Mobile Networks and Applications
DOI 10.1007/s11036-006-7794-9

Efficient Resource Allocation for Policy-Based Wireless/Wireline Interworking

Yu Cheng · Wei Song · Weihua Zhuang · Alberto Leon-Garcia · Rose Qingyang Hu

Published online: 23 May 2006

Abstract This paper proposes efficient resource allocation techniques for a policy-based wireless/wireline interworking architecture, where quality of service (QoS) provisioning and resource allocation is driven by the service level agreement (SLA). For end-to-end IP QoS delivery, each wireless access domain can independently choose its internal resource management policies to guarantee the customer access SLA (CASLA), while the border-crossing traffic is served by a core network following policy rules to meet the transit domain SLA (TRSLA). Particularly, we propose an engineered priority resource sharing scheme for a voice/data integrated wireless domain, where the policy rules allow cellular-only access or cellular/WLAN interworked access. By such a resource sharing scheme, the CASLA for each service class is met with efficient resource utilization, and the interdomain TRSLA bandwidth requirement can be easily determined. In the transit domain, the traffic load fluctuation from upstream access domains is tackled by an inter-TRSLA resource sharing technique, where the spare capacity from underloaded TRSLAs can be exploited by the overloaded TRSLAs to improve resource utilization. Advantages of the inter-SLA resource sharing technique are that the core network service provider can freely design the policy rules that define underload and overload status, determine the bandwidth reservation, and distribute the spare resources among bandwidth borrowers, while all the policies are supported by a common set of resource allocation techniques.

Keywords resource allocation · service level agreement · policy-based networking · wireless/wireline interworking · cellular/WLAN integration · call admission control

Y. Cheng (✉) · A. Leon-Garcia
Department of Electrical and Computer Engineering,
University of Toronto,
10 King's College Road,
Toronto, Ontario M5S 3G4, Canada
e-mail: y.cheng@utoronto.ca

A. Leon-Garcia
e-mail: alberto.leongarcia@utoronto.ca

W. Song · W. Zhuang
Department of Electrical and Computer Engineering,
University of Waterloo,
200 University Avenue West,
Waterloo, Ontario N2L 3G1, Canada

W. Song
e-mail: wsong@bbcr.uwaterloo.ca

W. Zhuang
e-mail: wzhuang@uwaterloo.ca

R. Qingyang Hu
Department of Electrical and Computer Engineering,
Mississippi State University,
239 Simrall Building, Hardy Road,
Mississippi State, MS 39762, USA
e-mail: hu@ece.msstate.edu

1. Introduction

Provision of quality of service (QoS) guaranteed multimedia applications to mobile users is the main objective of the next-generation wireless networks, which are expected to be IP-based and able to interwork with the Internet backbone seamlessly. Currently, various air interface techniques coexist to provide Internet access for mobile users, e.g., the general packet radio service (GPRS), wideband code-division multiple access (WCDMA), CDMA2000, and IEEE 802.11 wireless LAN (WLAN). Further, end-to-end communications are likely to involve multiple

wireless/wireline administrative domains that apply different QoS technologies.

Policy-based networking [34–36] is a promising approach to integrate the heterogeneous wireless access networks and the Internet to provision broadband access, seamless global roaming, and end-to-end QoS support. Basically, service level agreements (SLAs) are negotiated between customers and the Internet access service providers (SPs) and between the neighboring administrative domains along the end-to-end path, which describe the sets of IP services and associated QoS levels that the network domains have mutually contracted to provide. Individual domains can independently choose proper internal resource management policy rules to enforce the contracted SLAs. These policy rules describe the amount of network resources required to guarantee QoS without going into the details of how to configure the network devices [36].

Although the policy-based architecture (PBA) greatly facilitates network management by presenting an abstract view of network resources to its operator, this architecture needs to be solidified and materialized through specific resource allocation techniques, which are the focus of this paper. The differentiated services (DiffServ) scheme [5] has been extensively accepted for both wireless and wireline domains as the network-layer QoS mechanism due to its scalability and convenience for SLA-based network management [11, 24, 26], we therefore discuss resource allocation in the DiffServ context.

In PBA, resource allocation is driven by the SLAs. A wireless access domain behaves as the SP to individual customers, and the customer access SLA (CASLA) is contracted between the two parities. The main contents of the CASLA are the specifications of the applications provided to the customers, the QoS measures associated with each application, and the corresponding billing information. Here, we investigate call or connection admission control (CAC) in a wireless domain, with the objective to serve as many customers (i.e., to obtain as much revenue) as possible while SLA compliance is guaranteed. The derived admission region and resource reservation thresholds can then be used as policy rules to enforce the SLA. On the other hand, the wireless access domains behave as the customer to the backbone transit domain, which is usually an Internet service provider (ISP). The border-crossing traffic is served by the ISP according to transit domain SLAs (TRSLAs). The main contents of TRSLA include the mapping relationship between wireless service classes and standard DiffServ classes, the aggregate traffic load of each service class, QoS performance and billing information in the transit domain. Efficient resource allocation over the transit domain is beneficial to the ISP in terms of revenue. Novel resource allocation techniques for policy-based inter-SLA resource sharing are proposed in this paper.

We consider CAC for wireless access domains supporting multiple QoS service classes. The existing resource sharing schemes can be broadly classified into three categories: complete sharing (CS) [20, 27], complete partitioning (CP) [16, 20], and virtual partitioning (VP) [5, 33]. CS achieves the highest resource utilization among the three categories, but the individual QoS of a certain class cannot be guaranteed in this scheme. On the other hand, CP can guarantee the isolation among traffic classes, but may underutilize the resources when underloaded traffic classes exist. VP is a tradeoff between CS and CP, where the free capacity from underloaded traffic classes can be utilized by the overloaded traffic classes, and trunk reservation [14] is implemented to protect underloaded traffic classes from resource starvation.

In the SLA-centric PBA, resource allocation in a wireless access domain on one hand should fully exploit its wireless resources and, on the other hand, should facilitate the TRSLA resource negotiation to support its border-crossing traffic. The CP and CS schemes are not appropriate choices due to their obvious disadvantages on resource utilization and service isolation, respectively. VP is an efficient resource sharing scheme, but VP is not convenient for applications with an elastic transmission rate. In a wireless access domain, priority-based service differentiation schemes are widely used [23, 30], where the low loss, low delay realtime connections are served with high priority and the leftover bandwidth can be fully exploited by the elastic non-realtime connections using the Transmission Control Protocol (TCP). One contribution of this paper is that we study the *engineered priority scheme* [27] and the associated optimal CAC policies for a voice/data integrated wireless domain, where the CASLAs are met with efficient resource utilization and the boundary-crossing bandwidth requirement can be easily determined for TRSLA negotiation.

Considering the high costs of acquiring the necessary radio spectrum and software/hardware upgrading to provide 3G cellular communications, mobile network operators are increasingly interested in the cellular/WLAN integration solution for more network revenue [1, 4, 36]. Resource allocation in a cellular/WLAN interworked domain is also investigated in this paper. Particularly, we demonstrate that the proposed engineered priority scheme and associated mathemat-

ical modeling can be applied to analyze either a cellular-only domain or a cellular/WLAN interworked domain, which considerably facilitates the policy-based network management. When the network operator needs to change the interworking policy or vertical handoff management policy, a uniform set of mathematical tools can be used for the capacity replanning.

In a DiffServ transit domain, the network resources are shared by all the TRSLAs. That is, all the TRSLA resource commitment at boundaries should be mapped to the bandwidth allocation of each service class at each link. DiffServ itself only defines the data plane schemes, i.e., the per-hop behaviors (PHB) and the edge traffic conditioning, which are augmented in this paper by the traffic engineering functionalities using the multiprotocol label switching (MPLS) technique [2] to facilitate accurate internal resource mapping. At the present time, static inter-domain SLAs are mainly used, which are negotiated based on estimation of the long-term (e.g., in days or weeks) average traffic volume (i.e., the *engineered traffic load*) from the upstream domain. In reality, the traffic from the upstream access domain may change dynamically at a short time scale (e.g., in minutes or hours) due to customers' random behaviors or inter-class resource sharing. Such traffic load dynamics lead to overloaded or underloaded TRSLAs, where the static resource partitioning results in inefficient resource utilization.

Another contribution of this paper is that we propose a *policy-based inter-SLA resource sharing* (PBISRA) scheme to efficiently exploit the spare capacity from those underloaded TRSLAs. In the PBISRA scheme, if a TRSLA is in the underload status, a *protection bandwidth* smaller than its nominal capacity is determined according to the resource sharing policy defined in the SLA. The protection bandwidth is guaranteed for the underloaders to satisfy their QoS requirements during the underloaded periods, and the available spare capacity is then properly distributed to related links to be borrowed by others according to a *call-level differentiation* policy. The inter-SLA resource sharing concept was first proposed in our work [10] to guarantee the QoS requirement on call blocking probability for all the SLAs involved in the resource sharing. In this paper, we probe further to reveal the potentials of such a resource sharing approach for PBISRA, where the network manager can flexibly design or modify the policy rules that define underload and overload status, determine the bandwidth reservation, or distribute the spare resources among bandwidth borrowers, which are all supported by a common set of resource allocation techniques to enforce the SLA.

The remainder of this paper is organized as follows. Section 2 describes the policy-based wireless/wireline interworking architecture. Section 3 presents the engineered priority scheme for a wireless domain which provides either the cellular-only or the cellular/WLAN interworked access. Section 4 presents the PBISRA scheme for the transit domain. Some numerical and simulation results are shown in Section 5 to demonstrate the performance of the proposed resource allocation schemes. Section 6 gives the concluding remarks.

2. Policy-based wireless/wireline interworking

2.1. All-IP policy-based interworking architecture

The policy-based interworking architecture under consideration is shown in Fig. 1, which consists of multiple wireless/wireline administrative domains. The DiffServ model is applied in each domain to provision QoS. In the architecture, a certain number of homogeneous radio access networks (RANs) using the same air interface technique or heterogeneous RANs interworked according to policies (e.g., cellular/WLAN interworking) are grouped into a DiffServ wireless access domain. All wireless access domains are connected through the DiffServ transit domain(s) (Internet backbone) to provide end-to-end Internet services to a mobile station (MS). In a DiffServ wireless domain, all the network elements are enhanced to combine the functions of a DiffServ IP router. The base stations (or access points in the WLAN context) and the gateway are edge routers of a domain and connected through the core routers. The base stations provide MSs the access points to the Internet according to the CASLA. Here, we consider unidirectional SLAs, where CASLAs for uplink and downlink traffic are negotiated separately. The gateway is the interface connecting to the DiffServ transit domain. For example, the gateway GPRS support node (GGSN) in the WCDMA domain is the gateway of the domain to the external DiffServ Internet. Between the gateway and the connected transit domain ingress router, TRSLAs are negotiated to specify the resources allocated by the ISP to serve the aggregate traffic flowing from/into the access domain.

In the interworking architecture, policy-based resource management is applied, where individual operators transform the QoS requirements specified in the SLAs into a set of policy rules that are applied to their network domains to enforce the SLA. According to the policy framework [34] defined by the Internet Engineering Task Force (IETF), each

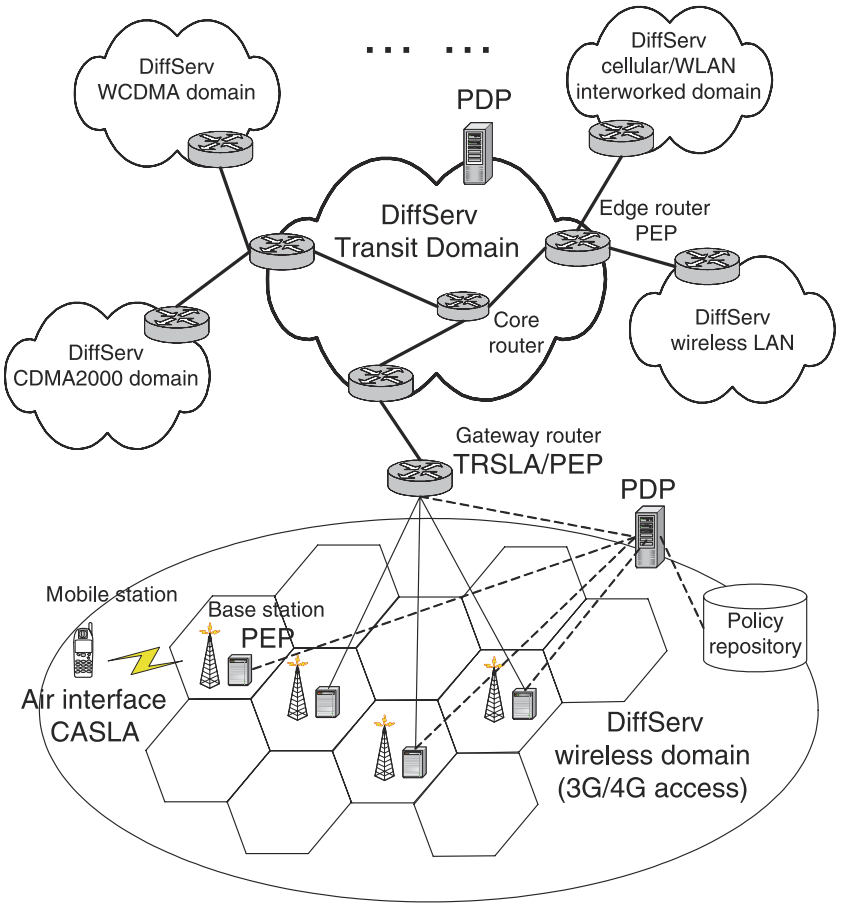


Figure 1 The all-IP policy-based wireless/wireline interworking architecture

domain has a *policy repository* to store the policy rules. A logically centralized *policy decision point* (PDP) is responsible for retrieving the appropriate policy rules from the repository in response to policy events that are triggered by the contracted services, such as the reception of out profile traffic at a certain edge router that serves as a *policy enforcement point* (PEP) in the policy framework. The retrieved policy rules are then translated by the PDP into network device configuration actions and sent as policy decisions to the associated PEP to handle the policy events for QoS guaranteed IP services. The critical point to enable policy-based networking is the resource allocation techniques that properly map the SLA QoS requirements and resource commitments into policy rules that determine the admission and service of a traffic flow. In the IETF policy framework, it is assumed that a *policy management tool* (PMT) is available for the automatic SLA-to-policy translation, but the details of the PMT are still left open. In this paper, we develop SLA-centric resource allocation techniques for both the access domains and the transit domains, which can be used to facilitate the development of automatic PMTs.

It is noteworthy that in each domain, although QoS is provisioned according to the DiffServ architecture, the specific definition and QoS handling of service classes may be different from each other. Therefore, a proper service class mapping between neighboring domains is very important for maintaining a consistent level of end-to-end QoS. For example, the Universal Mobile Telecommunications System (UMTS) based on WCDMA technology defines four QoS classes, i.e., conversational, streaming, interactive, and background, for 3G wireless communications [24]. Conversational and streaming classes are intended for realtime multimedia traffic, and the other two classes for delay tolerant data traffic. On the other hand, DiffServ defines the expedited forwarding (EF) [17] PHB for the premium service and the assured forwarding (AF) PHB [15] for the assured service, in addition to the classic best-effort service. To extend IP services to the wireless domain, the UMTS QoS classes must be mapped into DiffServ classes. Normally, the conversational class can be mapped to the EF PHB for a very low delay and low loss service, the streaming and interactive traffic to the AF PHB, and the background traffic to the best-effort service [24].

2.2. Path-oriented transit domain

The transit domain in the interworking architecture has a path-oriented environment, where *per-class per-ingress/egress pair* TRSLAs are negotiated at network boundaries [9]. The path-oriented environment can greatly facilitate the DiffServ network planning and the PBISRA, to be discussed in Section 4.

The MPLS technique is used to establish a path-oriented environment. We assume there exists an off-line routing algorithm which sets up several parallel paths for each ingress/egress pair. These paths are fixed by label switching and referred to as virtual paths (VPths). All traffic traversing an ingress/egress pair is distributed among the VPths. An MPLS *traffic trunk* is defined as a logic pipeline within a VPth, which is allocated a certain amount of capacity to serve the traffic associated with a certain SLA. Therefore, a VPth between an ingress/egress pair may include multiple traffic trunks for different SLAs. In the path-oriented environment, boundary SLA resource commitments are properly mapped to the bandwidth allocation at each traffic trunk by a network dimensioning (planning) procedure, with the objective to maximize the long-term resource utilization [25, 32]. Let s denote a service class, σ an ingress/egress pair, and r a route. We use (s,σ) to identify an SLA, (s,r) a traffic trunk, and $\mathcal{R}(s,\sigma)$ the route set or trunk set of an SLA. The nominal capacity of SLA (s,σ) is denoted as $C_{s,\sigma}$, and the bandwidth allocated to traffic trunk (s,r) (determined by the network dimensioning) is denoted as $C_{s,r}$, with $C_{s,\sigma} = \sum_{r\in\mathcal{R}(s,\sigma)} C_{s,r}$.

In the DiffServ context, the logical central entity responsible for the network resource management is termed as a *bandwidth broker* [24, 28, 31, 32]. In practice, the bandwidth broker also adopts the policy-based management framework, where the PEPs/PDPs are not separate components but are collocated with the resource management function blocks, e.g., SLA subscription, network dimensioning, and dynamic resource management, which composes the bandwidth broker [32].

2.3. Call admission control

Call admission control is a necessary part for QoS guarantee, and a properly determined admission region can be considered as an abstract representation of the network resources to facilitate policy-based resource allocation. In the wireless/wireline interworking architecture, CAC should be implemented in each domain to achieve the end-to-end QoS. Section 3 is to present a CAC scheme for a wireless access domain supporting voice/data integrated services. If a new call is accepted by the access domain, the gateway router will then forward the new request to the connected transit domain ingress router. The PEP installed at the ingress router first tries to make an admission decision according to local policy rules; if the PEP finds that the request incurs a new policy event, e.g., more bandwidth reservation is required, the PEP then sends the new policy event to the PDP. The PDP will retrieve appropriate policy rules from the policy repository and pass them to the PEP to deal with the new event. If no suitable policy can be found in the policy repository, the PMT may contact the bandwidth broker to adjust the resource allocation, by which some new policy rules are generated and added to the repository. The CAC in the transit domain is to be discussed in Section 4.

3. Access domain resource allocation

In this section, we discuss the resource allocation in a cellular access domain supporting voice/data integrated services. An engineered priority scheme is proposed for efficient resource sharing, where the associated mathematic modeling is developed and the admission regions for both voice and data calls are solved to support policy-based resource allocation. The engineered priority scheme is also extended to a cellular/WLAN interworked wireless domain.

3.1. Engineered priority scheme

We consider a wireless cellular system supporting both realtime and non-realtime services and focus on the uplink traffic. For convenience, we use two typical applications, voice and TCP data, to represent the two service classes, respectively. The term “call” at air interface refers to a voice call or a connection for data service.

In a cell with capacity C, each voice call consists of a constant-rate packet stream. Due to the realtime nature, voice calls are given preemptive priority over data calls in obtaining resources up to a certain amount, denoted by $\Gamma(< C)$. Each admitted voice user is allocated the required bandwidth, denoted by γ_v. Admitted data calls, at any time, equally share only the leftover bandwidth by voice calls. This method of allocating resources to different classes of traffic is first proposed in [27]. The scheme guarantees a certain amount of bandwidth (i.e., $C-\Gamma$) always available to low-priority data calls, therefore referred to as *engineered priority scheme*. The value of Γ should be large enough to satisfy the QoS requirements for voice calls

but small enough to give data calls as many leftover resources as possible. When there are i voice and j data users in the cell, the serving capacity available to each data user is

$$\gamma_{\mathrm{d}} = \frac{C - i\gamma_{\mathrm{v}}}{j}. \quad (1)$$

The value of γ_{d} is non-constant, dependent on the instantaneous numbers of voice and data users in the cell. Consequently, there are chances that γ_{d} may drop below a critical threshold, denoted by c_{d}, which is the minimum rate required to maintain the bottom-line service quality of a data call. This phenomenon is called *overload*, and the probability of its occurrence should be kept low by restricting the number of data users admitted to the cell.

3.1.1. TRSLA negotiation

If the cellular domain includes N cells, the total voice and data traffic loads in the domain are $N\Gamma$ and $N(C-\Gamma)$, respectively. According to the transit domain's per-class per-ingress/egress SLA format, the access domain needs to divide the total boundary-crossing traffic load using the destination distribution information (which can be obtained from address analysis and Border Gateway Protocol (BGP) routing analysis). Let α_σ denote the load portion passing the ingress/egress pair σ;[1] the traffic load distribution satisfies $\alpha_0^{\mathrm{v}} + \sum_\sigma \alpha_\sigma^{\mathrm{v}} = 1$ for voice traffic, and $\alpha_0^{\mathrm{d}} + \sum_\sigma \alpha_\sigma^{\mathrm{d}} = 1$ for data traffic, where α_0 represents the fraction of intradomain communications. In the TRSLA negotiation, an SLA for voice can then be defined as "realtime application being mapped to the premium service in the transit domain with bandwidth requirement of $\alpha_\sigma^{\mathrm{v}} N\Gamma$," and for data as "non-realtime application being mapped to the assured service with bandwidth requirement of $\alpha_\sigma^{\mathrm{d}} N(C-\Gamma)$, and bursty arrivals higher than the contracted rate being marked as out profile traffic."

3.2. Mathematic modeling

For simplicity, we study a homogeneous system in statistical equilibrium, where any cell is statistically the same as any other cell and the mean handoff arrival rate to a cell is equal to the mean handoff departure rate from the cell. Hence, we can evaluate the system performance by analyzing the performance of one cell. Such single-cell analysis has been extensively used for CAC [18, 33].

There are four types of call arrivals in a cell: new voice and data calls originating within the cell, handoff voice and data calls coming from adjacent cells. The voice and data call arrival processes are assumed to be independent of each other. For mathematical tractability, we adopt the Markovian modeling of voice/data call behaviors. For voice calls, new call arrivals are Poisson with the arrival rate λ_{v}. Cell residence time X_{v} and call lifetime Y_{v} are both exponentially distributed with the mean of $(\mu_{\mathrm{v}}^X)^{-1}$ and $(\mu_{\mathrm{v}}^Y)^{-1}$, respectively. The channel holding time of a voice call in a cell is then also exponentially distributed with mean $(\mu_{\mathrm{v}}^X + \mu_{\mathrm{v}}^Y)^{-1}$, leading to a Poisson handoff call arrival process.

For data calls, the new call arrivals are also Poisson with the average rate of λ_{d}. The cell residence time X_{d} is exponentially distributed with mean $(\mu_{\mathrm{d}}^X)^{-1}$. Assume that the total length of a data call in packets (the data file size), denoted by L_{d}, is exponentially distributed with mean $(\mu_{\mathrm{d}}^L)^{-1}$ in the mathematical analysis. At state (i,j) the data call has a state-dependent exponential lifetime $Y_{\mathrm{d}}(i,j)$ with mean $\left[\frac{(C-i\gamma_{\mathrm{v}})\mu_{\mathrm{d}}^L}{j}\right]^{-1}$, and hence a state-dependent exponential channel holding time with mean $\left[\mu_{\mathrm{d}}^X + \frac{(C-i\gamma_{\mathrm{v}})\mu_{\mathrm{d}}^L}{j}\right]^{-1}$. The state-dependent exponential property does not mean that the total channel holding time averaged over all the states is also exponential. To facilitate further analysis, we assume that the data call channel holding time is exponentially distributed, which then leads to a Poisson handoff arrival process of data calls. Note that the exponential distribution assumption has been widely used in the literature [6, 16, 18, 20, 27, 33] to provide approximate solutions for cellular systems.

In a cell, the handoff calls are given higher priority to access resources than the new calls by the limited fractional guard channel policy (LFGCP) [29]. The LFGCP can be denoted by $g_{T,M}^{\beta}$, where M is the call-level channel capacity, $T(<M)$ is the channel occupancy threshold over which no new calls are accepted, and β is the probability of accepting a new call when the channel occupancy is T calls. Handoff calls can occupy the channel up to M. Here, the LFGCP is extended to the integrated voide/data system. We use $g_{T_{\mathrm{v}},M_{\mathrm{v}}}^{\beta_{\mathrm{v}}}$ and $g_{T_{\mathrm{d}},M_{\mathrm{d}}}^{\beta_{\mathrm{d}}}$ to denote the LFGCPs for voice and data calls, respectively.

From the perspective of policy-based management, the LFGCPs with the six parameters $\{M_{\mathrm{v}}, T_{\mathrm{v}}, \beta_{\mathrm{v}}, M_{\mathrm{d}}, T_{\mathrm{d}}, \beta_{\mathrm{d}}\}$ are used by the PEP as policy rules for admission control. These parameters should be determined in such a way that, given the traffic load, the

[1] All the transit-domain ingress/egress pairs associated with the boundary-crossing traffic from an access domain have the same ingress point, which is connected to the access domain gateway.

system under the CAC policy can guarantee the customers' QoS requirement defined in the CASLAs and achieve a high resource utilization. Note that M_v, T_v, M_d, and T_d represent the numbers of the calls. Here, we consider the following connection-level QoS measures:

- $B_{nv}(D_{hv})$: the new call blocking (handoff call dropping) probability, abbreviated as NBP (HDP), for voice calls;
- $B_{nd}(D_{hd})$: the NBP (HDP) for data calls;
- Π_{od}: the overload probability (OP) for data calls.

The QoS requirements are specified by the upper bounds of $\{B_{nv}, D_{hv}, B_{nd}, D_{hd}, \Pi_{od}\}$, denoted by $\{Q_{nv}, Q_{hv}, Q_{nd}, Q_{hd}, Q_{od}\}$. The HDP upper bounds ($Q_{hv}$ and Q_{hd}) are normally one-order lower than the corresponding NBP upper bounds (Q_{nv} and Q_{nd}) to protect the handoff calls.

Since both voice and data calls share the total resources of the cell, the two LFGCPs are not independent. With Poisson arrivals and exponential channel holding times, the traffic flows under the control of the proposed CAC policy can be modeled by a two-dimensional (2-D) continuous-time birth-death process with state (i,j). The steady state probabilities $p(i,j)$ $(0 \le i \le M_v$ and $0 \le j \le M_d)$, can be obtained by solving the balance equations of the 2-D model, based on which the QoS measures can be readily calculated.

3.3. Optimal CAC parameters

In order to achieve the maximal resource utilization, the optimal CAC policy parameters $\{M_v, T_v, \beta_v, M_d, T_d, \beta_d\}$ need to be searched. For voice calls, the objective is to find the minimum value of M_v (namely the minimal $\Gamma = M_v r_v$) and the corresponding values of T_v and β_v to satisfy the specified voice QoS upper bounds. The algorithm `Min M` proposed in [29] can be used to determine such optimal CAC parameters. For data, the objective is to utilize the leftover capacity to accept as many data calls as possible with QoS guarantee. However, the parameters for data LFGCP cannot be determined in a straightforward manner because of their lower priority and non-constant bandwidth allocation. Moreover, given the total cell capacity, it might be impossible to satisfy all the QoS upper bounds especially when the traffic load is high. In that case, we have to sacrifice the NBP to guarantee the HDP and OP of the already accepted calls. A procedure to determine the optimal CAC parameters for data traffic is given in [22]. The procedure guarantees the four upper bounds $\{Q_{nv}, Q_{hv}, Q_{hd}, Q_{od}\}$ and takes necessary steps to minimize B_{nd}.

3.4. Adaptive CAC

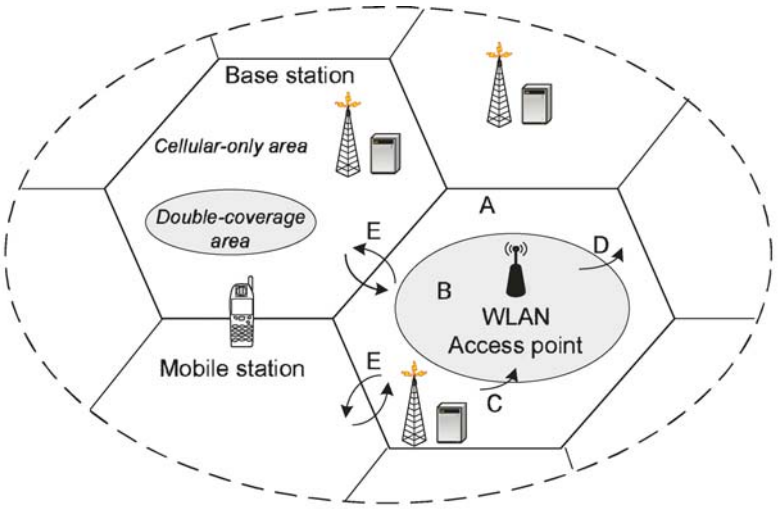
The CAC parameters are normally designed for a target traffic load condition. If the actual traffic condition deviates from the target one, the CAC may result in dissatisfactory service quality or under-utilized resources. A natural *load adaption* solution is to recalculate the CAC policy parameters according to traffic load changes. Considering that voice calls can tolerate a certain amount of reduction in transmission rate before the service quality drops to an unacceptable level, a more aggressive *bandwidth adaption* CAC can be used to strengthen the call-level performance, where CAC parameters are recalculated with a slightly reduced γ_v.

The adaptive CAC (for inter-class resource sharing) in the access domain transfers the traffic load dynamics to the transit domain. As the total capacity in an access domain is always fixed, the adaptive inter-class sharing will simultaneously leads to overloading and underloading of related TRSLAs that are precontracted according to a target load. Further, the load deviation from SLA resource commitment is not a rare case considering that a lot of access domains are connected to the transit domain. In next section, we present an inter-SLA resource sharing technique for the transit domain to exploit the spare capacity from underloaded TRSLAs to serve the traffic from overloaded TRSLAs, which helps to achieve end-to-end efficient resource sharing.

The mathematical details of the optimal CAC with load adaption and bandwidth adaption are given in [21, 22]. Next, we show that the mathematical modeling presented in this section can be readily extended to analyze a cellular/WLAN integrated wireless domain.

3.5. Cellular/WLAN interworking

In order to support more subscribers, the wireless access network operator may choose to deploy WLANs in the hot spots, while the cellular network is used to support high mobility and large cover area. For simplicity, we consider a cellular/WLAN integrated network with one overlaying WLAN in each cell as shown in Fig. 2. Since both cellular and WLAN accesses are available to a dual-mode MS within the WLAN coverage, the area is referred to as *double-coverage* area; the area with only cellular access is referred to as *cellular-only* area. Moreover, the WLAN is *tightly coupled* with the



A: new voice (data) call arrivals to the cellular-only area
B: new voice (data) call arrivals to the double-coverage area
C: vertical handoff voice (data) call arrivals from the cellular to WLAN
D: vertical handoff voice (data) call arrivals from WLAN to the cellular
E: horizontal handoff voice (data) call arrivals between neighboring cells

Figure 2 A cellular/WLAN interworked access domain

cellular system, i.e., the WLAN access point (AP) is connected to the serving GPRS support node (SGSN) and treated as another radio network controller (RNC) [36]. The advantages of the tightly coupled architecture are that the WLAN overlay has no effect on the access control and billing/charging entities of the cellular network and only incurs a very light signalling overhead for vertical handoff.

In the cellular/WLAN interworked scenario shown in Fig. 2, there are five types of arrivals for both voice and data, i.e., new calls in the cellular-only area and in the double-coverage area, vertical handoff calls from WLAN to the cellular and from the cellular to WLAN, and horizontal handoff calls between neighboring cells. Depending on the interworking policies, a new call in the double-coverage area can be directed to the cellular network or the WLAN, a vertical handoff call into the double-coverage area can be switched to the WLAN or stay in the cellular network, and a vertical handoff call into the cellular-only area can be switched to cellular network or be dropped. Independent of the specific interworking policies, the engineered priority scheme, LFGCP CAC policies, and associated mathematic modeling can still be applied in the cellular network. For that, the key point is to determine the average channel holding time of a call that may experience one or more times of vertical handoff in a cell.

Another important issue is the differentiation between the user-mobility characteristics in the hot spots (i.e., the double-coverage area) and those in the other (i.e., the cellular-only) areas. A hot spot (e.g., an office building or a hotel) is usually an indoor environment, where user mobility is very low compared to that in the cellular-only area. Let T_r^{co} denote the residence time that a user stays within the cellular-only area before moving to neighboring cells with probability p^{c-c} or to the overlaying WLAN with probability p^{c-w}, and T_r^{dc} the user residence time in the double-coverage area. T_r^{co} and T_r^{dc} are assumed to be exponentially distributed with mean time $(\eta^{co})^{-1}$ and $(\eta^{dc})^{-1}$, respectively. For a user initiating a new call in the cellular-only area or carrying a handoff call to the cellular-only area, let T_{r1}^c denote its residence time within the cell, which follows a phase-type distribution shown in Fig. 3. The sum of T_r^{co} and T_r^{dc}, which are assumed to be independent of each other, follows a Gamma distribution with the moment generating function (MGF) given by

$$\phi(s) = E[e^{s(T_r^{co}+T_r^{dc})}] = \frac{\eta^{co}}{\eta^{co}-s}\,\frac{\eta^{dc}}{\eta^{dc}-s}. \tag{2}$$

The MGF of T_{r1}^c can then be obtained according to Fig. 3 as

$$G_1(s) = \sum_{i=1}^{\infty} p^{c-c}\,\frac{\eta^{co}}{\eta^{co}-s}\,[p^{c-w}\phi(s)]^{i-1}. \tag{3}$$

Similarly, the residence time of a user initiating a new call in the double-coverage area, denoted by T_{r2}^c, also

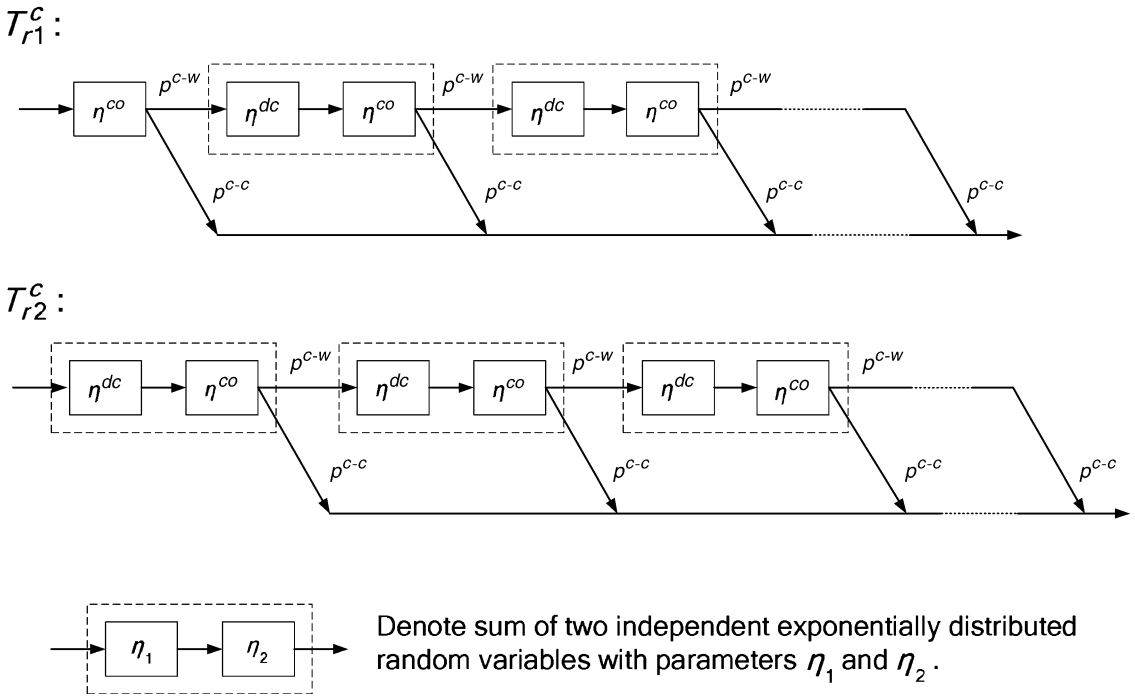


Figure 3 Modeling of user residence time in a cell

follows a phase-type distribution as shown in Fig. 3 with the MGF

$$G_2(s) = \sum_{i=1}^{\infty} p^{c-c}\phi(s)[p^{c-w}\phi(s)]^{i-1}. \tag{4}$$

In general, for two independent random variables A and B where A is exponentially distributed with mean $\frac{1}{a}$ and the MGF of B is $G_B(s)$, it is can be shown that

$$E[\min(A,B)] = \frac{1}{a} - \frac{1}{a}G_B(-a). \tag{5}$$

The result of equation (5) can be used to obtain the average channel holding time.

In the cellular network, the average channel holding time for a voice call in a cell is determined as follows:

- For a call in the cellular-only area, the conditional average channel holding time is $E\left[\min(Y_{\mathrm{v}}, T_{r1}^{c})\right]$ given that the WLAN is full and cannot accept a vertical handoff; otherwise, the conditional average is $E\left[\min(Y_{\mathrm{v}}, T_{r}^{co})\right]$ given that the WLAN can accept a vertical handoff.
- For a call in the double-coverage area, the conditional average channel holding time is $E[\min(Y_{\mathrm{v}}, T_{r2}^{c})]$ given that the WLAN is full and cannot accept a vertical handoff; otherwise, the conditional average is $E\left[\min(Y_{\mathrm{v}}, T_{r}^{dc} + T_{r}^{co})\right]$ given that the WLAN can accept a vertical handoff.[2]
- By finding the probability that the WLAN is full, the total average channel holding time can then be obtained from the conditional averages.

The average channel holding time of a data call can be determined in a manner similar to that of a voice call, except that the data call lifetime Y_{d} is state dependent. At the state of i voice calls and j data calls, $E[Y_{\mathrm{d}}(i,j)] = \left[(C - i\gamma_{\mathrm{v}})\mu_{\mathrm{d}}^{L}/j\right]^{-1}$. $Y_{\mathrm{d}}(i,j)$ is then used instead of Y_{v} to determine the state-dependent channel holding time of a data call. With the knowledge of the call arrival and departure rate, the 2-D Markov chain can be established and the CAC policy parameters for the voice/data LFGCPs can then be solved under the QoS constraint of $\{Q_{\mathrm{nv}}, Q_{\mathrm{hv}}, Q_{\mathrm{nd}}, Q_{\mathrm{hd}}, Q_{\mathrm{od}}\}$.

In the cellular/WLAN interworked scenario, the LFGCPs in the cell are coupled with the resource allocation and sharing in the WLAN. The coupled resource allocation problems can be jointly solved by considering the statistical equilibrium point where both the horizontal handoff calls crossing a cell boundary and the vertical handoff calls crossing a WLAN boundary attain balance in their respective in/out directions. We are able to obtain the call-level capacity of a WLAN, in terms of the maximum number of voice and data calls that the WLAN can simultaneously accommodate while meeting the packet-level throughput constraint of a voice/data call, by analyzing the operations of the media access control (MAC) protocol [8]. Thus, the call-level resource

[2] We assume that a vertical handoff occurs only when a mobile station crosses the boundary of the double-coverage area.

allocation (which can also be an LFGCP policy) is also applicable to the WLAN to control the NBP and HDP in the double-coverage area. Whatever call-level policies are applied to the WLAN, it is preferred that the interworking QoS performance over the whole cell still meets the specifications of $\{Q_{nv}, Q_{hv}, Q_{nd}, Q_{hd}, Q_{od}\}$, so that the impact of cellular/WLAN interworking is transparent to the mobile users.

The TRSLA negotiation also needs to count for the traffic from WLAN. With the tightly coupled architecture, both the cellular traffic and the WLAN traffic go through the same gateway to the transit domain. The resource requirement in a voice/data TRSLA can then be estimated as $\alpha_\sigma^v N(\Gamma^c + \Gamma^w)$ and $\alpha_\sigma^d N[(C^c + C^w) - (\Gamma^c + \Gamma^w)]$, respectively. The superscriptions of c and w are used to distinguish the channel capacity and the voice allocation in the cell and in the associated WLAN, respectively.

4. Policy-based inter-SLA resource sharing

For maximal resource utilization, when the access domain has a traffic load fluctuation, adjusts its inter-class resource sharing, or detects a change of the destination distribution, it should renegotiate the TRSLA with the transit domain to ensure that the border-crossing traffic in the new load condition can be supported. Correspondingly, the transit domain should then replan or reconfigure its internal resources upon a TRSLA renegotiation request to efficiently adapt to the access domain resource allocation. However, replanning requires optimization over the whole domain and often incurs a long overhead time. The frequent replanning in practice is impossible, considering that the access domain resource adaption usually happens at a short time scale, the destination distribution changes often with the application, and the adaption in different access domains normally happens asynchronously. With the *policy-based inter-SLA resource sharing* (PBISRA) scheme proposed in this section, the load fluctuation of the border-crossing traffic will be efficiently dealt with by the transit domain according to predefined resource sharing policies. With the clear context in this section, we use SLA to represent TRSLA for convenience.

4.1. SLA definition and resource-sharing policies

To facilitate resource sharing, the SLA contents are extended to include the QoS and resource commitments for the underloaded period according to a *call-level differentiation* concept. The SLA specifications are as follows:

- A nominal capacity C is allocated to the SLA in accordance with the target border-crossing traffic arrival rate to satisfy the target QoS performance.
- During operation, according to the actual traffic arrival rate for the SLA (informed by the access domain or measured by a traffic monitor at the ingress router), two resource utilization states are defined for the SLA: *lendable* state if the actual rate is smaller than the target rate,[3] and *unlendable* state otherwise.
- In the lendable state, a *protection bandwidth* $R(\leq C)$ is calculated according to the QoS specifications for the underloaded period. The amount of bandwidth, R, is reserved for the SLA. The spare bandwidth, $C - R$, can be exploited by related SLAs (including both lendable and unlendable ones), by the CS paradigm.
- In the unlendable state, the nominal capacity is guaranteed. The SLA may accept overloaded traffic, by borrowing bandwidth from the lendable SLAs. The traffic flows accepted with borrowed bandwidth are tagged as *out* profile calls.
- When the SLA changes back to the unlendable state from the lendable state, the protection bandwidth is increased to the nominal capacity to claim back resources of the SLA. Some tagged traffic flows from the borrower SLAs may be preempted during the bandwidth claiming.

In the above SLA definition, the traffic arrival rate under consideration can be at the packet level or at the call level. Normally, the traffic load and QoS performance at both levels are of importance. A popular and efficient approach to facilitate the QoS control and resource allocation is the notion of *effective bandwidth* [3, 19]. Each traffic flow is allocated an effective bandwidth which encapsulates various packet level issues, such as burstiness and QoS (delay, jitter, loss) at network elements. The SLA capacity, at call level as the maximum number of bandwidth guaranteed calls, is then properly contracted to satisfy customers' call level QoS requirements at an engineered call arrival rate [7].

We use the term "lendable state" instead of the common term "underloaded state" to emphasize that

[3] To avoid unnecessary adjustment of the resource allocation due to small statistical fluctuations in the measurement result, a message indicating a new arrival rate is generated only when the measured rate deviation exceeds a predefined threshold, e.g., 10% of the target rate, and remains stable for a while.

an underloaded SLA can share its spare capacity with other SLAs under the following policy constraints.

1) *Spare bandwidth detection (SBD) policy:* the amount of spare bandwidth is determined by the protection bandwidth R required to guarantee the QoS in the underloaded period. Each SLA may apply different policy rules to determine the protection bandwidth and the associated spare capacity. For example, if the SLA is negotiated to guarantee the call blocking probability, the Erlang-B formula can then be used to calculated the protection bandwidth [10]. For a stable SLA management system, sufficient protection bandwidth should be reserved so that the QoS performance of an underloaded SLA is better than or at least as good as the target QoS specification; otherwise, malicious overloading would be encouraged.
2) *Call-level differentiation (CLD) policy:* traffic flows associated with an overloaded SLA are tagged as call-level *out* profile traffic if they are accepted with the borrowed bandwidth. The possibility of preemption of the *out* calls can be considered for the QoS differentiation between the *in* traffic and the *out* traffic; the counterpart differentiation scheme at the packet level is the AF PHB. Such a call-level differentiation policy efficiently utilizes the spare capacity as well as avoids SLA violations. The call-level differentiation can bring a more customer-friendly service model by sending a message of the SLA load status and flow admission status to the customer before the data transmission. The customer can then determine to continue or try at a later time, or send the most important information first. Such a service model is further validated by the fact that per-flow signalling, e.g., session initial protocol (SIP) [35], is supported in the next-generation all-IP services.
3) *Spare bandwidth sharing (SBS) policy:* between an ingress/egress pair, the available spare bandwidth may be shared by multiple borrowers. The network operator needs to define some policy rules for a fair spare bandwidth distribution. A reasonable sharing policy is to give those borrower SLAs associated with wireless domains a higher priority over those with the wireline domains, because the wireless spectrum is much more precious than the wireline bandwidth and the wireless calls already in service should be protected. A fair sharing policy based on properly designed billing schemes is also applicable.
4) *Dynamic spare bandwidth distribution (DSBD) policy:* the spare bandwidth associated with an SLA is determined at the edge, which should be properly distributed to the paths and links where the bandwidth can be exploited by the borrowers. With the dynamics of the loading status between different ingress/egress pairs, the locations within the network where bandwidth borrowing happens also change dynamically. Therefore, dynamic spare bandwidth distribution schemes are required for the maximum resource utilization.

4.2. Network status record for resource sharing

In [9, 10], we refer to the inter-SLA resource sharing as *bandwidth borrowing*; both inter-SLA sharing and bandwidth borrowing are used interchangeably in the following for convenience. When bandwidth borrowing happens, the related unlendable and lendable SLAs are termed as *borrower SLAs* and *lender SLAs*, respectively. Within the path-oriented transit domain, all traffic trunks of a lendable SLA are termed as *lender trunks*. A traffic trunk associated with a borrower SLA is termed as a *borrower trunk* when the trunk runs out of its nominal capacity and borrows bandwidth to service traffic flows. A *spare route (path)* is a route (path) along which an *out* profile flow can be successfully accepted via bandwidth borrowing. For a lendable SLA, the protection bandwidth $R_{s,\sigma}$ (correspondingly the spare bandwidth) is first evenly distributed to each traffic trunk, denoted as $R_{s,r}$, which will be dynamically adjusted in operation according to the DSBD policy.

For bandwidth borrowing, data structures are designed to record the traffic load information, flow information and trunk resource information in edge routers and the bandwidth broker.

4.2.1. Data in the bandwidth broker

All the information kept in the bandwidth broker is organized into three tables: *Route Table, Trunk Status Table*, and *SLA Status Table*. The topology and trunk deployment information is organized into a route table, where we can find all the links of a traffic trunk and all the trunks crossing a certain link. The network planning results, current network resource usages (denoted as $U_{s,\sigma}$ for an SLA and $U_{s,r}$ for a trunk), and bandwidth borrowing information are organized into the trunk status table. Particularly, a trunk utilization status (TUS) flag is used to indicate the current trunk utilization status. TUS flag is set as *notfull* when $U_{s,r} < R_{s,r}$, as *full* when $R_{s,r} \leq U_{s,r} \leq C_{s,r}$, and as *borrowing* when $U_{s,r} > C_{s,r}$. The SLA status table is

an array indicating the status of each SLA, *lendable or unlendable*.

4.2.2. Data in the edge routers

Each ingress router records the per-flow information in a *Flow Record Table* which is organized according to the traffic trunks where the traffic flows are placed. A flow information record includes the traffic flow ID, the effective bandwidth allocation, and the flow admission status (FAS) flag. This flag indicates what bandwidth is used to admit the traffic flow. FAS flag is set as *nor-admit* (normally admitted) if the traffic flow is admitted into a *notfull* trunk, and set as *bor-admit* if admitted through bandwidth borrowing. A *bor-admit* flow is considered as an *out* profile flow.

4.3. Bandwidth borrowing

In the PBISRA scheme, the network operator can freely choose its SBD, CLD, SBS, and DSBD policies. In any scenario, the network status and resource utilization information are mapped to a uniform set of information records as presented in the previous subsection, based on which the bandwidth borrowing scheme [9, 10] can then be applied to implement the inter-SLA resource sharing.

In the bandwidth borrowing, the dynamic resource sharing is implemented at the trunk level. When a class s traffic flow arrives, a trunk (s, r) between the ingress/egress pair is selected according to a routing algorithm. If the nominal capacity or protection bandwidth is run out of, the traffic trunk then tries to grab the spare capacity from the lender trunks. Let $r \supset \ell$ represent a route passing link ℓ. An out profile flow with bandwidth allocation[4] e_s can be accepted at link ℓ by exploiting the spare bandwidth, if

$$U_{s,r} + e_s \leq C_\ell - \sum_{(s',r'):(s',r')\neq(s,r),r'\supset\ell} \max(U_{s',r'}, R_{s',r'}) \qquad (6)$$

where C_ℓ is the link capacity and $\max(U_{s',r'}, R_{s',r'})$ guarantees that the capacity of $R_{s',r'}$ is dedicated to trunk (s', r'). For convenience of expression, the algorithm for trunk resource sharing at a link is referred to as the `trunkshare` algorithm.

An out profile flow can be accepted only when bandwidth borrowing via `trunkshare` is successful at all links along the selected path, which is termed as a *lendable route*. If multiple lendable routes exist, the route with the largest lendable bandwidth is picked to service the new flow with the objective to protect the out profile call from future preemption as much as possible. It is noteworthy that the hop-by-hop checking of resource availability here is not through signaling, but through looking up the route table and resource usage information recorded in the bandwidth broker. Hence, there is no scalability problem and the CAC overhead time is expected to be small. The `trunkshare` checking at each link and the route selecting procedure are summarized as the `sparestroute` subroutine to be used in the CAC procedure.

4.4. Policy-based routing/CAC

The flowchart of the proposed routing/CAC algorithm to implement the PBISRA is given in Fig. 4. The algorithm follows the principle of "borrow bandwidth only when necessary" and works as follows.

1) When the protection bandwidth $R_{s,\sigma}$ (which is set as $C_{s,\sigma}$ for normally loaded or overloaded SLAs) dedicated to an SLA is available, the new call will be treated as in profile and its acceptance is guaranteed. In the underloaded case, the dedicated protection bandwidth is determined according to the SBD policy.
2) Upon a *nor-admit* (i.e., in profile) admission, the need of bandwidth claiming is checked in the subroutine `claimbackband`. The reason is that the SLA under consideration may just increase its dedicated protection bandwidth due to the increase of traffic load, where the previous spare capacity (but not available now) may have been borrowed by other SLAs and should be claimed back if itself needs to use it. During the resource-claiming procedure, some *bor-admit* calls have to be preempted according to the CLD policy. The CLD policy defines that, upon a preemption request, how many and what types of out profile calls should be preempted.
3) When the dedicated bandwidth is run out of, out profile calls may be accepted via bandwidth borrowing. If multiple borrowers exist to share the spare bandwidth between an ingress/egress pair, the allocation of the spare capacity to each borrower is controlled by the SBS policy. Note that the lender trunk is not allowed to go to the *borrowing* state ($U_{s,r} > C_{s,r}$) for stable resource management.
4) Upon each admission via inter-SLA resource sharing, the subroutine `pushprotband` is called to dynamically adjust the protection bandwidth distribution among related lender trunks, so that

[4] we assume that all the flows belonging to the same service class have the same bandwidth allocation.

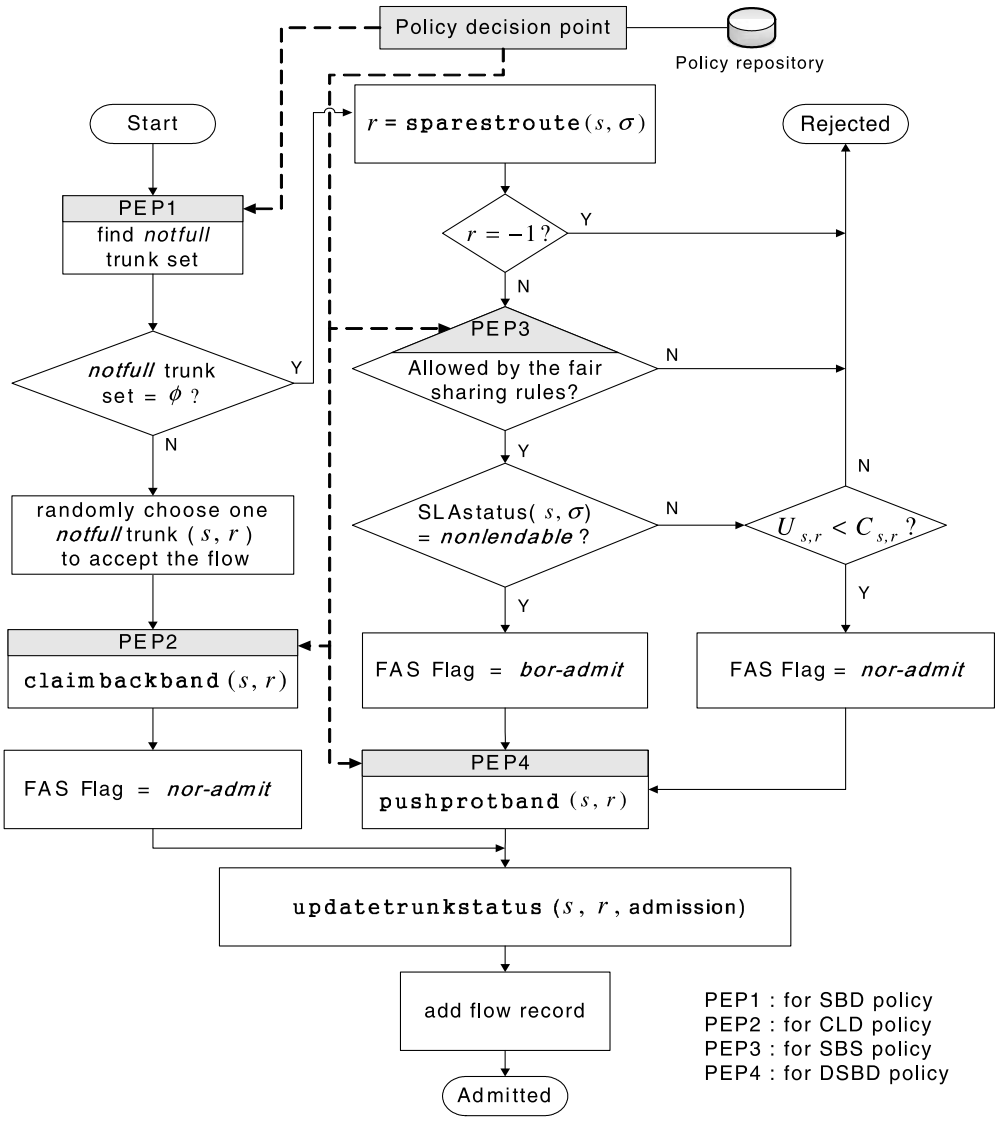


Figure 4 The routing/CAC procedure for PBISRA

the spare bandwidth can concentrate on the routes where it can be more efficiently utilized. The specific algorithm for such distribution adjustment should be designed according the DSBD policy, which defines the adjustment frequency, overhead time, and centralized/distributed implementation criteria that need to be met by the adjustment algorithm.

5) The network manager has the freedom to change the policy rules to control the tradeoff among resource utilization, resource sharing fairness, and operation cost. However, the same routing/CAC procedure can always be applied under different policy combinations.

The CAC procedure and associated implementation of all the subroutines presented in [10] are an example of the general procedure given in Fig. 4, where the SLA is to guarantee the call blocking probability (CBP). To map the general PBISRA CAC procedure to the specific case considered in [10], the corresponding policies can be implemented as follows. The SBD policy determines the protection bandwidth for an underloaded SLA, which should be sufficient to guarantee its CBP not worse than the target QoS specification. The CLD policy defines that the preemption is executed by three rules: (a) enough bandwidth should be returned to the original owner to service the new in-profile flow; (b) at a link, the bandwidth is claimed from multiple borrower trunks where the fraction from each trunk is proportional to its spare bandwidth usage; (c) within a borrower trunk, the latest *bor-admitted* arrival is preempted first. The DSBD policy defines that the dynamic spare band-

width distribution adjustment should be implemented in a distributed manner for good scalability and small overhead, and the adjustment should be operated at the inter-arrival time scale to exploit the call-level statistical multiplexing gain.

The fairness issue has not been considered in [9, 10]. In the PBISRA CAC procedure, the SBS policy is used to control the resource sharing according to some fairness rules. Specifically, in the next section, we evaluate a *proportional SBS policy*.

5. Performance evaluation

In this section, we present results from some case studies to demonstrate the performance of the proposed access domain and transit domain resource allocation techniques for the policy-based wireless/wireline interworking.

5.1. Adaptive CAC in the access domain

This example presents numerical results based on the mathematical analysis given in Section 3 to demonstrate the adaptive adjustment of the CAC parameters for efficient resource utilization. Both cellular-only and cellular/WLAN interworking access domains are considered. With a change of the voice call arrival rate, the CAC parameters are determined adaptively according to Section 3.3. The main system parameters used in the numerical analysis include: $C^c = 2$ Mbps, $C^w = 11$ Mbps, $(\mu_v^X)^{-1} = 10$ min, $(\mu_v^Y)^{-1} = 140$ sec, $l_{di} = 512$ KB. When the voice call arrival rate λ_v changes from 0.3 to 0.4 calls/sec, the maximum number of data calls accommodated by a cell is calculated and compared in Fig. 5 with respect to separate cellular and WLAN operations and cellular/WLAN interworking. In each run of calculation, the voice and data handoff rates, h_v and h_d, are solved by the rule that the mean handoff arrival rate to a cell is equal to the mean handoff departure rate to neighboring cells.

From Fig. 5, we can have the following observations: (1) the data admission region under cellular/WLAN interworking is improved as compared with that when cellular networks and WLANs operate separately. This is due to the effective exploitation of the complementary strengths of the cellular network and WLANs in supporting voice and data services; (2) The improvement becomes more significant with a larger voice traffic load. Because the capability of WLANs in supporting realtime voice traffic is very limited, with a larger voice traffic load, the achievable WLAN throughput is jeopardized to a higher degree. When cellular/WLAN interworking is applied, the WLAN bandwidth can be properly shared between voice and data services by controlling the voice and data admission regions. Then, the WLAN can be prevented from operating in its inefficient areas of carrying too much voice traffic. The overall resource utilization is maximized when a balance is achieved between the bandwidth allocations to voice and data services.

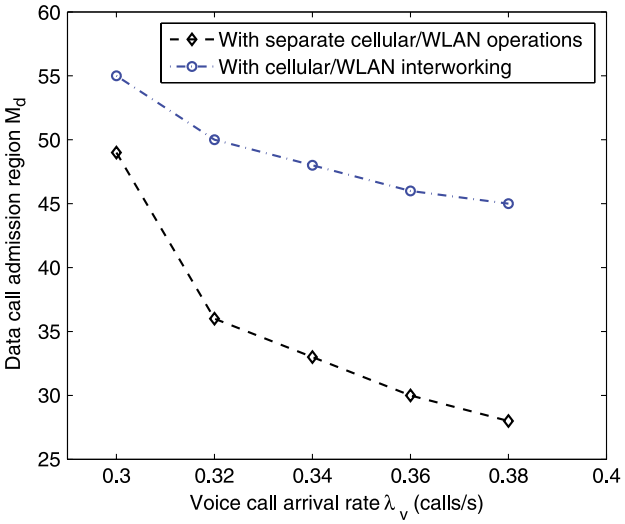


Figure 5 The CAC parameter M_d of data calls

5.2. Efficient and fair inter-SLA resource sharing

In this case study, we present the computer simulation of a small DiffServ/MPLS transit domain to illustrate the efficiency of the PBISRA to exploit the spare capacity from the underloaded TRSLAs. We consider SLAs to guarantee a CBP[5] as those considered in [10], with the SBD/CLD/DSBD policies being properly mapped.

The topology of the network is shown in Fig. 6. Five SLAs are supported in this network, and each SLA is served with parallel traffic trunks. The units used for related measures are second for time, capacity unit (c-unit) for link/trunk/SLA capacity, call/second for call arrival rate. Assume Poisson arrivals for each SLA, and exponentially distributed call holding times with mean $\frac{1}{\mu} = 1$. Each call for each SLA is assigned a normalized bandwidth of 1 unit. Erlang-B formula $E(\lambda, C)$ is used to calculate the CBP. The target CBP for each SLA is 10^{-2}. The target call arrival rate for the five SLAs is $(\lambda_p^1, \lambda_p^2, \lambda_p^3, \lambda_p^4, \lambda_p^5) = (46.9, 29, 29, 29,$

[5] In this section, the SLA and associated QoS are of importance in the transit domain.

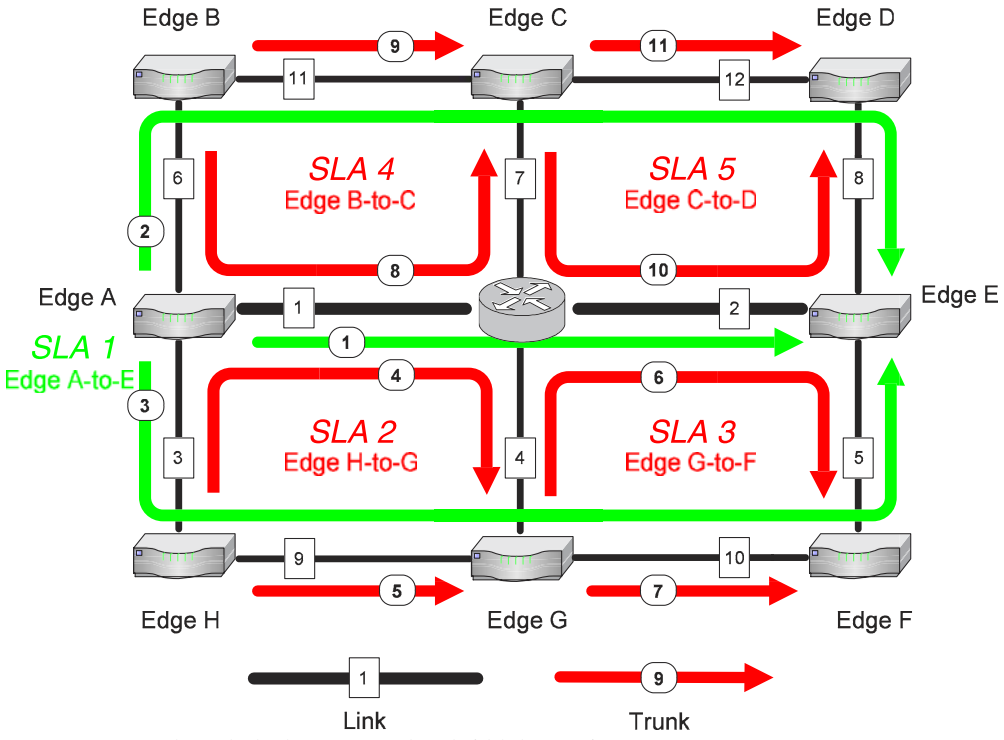


Figure 6 Topology, TRSLAs and trunk deployment for bandwidth borrowing

29), and corresponding capacity planning is $(C_1, C_2, C_3, C_4, C_5) = (60, 40, 40, 40, 40)$. The SLA capacity is evenly distributed to the related traffic trunks. The capacity of each link is 60 for links 1 and 2, and 40 for the other links, which have been properly tailored to obtain an accurately dimensioned network.

5.2.1. Resource utilization improvement:

We first briefly summarize the example-1 presented in [10] to demonstrate the resource utilization improvement from bandwidth borrowing, and then we extend the example to achieve a fair spare-bandwidth distribution under a proportional SBS policy. In the example, traffic for each SLA starts with the specified call arrival rate, and the call arrival rates for some SLAs are changed at certain time points to create the overloaded and underloaded periods. The actual call arrival rate for each SLA, $\lambda_d^i (i = 1, \ldots, 5)$, and the corresponding protection bandwidth for each trunk are given in Table 1. The measured call blocking probability for each SLA is presented in Fig. 7.

All SLAs start from the engineered load and correspondingly the target CBP. During the time period of $(6000, 12000)$, SLA-1 has the CBP of $E(62.6, 60) \approx 0.1208$ due to the overloading. During the time period of $(12000, 36000)$, SLA-2 and SLA-5 become underloaded. According to Fig. 6, we can see that SLA-1 can utilize the spare capacity from SLA-2 and SLA-5 via the bandwidth borrowing along trunk-1. In operation, the dynamic spare bandwidth distribution algorithm for SLA-2 and SLA-5 then pushes 16 units of their 17-unit spare capacity to trunk-4 and trunk-10, respectively, so that the spare capacity can be used by SLA-1 along link-1 and link-2. The 1 unit of spare capacity reserved on trunk-5 and trunk-11 is to maintain the "spare" property of the routes. Along link-1 and link-2, the heavily loaded trunk-1 grabs almost all the spare capacity and achieves the CBP of $E(62.6, 60 + 16) \approx 0.0125$. After $t = 36000$, SLA-5

Table 1 The call arrival rate and protection bandwidth for each SLA

t	$\lambda_d^1(R_1, R_2, R_3)$	$\lambda_d^2(R_4, R_5)$	$\lambda_d^3(R_6, R_7)$	$\lambda_d^4(R_8, R_9)$	$\lambda_d^5(R_{10}, R_{11})$
0	46.9 (20,20,20)	29 (20,20)	29 (20,20)	29 (20,20)	29 (20,20)
6,000	62.6 (20,20,20)	29 (20,20)	29 (20,20)	29 (20,20)	29 (20,20)
12,000	62.6 (20,20,20)	14.4 (11,12)	29 (20,20)	29 (20,20)	14.4 (12,11)
36,000	62.6 (20,20,20)	14.4 (11,12)	29 (20,20)	29 (20,20)	29 (20,20)

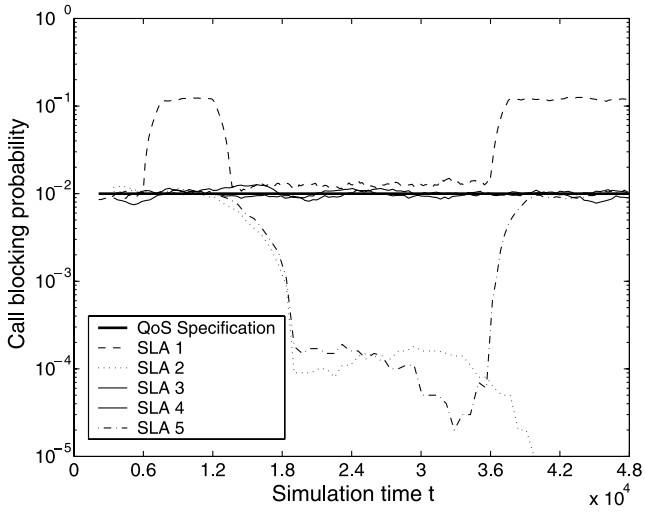


Figure 7 The call blocking probability of each SLA by bandwidth borrowing

changes back to the *unlendable* state and the spare capacity on link-2 is not available anymore, the bandwidth borrowing along trunk-1 stops.

Note that by bandwidth borrowing the efficient resource utilization is achieved with SLA compliance. That is, the CBP in underloaded and normally loaded cases does not exceed the target specification of 0.01.

5.2.2. A proportional SBS policy

In the previous example, only one borrower SLA is considered; but in practice, multiple borrower SLAs may exist between a certain ingress/egress pair. How to fairly distribute the spare capacity among those borrowers is an important issue, which is controlled by the SBS policy. In this example, we use a proportional spare bandwidth distribution scheme as an implementation example of the SBS policy control. For simplicity, we consider the case of two borrower SLAs, i.e., W_1 and W_2, residing between an ingress/egress pair. With the proportional SBS policy, each borrower is assigned a weighting factor, i.e., w_1 and w_2, respectively, for spare bandwidth distribution. The spare bandwidth occupied by each borrower SLA should satisfy

$$Z_{W_i} \geq \frac{w_i}{w_1 + w_2} Z_\sigma \quad i = 1, 2 \tag{7}$$

where Z_σ denotes the spare bandwidth available between an ingress/egress pair and the "$\geq$" sign is due to the statistic multiplexing effect.

How to map the proportional SBS policy to specific admission control schemes that can be used by the PEP-3 in Fig. 4 depends on the statistical characteristics of the traffic arrival process. Following the previous example where the Erlang-B formula can be applied, the proportional distribution control is relative simple. Specifically, for an borrower SLA that successfully occupies the spare capacity of Z in the steady-state, the average channel occupancy (i.e., the efficient bandwidth usage (EBU) considered in [7]) in terms of call-numbers can be calculated as

$$U = \lambda_d[1 - E(\lambda_d, C + Z)] \tag{8}$$

where the normalization of $\frac{1}{\mu} = 1$ is applied. Equation 8 can then be used for the spare-bandwidth distribution control in the following three steps:

1) For a borrower SLA, the target spare bandwidth allocation is determined according to equation 7 as $\frac{w_i}{w_1+w_2} Z_\sigma$.
2) The target EBU is then calculated from equation 8. According to the PASTA (Poisson arrivals see the time average) theory, such an EBU can be observed by each arrival, which leads to a VP-based [6, 33] spare bandwidth distribution scheme.
3) For a to-be-accepted out profile call (for which a spare trunk has been searched via the bandwidth borrowing), it is admitted if the instantaneous SLA bandwidth usage has not reached the target EBU yet. However, if the instantaneous SLA bandwidth usage reaches or exceeds the target EBU, the call can only be admitted when the leftover spare bandwidth after the admission is enough for an extra *trunk reservation* [6, 7].

To demonstrate the performance of the proportional SBS policy control, we extend the previous bandwidth

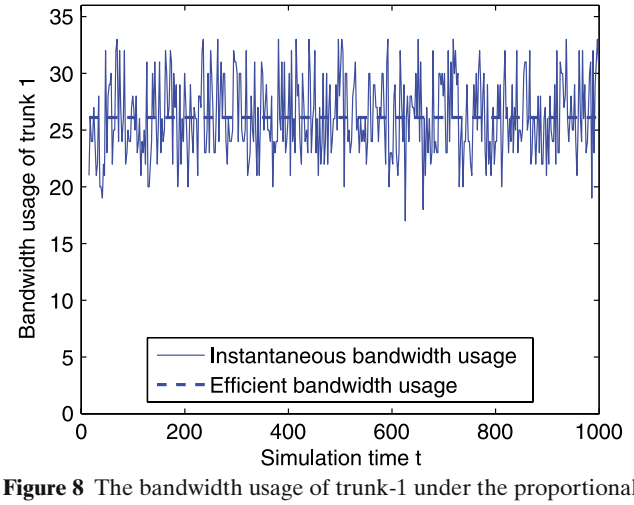


Figure 8 The bandwidth usage of trunk-1 under the proportional SBS policy

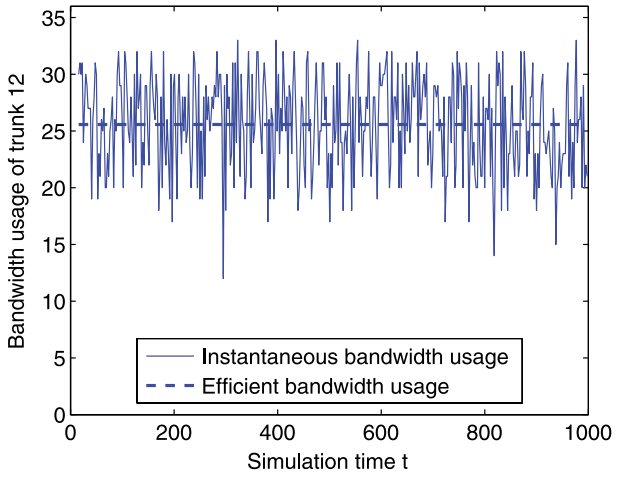


Figure 9 The bandwidth usage of trunk-12 under the proportional SBS policy

borrowing example. We use the same network dimensioning and configuration, with the following changes:

- An SLA-6 between ingress A and egress E is added. SLA-6 is supported by three trunks, i.e., trunk-12, 13, and 14 that share the same paths with SLA-1's trunk-1, 2, and 3, respectively. The engineered call arrival rate for SLA-6 is also 46.9, with a nominal SLA capacity of 60 to guarantee the target CBP of 0.01.
- The link capacity is adjusted correspondingly to maintain a well-dimensioned network.
- We focus on the bandwidth borrowing scenario where $(\lambda_d^1, \lambda_d^2, \lambda_d^3, \lambda_d^4, \lambda_d^5) = (46.9 * 2, 14.4, 29, 29, 14.4, 46.9 * 1.5)$. The SLA-1 traffic is from a wireline access domain and SLA-6 traffic is from a wireless access domain, and they compete for the spare capacity between edges A and E with weighting factors w_1:w_6 = 1:3. Such a weighting factor assignment reflects that the SBS policy allocates at least 75% of the spare capacity to the wireless traffic. The trunk reservation is set as 3 calls.

The simulation results of the spare bandwidth distribution under the proportional SBS policy are plotted in Figs. 8 and 9. Because the bandwidth borrowing can only be implemented along link-1 and link-2 according to the network configuration, the total spare bandwidth of 16 units from SLA-2 and SLA-5 is shared by trunk-1 and trunk-12 along that path. Both the instantaneous bandwidth usage (IBU) and the EBU are shown in the figures. From the results we can see that the IBUs of both trunk-1 and trunk-12 does not exceed $20 + 16 - 3 = 33$, due to the trunk reservation scheme. The EBUs of trunk-1 and trunk-12 are 26.11 and 25.58, respectively. Theoretically, the target EBUs of SLA-1 and SLA-6 are 62.18 and 65.02, respectively, that is, the corresponding target EBUs of the two borrower trunks (trunk-1 and trunk-12) are 22.18 and 25.02. The result that the achieved EBUs of the borrower SLAs are larger than their targets implies that the VP-based bandwidth distribution can efficiently utilize the statistical multiplexing gain and satisfy the proportional SBS policy defined in equation 7.

It is noteworthy that, although we only consider the spare bandwidth distribution between two borrowers, the VP based scheme can be readily extended for fair bandwidth sharing among multiple (more than two) borrowers, for example by the *capacity resizing approach* presented in [13]. In some scenarios where the Erlang-B formula fails, new bandwidth distribution schemes should be designed to execute the proportional SBS policy. However, the PBISRA architecture and the routing/CAC procedure are still applicable.

6. Conclusions

We have proposed efficient resource allocation techniques for a policy-based wireless/wireline interworking architecture. Specifically, an engineered priority scheme and associated mathematical modeling are proposed for a wireless access domain supporting voice/data integrated services, where the optimal CAC parameters are adaptively calculated to meet the CASLAs with efficient resource utilization. The admission region can be used as an abstract representation of the network resources to facilitate policy-based resource allocation. The engineered priority scheme and the proposed CAC techniques can also be extended to analyze a cellular/WLAN integrated wireless domain.

The border-crossing traffic from access domains is served by a DiffServ/MPLS core according to TRSLAs. A dynamic inter-TRSLA resource sharing technique is developed for the transit domain to deal with the traffic load fluctuations from upstream access domains. The spare capacity of underloaded TRSLAs can be efficiently exploited by the overloaded TRSLAs under the SBD, CLD, SBS, and DSBD policies, while the SLA compliance is always guaranteed.

Acknowledgement This research was supported by a Postdoctoral Fellowship from Natural Science and Engineering Research Council of Canada (NSERC).

References

1. Ahmavaara, K., Haverinen, H., & Pichna, R. (2003, November). Interworking architecture between 3GPP and WLAN systems. *IEEE Communications Magazine*, *41*, 74–81.

2. Armitage, G. (2000, January). MPLS: the magic behind the myths. *IEEE Communications Magazine, 38*, 124–131.
3. Berger, A. W., & Whitt, W. (1998, August). Effective bandwidth with priorities. *IEEE/ACM Transactions on Networking*, *6*, 447–460.
4. Bernaschi, M., Cacace, F., Iannello, G., Za, S., & Pescape, A. (2005, June). Seamless internetworking of WLANs and cellular networks: architecture and performance issues in a mobile IPv6 scenario. *IEEE Wireless Communication, 12*, 73–80.
5. Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., & Weiss, W. (1998, December). An architecture for differentiated services. *Internet RFC, 2475*.
6. Borst, S. C., & Mitra, D. (1998, June). Virtual partitioning for robust resource sharing: computational techniques for heterogeneous traffic. *IEEE Journal on Selected Areas in Communications*, *16*, 668–678.
7. Bouillet, E., Mitra, D., & Ramakrishnan, K. G. (2002, May). The structure and management of service level agreements in networks. *IEEE Journal on Selected Areas in Communications*, *20*, 691–699.
8. Cai, L., Shen, X., Mark, J. W., Cai, L., & Xiao, Y. (2005, August). Voice capacity analysis of WLAN with unbalanced traffic. In *Proceedings 2nd International Conference QoS in Heterogeneous Wired/Wireless Networks (QShine'05)*. Orlando, Florida.
9. Cheng, Y. (2003). Efficient resource allocation in differentiated services networks. PhD thesis, Unviersity of Waterloo.
10. Cheng, Y. & Zhuang, W. (2006, June) Dynamic inter-SLA resource sharing in path-oriented differentiated services networks. *IEEE/ACM Transactions on Networking*, to appear.
11. Cheng, Y., & Zhuang, W. (2002, May). DiffServ resource allocation for fast handoff in wireless mobile Internet. *IEEE Communications Magazine*, *40*(5), 130–136.
12. Downey, A. B. (2001, June). The structural cause of file size distributions. *ACM SIGMETRICS Performance Evaluation Review*, *29*, 328–329.
13. Garg, R., & Saran, H. (2000). Fair bandwidth sharing among virtual networks: a capacity resizing approach. In *Proceedings IEEE INFOCOM'00*, 255–264.
14. Gibbens, R. J., & Kelly, F. P. (1995, September). Network programming methods for loss networks. *IEEE Journal on Selected Areas in Communications*, *13*, 1189–1198.
15. Heinanen, J., Baker, F., Weiss, W., & Wroclawski, J. (1999, June). Assured forwarding PHB group. *Internet RFC 2597*.
16. Huang, Y.-R., Lin, Y.-B., & Ho, J. M. (2000, March). Performance analysis for voice/data integration on a finite mobile systems. *IEEE Transactions on Vehicular Technology*, *49*, 367–378.
17. Jacobson, V., Nichols, K., & Poduri, K. (1999, June). An expedited forwarding PHB. *Internet RFC 2598*.
18. Jeon W. S., & Jeong, D. G. (2001, January). Call admission control for mobile multimedia communications with traffic asymmetry between uplink and downlink. *IEEE Journal on Selected Areas in Communications*, *50*(1), 59–66.
19. Kelly, F. P. (1996). Notes on effective bandwidth. In F. P. Kelly, S. Zachary, & I. Ziedins (Eds.), *Stochastic networks: theory and applications,* (pp. 141–168).Oxford, U.K.: Oxford University Press.
20. Lai, F. S., Misic, J., & Chanson, S. T. (1998, October). Complete sharing versus partitioning: quality of service management for wireless multimedia networks. In *Proceedings of the 7th International Conference on Computer Communications & Networks*, pp. 584–593.
21. Leong, C. W. (2001). CAC for voice and data services in wireless communications. MASc thesis, University of Waterloo.
22. Leong, C. W., Zhuang, W., Cheng, Y., & Wang, L. (2006, March). Optimal resource allocation and adaptive call admission control for voice/data integrated cellular networks. *IEEE Transactions on Vehicular Technology*, *54*, 654–669.
23. Li, B., Li, L., Li, B., Sivalingam, K. M., & Cao, X.-R. (2004, May). Call admission control for voice/data integrated cellular networks: performance analysis and comparative study. *IEEE Journal on Selected Areas in Communications*, *22*, 706–718.
24. Maniatis, S. I., Nikolouzou, E. G., & Venieris, I. S. (2002, August). QoS issues in the converged 3G wireless and wired networks. *IEEE Communications Magazine*, *40*(8), 44–53.
25. Mitra, D., & Ramakrishnan, K. G. (1999). A case study of multiservice, multipriority traffic engineering design for data networks. In *Proceedings of IEEE GLOBECOM'99*, *1B*, 1077–1083.
26. Moon, B., & Aghvami, H. (2003, October). DiffServ extension for QoS provisioning in IP mobility environments. *IEEE Wireless Communications Magazine*, *10*(5), 38–44.
27. Naghshineh, M., & Acampora, A. S. (1995). QoS provisioning in micro-cellular networks supporting multimedia traffic. In *IEEE Proceedings INFOCOM'95*, 1075–1084.
28. Nichols, K., Jacobson, V., & Zhang, L. (1999, July). A two-bit differentiated services architecture for the Internet. *IETF RFC 2638*.
29. Ramjee, R., Towsely, D., & Nagarajan, R. (1997). On optimal call admission control in cellular networks. *Wireless Networks*, *3*(1), 29–41.
30. Ramos, N., Panigrahi, D., & Dey, S. (2005, July–August). Quality of service provisioning in 802.11e networks: challenges, approaches, and future directions. *IEEE Network*, *19*, 14–20.
31. Terzis, A., Wang, L., Ogawa, J., & Zhang, L. (1999). A two-tier resource management model for the Internet. In *Proceeding IEEE GLOBECOM'99*, *3*, 1779–1791.
32. Trimintzios, P., et al. (2001, May). A management and control architecture for providing IP differentiated services in MPLS-based networks. *IEEE Communications Magazine*, *39*, 80–88.
33. Yao, J., Mark, J. W., Wong, T. C., Chew, Y. H., Lye, K. M., & Chua, K.-C. (2004, May). Virtual partitioning resource allocation for multiclass traffic in cullular systems with QoS constraints. *IEEE Transactions on Vehicular Technology*, *53*, 847–864.
34. Yavatkar, R., Pendarakis, D., & Guerin, R. (2000, January). A framework of policy-based admission control. *IETF RFC 2753*.
35. Zhuang, W., Gan, Y.-S., Loh, K.-J., & Chua, K.-C. (2003, May/June). Policy-based QoS architecture in the IP multimedia subsystem of UMTS. *IEEE Network*, *17*, 51–57.
36. Zhuang, W., Gan, Y.-S., Loh, K.-J., & Chua, K.-C. (2003, November). Policy-based QoS management architecture in an integrated UMTS and WLAN environment. *IEEE Communications Magazine*, *41*, 118–125.

Yu Cheng received the B.E. and M.E. degrees in Electrical Engineering from Tsinghua University, Beijing, China, in 1995 and 1998, respectively, and the Ph.D.degree in Electrical and Computer Engineering from the University of Waterloo, Waterloo, ON, Canada, in 2003. From September 2003 to August 2004 he was a postdoctoral fellow in the Department of Electrical and Computer Engineering, University of Waterloo. Since September 2004, he has been a postdoctoral fel-lowin the Department of Electrical and Computer Engineering, University of Toronto, ON, Canada. Hisresearch interests include service-oriented networking, autonomic network management, Internet perfor-mance analysis, quality of service provisioning, wireless networks, and wireless/wireline interworking. He received a Postdoctoral Fellowship Award from the Natural Sciences and Engineering Research Councilof Canada (NSERC) in 2004. He is a Member of IEEE.

Wei Song received a B.S. degree in electrical engineering from Hebei University, Baoding, China, in 1998 and an M.S. degree in computer science from Beijing University of Posts and Telecommunications, China, in 2001. She is currently working toward a Ph.D. degree at the Department of Electrical and Computer Engineering, University of Waterloo, ON, Canada. Her current research interests include resource allocation and QoS provisioning for the integrated cellular networks and wireless local area networks (LANs).

Weihua Zhuang received the B.Sc. and M.Sc. degrees from Dalian Maritime University, Liaoning, China, and the Ph.D. degree from the University of New Brunswick, Fredericton, NB, Canada, all in electrical engineering. Since October 1993, she has been with the Department of Electrical and Computer Engineering, University of Waterloo, ON, Canada, where she is a Professor. She is a co-author of the textbook Wireless Communications and Networking (Prentice Hall, 2003). Her current research interests include multimedia wireless communications, wireless networks, and radio positioning. Dr. Zhuang is a licensed Professional Engineer in the Province of Ontario, Canada. She received the Outstanding Performance Award in 2005 from the University of Waterloo and the Premier's Research Excellence Award (PREA) in 2001 from the Ontario Government. She is a Senior Member of IEEE. She is an Editor/Associate Editor of IEEE Transactions on Wireless Communications, IEEE Transactions on Vehicular Technology, and EURASIP Journal on Wireless Communications and Networking.

Alberto Leon-Garcia received the B.S., M.S., and Ph.D. degrees in electrical engineering from the University of Southern California, in 1973, 1974, and 1976 respectively. He is a Full Professor in the Department of Electrical and Computer Engineering, University of Toronto, ON, Canada, and he currently holds the Nortel Institute Chair in Network Architecture and Services. In 1999 he became an IEEE fellow for For contributions to multiplexing and switching of integrated services traffic. Dr. Leon-Garcia was Editor for Voice/Data Networks for the IEEE Transactions on Communications from 1983 to 1988 and Editor for the IEEE Information Theory Newsletter from 1982 to 1984. He was Guest Editor of the September 1986 Special Issue on Performance Evaluation of Communications Networks of the IEEE Selected Areas on Communications. He is also author of the textbooks Probability and Random Processes for Electrical Engineering (Reading, MA: Addison-Wesley), and Communication Networks: Fundamental Concepts and Key Architectures (McGraw-Hill), co-authored with Dr. Indra Widjaja.

Rose Qingyang Hu received her B.S.E.E from the University of Science and Technology of China, her M.S from Polytechnic University, Brooklyn, NY, and her Ph.D in EE from the University of Kansas. After receiving her Ph.D., she had worked for Nortel Networks as a senior member of scientific staff for approximately three years and then worked for a start-up company, Yotta Networks, as a senior systems engineer for another one year. Since January 2002, she has been with Electrical and Computer Engineering Department of Mississippi State University as an assistant professor. She has published about 30 journal and conference papers and has been awarded 1 USA patent. Her current research interests include mobile broadband wireless access, wireless sensor networks, optical network mesh restoration, QoS and performance evaluations. She is a member of IEEE and a Member Phi Kappa Phi and Epsilon Pi Epsilon.