Autonomic hierarchical reconfiguration for wireless access networks

Konstantinos Samdanis *, Vasilis Friderikos, A. Hamid Aghvami

Centre for Telecommunications Research, King’s College London, Strand, London WC2R 2LS, UK

ARTICLE INFO

Article history:
Received 19 November 2007
Received in revised form 4 June 2008
Accepted 5 July 2008

Keywords:
Network management
Hierarchical QoS
Wireless access networks

ABSTRACT

A novel autonomic area reconfiguration mechanism is proposed to manage the hierarchical arrangement of routing protocols in wireless access networks. The reconfiguration mechanism monitors the traffic and mobility patterns to identify symptoms, which indicate a poor existing area partition. A re-partition algorithm is invoked on demand with the objective to revise the border between neighboring areas in order to achieve more efficient resource management. A centralized implementation is proposed and evaluated by comparison with the equivalent one without the autonomic area reconfiguration capability. Simulations illustrate that the autonomic area reconfiguration mitigate both routing and mobility overhead as well as dropping and blocking rates, suggesting better network performance with rapid response to traffic alternations.

1. Introduction

The fast adoption of the pure IP-based communications for wireless and mobile networks is creating new challenges for the Internet evolution. IP provides a common incorporation platform, which delivers ubiquitous computing integrating existing and emerging wireless as well as fixed wired networks. WLANs proposals like WiFi and HiperLAN are expected to complement existing 3G wireless cellular systems including UMTS and cdma2000 to provide high bit rates at particular hotspots. WMAN proposals like WiMAX, WiBro and HiperMAN aim to extend the limited coverage of hotspots to provide high speed wireless broadband services to either fixed or nomadic mobile users. More information about the emerging wireless broadband technologies is available in Kuran and Tugcu (2007).

The demand for higher rates and seamless mobility augment the desire for wireless coverage, causing expansion of wireless access networks. Such expansion introduce scalability concerns especially when the network is configured as a single domain. To achieve scalability, large size networks are divided into an hierarchy of smaller autonomous areas, selected to satisfy certain criteria but manually configured defining a static network arrangement. In wireless access networks, the static nature of an existing optimal formation may fail to reflect evolving traffic and mobility patterns. Traffic concentration in specific areas might cause poor network utilization due to unbalanced resource consumption between neighboring areas. Locations that attract population movements might be formed by the border between different areas causing excessive inter-area handover rates and increasing the routing and mobility overhead. Such incidents might occur on a temporary or permanent basis at an instant that is least expected. Therefore, the efficient configuration of a wireless hierarchical network is complex to manage and the manual network re-arrangement might prove costly as well as prone to human errors.

Our work concentrates on the design and evaluation of a novel autonomic area reconfiguration mechanism to manage the hierarchical arrangement of wireless access networks. The aim of the proposed scheme is to accommodate flexibility and improve responsiveness on refining the border between neighboring areas according to the perceived change in the traffic and mobility state. Its objectives are concentrated on the provision of load balancing among neighboring areas and on controlling the inter-area handover rates while introducing minimal changes into the existing hierarchical configuration. The benefits of the proposed approach are firstly the reduction of the routing and mobility overhead related to excessive inter-area handover rates and secondly the mitigation of the blocking as well as dropping associated with unbalance concentration of users and bandwidth consumption between adjacent areas. Ultimately, the hierarchical area reconfiguration increases the QoS perceived by end users while providing an efficient network resource management.

The means of identifying the network elements to be reconfigured is through a re-partition algorithm, which is invoked on demand based on inter-area handover rates and on per area resource consumption considering the bandwidth and user population. Its purpose on revising the partition border is to keep together Access routers (ARs) involved into high handover rates and assign into different areas ARs that create a high intra-area traffic concentration hotspot. The remaining of this paper is
organized as follows. Section 2 states our motivation, while Section 3 presents the related work. Section 4 illustrates the interactions between mobility and QoSR in wireless access networks considering a hierarchical arrangement. The proposed reconfiguration mechanism is described in Section 5, while Section 6 presents the simulation setup and illustrates the results. Finally, Section 7 summarizes our research and provides the future directions.

2. Motivation

Current routing protocols scale poorly when handling a large amount of routers as a single domain. Configuring hierarchy is a common practice to scale large size networks, where areas are defined as a group of routers, which are strongly related according to some sense. In wireless access networks this corresponding degree of association between members of the same area is affected significantly by mobility. Specifically, mobility defines the handover rates between ARs and affects the concentration of traffic load of certain network areas. Considering the dynamics of mobility, it is realized that an optimal hierarchical configuration is maintained through regular adjustments. Since such adjustments are required even on a temporary basis, addressing them manually is complex, costly and prone to error.

Motivated by the demand of the dynamic hierarchical reconfiguration and the problems related to manual delivery we proposed a novel autonomic area reconfiguration mechanism. The objective of the proposed mechanism is to autonomously manage the hierarchical configuration in order to reduce the overhead and avoid congestion on per area basis. Revising the hierarchical configuration is not simple, since re-partitioning the AR plane may cause further consequent adjustments on the core network. Fig. 1(b) demonstrates an example where the original partition illustrated in Fig. 1(a) is revised in such a way that the marked AR is disconnected. To establish connectivity such AR or any of its neighboring routers need to employ area border router (ABR) functionality. This increases the overall routing overhead and therefore is considered as an illegal operation. Since the objective is to reduce the overhead related to routing and mobility, the reconfiguration prohibits formations that require the creation of new ABRs.

A typical reconfiguration example is presented in Fig. 1(c) where the re-partition considers only ARs connected to ABRs between the two areas. A similar yet different case is illustrated in Fig. 1(d), where the reconfiguration encourage the shift of the ABR functionality. Such process is considered as legal only when it does not violate any other core network criteria like balance or capacity. Nevertheless, due to its complexity, this case is considered to be illegal in the remaining of this paper.

3. Related work

Network scalability through hierarchical clustering is a widely studied problem since the landmark paper of Kleinrock and Kamoun (1977) and various solution have been proposed as summarized in Amer and Lien (1988). Common implementations include the ATM PNNI (ATM Forum Technical Committee, 1996) and the OSPF (Moy, 1998), which is adopted as the hierarchical paradigm in our work. The best practices of OSPF hierarchy from the existing literature have been consistently organized as a framework in Kim (2003). Such framework provides optimization based exclusively on network characteristics without taking into account traffic patterns.

Fig. 1. An hierarchical reconfiguration example of a wireless access network with two areas and two ABRs having the original formation shown in (a). Solid lines present communication links between routers, while dotted lines between ARs show possible mobility movements. (b) Shows an illegal reconfiguration since the marked AR requires the creation of a new ABR increasing the routing overhead. A typical reconfiguration is presented in (c), while (d) presents a reconfiguration which is acceptable only under the condition that it cause no further changes in the core network.
Network topology division based on traffic patterns is considered in Deb and Woodward (2004), which introduces the concept of stochastic partitioning. In stochastic partitioning, the network is divided following a probabilistic way associated with the frequency of connection requests between network nodes. The proposed area reconfiguration adopts a grouping concept based on traffic and mobility patterns, which is applied on an ARs plane. In particular, it arranges together ARs involved into frequent handovers, while satisfying the area resource availability. Alternatively, it distributes ARs with high resource consumption into different areas to avoid congestion on per area basis while keeping the inter-area handover overhead minimum.

Since the grouping criteria are inherently dynamic, the area arrangement is desired to be equipped with a self-organized capability. Such self-organized concept is introduced in Krishnan et al. (1999), where a self-structure graph partition based algorithm is presented to reconfigure hierarchical networks which evolve in a way that priori determination of inter-connection is difficult to be foreseen. Graph partitioning heuristics and self-organized methods are also proposed for ad hoc networks (Galli et al., 2005; Manousakis et al., 2005), where the primal objective is to provide balance in terms of area size and mobility characteristics as well as minimize the inter-area traffic and suboptimality on path selection. The proposed area reconfiguration mechanism adopts a similar vision having different objectives, which mainly concentrate on maintaining a near-optimal hierarchical configuration in the presents of mobility and traffic fluctuations, which cannot be pre-determined.

The operation of the proposed area reconfiguration is not to apply certain policies but incorporate self-learning and develop self-managing capabilities as described in IBM (2005). In particular, the proposed mechanism employs a method to systematically collect and analyze the required information in order to identify whether changes are necessary for the network to maintain a near-optimal configuration. Upon an indicated need for adjusting the network, a re-partition algorithm revises the hierarchical formation and specifies the ARs, which need to change area orientation. The reconfiguration mechanism performs the designated changes and updates the network to incorporate the new hierarchical arrangement.

4. HQoSR in wireless access network

This section presents a framework for integrating hierarchical QoSR (HQoSR) and mobility considering HMIPv6 (Soliman et al., 2005) as the local mobility protocol. Such framework constitutes the basis of the proposed reconfiguration mechanism and provides the implementation details of the simulator components.

4.1. QoSR in wireless access networks

QoSR (Apostolopoulos et al., 1998) is a mechanism developed primarily for fixed networks with the purpose of identifying and selecting paths with certain QoS requirements based on the knowledge of network resource availability. QoSR improves the service received by end users, while providing a more efficient network resource usage especially in case of congestion. In wireless access networks such QoSR attributes are particularly useful to handle congestion on particular hotspots (Samdanis et al., 2004). Its incorporation with HMIPv6 is relatively simple, since HMIPv6 relies on IP routing to forward data packets inside the local domain. Nevertheless, QoSR paths are obliged to transverse certain network locations that accommodate mobility functionality. Such process limits the scope of QoSR and may impact its performance depending on the particular implementation.

The deployment of QoSR in wireless access networks has different requirements compared to fixed networks due to mobility. Specifically, the search for feasible paths requires flexibility due to handover of ongoing sessions, which originate the need for adjusting an existing path. To accommodate flexibility a distributed hop-by-hop routing strategy is employed, to avoid the source routing computational redundancy and the concentration of path requests on centralized locations, which impacts the handover performance. The handover performance is also affected by the time fashion of invoking the routing algorithm. Since the on-demand approach has a major impact on processing cost and handover delay, path pre-computations are preferred with controlled inaccuracy defined by the re-computation and the resource update policy introduced in (Apostolopoulos et al., 1998).

The maintenance of the initial QoS negotiation is a desirable but challenging issue due to resource fluctuation caused by mobility. Distributing incoming sessions among routes with maximum resource availability mitigates the bandwidth dropping rate. The routing algorithm adopted to accommodate such requirement is the Wang–Crowcroft (1996) distributed algorithm. Such algorithm employs a shortest-widest path (SWP) model, which selects paths with the maximum bandwidth breaking ties based on shortest propagation. SWP ensures loop free path selections especially when routing is performed in a distributed fashion as proved in Wang and Crowcroft (1996). In addition, when resources are imprecise, SWP algorithm is more likely to find a feasible path reducing the overall call blocking rate.

4.2. Inter-area routing, local mobility and ABRs

OSPF defines two levels of hierarchy with one area always configured as the backbone. Every router maintains link state information about its own area and summarized information about the remaining areas. ABRs are routers that belong to more than one area providing backbone connectivity to non-backbone areas and being responsible for inter-area routing. ABRs summarize internal routing information, which advertise to the remaining network and provide summarized information about other areas to routers inside their own area.

In a wireless access network, every OSPF area forms a cluster of ARs, which is geographic continuous and reflects estimated handover patterns. Each AR cluster needs to be equipped with one or more MAPs to provide local mobility. HMIPv6 specification does not restricts the location of MAPs and provides flexibility in the protocol deployment. Assuming the intra-area handover rates are much higher compared to the inter-area, it improves the handover performance to place MAPs either inside each area or at the border with the backbone. Usually, a multi-level arrangement of MAPs is desired for scalability purposes, but it has a minor effect for the proposed reconfiguration scheme. Therefore, for simplicity reasons it is assumed a single level MAP arrangement, with MAPs located at the ABRs with the backbone.

Non-backbone areas are usually not directly connected and routing between them is performed through the backbone. However, from the perspective of an autonomic hierarchical configuration, connectedness between neighboring areas is essential as pointed out in Krishnan et al. (1999). Otherwise, new configurations might create areas where ARs have to communicate via external means, which contradicts the purpose of forming an area on the first place. The direct connection between neighboring non-backbone areas can also minimize the
risk of session disruption in case of an inter-area or in other words inter-MAP handover. Specifically, it enables the use of fast handovers (Jung et al., 2004) and also ensures that packets forwarded from the previous MAP to the new AR located in a different area travel through the shortest distance. Therefore, introducing direct connectivity ensures smooth inter-area handovers.

Direct connectivity between OSPF non-backbone areas is provided through the use of ABRs. The OSPF specification defines ABRs as routers attached to multiple areas without necessarily requiring backbone connectivity. However, without backbone connectivity such ABRs get summarized routing information about distant areas, through the attached neighboring areas. This follows a distant vector paradigm with all the performance degradation consequences as explained in Zinin et al. (2003) and therefore, such ABRs deployment is prohibited. As an alternative implementation (Zinin et al., 2003) introduces ABRs, which only maintain a separate routing table for each area they belong to, without including summarized information. Such ABRs do not belong to the backbone area and act like internal routers to both areas. Specifically, they do not advertise neighboring area connectivity to any other routers, but forward all bypass traffic designated to their neighboring area directly. The operation of “alternative” ABRs exploit the direct area connectivity mainly for performing inter-area handovers while sessions between two neighboring areas usually are established through the backbone. Since the use of such ABRs serves the purpose of the direct area connectivity and at the same time reduces the routing overhead related to advertising summarized information, they are the choice of preference.

4.3. Aggregation and QoS

Hierarchical aggregation summarizes area internal routing information using specified metrics and advertise it externally according to the routing policy in use. The performance of hierarchical routing depends on the adopted aggregation method, the definition of aggregation metrics and the re-aggregation update policy. Accuracy of the aggregated state controls the blocking rate of incoming sessions and dropping rate of inter-area handovers at the cost of increased routing overhead.

The aggregation method of preference is the asymmetric simple proposed in Chang and Hwang (2001) due to its performance characteristics in terms of accuracy and overhead. In asymmetric simple, each ABR advertises a different aggregated state in relation to its position and associated ARs. The figure also demonstrates the interaction of HMIPv6 with QoSR, where the QoS path transverse the appropriate MAP and the path flexibility through use of hop-by-hop routing strategy upon a handover execution.

The QoS state is re-aggregated each time the residual bandwidth crosses a class boundary. To avoid the undesirable effect of generating frequent and meaningless updates when the value of the residual bandwidth oscillates between class boundaries, a hysteresis mechanism is also employed. The functionality of the hysteresis mechanism follows the paradigm introduced in Apostolopoulos et al. (1998). In particular, an update is triggered when the increasing available bandwidth crosses a class boundary or when the available bandwidth decreases more than the middle value of the new class. Fig. 2 illustrates an example summarizing the major elements of the proposed hierarchical configuration and demonstrating the interaction of QoSR with HMIPv6.

5. Autonomic hierarchical reconfiguration

The section presents the major elements of the proposed area reconfiguration, including the architecture, the mechanism maintaining re-partition related information, the re-partition algorithm, and the area reconfiguration process.

5.1. Autonomic architecture

The hierarchical configuration is assumed to be managed on a centralized location, referred as the re-partition server, which maintains resource and mobility information related to the scope of its use. Since mobility information is retrieved from MAPs, which are associated with each area separately, it is a more efficient and robust solution to introduce a network entity called the mobility server to monitor such information on per area basis. The mobility server improves the network efficiency since it processes the corresponding MAP information to inspect the quality of the current partition. In case of identifying a need for re-partition, it notifies the re-partition server by supplying
the appropriate data. Therefore, its presence regulates the re-partition overhead, while providing fault isolation through the distribution of the re-partition data. The server architecture is illustrated in Fig. 3. The re-partition server is positioned on the backbone area, which communicates with every mobility server in each area.

The functionality and relation of the re-partition and mobility server is based on the control plane paradigm introduced in Shaikh et al. (2002) having the following differences. The re-partition server on the backbone area establish adjacent relation with connected routers and snoop Summarized Link State Advertisement (SumLSA) packets. In this way it gets informed about the bandwidth consumption per area and per AR instead of establishing a TCP connection retrieving these information from servers located into each area. The re-partition server employs a reliable TCP connection to retrieve exclusively the mobility related information from the mobility servers. Mobility servers maintain primarily user and handover information by collecting MAP data concerning their serving ARs on a periodic basis. In addition, they maintain bandwidth consumption information of the associated area by snooping the Link State Advertisement (LSA) packets. Such data can serve as a recovery source in case the re-partition server fails and another backup server takes over its functionality.

The proposed HQoSR protocol follow the autonomic architecture defined in IBM (2005). The re-partition server act as an autonomic manager, which sensors the network configuration through mobility servers and ABRs, while activates the selected network changes through OSPF and MAPs. In particular, mobility servers and ABRs incarnate the touchpoint manager functionality with OSPF and MAPs as touchpoints providing interfaces for accessing and controlling the hierarchical configuration. The re-partition server can also be equipped with a manual manager to provide human interface for network administrator purposes. The autonomic server architecture provides self-optimization of the hierarchical arrangement. It is self-configured since the reconfiguration of the network hierarchy as well as the computation of the re-partition thresholds are carried out by the system itself. The information required for the re-partition process is duplicated and distributed among the mobility servers, providing a degree of robustness and self-healing. However, the provision of a complete autonomic system requires the incorporation of a more sophisticated self-healing mechanism as well as the introduction of self-protection properties. Such details are out of the scope of the current paper, which concentrates on the self-optimization and self-configuration aspects.

5.2. Monitoring and analyzing the network configuration

The re-partition server maintains a knowledge source, which includes various information about the hierarchical configuration of the network. A part of such information is provided to the server on a preparation step, while the remaining is either obtained through monitoring or is created by itself. Specifically, on the preparation step the re-partition server is equipped with the AR adjacency graph, the original network configuration, as well as the utilization per area and the maximum user limit per MAP. The initialization as well as monitoring information is collected through the LSA distribution of the QoS extended OSPF, which offers details about the original network configuration, its utilization as well as the bandwidth consumption per area and per AR. Additionally, MAPs supply information through mobility servers about the AR adjacency graph, the amount of attached users per MAP and per AR as well as the handover rates between ARs.

The re-partition server creates the threshold values used to invoke the re-partition algorithm, which evolve with the hierarchical configuration. The aim of this is to control the triggering of the re-partition algorithm. The inter-area handover thresholds are initially defined as \(1 \pm 0.5 R_{\text{Handover}}\), where \(R_{\text{Handover}}\) is the inter-area handover rate assuming uniform mobility. Handover rates that surpass the \((1 \pm 0.5)R_{\text{Handover}}\) limits are considered as an indication of a significant handover rate change, which triggers the re-partition process. The value of \(R_{\text{Handover}}\) changes every time the inter-area handover rates surpass either limit of the previous threshold. Its value is adjusted based on the new handover data by the re-partition server. The balance thresholds invoke the re-partition algorithm in case of congestion on per area basis either in terms of bandwidth or MH population per MAP. They are initially defined as the expected percentage of bandwidth consumption and MH per MAP for each area estimated in the process of network design. Although such threshold values remained constant in the simulations carried out in this paper for simplicity reasons, their value is recommended to adapt according to statistical traffic measures in order to capture topological changes.

The re-partition server monitors the hierarchical configuration from the supplied information in order to identify symptoms that suggest revision of the areas border. Specifically, it calculates and compares the inter-area handover rate as well as the resource consumption in terms of bandwidth and MH with the pre-defined corresponding thresholds. Since re-partition is applied to optimize the use of the hierarchical formation for the future state, the calculation process uses forecasting values of the monitoring information. The forecasting method follows an exponential smoothing approach to avoid rapid fluctuations. Upon the identification of a need for a network hierarchy revision, it invokes the re-partition algorithm.

5.3. Re-partition algorithm

Let \(G(V,E)\) be a weighted undirected graph. Each node \(u \in V\) represents an AR with size \(w_u \geq 0\) associated to the number of resident MHs, and capacity \(c_u \geq 0\) related to the requested bandwidth. Every edge \((u,v) \in E\) corresponds to AR adjacency and has a weight \(r_{(u,v)} \geq 0\), representing the handover rate. Assuming \(G\) is optimally partitioned into \(V_1, \ldots, V_k\), where \(V = \bigcup_{i=1}^{k} V_i\) and \(V_i \cap V_j = \emptyset, i \neq j\) but under certain mobility and traffic conditions is no longer optimal either in terms of the inter-area handover or balance considering both bandwidth and MH population. The re-partition algorithm revises the border with the aim of re-establishing an ideal partition. Its purpose is to increase

![Fig. 3. Re-partition server architecture.](image-url)
the overall network utilization and improve the inter-area handover introducing minimal AR reconfiguration.

Since the partitioning problem is known to be NP-Hard (Garey and Johnson, 1979), polynomial time heuristics are employed to solve it with reasonable effectiveness. Although fast heuristics exist to partition from scratch ensuring a low inter-area handover, they might lead to an excessive migration of ARs, which increases the transition overhead. A heuristic algorithm that assembles a re-partition by modifying a previous optimal partition can potentially introduce small modifications and be much faster.

The algorithm used for the re-partitioning process is based on Sanchis (1989) with the appropriate extensions to handle multi-constrains as introduced in Karypis and Kumar (1998). The initial purpose of the algorithm is to establish balance by keeping the MHs number below the MAP threshold limits while satisfying the imposed per area bandwidth threshold. Meeting the MAP capacity requirements first is important, otherwise the MHs that enter the network cannot use the potential bandwidth availability. The aim of the algorithm is not to achieve a strict equal division of the bandwidth and MH population among adjacent areas but to maximize the number of non-congested partitions in terms of both constraints.

Once balance is established the algorithm maintains it by restricting ARs moves that cause imbalance. At each iteration it identifies ARs with the highest contribution on the reduction of the inter-area handover rate, which satisfy the balance conditions. Tie breaking among ARs with the same performance is accomplished through the best balance contribution. Besides the initiation from the previous optimal state, the algorithm ensures further minimum alternation of the original partition by allowing only moves that have a major contribution on either balance or inter-area handover rate. A major contribution considering balance is regarded a re-partition that accommodates at least more than the minimum bandwidth per request, otherwise there is no practical benefit. In terms of the handover rate, a major contribution is considered a border revision were the reduction of the routing and mobility overhead is higher than the reconfiguration one. Such point beyond which there is a re-partition benefit is determined experimentally around 3% of the inter-area handover reduction.

5.4. Area reconfiguration process

The re-partition algorithm identifies ARs required to change area orientation and be reconfigured as members of the neighboring area. The re-partition server completes the process by remotely assigning them a new IP address depending on the area they move accordingly, using a signaling protocol. New area members discover their neighbors using the Hello protocol of OSPF (Moy, 1998), which triggers a routing update to synchronize the routing tables of the area. In addition, the Hello protocol makes routers notice the absence of their neighbors, which issue a new LSA to inform the remaining area router about their new local state.

ARs shifted into a new area establish routes through the new ABRs and configure themselves to new MAPs dynamically using the information related to the MAP options as described in Soliman et al. (2005). Before area reconfiguration takes place, transit ARs may handle ongoing sessions towards attached MHs. Re-arranging all these sessions simultaneously, according to the new formation can potentially cause excessive temporary overhead and possible disruption. To avoid this undesirable effect, the re-partition process prevents such simultaneous MH shift by maintaining constant the existing routes of ongoing sessions until the session terminates or the MH performs a handover. Since a part of the ongoing sessions terminate without a need for handover, the transition overhead is declined. The remaining MHs perform a handover at different times and therefore the concentration of signaling and the possibility of session disruption is reduced. Fig. 4 illustrates an example of session management related to an AR, which is shifted into a new area pointing out the route of ongoing and new incoming sessions, before and after reconfiguration.

6. Result analysis

This section presents the evaluation analysis of the area reconfiguration scheme through a comparison with the manual approach, which maintains a static hierarchical formation. The analysis concentrates on illustrating a framework for evaluating the area reconfiguration mechanism and its purpose is to demonstrate through simple simulation scenarios the contribution of our proposal.

6.1. Simulation setup

The aim of the evaluation is to capture the difference of the two approaches when traffic and mobility fluctuations are beyond expectation, which eliminates the drive of manual control. The evaluation criteria include the reduction of routing and mobility overhead as well as blocking and dropping rate in terms of user

Fig. 4. An example of session management related to a reconfigured AR. (a) Shows the route of an active session before reconfiguration. Since the destination AR is associated with area 1 the route traverse the corresponding area 1 MAP/ABR. After reconfiguration the ongoing session is kept constant using area 1 while new incoming sessions are routed according to the new arrangement through area 2, as illustrated in (b).
applications. The MH capacity limit per MAP is defined as the interval $[6.7 \text{ Kbps}, 1 \text{ Mbps}]$, representing from voice to video traffic.

The session holding time is assumed to be an exponential random variable with mean $\mu = 3 \text{ min}$. The cell residence time is also assumed to be exponentially distributed with a mean $\mu = (0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0) \text{ min}$. The requested bandwidth is uniformly distributed in the interval $[6.7 \text{ Kbps}, 1 \text{ Mbps}]$, representing from voice to video applications. The MH capacity limit per MAP is defined as $L_{\text{MAP}} = 25$ and the area bandwidth capacity is equal to $C_{\text{area}} = 20 \text{ Mb}$. To define the inter-area handover rate threshold, the inter-area handover rate is measured for the range of cell residence times under uniform mobility. The threshold low and high values are then $\text{thr}_{\text{low}} = (0.18, 0.56)$. The bandwidth and user thresholds are defined as $\text{thr}_{\text{bw}} = 0.8(C_{\text{area}}) = 16 \text{ Mb}$ and $\text{thr}_{\text{MAP}} = 0.8(L_{\text{MAP}}) = 20$ assuming that $80\%$ is the expected traffic percentage. The routing overhead is measured in terms of the resource consumption including bandwidth and processing cost. The bandwidth or communication cost is measured as $(\text{number of LSAs}) \times (\text{LSA size})$. The processing cost indicates the LSA processing consumption of each router and is defined as $(\text{number of LSAs})$ per interface. The packet delivery cost of the BU is defined as $(k(\text{bytes}) \times \text{number of hops})$, where the minimum length of the BU and BU-Ack in HMIPv6 specification is 112 and 96 bytes, respectively.

6.2. Re-partition and inter-area handover rate

This section presents the simulation results under the scenario, in which the inter-area handover rate increases and surpasses its threshold value. To create such scenario incoming users enter the network by the border between the two areas with probability $p = 0.7$ and execute and inter-area handover with probability $\rho_{\text{inter}} = 0.6$. The average maximum number of users in the network is around 40. The comparison of the manual and autonomic approach in terms of the inter- and intra-area handover rates is illustrated in Fig. 6. The proposed area reconfiguration re-arranges the network and reduces the inter-area handover rate at the cost of increasing the intra-area one. The degree of the inter-area handover reduction is higher with smaller cell residence times. The contribution of the handover rate adjustment to the network performance is presented in Figs. 7–9. It is worth noting that both routing and mobility overhead of the proposed approach contain also the correspondent overhead for performing the area reconfiguration process. This includes the routing update and mobility overhead introduced by shifting the selected ARs and their attached MHs towards the new areas.

Fig. 7 illustrates the BU communication overhead per minute in relation with the cell residence time. The autonomic area reconfiguration decreases the BU communication overhead with nearly constant rate around $9\%$ independent of the cell residence time. If the gateway of the backbone area is not equipped with MAP functionality, an inter-area handover also means a global handover using a protocol like MIPv6. In this case the proposed
area reconfiguration, also mitigates the global mobility signaling making HMLIPv6 even more efficient in terms of performance and scalability. Figs. 8 and 9 present the reduction of the routing overhead in terms of the communication and processing cost per minute. The autonomic approach reduces both communication and processing cost especially for smaller cell residence times. Specifically, the reduction of communication overhead is from 2% to 5.5% while the correspondent processing overhead reduction is from 2% to 7%, respectively. The communication overhead indicates the overall resource consumption per minute, while the processing cost denotes the mean number of LSA packets per interface, which is particularly significant for routers that contain more interfaces. The overall contribution of the proposed autonomic scheme in the reduction of total communication overhead considering both routing and mobility is from 4% to 6%.

6.3. Re-partition and area imbalance

This section presents the simulation results under the scenario in which the bandwidth or user resource consumption in a specified area surpasses the imposed thresholds while the equivalent resource consumption in neighboring areas are relatively low. To cause area congestion and at the same time imbalance between neighboring areas incoming users enter a specified area of the network with probability \( p = 0.8 \). The comparison of the two approaches is carried out based on two balance metrics. For simplicity simulations are performed with MHs having zero mobility and varying the arrival rate in the range \( \lambda = (5.0, 5.5, 6.0, 6.5, 7.0, 7.5, 8.0, 8.5, 9.0, 9.5, 10.0) \). The proposed autonomic approach re-arranges the network formation and enhance the network resource availability in terms of both parameters. This is clear in Figs. 10 and 11 where a comparison of the blocking bandwidth and blocking MH rates is presented for various arrival rates. Specifically, the autonomic approach reduces the bandwidth blocking rate from...
15% to 23% and the MH blocking rate from 22% to 31%, respectively.

The bandwidth blocking rate refers to the bandwidth rejection rate defined as the bandwidth sum of the discarded incoming requests divided by the total amount of bandwidth. Similarly, the MH blocking rate is defined considering the number of incoming users. Lower values of blocking rates indicate less congestion and therefore better performance. The performance difference of the two approaches increases with the arrival rates, but up to a certain limit. Such a limit is around $\lambda = 7.5$ beyond which both blocking rates increase while their difference remains nearly constant. This is an indication that the input traffic cause resource consumption close to the maximum limits.

Revising the network hierarchy in an attempt to increase the network utilization introduces overhead related to the routing and binding update of the reconfigured ARs and their associated MHs. The processing and communication overhead illustrated in Figs. 12 and 13 are produced under zero mobility and therefore is solely caused from the update of the reconfiguration state. The proposed approach introduces a negligible communication and processing routing overhead, while providing a significant contribution in the establishment of extra sessions. Specifically, the maximum increase of bandwidth consumption for the autonomic scheme related with the reconfiguration overhead is only 6.32 Kb per min, which is less than the minimum bandwidth request per session. Both processing and communication overhead increase with the arrival rate, while the difference between the two approaches also increases but up to $\lambda = 7.5$ beyond which remains constant. The increase in the overhead difference is associated with introducing more AR reconfiguration, which consequently cause a greater amount of MHs to be shifted from one area to another.

The proposed autonomic reconfiguration apart from increasing the overall network capacity it also provides balancing of the network resource consumption as a side effect. In particular, to demonstrate its balance capability three different balance parameters are computed including the MAP occupancy, the bandwidth consumption per area and the user distribution per area to
compare the two different approaches. The average maximum number of MH per MAP is illustrated in Fig. 14. It is clear that the proposed approach distributes more evenly the incoming MHs between both areas MAPs, especially as the MAP capacity approach its limits. This keeps the maximum MAP capacity below its limits for higher \( \lambda \) values. In particular, under the static approach the MAP reach its limit at \( \lambda = 6.5 \), while the autonomic approach requires \( \lambda = 7.5 \) to reach the same state.

The percentage of bandwidth consumption per area, for each session \( i \) with duration \( S_{di} \) and requested bandwidth \( B_{wi} \) is defined as

\[
\frac{\sum_{i \in S_{di}} S_{di} \times B_{wi}}{\sum_{i \in S_{di}} S_{di} \times B_{wi}}
\]

where \( S_{di} \) is the set of incoming sessions in areas \( k = \{1, 2\} \) and \( S_{i} \) is the set of the overall incoming sessions towards the network.

In a similar way it is also defined the percentage of MH per area. Figs. 15 and 16 illustrate the percentage of bandwidth consumption and the MH per MAP in each area, respectively. Clearly, the difference of resource occupation between the two areas under a static manual approach has nearly constant values for both bandwidth consumption and MH per MAP and is much greater compared to the equivalent autonomic reconfiguration one. It is worth noting that under the autonomic approach such difference of both resource consumption parameters declines as the arrival rate increases. This is caused from the re-partition balance criteria, which shift resources concentrated in a certain area gradually as the incoming rate increases.

6.4. Re-partition operation cost

The operation of the proposed autonomic approach introduces a cost in terms of the area transition and the re-partition information maintenance. The area transition cost involves the increase of routing and mobility overhead. The overall effect of the reconfiguration overhead is already computed in the previous section. However, the evaluation process did not take into account the effect of the smooth MH transition mechanism, which avoids the simultaneous handovers of MHs attached on the reconfigured ARs. Given the call holding time with mean \( \mu \) and the MH cell residence time with mean \( n \), the MH handover probability to a neighboring cell is defined as \( 1/(1 + \rho) \), where \( \rho = \mu/n \) according to Fang and Chlamtac (2002). Due to the memoryless property of the exponential distribution, this also indicates the handover probability of an ongoing session from the time it is kept constant on a reconfigured AR. Otherwise such session terminates inside the cell. Fig. 17 illustrates the handover probability for a variety of session holding times and cell residence times. Clearly, the smooth MH transition mechanism reduces the number of handovers that would be carried out otherwise if the MH were not kept constant. In particular, its effect is significant even when the handover probability is high, i.e. the session holding time is high and the cell residence time is low.

Maintaining the required re-partition information increases the network overhead. The re-partition server, attached to the backbone area establish adjacent relation with connected routers to snoop SumLSAs. Since such process introduces negligible overhead, the overhead related to maintaining re-partition information is mainly concentrated on collecting mobility information. The re-partition server retrieves mobility information from mobility servers using a reliable TCP connection. Assuming that the ARs adjacency graph is advertised in a similar
way like OSPF, where each Router/Link id \((R_{id}, L_{id})\) and related data \((R_{data}, L_{data})\) like resource consumption or inter-area handover rate, consume 4 bytes each, the overhead needed to advertise the mobility information per AR is

\[
AR_{\text{adj}} = (R_{id} + R_{data}) + (AR_{\text{adj}})(L_{id} + L_{data})
\]

where \(AR_{\text{adj}}\) is the average number of adjacent ARs. Assuming that the mobility information is carried using the default maximum TCP packet length of 576 bytes, the mobility overhead per server update in relation with the access network size and ARs adjacency is illustrated in Fig. 18. In particular, a network is considered containing up to 200 ARs with ARs having an average number of 3–6 adjacent ARs and an average hop distance between the server and MAFs from 1 to 5 hops. The increase of overhead is linear in both number of ARs and average number of hop planes.

Besides the overhead related with maintaining the re-partition information, the algorithm itself also introduces complexity in terms of time and space. The time complexity is defined as \(O(\text{p}(\log k + g_{\text{max}})\text{ per pass})\) (Sanchis, 1989). \(k\) is the number of partitions, \(p = \sum_{i=1}^{n} d_i\) where \(d_i\) is the degree of a node or AR \(u_i\), which represents possible mobility movements and \(g_{\text{max}}\) is the maximum gain related with the cost reduction of the partition cut by moving an AR from one partition to another. The space complexity indicates the amount of storage required to maintain the information related to the re-partition algorithm. Assuming that each AR has potential to be moved to the remaining \(k - 1\) partitions, the re-partition algorithm requires the maintenance of the equivalent amount of gain values associated with each move. Since each block has to maintain \(k - 1\) buckets according to Sanchis (1989) the space complexity of the re-partition algorithm is \(O(Nk^2 - 1)\), where \(N\) is the total amount of ARs.

The re-partition process is invoked upon a significant change of mobility and traffic patterns. The notion of significance is relative to the network topology and traffic. However, frequent executions of the re-partition algorithm introduce overhead associated with supplying the required information and may also cause bottleneck at the re-partition server. To avoid such situation especially when the required data for the re-partition process is unstable and traffic fluctuates rapidly, hold down timers are recommended to enforce a minimum fixed interval between consecutive re-partition triggers. Apart from providing a mean of control for the overhead and the efficient operation of the re-partition server, the use of hold down timers may also improve the performance of the re-partition algorithm since it ensures that the algorithm input data is more stable.

7. Conclusions

This paper addresses the problem of hierarchical area reconfiguration in wireless access networks and introduces an autonomic mechanism to revise the border between neighboring areas based on mobility and resource consumption. The proposed autonomic area reconfiguration approach is evaluated through simulation and compared with an equivalent manual one. The results demonstrated the superiority of the proposed approach for the following two main reasons. Firstly, it reduces the routing and mobility overhead, when an unexpected hotspot by the border of neighboring areas increases the inter-area handover rate beyond the estimated rates. Secondly, it revises the area borders to provide resource flexibility in terms of the number of users and bandwidth availability between adjacent areas. It is worth noting that the proposed autonomic scheme reconfigures the hierarchical areas only when there is a benefit. However, its effect depends on the topology as well as on the traffic and mobility parameters.

Further research is conducted on implementing a distributed version where the functionality of the re-partition server is divided among several network entities. The main benefit of such approach is the reduction of the computation and storage complexity, while its distributed nature isolates failures increasing fault tolerance. ABRs between neighboring areas comprise the ideal location for hosting the re-partition functionality since they can retrieve the required re-partition information easily from all areas of interest. The distributed approach reduces the overhead related with retrieving all necessary information due to the position of the re-partition entities. Nevertheless, the necessity for synchronization between the re-partition members requires information exchange. An evaluation study and comparison with the current centralized approach in terms of the communication overhead and the quality of re-partition algorithm is essential considering various topologies as well as traffic and mobility scenarios. Since the proposed mechanism operates in a heterogeneous and dynamic environment, the incorporation of cognitive radio into the reconfiguration process constitutes an essential progress, which requires further investigation. Cognitive radio is expected to enrich the context scope of the re-partition algorithm.

References


Kurban MS; Tugcu T. A survey on emerging broadband wireless access technologies. Comput Networks 2007;51(1).


