Moving Core Services to the Edge in NGNs for Reducing Managed Infrastructure Size

Roberto Baldoni, Roberto Beraldi, Giorgia Lodi, Marco Platania, Leonardo Querzoni
Dipartimento di Informatica e Sistemistica “A. Ruberti”
Sapienza - Università di Roma
via Ariosto 25, 00185, Rome, Italy

***Abstract*—Telco providers are in the phase of migrating their services from PSTN to so called Next Generation Networks (NGNs) based on standard IP connectivity. This switch is expected to produce a cost degression of 50% for CAPEX, while OPEX remains fairly stable due to network management and energy costs. At the same time we are expecting a big increase of the load of a telco provider at the core level due to the istantiation of new telco services (VoIP, video conferencing etc) and to the support of third parties services (such as support to smartphone applications, etc.). The goal of this work is to show how management and energy costs can be effectively reduced by leveraging autonomic approaches to move some NGN services toward the telco network edge while still providing QoS levels comparable with those provided by a traditional fully-managed infrastructure.**

I. Introduction

Next Generation Networks (NGNs) are packet-based networks able to provide Telecommunication Services to users and make use of multiple broadband, QoS-enabled transport technologies in which service-related functions are independent of the underlying transport-related technologies [2].

Worldwide telco operators are striving to develop new solutions over NGNs, which exhibit a telephone traffic cost reduction and provide a higher number of services such as web call center, e-commerce, home banking, video conference, as well as supporting third party applications such as smartphone applications, social network applications etc. These applications are expected to increase the resources requested to the telco core network of one order of magnitude in the close future. Despite this raising load, almost all NGN communication protocols such as H.323 and SIP (Session Initiation Protocol) [1] rely on a collection of central servers that manage all clients requests in case, for example, a user contact information has to be retrieved and a connection with that contact has to be established. These centralized architectures grow up telco operational costs by adding servers in order to cope with the increased load as well as introducing scalability issues in the long run.

Usually, the network infrastructure of a large telco provider (e.g. Telecom Italia, AT&T, France Telecom) is organized as a static tree with servers installed in different geographic domains (leaves of the tree). These servers are used to respond to regional user contacts requests and forward those requests to a cluster of other servers (root of the tree) made available for user contact retrieval purposes. In these networks, the majority of the system costs are represented by maintenance and configuration, due to the employment of dedicated system administrators in each domain [3]. If the load increases, servers are added to the system to prevent the creation of bottlenecks.

In this paper we describe the outcome of a project, funded by *Telecom Italia* [5], which focuses on the design and evaluation of a load balancing architecture whose goal is to efficiently support core telco services for millions of users. This architecture, on one hand, aims to reduce operational costs and, on the other hand, it must cope with a load of the core services of one order of magnitude higher than the one handled by the current telco systems. More specifically, in this paper we evaluate the behavior of the architecture applied to a contact management service, that is, considering a unique client ID, the service returns the *current* IP address of the users. This service is a basic building block of many applications deployed on the top of the telco provider network such as VoIP, video conference, Instant Messaging. Moreover, the contact service also supports a variety of applications that are developed by third parties, from smartphone applications to social network ones. We consider a telco infrastructure constituted by a core network made up of several servers and an edge network. The latter is a DHT[4] formed by devices installed at users' premises, namely *Home Gateways*. The load of contact requests is then balanced between the managed network and the edge in order to both respect the QoS of the specific service and reduce the management costs.

The performance analysis shows that our hybrid architecture can reduce the number of servers of one order of magnitude passing from a core network of thirty servers (actual size of the inter-regional core infrastructure for a mid-large size telco operator) to a few units when this small core network is combined with an edge network of four millions home gateways. From the point of view of the telco provider, this entails a substantial reduction of (i) the OPerating EXpences (OPEX) for running the system, i.e., maintenance, management, and energy costs, and (ii) the CApital EXpenditure (CAPEX) due to a smaller number of servers forming the core network.

II. System model

The internal architecture of a NGN can be described in terms of two Network Access Points (NAPs): Access Routers (ARs) and Home Gateways (HGs). An AR is a powerful server

This work has been supported by Telecom Italia.

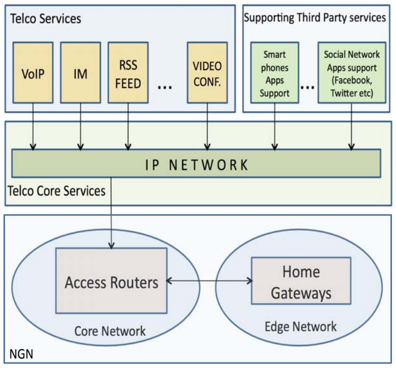

Fig. 1. Expected Services offered by a telco provider

playing the role of NAP for a wide geographical area. A HG is a small device installed at users' homes by telco providers. It is characterized by limited computational resources and scarce available memory. We consider the IP architecture of a NGN depicted in figure 1, where core services are identified and applications supporting third party services are hosted on a telco network.

Core service model: We model a *location service* that is accessed through a *lookup* and an *update* primitive. Specifically, the $lookup(UserID)$ primitive retrieves the user profile associated with a given UserID. The user profile contains at least the current IP address of the user. For example, in a VoIP application UserID corresponds to the phone number and the lookup procedure provides the *phone-number IP address* mapping which is exploited on a per-call basis. The lookup service should guarantee short lookup time for QoS requirements enforcement (i.e., low latency), limited control message overhead, low memory and processors consumption on the NAPs, dependability (i.e., no false responses on lookup requests).

The $update(UserID, UserProfile)$ procedure is used to periodically refresh the user profile associated with the UserID. It is issued by a user every time his/her IP address changes. However, a user renews his/her data once per hour even if the associated IP address does not change; this is a mechanism similar to the one presented in [1].

User model: We consider a system composed by 10 millions of users, 40% of which fixed and 60% mobiles. This model has been specifically required by *Telecom Italia* as it reflects the real user scenario in our country. Fixed users access the network via their own HG or through an AR (for example because they are using dial-up connections). Mobile users always get the access through an AR. Once a user has joined the IP network, he/she sends his/her new profile to the location service and refreshes it periodically.

III. Hybrid architecture for balancing users requests

In the proposed solution we move part of the location service load on a service that is implemented exploiting the HGs subsystem. In the NGN the location service is likely to increase in volume and this can cause performance penalties that can in turn entail a service scalability problem. By moving part of this load on the HGs we aim at mitigating this issue. Specifically, the system architecture we propose is organized in two levels: a *managed core* level, populated by ARs connected through a clique, and an *unmanaged edge* level, populated by HGs, connected through a DHT. Due to its simple structure and popularity, we have chosen Chord [4] as DHT. Both the clique and DHT are implemented over an IP network. In addition, each HG maintains an IP connection with a single AR, assigned at random by the telco provider.

Both ARs and HGs store user profiles; the profiles are replicated in k copies, with the constraint that at least one of them is stored in both the managed and unmanaged networks. We assume that each AR stores a fraction of the user profiles and uses an in-memory hash table to rapidly redirect a lookup request towards the AR that contains the full user profile. In addition, ARs and HGs are equipped with an *Incoming Messages* queue for messages incoming from other NAPs, and an *Access to memory* queue for accessing the user profile repository.

The service works as follows: all lookup requests generated by users connected through ARs and HGs are sent to the core network. When an AR extracts a lookup message from the *Incoming Messages* queue, it accesses the hash table to find the destination. If the destination is itself, then it enqueues the message into the *Access to memory* queue; otherwise, with probability p, that message is redirected to the DHT subsystem, and with probability $1 - p$ to the destination AR.

Update messages are sent to the core network too. When an AR extracts an update message from the *Incoming Messages* queue, it accesses the hash table to find the destination AR. Hence, the *Access to memory* queue contains all the update requests for profiles stored in the local disk and only a fraction of lookup requests that it can directly serve. Updates are not redirected over the DHT. However, in order to guarantee profiles consistency, when an AR updates a user profile, it executes an update procedure over the DHT. The motivations behind the choice of this architecture are twofold: (i) we exploit ARs' memory and bandwidth availability to connect ARs through a clique. This ensures a fast lookup latency; (ii) we use a distributed system among HGs in order to download part of the traffic from the clique, thus reducing the time spent by a message in the AR's *Access to Memory* queue; this helps us in meeting QoS requirements such as latency and message losses. The use of a DHT as distributed system contributes to minimize the time necessary to localize a resource.

IV. Theoretical model

In this section we describe a theoretical model that can be used in order to predict the mean messages delay as function of the load imposed on ARs and HGs. We express the load on a device as the ratio between incoming and processed messages in a time unit. Without loss of generality, we consider NAPs equipped only with the *Access to Memory* queue; the motivation is that the disk access time is two order of magnitude greater than the extracting message time. Thus,

the time spent by a message in the *Incoming Messages* queue can be considered negligible. The *Access to Memory* queue is modelled as a M/M/1 message queue [7], that is a single-server FIFO queue model in which arrivals are a Poisson process, the service time is exponentially distributed, and the buffer length is infinite. The model is characterized by two main parameters: μ, i.e., the number of processed messages in a time unit; and λ,i.e., the number of incoming messages in a time unit. The ratio $\rho = \frac{\lambda}{\mu}$ is called *traffic intensity*. In order to guarantee queue stability, it must be $\rho < 1$.
As our aim is to define a theoretical model that computes the average message latency for both ARs and HGs, we use the well-known Little's theorem in the queue model:

$$W = 1/(\mu(1-\rho)) \tag{1}$$

where W is the average time spent by a message in a queue.

For the sake of simplicity, in our analysis we consider the managed network consisting of a single AR: a lookup message is queued just in the AR containing the requested profile, which is reached at most in two hops (we neglect transmission delay in the core network). Thus, the time spent by a message in the queue is determined by:

$$W_{AR} = 1/(\mu_{AR}(1-\rho_{AR})) \tag{2}$$

In contrast, for the unmanaged network we consider intermediate steps before reaching the target HG: the average number of hops in Chord is $\frac{1}{2}\log_2 N$ [4], with N that represents the network size. Thus, formula (1) in case of HGs becomes:

$$W_{HG} = \frac{1}{2}\log_2 N_{HG}\Delta t_{hop} + \frac{1}{\mu_{HG}(1-\rho_{HG})} \tag{3}$$

where N_{HG} is the number of HGs in the network and Δt_{hop} the mean transmission delay per DHT hop.

ρ_{AR} and ρ_{HG} represent the messages load over an AR and a HG, respectively. In particular, ρ_{AR} depends on the update and fraction $1-p$ of lookup in a time unit. Since users are randomly assigned to an AR, we can consider the overall load balanced uniformly over ARs. Thus, ρ_{AR} is:

$$\rho_{AR} = \frac{\alpha(1-p)+\beta}{N_{AR}\ \mu_{AR}} \tag{4}$$

where α is the total number of lookups, β is the total number of updates and N_{AR} is the number of ARs in the managed network.

Similarly, ρ_{HG} depends on the fraction p of lookup messages redirected by AR connected to it, and on the fraction $\frac{1}{N_{HG}}$ of updates coming from that AR (an AR issues an update procedure over the DHT by choosing at random the target HG). Thus, ρ_{HG} is:

$$\rho_{HG} = \left(\frac{\alpha}{N_{AR}N_{HG}}p + \frac{\beta}{N_{HG}}\right)\frac{1}{\mu_{HG}} \tag{5}$$

Finally, the average time spent by a message in system queues is determined as follows:

$$W = (1-p)\ W_{AR} + p\ W_{HG} \tag{6}$$

In the extended version of this paper [11], we provide a validation of the theoretical model.

V. Evaluation

We run a simulation study aiming at assessing the trade-off between the number of ARs and HGs that could lead to a significant managed network size reduction, while providing services within QoS constraints in terms of messages latency.

Test details: We used realistic data collected in an experimental analysis to simulate our model on a large scale system. These data concern: (i) the *service time* (i.e., the time needed for extracting a message from the queue and accessing user information); (ii) the transmission delay in the provider managed network. The service time follows a Gaussian distribution with mean value equal to 4.2323 milliseconds and standard deviation equal to $3.91626\ 10^{-3}$ milliseconds for ARs, and mean and standard deviation for HGs equal to 2.05195 and $1.0759\ 10^{-1}$ milliseconds, respectively. Transmission delay in the managed network is modelled as a Gaussian distribution with mean value and standard deviation of 2.44949 and $2.19\ 10^{-1}$ milliseconds respectively. The mean value and standard deviation for transmission delay in DHT channels were set to 30 and 2.569 milliseconds, as assessed in a previous study on a WAN environment [8]. More details can be found in [11]. The architecture and VoIP service are implemented using OMNeT++ v. 4.0 64-bits, a C++ component-based simulator [9]. We fixed the number of HGs to 4 millions. Thus, we simulate a scenario in which all fixed users are equipped with an HG. Currently, not all customers are provided with a home gateway. However, due to the continuous effort telco providers are making toward the reduction of management costs, we expect that in few years the simulated scenario will be adopted in practice. For each configuration *number of ARs - number of HGs*, we evaluated the mean message latency.

All the following results are the average of 5 different tests on the same scenario; that is, the number of unique profiles in the network is 10 millions; each profile has $k = 2$ copies, one stored in the managed network and one in the unmanaged network.

We simulated the service execution for one hour and fixed the maximum request rate RR to 3000 lookup/sec. This value is obtained by considering the number of phone calls in a peak hour for 1 million population of fixed [10] and mobile [6] users. For a population of 10 millions users, we assume the number of calls growing approximately up to 300. In order to accommodate the growing of applications using such services as depicted in figure 1, we estimate an aggregate RR of an order of magnitude greater than the one expected for phone calls.

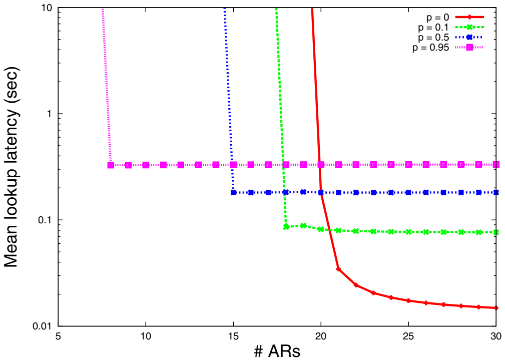

Fig. 2. Mean lookup latency with 4M HGs varying the number of ARs

Results: Figure 2 shows the mean lookup latency varying the number of ARs in the set $\{5-30\}$ and the probability p of traffic redirection in the set $\{0, 0.1, 0.5, 0.95\}$. The parameter that helps in reducing the managed network size is p. Redirecting a lookup message avoids it to wait for a long time in the AR's *Access to Memory* queue, due to the presence of update and other lookup messages. Augmenting p from 0 to 0.95, the number of required ARs in the managed network is more than halved. In addition, the figure shows an interesting trade-off between number of ARs and p: an higher p value helps to reduce the managed network size; however, it imposes an overhead due to lookup procedures over the DHT. Hence, a telco provider should properly set p accordingly with the managed network size that is willing to maintain and the expected QoS standard.

Figure 3 shows how experimental results are confirmed by the theoretical model. Formula (6) represents a powerful tool for a telco provider to determine the composition of the system internal infrastructure (number of ARs versus number of HGs) in order to guarantee services within the desired QoS constraint. In (6), μ_{AR} and μ_{HG} are set to 238 and 130 messages/sec. respectively, as assessed in [11].

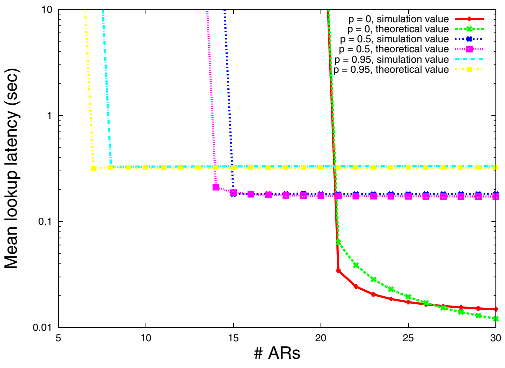

Fig. 3. Comparison between simulation and theoretical results with 4M HGs varying the number of ARs

Finally, we use function (6) to compute the minimum number of ARs and HGs that guarantee a service with mean lookup latency of 200 milliseconds [3]. Results are reported in table I. They show that when the probability p is very high (i.e., $p = 0.95$), the number of ARs in the network is significantly reduced. In this case, the size of the DHT has a big impact on the mean lookup latency, due to the high number of lookup procedure performed over it. Hence, a small DHT has to be used (i.e., $10k$ HGs), in order to decrease the mean number of hops to retrieve a user profile and, in turn, the lookup latency. In contrast, when p is lower (i.e., $p = 0.1$, $p = 0.5$) the lookup latency is mainly determined by ARs; thus a higher number of servers in the managed network is required. In this case the DHT size has much less impact due to the limited number of lookup procedures performed over it.

p	**HGs**	**ARs**
0.1	$100k$	19
0.5	$100k$	14
0.95	$10k$	8

TABLE I
POSSIBLE ARS-HGS CONFIGURATIONS THAT GUARANTEE A MAXIMUM LOOKUP LATENCY OF 200 MS

VI. CONCLUSION

In this work we described a hybrid architecture for supporting telco and third party services in NGNs. In particular, to embrace the expected developments of these services in the close future, we considered an environment in which the number of requests of lookup per second is one order of magnitude greater than the one generated by current VoIP services. This load is balanced between a core network consisting of a set of servers and an edge network constituted by home gateways arranged as a DHT. We have shown that the DHT can be used to significantly reduce the number of servers while meeting specific QoS requirements. This in turn implies a reduction of the operational (energy and management) cost from the operator despite the increased load.

REFERENCES

[1] Rosenberg, J., Schulzrinne, H., Camarillo, G., Johnston, A., Peterson, J., Sparks, R., Handley, M., Schooler, E.: SIP: Session Initiation Protocol. RFC 3261 (Proposed Standard) (2002)

[2] ITU-T Rec. Y.2001: General Overview of NGN (2004)

[3] Singh, K., Schulzrinne, H.: Peer-to-Peer Internet Telephony Using SIP. NOSSDAV '05: Proceedings of the International Workshop on Network and Operating Systems Support for Digital Audio and Video (2005)

[4] Stoica, I., Morris, R., Karger, D., Kaashoek, M.F., Balakrishnan, H.: Chord: A scalable peer-to-peer lookup service for internet applications. Proceedings of the 2001 conference on Applications, technologies, architectures, and protocols for computer communications, page 160 (2001)

[5] Telecom Italia, http://www.telecomitalia.it

[6] Bregni, S. and Cioffi, R. and Decina, M.: An Empirical Study on Time-Correlation of GSM Telephone Traffic. IEEE Transactions on Wireless Communications, volume 7, number 9, pages 3428–3435 (2008)

[7] Kleinrock, L.: Queueing Systems: Volume 1: Theory, 1975, John Wiley & Sons New York.

[8] Baldoni, R., Marchetti, C., Virgillito, A.: Impact of WAN Channel Behavior on End-to-end Latency of Replication Protocols, In Proceedings of European Dependable Computing Conference, 2006

[9] OMNeT++, http://www.omnetpp.org/

[10] Iversen, V.B., Glenstrup A.J., Rasmussen, J.: Internet Dial-Up Traffic Modelling, NTS-15, Fifteenth Nordic Teletraffic Seminar, Lund, Sweden, August 22-24, 2000

[11] Baldoni, R., Beraldi, R., Lodi, G., Platania, M., Querzoni, L.: Moving Core Services to the Edge in NGNs for Reducing Managed Infrastructure Size, Technical Report 2010, http://www.dis.uniroma1.it/~midlab/ngn_tech_rep_10.pdf