internat. Conf. on Broadband Commun., Networks and Syst. (Broadnets '06) (San Jose, CA, 2006).
[39] M. Zorzi and R. R. Rao, "Throughput of Selective-Repeat ARQ With Time Diversity in Markov Channels with Unreliable Feedback," Wireless Network, 2:1 (1996), 63–75.

(Manuscript approved March 2007)

MARKUS BAUER is a member of technical staff at



Bell Labs Europe in Nuremberg, Germany. He received his Dipl.-Ing. (M.Sc.) degree in electrical engineering from the University of Erlangen-Nuremberg, Germany. His current work is in converged next-generation network architectures. His main research interests are in generalized and unified IP-driven resource and mobility management schemes in flat and converged fourth-generation access networks.

ARMIN DEKORSY is a distinguished member of



technical staff at Bell Labs Europe in Nuremberg, Germany. He received his Dipl.-Ing. (FH) (B.Sc.) degree from Fachhochschule Konstanz, Germany; his Dipl.-Ing. (M.Sc.) degree from University of Paderborn, Germany; and his Ph.D. degree from the University of Bremen, Bremen, Germany, all in communication engineering. Prior to joining Alcatel-Lucent, he was with T-Nova Deutsche Telekom, where he led research projects on smart antenna technologies. Dr. Dekorsy is currently leading Bell Labs internal and government-funded research projects on next-generation networks. His current research interests are multiple antenna technologies, cross-layer design, and resource allocation for wireless multihop/relaying networks.

PETER SCHEFCZIK is a member of technical staff at



Bell Labs Europe in Nuremberg, Germany. He received both an M.Sc. degree and a Ph.D. degree in applied mathematics from the University of Erlangen-Nuremberg in Germany. His current work is in fourth-generation wireless and wireline network architectures

and network convergence. Dr. Schefczik's main research interests include heterogeneous access networks for both wireless and wireline and architecture and performance evaluation of communication systems in general.



MICHAEL SOELLNER is a technical manager at Bell Labs Europe in Nuremberg, Germany. He received a diploma in mathematics from the University of Erlangen and a Ph.D. degree in applied mathematics from the University of Bochum in Germany. He has held various positions in communication technology in industry, where he focused on systems engineering and research for mobile radio systems. At present, Dr. Soellner is responsible for a research group dealing with radio access network solutions beyond 3G systems. Current focuses of his work are advanced technologies; protocols and architectures for future wireline/wireless converged networks, especially optimization of mobility and resource management based on a native IP approach; as well as the evolution to truly converged services. ◆