GSA: An Architecture for Optimising Gateway Selection in Dynamic Routing Groups

Michael Eyrich[1], Mikko Majanen[4], Eranga Perera[3], Ralf Toenjes[2], Roksana Boreli[3], Tim Leinmueller[5]

[1] TKN, TUB, eyrich@tkn.tu-berlin.de,
[2] Ericsson, ralf.toenjes@ericsson.com,
[3] National ICT Australia, {eranga.perera, roksana.boreli@nicta.com.au},
[4] VTT Technical Research Centre of Finland, Mikko.Majanen@vtt.fi,
[5] Daimler Chrysler, tim.leinmueller@daimlerchrysler.com

Abstract — **The proliferation of the increasing number of devices and technologies has brought about the extension of the concept of roaming users to roaming networks. This has inspired many researchers to investigate optimal mechanisms to enable communications for such roaming networks. Within the European 6th Framework Ambient Networks Project, we identify such roaming networks as dynamic Routing Groups, which consist of a number of different types of nodes with different capabilities. External communications from nodes within the routing group can be done via selected Gateway nodes. In this paper we present the Gateway Selection Architecture (GSA) which provides support for Gateway identification, management and selection for nodes within a RG. We describe the benefits of this architecture and compare it to other known approaches.**

***Index Terms*— Ambient Networks (AN), Routing Groups (RG), Gateway (GW), Gateway Selectors (GWS).**

I. Introduction

Increased mobility of users, devices and applications, rather than being a long term goal, is becoming a reality and a large number of different approaches are currently being pursued to enable full mobility and seamless connectivity in a heterogeneous network environment e.g. [1],[2]. Additionally, it is of uttermost importance to provide efficiency in the use of network resources and therefore reduce cost and also to provide the flexibility to include end-user preferences in regards to service capabilities, providers, security domains, etc.

The Ambient Networks (AN) project [3] considers a future networking architecture, with the aim of enabling the cooperation of heterogeneous networks belonging to different technologies or operator domains.

A common AN scenario includes a number of nodes moving together, e.g. within a personal or vehicular area network (PAN, VAN), but also as a group of unrelated devices which may be associated by their proximity and direction of movement, e.g. pedestrians in a city street. Within the AN, such network nodes can be linked into a cluster referred to as the Routing Group (RG) [4]. The increased demands for mobility and efficiency within the AN framework, together with the concept of the linked RG nodes moving together, have led us to devise a mechanism that provides increased flexibility for RG nodes, and enables each node within the RG to perform external communications outside of the RG not only directly but also via the most appropriate Gateway node. This is achieved by introducing the architecture which relies on Gateway Selectors, i.e. nodes within the RG which have specific capabilities which enable performing of the selection process, and knowledge about the topology of the RG including the capabilities of all individual nodes.

The remainder of this paper is structured as follows. In the next section we provide an overview of the related work. Following this, we give a brief insight into the Ambient Networking architecture and show how the RGs fit into the AN framework. In section IV we present the proposed Gateway Selection Architecture and section V presents the Gateway selection process. Security considerations are included in section VI. An evaluation of GSA architecture is given in section VII, and section VIII concludes the paper.

II. Related Work

The most popular way to provide Internet access to nodes within ad-hoc networks and in mobile networking scenarios seems to be extending the Mobile IP protocol. In this section, two existing mobility optimisation architectures are shortly introduced, namely IETF NEMO (Network Mobility) [5] and the MObile CommuniCations Architecture (MOCCA) by the FleetNet project [6].

NEMO [5] extends Mobile IPv6 to manage network mobility by allowing bindings between a network prefix and a Care of Address (CoA) indicating the current location of a Mobile Router (MR). These bindings are managed by a Home Agent (HA) (implementing a location management service) associated with the MR. Nodes within the moving network are able to connect to the Internet without having to participate in mobility management, because the MR updates the HA for the entire network, not only for itself.

Nodes within the moving network are allocated an address from the prefix, and traffic destined towards that IP address is

intercepted at the HA and tunnelled to the MR, which then forwards the data to the correct node within the moving network. In the reverse direction, traffic from nodes within the moving network is tunnelled by the MR to the HA, where it is then forwarded onto the appropriate destination.

Whilst this proposal solves the basic problem of network mobility, it has some of the inherent disadvantages of Mobile IPv6 (introduction of a single point of failure on the routing path, tunnel overhead and dog leg routes), which become more extreme in case when nodes within the moving network use also Mobile IPv6 to manage their own mobility or if the moving networks are nested. Solutions investigating route optimisation are under way to mitigate these problems but they will add additional complexity to the solution.

The MOCCA architecture [6] from FleetNet covers both network and transport layer protocols. It addresses the interoperability and efficient communication between Internet and FleetNet, supports the vehicles' mobility, and provides Internet Gateway (IGW) discovery and selection.

MOCCA uses a modified version of Mobile IP (called Mobile IPv6*) to support the mobility of vehicles. A Proxy (the exchange point between Fleetnet and the Internet) maintains the vehicles' home agents (HAs), IGWs function as foreign agents (FAs) and the vehicles represent the Mobile Nodes (MNs). The Correspondent Node (CN) in the Internet sends its data packets to the MN's home address (i.e. the HA in the Proxy). The Proxy tunnels them to the FA, which decapsulates and forwards them to the MN. Vehicles are also able to access IPv4-based Internet services transparently by using the Proxy's NAT-PT protocol.

In MOCCA's service discovery protocol, the service agent is divided into two distributed functional units. The first unit is situated on the IGWs and announces its service provision of Internet access periodically. The service announcements are broadcasted only in a geographically restricted area using FleetNet's geocasting capabilities. The second functional unit resides locally in the vehicles. It extracts the information from the service announcements and caches it to local database. The user agent (UA) within the vehicle queries the database and configures Mobile IPv6* to use one of the IGWs in the database as its FA. In case multiple IGWs are available, the UA selects the IGW that fits best to the requirements of the applications. The selection is made by a fuzzy-based algorithm according to the information provided by service announcements.

[7] and [8] present architectures where Mobile IP is combined with the Ad hoc On-Demand Distance Vector (AODV) protocol to make Internet connections available for the ad hoc network nodes. [9] proposes a more sophisticated hybrid GW advertisement scheme for connecting ad hoc networks to the Internet. [10] presents a performance analysis of proactive, reactive, and hybrid IGW discovery protocols.

III. ROUTING GROUPS WITHIN THE AMBIENT NETWORK FRAMEWORK

The AN architecture introduces the Ambient Control Space (ACS) which includes common control functions for all relevant networks which may be part of the AN. The ACS is divided into different functional entities with the mobility related control functions aggregated into the Mobility Control Space (MCS).

The MCS incorporates four different functional areas, namely:

- Handover and Locator Management, which enables the movement of nodes between different locations in the AN and ensures the correct location for packet deliveries;
- Reachability Management, which enables two nodes to establish communication with each other;
- Moving Network Support, which manages groups of nodes, i.e. Routing Groups;
- Triggering, which collects and interprets all kinds of events which are related to mobility management and the formation of RGs.

The most primitive building block of an AN is a physical cluster, which can be defined as a group of nodes that are physically close to each other, are likely to stay near each other and are able to communicate. If the nodes of such a physical cluster are aware of each other then it is considered as Routing Group within the AN architecture.

A Routing Group is formed by a number of network nodes (user devices) which are in physical proximity, with a purpose of optimising mobility management and routing functionality of the group, as opposed to the same functionality being implemented for individual nodes. Regarding mobility management, as multiple devices are all moving together, the handover (be it 3G/4G cell, radio access network, etc.) process for those devices can be aggregated, thus reducing the amount of required signalling overhead. Also, devices that do not have inherent mobility support may in this way make use of the mobility capabilities of other devices in the moving network. In regard to routing, given the physical proximity of devices, it is possible to provide more efficient routes for traffic, be it local routing between nodes or routing in the network.

RG issues in regard to formation, maintenance, etc. are covered in the [4].

IV. GATEWAY SELECTION ARCHITECTURE

The proposed architecture implements control functionality which handles routing and mobility for the RG in an efficient and flexible way.

A node that is able to provide external connectivity to one or more nodes within the RG is considered as a Gateway within the AN architecture. Thus we consider a Mobile Router to be a more capable Gateway that additionally provides mobility management for the entire RG.

Gateway Selectors (GWS) are specific nodes which include the functionality to provide Gateway identification, updates, and selection of the most appropriate Gateways for specific nodes within the RGs, which are transmitting or receiving traffic. The information oriented architecture envisioned within AN augmented the notion of Gateway Selectors for RGs. For

example the context provisioning functional area within the ACS provides mechanisms to collect, store and disseminate context information. Further one of the design principles of the AN architecture is that whatever information that is available within one AN can be used by any other AN consistently irrespective of the mobility environment. This allows context information to be disseminated within ANs readily paving the way for an architecture such as the GSA.

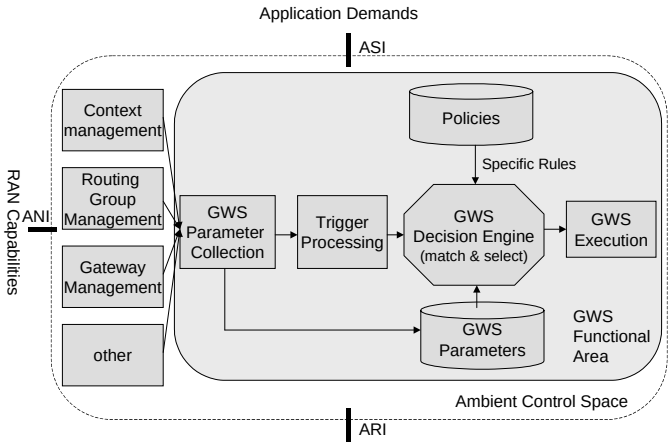

Figure 1. Gateway selection architecture

Figure 1 shows the proposed Gateway selection architecture. The GWS functional area is part of the Ambient Control Space (ACS). The ACS exchanges information with neighbouring Ambient Networks over the Ambient Network Interface (ANI), for example, to collect their radio access network (RAN) capabilities. Within the Ambient Network the Ambient Service Interface (ASI) provides the interface from the ACS to the services. Applications use this interface to inform the GWS about their demands. The ACS controls the user plane over the ARI and configures the Mobile Router according to the GWS decision.

The central entity of the GWS architecture is the GWS decision engine. To allow easy customisation and exchange of policies for GWS an explicit representation is recommended. The policies are presented by a set of rules. An inference engine, the GWS decision engine, compares, i.e. matches, the policy rules with the current situation represented by the GWS parameters, selects the best matching rule and executes the rule to select the GW. Please note that the inference engine may apply several match-select-execute cycles, i.e. fire several rules, till the decision is found.

The GWS depends on several parameters, such as the context, the nodes in the routing group, the GW capabilities, the RAN capabilities or service demands. The Parameter Collection entity compiles the parameters influencing the decision and stores them in a database.

The GWS is either triggered by an initial request from a node asking for a Gateway or a change indication due to altered parameters, for example, loss of RAN connection. The initial trigger sets up the connection, the change trigger ensures actions to maintain the connection. For example, a service could store the required QoS in the trigger database. The GWS Parameter Collection monitors the parameters (by polling or interrupts from the context, routing group, GW or other management systems). If the parameters do not fulfil the minimum QoS threshold required by the service, the trigger processing starts a new GWS decision.

V. The GW selection process

The Gateway selection process introduces mechanisms for discovering GWs within the Ambient Network framework. This architecture was developed after analysis of current GW selection mechanisms such as reactive or on-demand discovery processes, proactive processes and hybrid approaches, which have been adapted to suit the specific requirements introduced by AN.

Three components need to be in place for a GW selection process: parameters for driving a selection, the decision making point and or at least the result of a computed proposal and the points/nodes utilizing the decision results. Also, the protocol to support transfer of information between devices has to be defined.

GWSs and GWs provide a service to the RG nodes, and in the process of selecting the Gateways or routing traffic use their own computation or battery power, and potentially incur traffic cost by forwarding traffic from other nodes. Similar arguments to the ones which are valid in peer-to-peer networks (P2P) may apply in the AN space, and various incentive mechanisms in P2P are described in [1].

A. Parameters driving a selection

There are a number of criteria that can be regarded as possible input parameters that would assist and influence the GW selection process. These can be classified as TOPology related inputs (TOP), PHYsical transmission related inputs (PHY) and SERvice demands (SER). These can be divided into mandatory (necessary for the selection) and optional, i.e. inputs which assist with various optimisations within the selection process.

Mandatory Inputs:

- I_TOP_01: Organization of the RG (managed (contains MR) or unmanaged (no MR))
- I_PHY_01: Availability of radio access networks (RANs)
- I_SER_01: Policy configuration of the node (i.e. whether node is allowed to provide Gateway service)
- I_SER_02: Request for Gateway service from node or application

Optional Inputs:

- I_TOP_02: Managing node (cluster-head)
- I_TOP_03: RG change event (node join or leave)
- I_TOP_04: GW change event (eg. GW unreachable)
- I_SER_03: QoS requests from RG members
- I_SER_04: Pricing requests from RG members
- I_SER_05: QoS offers by RANs
- I_SER_06: Pricing offers by RANs
- I_SER_07: Credentials for RAN connection

B. Functional view

The GW Selector may have additional functionality within the RG. For RGs that have elected cluster-heads, the Gateway Selector is usually collocated on the same node as the cluster-head role. However, this does not have to be the case; it may be that the Gateway Selector functionality has to carry out an election procedure of its own to identify a node to control Gateway selection for RGs where cluster-heads are not elected during the formation process and are not needed to manage the RG.

The RG formation framework provides nodes of a RG with the necessary information to contact Gateway Selectors.

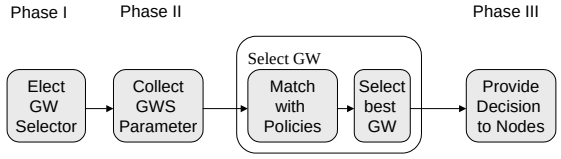

Figure 2. Phases of Gateway selection process

The Gateway discovery process consists of three phases (see **F**igure 2):

Phase one is to trigger the election of a Gateway Selector. This phase is usually handled by the RG formation and is executed only on cluster reformation. During the formation process the nodes also discover whether they have abilities to act as a Gateway or Mobile Router.

Phase two describes the communication necessary to dynamically disseminate information about available Gateways to the Gateway Selector. Information is transported by the same communication mechanism and messages used for RG formation. Thus, the communication between GWS and GWs/MRs is based on advertisements, i.e. a proactive approach. The parameters needed for GWS are collected and matched with the policies to select the best GW.

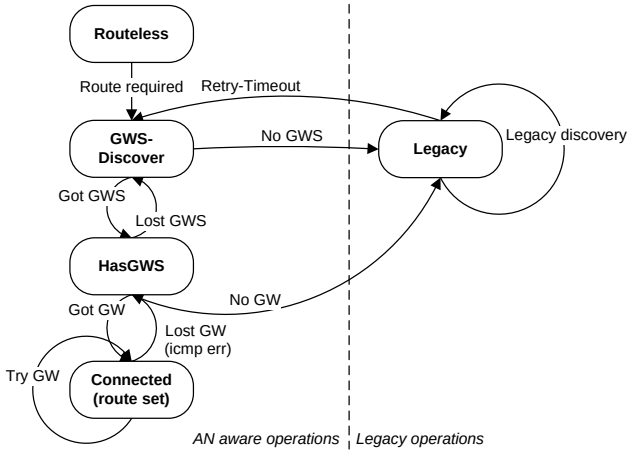

Figure 3: State machine for a RG node

Phase three provides Gateway information to RG nodes, where nodes request available Gateway information from the Gateway Selector. Here the communication follows the reactive or on-demand approach and is primarily based on the assumption that the client is not capable or willing to calculate its best Gateway by itself but depends on the infrastructure to make an informed decision by providing its personal profile to the Gateway Selector. Since the GWS has potentially more information it can select a Gateway that is more appropriate for the given node.

A RG node passes various states (Figure 3) starting with the Routeless state before reaching a connected state. Following the assumption that not every visited network will be AN-aware, the selection process is able to fall back to a legacy discovery process if any of the discovery subprocesses fails. Under regular operation, it first tries to get an appropriate GWS. Usually, a list of active GWSs can be obtained from a cluster- head managing the RG, otherwise it needs to be actively discovered. On success state HasGWS is entered. When a special route is required (e.g., depending on certain service needs), a GW service request with appropriate parameters is sent to the GWS. A route reply results in the node entering connected state as a sub-state of HasGWS. Failures using a given route move the state back to HasGWS. If no other route is available the node enters standalone state but tries at regular intervals to restart the whole selection process.

C. Protocol considerations

The requirements for the protocols need to be tailored to a highly dynamic environment. Gateway nodes can as easily become unavailable, as can Gateway Selectors, thus, there is no stable environment to base on. Therefore, the protocol needs to cleanly handle unavailability of Gateways in an efficient way and also address the same issue for Gateway Selectors. The proposed protocol handles the changes in the availability of Gateway Selectors by supporting primary and secondary instances that synchronize at regular intervals and on new events. The change in the availability of Gateways will first be recognized by nodes receiving a 'host/route unreachable' notification. This will trigger a new request for Gateway service combined with a 'GW change event'. As a result the Gateway Selector will re-evaluate the respective Gateway.

VI. Security considerations

Since security of the GW selection is not the primary scope of this paper, this section will only briefly identify and summarize potential security risks. Basically, GWS introduces two trust related security issues, one between GWs and GWSs and the other between GWSs and other network nodes.

Nodes have to be able to discover a trustworthy GWS, which will provide them with responses to their GW requests.

In order to provide the desired selection process for the network, GWSs must be able to identify trustworthy Gateways that provide correct information on the Gateway's capabilities and current status.

VII. Evaluation of GSA

GWs are used to provide additional connectivity for nodes which may only be able to connect to a limited number of networks by relaying the traffic from specific nodes. MRs are

used to increase the overall efficiency combining traffic, to provide additional capacity, coverage or enhanced capabilities (including QoS), also potentially reduced transmission costs. MR can also combine handover traffic from a number of nodes, provide mobility management to nodes which inherently do not have this capability, etc.

The motivation for including the GWSs includes simplifying the dissemination and information updates regarding GW and MR nodes and their capabilities, minimising the amount of signalling overhead (i.e. star configuration, rather than each node finding out about GWs independently). It also allows majority of the nodes to have limited computational capabilities and battery power by keeping the intelligence in the GWSs, hence no need to spend resources (battery life, computation power) in the other nodes.

A moving network such as a NEMO network can be deemed as a RG with only GWs. Consider a mobility scenario where a group of nodes formed into a RG boards a train and integrates with the on-board WLAN network. These nodes would rely on the MR within the integrated network for external communications. Using the MR might not be the most lucrative option for some of the nodes within the RG. In such a situation our proposed GSA would be able to direct the node to the cost effective GW as opposed to using the MR.

Further the failure of the MR would add a single point of failure to the RG. With the GSA architecture the GWS nodes would be able to direct the nodes to another GW if the MR fails, mitigating this problem.

As identified in section 2 dog leg routing is an issue with the NEMO architecture and this is more of an issue in the case of nodes belonging to another Home Network as opposed to that of the MR's home network. This is due to all packets needing to go through two Home Agents from and to such a node within the RG. Suppose there are a few nodes from the same home network as opposed to the MR's home network and one of the few nodes is able to act as a GW. In such a situation GWSs, by having context information, would be able to direct the nodes from the same home network to the GW from the same network.

The main drawback of GSA is that GWSs become potential failure points, and even with allowing for primary and secondary GWSs there is an increased probability of failure in this configuration, when compared to an architecture based on P2P connectivity. In order to avoid this, our protocols will be designed with a fallback mechanism such as flood based neighbour discovery.

As identified in Section 2, the FleetNet architecture has roadside proxies (IGWs) that connect a FleetNet cloud to the Internet. Similar to the GSA's GWS nodes the User Agent of the FleetNet architecture has the capability to choose the best IGW when more than one IGW is available. Since the IGWs are fixed GWs as opposed to mobile GWs the FleetNet architecture handles a more fixed infrastructure. But in the case of GSA there is no priori knowledge of GWs until these nodes join the RG. Therefore the GSA is able to handle more unstable GWs within a RG.

VIII. CONCLUSION

This paper addresses the optimum way to provide connectivity from a group of associated network nodes which form a Routing Group within the Ambient Network. This is achieved by introducing three types of nodes: Gateways, which can forward traffic on behalf of other network nodes, Mobile Routers, which have the ability to provide additional optimisation of traffic and signalling for a number of nodes, and Gateway Selectors, control nodes which provide identification and selection of Gateways and Mobile Routers. The advantages of this architecture include the ability to provide either additional connectivity, via simple Gateway nodes, or jointly optimised traffic and signalling, via Mobile Routers, and a simple control of the selection of appropriate Gateways or Mobile Routers via the Gateway Selector nodes.

ACKNOWLEDGEMENT

This document is a by-product of the Ambient Networks Project, partially funded by the European Commission under its Sixth Framework Programme. It is provided "as is" and without any express or implied warranties, including, without limitation, the implied warranties of fitness for a particular purpose. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Ambient Networks Project or the European Commission.

REFERENCES

[1] Cooperative Networks for the Future Wireless World, C. Politis et al, IEEE Communications Magazine, Sept 2004
[2] Hybrid Multilayer Mobility Management with AAA Context Transfer Capabilities for All-IP Networks, C. Politis et al, IEEE Communications Magazine, Aug 2004
[3] http://www.ambient-networks.org/
[4] Routing Group Formation in Ambient Networks, A. Surtees et al, 14th IST Mobile and Communications Summit, 19-23 June 2005, Accepted.
[5] Network Mobility (NEMO) Basic Support Protocol, V. Devarapalli et al, RFC 3963, Jan 2005
[6] Mobile Internet Access in FleetNet, M. Bechler et al, 13. Fachtagung Kommunikation in verteilten Systemen, Leipzig, Germany, April 2003.
[7] Internet Connectivity for Ad hoc Mobile Networks, Y. Sun et al, International Journal of Wireless Information Networks, special issue on Mobile Ad Hoc Networks (MANETs): Standard, Research, Applications, 9(2), April 2002
[8] A Hybrid Apprach to Internet Connectivity for Mobile Ad Hoc Networks, P. Ratanchandani et al, Proceedings of WCNC 2003, Volume 3, March 2003
[9] Hybrid gateway advertisement scheme for connecting mobile ad hoc networks to the Internet, J. Lee et al, Proceedings of VTC 2003, Volume 1, April 2003
[10] Performance Analysis of Internet Gateway Discovery Protocols in Ad Hoc Networks, M. Ghassemian et al, WCNC 2004 - IEEE Wireless Communications and Networking Conference, Vol. 5, no. 1, March 2004
[11] 'To Share or not to Share' An Analysis of Incentives to Contribute in Collaborative File Sharing Environments, K. Ranganathan et al, Workshop on Economics of Peer-to-Peer systems, Berkeley, CA, June 2003