

Autonomic Service Access Management for Next Generation Converged Networks

Monique Calisti, Roberto Ghizzioli, Dominic Greenwood

Abstract. This chapter presents the Living Systems Autonomic Service Access Management Suite, LS/ASAM, a comprehensive middleware solution enabling adaptive connectivity management of nomadic end hosts across heterogeneous access networks with autonomic optimization of network performance and availability.

1. Introduction

Next Generation Networks, NGN, are becoming increasingly open, shared and with infrastructure that is reliant on highly distributed components. This is largely being driven by the vision of ubiquitous broadband access that is continually evolving the way business and consumer customers interact. These networks must thus continue to improve in terms of performance through multiple dimensions, including for example service mobility, personalization, transparency and immediacy.

This evolution of network infrastructure offers operators the possibility to create many new forms of business. However, it also poses some significant new challenges in many areas of communications and service management, especially in resource-limited access networks. The NGN view is to rely upon an all-IP infrastructure, offering a clean separation between network and service layers and enabling QoS provisioning “out of the box”, which should be easier to manage and less expensive to maintain. However there are several factors which complicate the overall NGN picture.

To be published in Calisti, M., Meer, S. and Strassner, J. (eds.) <i>Advanced Autonomic Networking and Communication</i> - Whitestein Series in Software Agent Technologies and Autonomic Computing.
--

End users are increasingly demanding new services, and dynamic, case-specific service aggregations, to support a seamless and consistent experience across multiple access technologies, devices and locations. They expect to be always best-connected, i.e., to have anywhere and anytime access to the best available technology with the maximum capacity on offer, plus easy-to-use and problem-free services, all at ever lower prices.

Indeed, the proliferation of applications, services and heterogeneous technologies, including advanced multi-modal end users devices, enables a variety of ubiquitous deployment scenarios, but also poses significant challenges in terms of service usability and personalization. This is further complicated by the need to integrate new solutions with legacy systems, while optimizing resource-limited consumption (e.g., radio frequency in access networks).

In addition, the widespread expansion in the availability of high-speed broadband access technologies including cable, DSL, powerline, satellite, and wireless, is encouraging the entry of new service providers in both the fixed and mobile telecom sectors, thereby stimulating a competitive environment. In response operators need to identify means of lowering operating costs by optimizing service provisioning performance and connectivity management.

It is our belief that a fresh approach is required to achieve these objectives. We thus propose a comprehensive policy-driven, autonomic software solution spanning provider infrastructure and end-user devices that positions auto-adaptive control software directly within the devices. The majority of service and connection provisioning approaches in use today tend to operate on the traditional client/server model and are thus rather ineffective due to a common inability to handle the increasing dynamicity and diversity of heterogeneous access technologies. In this perspective, emerging solutions need to be “autonomic” by design; their components should be able to self-regulate and dynamically optimize their own behaviour according to detected changes in their host environment [1].

We call our approach the Living Systems Autonomic Service Access Management suite, LS/ASAM. It is a comprehensive and innovative solution that enables effective delivery of next-generation ubiquitous services by dynamically combining end user requirements and service provisioning policies with network-facing management and control functionality. By automating selected low-level processes on both the user and operator sides and introducing more “personal intelligence” (user context and behavior awareness) and “network intelligence” (network services, content and resources awareness) throughout the whole service delivery chain, the LS/ASAM solution realizes *Autonomic Service Access Management* (ASAM). The guiding ASAM vision is to use autonomic techniques that enable operators to efficiently manage and optimize resource utilization, performance and end user experience. This is achieved by transparently tuning service parameters while taking into account changes in both the client and network context.

This chapter continues with a discussion of the ASAM core principles before presenting the architecture and features of the LS/ASAM Suite as a means to

realise the ASAM vision. We then describe two key deployment scenarios coupled with a discussion of some of the most distinctive characteristics.

Subsequently we provide some data on experimental work conducted in the laboratory on performance analysis of the LS/ASAM suite prototype. The ASAM simulator is described before the presentation of selected results from recent experiments.

We conclude the paper with some discussion remarks, experimental conclusions and targets for ongoing work.

2. Autonomic Service Access Management

Due to the increasing deployment of multiple access technologies at the edges of networks, the management of ubiquitous communications and services is changing rapidly. Intelligence and specific management and control functions need to be migrated toward the edge of the network and even onto the customers' devices. In particular, *service access management*, i.e., the set of functions including the selection and maintenance of one of several available communication channels, is increasingly demanding:

- Fast and appropriate adjustment of the relevant connectivity parameters to a continuously changing network environment.
- The assurance of sufficient service quality and reliability, whose perception can vary from one user to another.
- In coordination with the aforementioned points, the optimisation of resource usage and reduction of operational costs.

Autonomic Service Access Management, ASAM, addresses these issues by dynamically and automatically adapting the configuration and utilization of available network access resources in a reliable and cost-efficient way. This is achieved by embedding specialized intelligence into complex multi-technology and multi-service access networks, including end user devices. The chosen approach is to deploy smart techniques allowing operators to efficiently manage and optimize resource utilization, performance and end user experience. This by transparently tuning service parameters (e.g., bandwidth, average delay), while taking into account changes in the context, including user preferences, Service Level Agreements (SLAs), user location, devices features, and network resources.

ASAM bases its adaptivity on the capability to autonomously observe, extract, understand and use context information to consequently modify its own functionality. Information exchange and correlation between client devices and access nodes, as well as between access nodes even of different technologies, is at the core of this approach. In particular, through dynamic mediation between (often conflicting) requirements on the client and network side, capacity for given connection requests is allocated by taking into account the status of the whole service provisioning chain. This requires accounting for a variety of parameters

that characterize the connection to be created, the consequently required network resources, and the policies existing both on the user and provider side.

For this to be realized, flexible and distributed monitoring, configuration and maintenance tools need to be smoothly interfaced and integrated within the evolving networking environment and pre-existing management systems. This is not an easy task, especially when considering that many operators must deal with a diverse mix of systems and processes that make it difficult to effectively monitor and tune service performance once already in the delivery phase. In this perspective, a new kind of management solution is needed. A comprehensive policy-driven and autonomic architecture, spanning basic infrastructures and end-user devices, which builds adaptive control functionality directly into the corresponding elements, enabling the shift of focus from technology to value-added services.

LS/ASAM is a comprehensive ASAM solution that addresses these challenges by making use of software agent technology [2]. Autonomous agents that adapt to changes in the environment, minimizing human intervention and service interruption, lie at the foundation of LS/ASAM and provide a powerful means to engineer a distributed and autonomic system that includes:

- Customizable and adaptive routines for automating and tuning repetitive information and control tasks.
- Coordination mechanisms enabling the spontaneous collaboration and dynamic aggregation of services.
- Abstraction of communication components to support context changes through adaptation of semantic grounding.

In this way, autonomous software agents acting as autonomic managers, see Figure 1, are enabling LS/ASAM to exhibit self-management capabilities that increase reliability and performance while reducing operational and management costs. This shifts the burden of many support and control tasks from users to the underlying solution, which assists, facilitates and empowers human decision making.

More specifically, LS/ASAM is a middleware solution empowered with autonomic self-management capabilities, including:

- Self-configuration: policy-based self-configuration of the Suites components according to changes in their usage and working environment.
- Self-optimization: proactive monitoring and control of resource usage, performance and end user experience to enforce optimal behavior.
- Self-healing: automatic fault discovery and correction, both on the end user devices and network elements.
- Self-protection: automatic detection of and protection from unauthorized system control changes.

Control over LS/ASAM components is expressed through policies bound to user preferences and business goals. The system senses, analyzes, plans and executes changes in the environment to ensure that business goals can be effectively met.

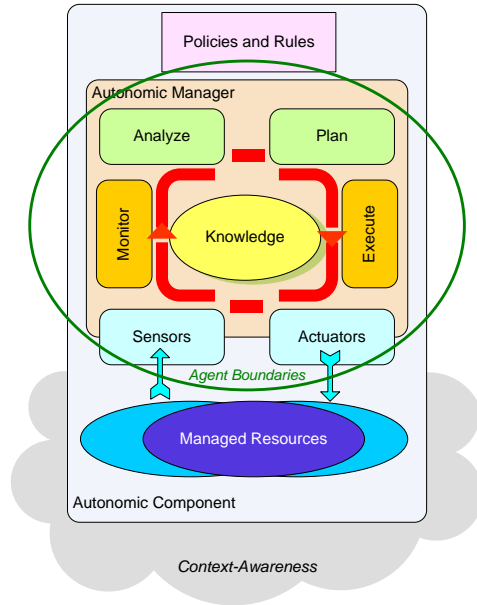


FIGURE 1. An autonomic component architecture.

Although other approaches have been proposed in the literature that address part of the ASAM challenges, none, to our knowledge, is able to dynamically mediate between network and client requirements and accommodate resource allocation and consumption accordingly. In particular, the solution presented in [3], which is the closest one to LS/ASAM, supporting vertical handover in radio access networks. In this system, a dedicated decision module, placed within a concrete provider system, can communicate with various network devices, including client devices, to determine radio access network selection based on QoS parameters. Some degree of negotiation takes place, but only between entities within the network and excluding the client devices that remain passive.

3. The LS/ASAM Suite Architecture

The LS/ASAM architecture includes two main types of autonomic software components, as depicted in Figure 2, which communicate by relying upon the use of common interaction protocols and a shared semantics-based ontology defining all LS/ASAM concepts. These components are:

- *LS/CA*, the *Living Systems Connection Agent*, is a client component that can run on a variety of mobile end user devices (e.g., laptops, PDAs, smart

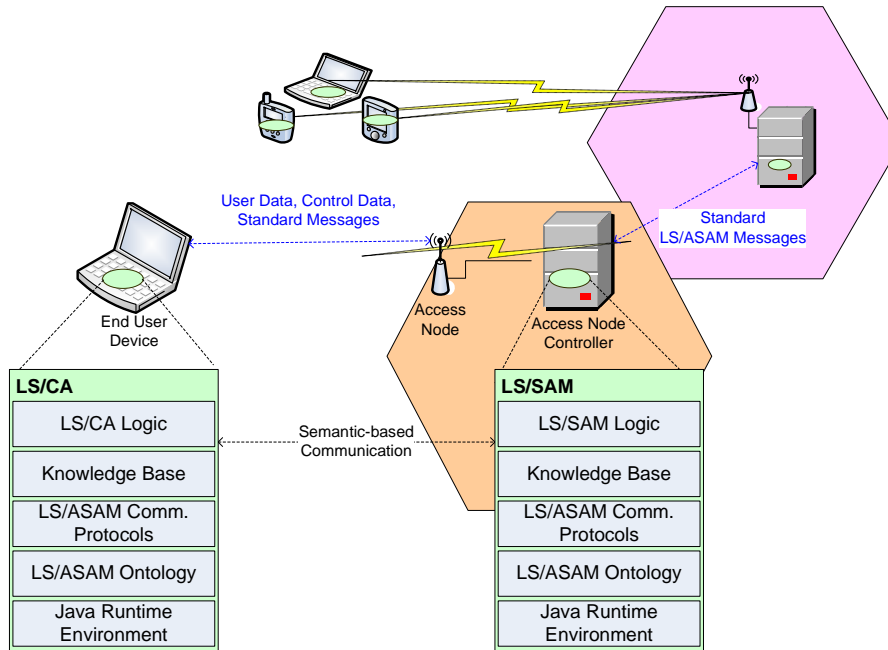


FIGURE 2. An overview of the LS/ASAM architecture.

phones) and provides mobile users with improved quality and reliability by optimizing service access through adaptive connection handover across multiple access technologies and dynamic mediation of service delivery parameters on behalf of the end user.

- *LS/SAM, the Living Systems Service Access Manager*, is a network component that can run on hardware located at the access nodes or at a network management facility. It dynamically optimizes resource allocation across heterogeneous network access domains with adaptive problem recovery and load balancing techniques.

These lightweight software components, i.e., they can live as processes in a Virtual Machine, can flexibly complement and extend many existing service management architectures, and are able to run on resource-limited devices and support asynchronous communication with intermittent network connections. By dynamically coordinating their actions and behavior, they enable adaptive communication service access by mediating between operator policies and end-users requirements and preferences.

3.1. The Living Systems Connection Agent

The LS/CA component provides adaptive service access by setting connectivity parameters according to the outcome of a mediation process to establish a service

access agreement based on the end user's requirements and the network provider's offering. This is determined by a set of factors including:

- Quality requirements of the applications and services running on the device the LS/CA is embedded in.
- Physical end user device status, e.g., battery power level, and properties, e.g., available network interfaces.
- Existing service provisioning conditions according to pre-defined subscription contracts/SLAs.

The LS/CA proactively manages and processes this information according to policies which capture end user preferences, e.g., minimising connection costs, maximising battery life when on-the-move, etc., and supports the following main features:

- *Seamless handover and session continuity.* This guarantees interruption-free service access across multiple technologies by allowing an LS/CA empowered device to maintain the same IP address for an entire session. This is achieved by making use of Mobile IP technology [4].
- *Secure communication.* Tight integration of the LS/CA with several third party VPN clients allows permanent secure connectivity. Furthermore, by integrating IPSec [5] and Mobile IP, the LS/CA ensures end-to-end encryption of all generated traffic (as an optional feature).
- *Connection adaptation.* This indicates automatic detection of available networks and selection of the preferred network adapter (access technology) based on service requirements and network conditions for improved reliability and QoS. This can trigger dynamic mediation between the LS/CA and the LS/SAM components.
- *Context-aware user support.* Through semantic service specifications, policy-driven decision making and dynamic information retrieval, the LS/CA improves end-user experience by directly addressing low-level issues (e.g., failure recovery, connection adaptation), while taking into account user policies and boundary constraints, i.e., context-based information and coordination with LS/SAM components as needed.

From the LS/CA perspective the mediation process is initiated by sending a Call For Proposal, CFP, to one or several LS/SAMs. Naturally any LS/SAM with a open connection established with the LS/CA may also receive the CFP so that it can also participate in the connectivity mediation process.

3.2. The Living Systems Service Access Manager

The LS/SAM component proactively monitors traffic and resources in the access node it controls, triggers appropriate actions (e.g., vertical handover, load balancing) according to the network status and current traffic conditions, processes incoming LS/CA calls for proposal and elaborate offers as appropriate - see Section 3.3. In particular, the two main distinctive features enabling LS/SAMs to optimize resource consumption at the access network level are:

- *Load-balancing.* Balancing traffic load across WLAN and cellular networks while considering the QoS needs of running services renders the network more resilient to traffic peaks. This is achieved by dynamic coordination between LS/SAMs that can hand over a certain number of connections to neighboring access nodes according to possibly several operator policies. The use of distributed constraint satisfaction algorithms [6] for LS/SAMs peer-to-peer orchestration enables effective load balancing by taking into account all existing constraints.
- *Congestion recovery.* Real-time and proactive detection, analysis and relief of congestion, reduces call dropping and increases service resilience and availability. Within an access node, once no new network connection can be accepted or the total requested bandwidth exceeds the total available one, i.e., packets are dropped, an LS/SAM can decide upon specific policies and existing SLAs (if any) whether and how to drop or hand over part of the traffic to neighboring access nodes.

LS/SAMs decisions and behavior are guided by the operator's policies that express service provisioning preferences with respect to a variety of aspects including, e.g., how to allocate traffic to balance out network utilization, how to treat specific users (i.e., connections) in case of congestion, how to adapt pricing schemes according to the user's subscription type. This requires dynamic management of information including:

- Traffic conditions and resources available within the access node the LS/SAM is controlling.
- Traffic conditions and resources available in other access nodes that a given portion of traffic can be handed over to, via dynamic LS/SAM-to-LS/SAM coordination.
- Existing service provisioning conditions according to pre-defined subscription contracts/SLAs.

3.3. Adaptive Coordination of the LS/ASAM Components

The mediation process conducted between the LS/CA and LS/SAM components consists of a sequential interchange formulated as a contract-net protocol [7] negotiation with the goal of determining the best connection parameters given the requirements of the end user, the offering of the network provider and the conditions of the transmission medium.

The requirements of the end user toward the provider are a combination of (i) the preferences of the end user formulated as user policies (e.g., minimising connection cost), (ii) the quality demands of the applications running on the end user device (e.g., a given application may require low end-to-end delay), (iii) the status of end user device resources (e.g., battery power, which can affect the selection of the transmission technology), (iv) the technologies supported by the end user device (e.g., only WLAN and UMTS network interfaces available), and

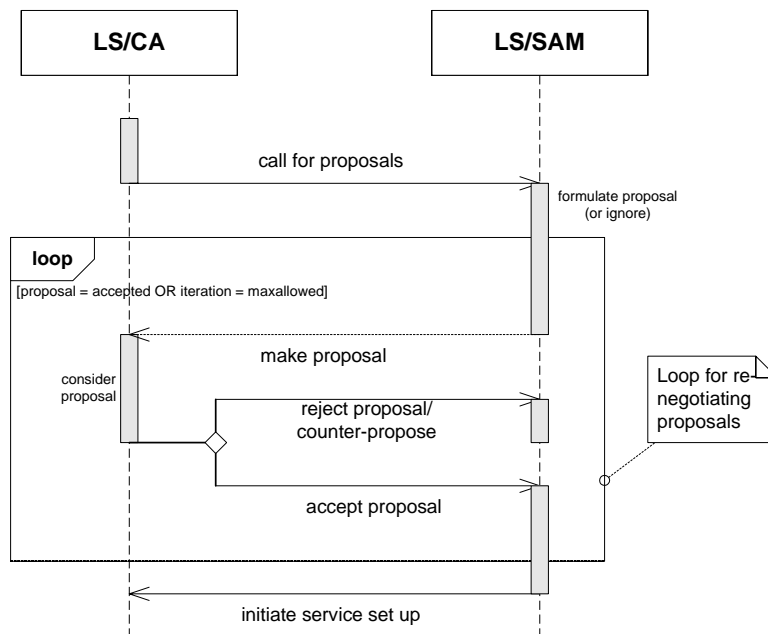


FIGURE 3. Mediation process between the client and network LS/ASAM components.

(v) the conditions stated in the subscription contract (e.g., costs for using certain technologies).

The offering of the provider toward the end user is determined by considering (i) the properties of the provider network (e.g., diversity of network access technologies), (ii) the network status (e.g., distribution of traffic load, delay times), (iii) the capabilities of the network (e.g., mobility support, QoS control) and (iv) the provider policies, including business rules, that relate to the use of its infrastructure, pricing schemes, traffic prioritization mechanisms, etc.

Figure 3 illustrates the typical message exchange during a proposal setup sequence. An LS/CA sends a Call For Proposal (CFP) to one or several LS/SAMs requesting offers to set up a connection with specified constraints including quality requirements, or connection characteristics.

An example of a simple CFP is:

```
(set up connection, (min. bandwidth: 100 KBit/s.
max. delay jitter: 50 ms))
```

Once sent to all prospective LS/SAMs, the LS/CA waits until some predefined deadline to receive proposals and/or rejections. Any LS/SAMs that have not sent a proposal or rejection by this deadline are considered to have been unable or

unwilling to respond to the CFP. A simple example of a proposal sent by a responding LS/SAM is:

```
(set up connection, (network: UMTS, min. bandwidth: 100 KBit/s, max.
bandwidth: 120 KBit/s, max. delay jitter: 40 ms, max. end-to-end
delay: 200 ms))
```

This proposal includes some additional connection parameters than those present in the original CFP. Although not mandatory to do so, these can be taken into account by the LS/CA when evaluating the suitability of the proposal.

The proposals are assessed by means of the Proposal Assessment Function (PAF) that takes as input (i) the set of quality requirements stated in the original CFP, (ii) the received proposal (or the relevant parameters stated in the proposal), (iii) optionally, the user preferences (that can be formulated as user policies), (iv) optionally, the status of the end user device (e.g., battery power level that can affect the selection of the transmission technology), (v) optionally, the properties of the end user device, (vi) optionally, the capabilities of the end user device and (vii) optionally, any Quality of Experience, QoE, metrics, (viii) optionally, the set of network operator policies including business rules.

The PAF computes a sum of weighted differences between the required quantitative parameters and their corresponding values in the proposal. Nominally, the PAF is normalised to a target value domain 0,1 where 0 indicates that the proposal does not satisfy any requirements and 1 indicates that the proposal is valid and fully acceptable. Intermediate results between these bounds indicate the degree to which the proposal meets the CFP requirements. Ancilliary annotations record if the proposal exceeded the CFP requirements for use with counter-proposal negotiations.

At this point the LS/CA must decide whether to make a counter-proposal to any number of selected LS/SAM's that responded favourably to the original CFP. This decision is made in accordance with how well a received proposal meets or exceeds the original CFP request. If selected, a counter-proposal can be issued to a responding LS/SAM in an attempt to initiate bilateral negotiation to revise the proposed offer. Multiple counter-proposal negotiations can be handled concurrently by an LS/CA with active PAF based comparison of each to determine variances between returned proposal updates thereby assisting with refining individual negotiations by taking into account all ongoing negotiations.

A counter proposal is created by modifying a received proposal in accordance with preferred characteristics. If the original CFP sent was:

```
(set up connection, (min. bandwidth: 100 KBit/s))
```

With a received proposal being:

```
(set up connection, (min. bandwidth: 70 KBit/s))
```

The PAF determines that this received proposal is close to its requirements, as expressed in the original CFP, and thus creates a counter proposal in order to initiate fine-grained bilateral negotiation with the sender of the proposal. The counter proposal in the instance of this example may be that the proposed 70Kbit/s bandwidth offer is iteratively increased to 80KBit/s:

Counter proposal: (set up connection, (min. bandwidth: 80 KBit/s))

This counter proposal is a compromise between the original bandwidth specified in the CFP and the bandwidth offered in the returned proposal.

It is important that the decision process exhibits a convergent behaviour to avoid continuous proposal revision. Several suitable algorithms can be found in the literature include that by Hofbauer et al. [8] and by Shamma et al. [9].

When, or if, a proposal is accepted the client device sends an accept-proposal message to the corresponding network provider. All other proposals that have been received are explicitly rejected by informing their source providers. The reason for rejection may be included in the message.

3.4. Technology Foundation

As networks grow increasingly larger and more complex, they become harder to manage efficiently and reliably. This is even more challenging in resource-limited access networks, which affects the capability to deliver true seamless mobility. Thus, network and service management solutions are required to exhibit autonomic behavior.

Their components detect, diagnose and repair faults, adapt their configuration and optimize their performance, while protecting and healing themselves according to changes in the network and operating environment.

The key idea is to assist, facilitate and empower humans (operators, network administrators, customers) by shifting the burden of many support and control tasks from them to the underlying solution components.

As anticipated in Section 2, the LS/ASAM Suite has been conceived and realized by embedding autonomic self-management capabilities at the core of its functionality. Its components autonomously observe, extract, understand and use context information to consequently modify their functionality, according to policies that are bound to business goals. The autonomic capabilities of the client components, LS/CA, and the network component, LS/SAM, are classified as follows:

Self-configuration. The LS/CA adjusts its own configuration according to changes in the working environment in which the user device is located. Policy-controlled profiles for different locations identify the configuration of features to be used, e.g., connection type, VPN, file shares. The LS/SAM performs self-configuration determining its own behavior to achieve high-level directives. This enables the network (namely the access resources the LS/SAMs control, e.g., base stations or access points) to respond dynamically to changes in operator policies and/or

network state. Different load balancing strategies may be adopted, depending on traffic conditions, resource availability and SLAs.

Self-optimization. The LS/CA selects a specific connection type according to user policies and in relation to changes in the context. This is particularly beneficial while roaming in partner networks where the nominal connection may not be the preferred, best or indeed cheapest option. The choice of alternative network adapters can also be triggered by the need of optimizing specific application performance in relation to device properties and network status. The LS/SAM efficiently manages access node resources to meet specified performance objectives under dynamic operating conditions. By proactively balancing load across distinct access nodes (via interaction with peer LS/SAMs) and triggering vertical handover of selected connections, it is possible to optimize network performance and availability according to existing operators policies.

Self-healing. The LS/CA detects faults in related system components (e.g., network cards, drivers, system interrupts) and transparently takes action to repair and circumvent the anomalous behavior. The LS/CA also attempts to re-establish lost connections or, if not possible, seamlessly transitions to a session over an alternative connection type. The LS/SAM is able to detect and repair unpredictable conflicts between service requirements and available network resources. If appropriate, it coordinates its behavior with other LS/SAMs. In particular, real-time and proactive detection, analysis and relief of congestion allows the LS/SAM to reduce call dropping and thereby increase service resilience and availability.

Self-protection. The LS/CA detects unauthorized alterations to obfuscated operator policies stored in the system registry. It stalls operations while replacing the policies with securely obtained replacements. The LS/SAM performs the necessary traffic analyses to detect potential security threats and informs peer LS/SAMs, the overall network management system and/or the network administrator. In particular, the LS/SAM supports identification of malicious nodes that attempt denial of service attacks and blacklists them, warning the complementary access network management components.

4. The LS/ASAM Suite in Action

Ubiquitous data connectivity and communications management are optimised transparently across multiple network access technologies by dynamic coordination of the LS/ASAM components according to the specific situations. In particular, different combinations of their features enable a variety of deployment scenarios. In the following, two of the most significant ones are presented including a discussion of the distinctive characteristics in relation to relevant work.

4.1. QoS Enforcement in Heterogeneous Access Networks

The notion of guaranteed data transmission quality with enforcement mechanisms, in particular for emerging QoS sensitive multimedia applications, e.g., voice or video over IP, is a key issue especially in converged networks [10]. While traffic

prioritization is often not of paramount importance in core networks due to over-provisioning, QoS is an essential differentiator in limited-capacity wireless access networks for capacity and/or delay sensitive traffic such as voice or video over IP. While for cellular access technologies belonging to 2.5G, 3G and 3.5G, appropriate standards for QoS have been defined, few operators yet make widespread use of them. In addition, the WLAN world is supporting its technologies with specifications that directly account for QoS management.

In particular, when integrating different access network technologies, e.g., WLAN and UMTS, the quality of a connection may be degraded during vertical handover where (i) the connection needs to be re-established at the new access node, which is time consuming and during which no data can be transmitted, and (ii) if too many IP packets are lost, they must be retransmitted which can also be time consuming in the case of a large number of packets - again leading to service interruption.

Various approaches have been developed and proposed to address this problem. In [11], a reservation-based QoS model for integrated cellular and WLAN networks is defined and an adaptive mechanism to ensure end-to-end QoS is proposed. However, this model can only work by making the assumption that cellular/WLAN interworking is realized by relying upon a common and uniform reservation-based QoS architecture, which is not (yet) the case for most real network scenarios. Similarly, Song et al. [12] proposed an admission control mechanism for integrated voice and data services in cellular/WLAN networks. The main limitation of this approach though is that it does not account for video traffic.

To effectively provision QoS and optimize resource utilization for a variety of possible heterogeneous network scenarios, the LS/ASAM Suite relies upon the dynamic combination of specific mechanisms both at the client side (i.e., seamless handover, session continuity and connection adaptation) and at the network side (i.e., congestion recovery and load-balancing) that are compliant with dominant industrial standards, e.g., mobile IP or SIP/IMS, when supported, or technology-independent, whenever possible.

Unlike legacy systems and hardware-based solutions, the LS/ASAM components accommodate high-level service and user needs and preferences (including QoS requirements) by implementing coordination mechanisms and resource allocation algorithms that hide low-level access technology dependent processes. This is achieved by deploying an agent-based middleware architecture that provides users with a common and higher level of abstraction, which makes low-level network access heterogeneity transparent.

On the client side, basic QoS in terms of service availability and continuity is enforced by the LS/CA through automatic and policy-driven vertical handover, i.e., all traffic is switched from one network interface, according to existing constraints and user policies. Moreover, by continuously monitoring network conditions and device status and properties, the LS/CA exerts QoS and context-aware resource management by selecting the most appropriate access technology to be

used for the running applications/processes. In addition, when appropriate, as detailed in Section 3.3, the LS/CA can also trigger negotiation with one or more LS/SAMs for different connectivity conditions.

On the network side, the key mechanisms deployed by the LS/SAM to enforce QoS provisioning are load-balancing and congestion recovery. Load-balancing can be triggered by LS/SAMs in order to redistribute traffic across several access nodes according to various criteria, including:

- Current utilization of resources at the access node, e.g., once the traffic overcomes a given threshold a certain portion of the supported connections might be handed over to neighbor LS/SAMs.
- QoS requirements of the running services, e.g., best-effort connections might be handed over to prioritize premium services for which charging might be based on service reliability guarantees (e.g., $\geq 95\%$ non-disruption).
- Predictions of the network resources usage to minimize the probability of congesting an access node.

Analogously, whenever congestion occurs a specific part of the traffic at a given access node might be handed over to other LS/SAMs or selected existing connections (e.g., the non-premium ones) might even be dropped as appropriate. This enables relief of congestion and increases service resiliency and availability.

For example, assume a user that launches an IP-based TV program (e.g., a news channel) on a smart phone. During the launch of the selected application to render the video stream, the LS/CA determines the connectivity parameters (typically bandwidth and delay) for interruption-free high quality service provision. Because different access technologies offer different QoS assurances, the LS/CA might try to switch to a specific technology, e.g., UMTS, that better supports the QoS level needed for the video down-streaming. In addition, in the case of an UMTS connection, the LS/CA would set up a new Packet Data Protocol context requesting the UMTS QoS streaming class [13].

Figure 4 depicts the deployment model for this case. Each end user device is installed with an LS/CA component able to enforce QoS. The LS/CA must be aware of the different traffic categories available in each network access technology. During a vertical handover, the QoS class of the active network is mapped into an appropriate QoS class of the target network. There is one LS/SAM agent being deployed per access node, i.e., each LS/SAM agent is in charge of a specific access node and thus is up-to-date at all times regarding the status of that node. When planning load balancing and congestion recovery, the LS/SAM agent must be aware of the QoS classes supported by the different access technologies to minimize the risk of degraded service quality. This involves LS/SAM-to-LS/SAM coordination first to exchange information on current traffic load (or resource availability) and then to possibly take or hand over part of the communications/traffic¹.

¹Peer LS/SAMs coordination is not described in this paper because of some pending patenting issues.

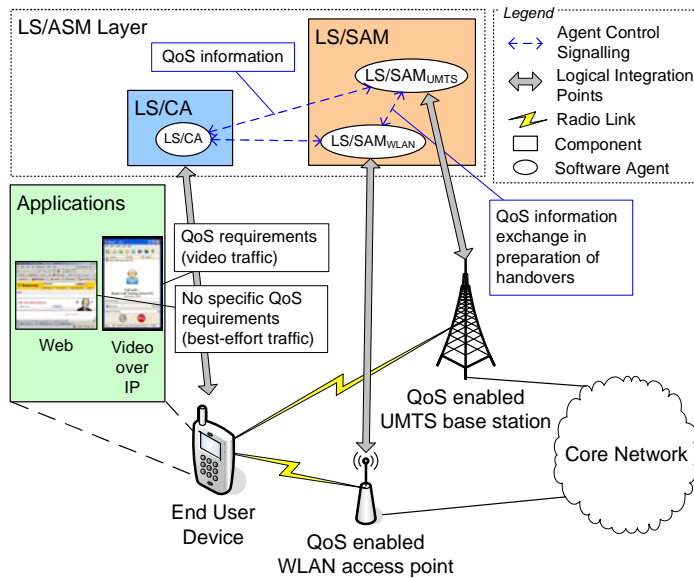


FIGURE 4. Deployment model of the LS/ASAM Suite for QoS enforcement.

4.2. Integration with an IMS/SIP Framework

IP Multimedia Subsystem, IMS, initially developed by 3GPP and 3GPP2 as an IP core network architecture for cellular/wireless-based access to Internet services, is now evolving into a standard that provides a common framework to create and offer next generation converged network services [14]. IMS builds on the Session Initiation Protocol, SIP, that is mainly responsible for delivering a session description to a user at its current location [15]. The key idea is to enable any kind of access (wireless or fixed) for any kind of media (including any combination of voice, text, image and/or video) supporting multiple devices and endpoints.

Because of the (at least initial) co-existence of IMS and non-IMS applications, the costs associated with moving to a full IMS-based network, and the inherent complexity of IMS (and its several standards, interfaces and protocols) most service providers and or operators are expected to migrate toward an IMS service framework iteratively.

One of the core issues to be addressed for successful adoption of IMS is the ability to face more aggressive bandwidth and latency demands, which implies increased QoS management and design capabilities on the bearer network [16]. In particular, IMS/SIP lacks traffic management capabilities and especially adaptive connectivity management and optimization mechanisms that can be regarded as key components for delivering ubiquitous quality-sensitive multimedia services.

In this perspective, the LS/ASAM Suite complements an IMS-based framework by ensuring the quality of delivered services at the bearer network level

through its adaptivity mechanisms, leaving IMS/SIP to cope with call control and service deployment issues. As depicted in Figure 5 the LS/CA component directly interacts with the SIP client installed on the end user device. In this way, the SIP client is able to obtain information on the quality of the connection which is helpful to determine, for instance, the appropriate codec to use, and to request the LS/CA component to ensure a certain quality level (in particular, when explicit QoS class enforcement is enabled). On the network side, an LS/SAM agent integrates with each access node and, by means of load balancing and congestion recovery enables to provide a high level of service quality.

A simple use case is when one considers the collaboration between a SIP client and the LS/CA component to guarantee a level of quality required by a user to perform a video call (or, similarly, to watch Mobile TV). Upon launch of the SIP-based video calling application, the SIP client assesses the connection quality by means of the LS/CA component. The SIP client is aware of the quality requirements imposed by the video call service that are also variable according to the size and quality of the video picture. The LS/CA component can, in collaboration with the respective LS/SAMs, discover the quality offering at alternative access nodes and, based on that decide whether a handover to another access node needs to be triggered. Both end devices that participate in the video call must also agree on the codecs to be used for encoding and decoding the voice and video data. The LS/CA component delivers the necessary information to the SIP client to make its choice. Once the video call is established and running, it is the LS/CA agent's responsibility, in cooperation with the active LS/SAM agent, to preserve the quality of the connection and take appropriate measures if tolerance thresholds are violated. Depending on the mobility profile of the user, but also on the evolution of the network conditions, handoffs are unavoidable and thus need to be well planned and efficiently executed to minimize quality breaches.

The LS/CA does not affect the SIP call itself nor infringe any of the IMS/SIP standards. SIP is concerned with controlling the call execution while LS/ASAM takes care of connectivity. LS/ASAM is therefore complementary to IMS/SIP and benefits result even if only a small proportion of the entire network infrastructure (namely the access part) and end user devices are LS/ASAM empowered.

5. Experimental Analysis

In order to give a measure of the concrete benefits brought to a telecom operator by the adoption and deployment of a solution based on LS/ASAM, several experimental tests have been performed. This section first introduces the ASAM simulator, an instrument built for validating the basic concepts and evaluating various autonomic service access strategies on a set of simulated network settings representing real scenarios. One particular scenario is then selected to illustrate performance when different service access algorithms have been deployed in the

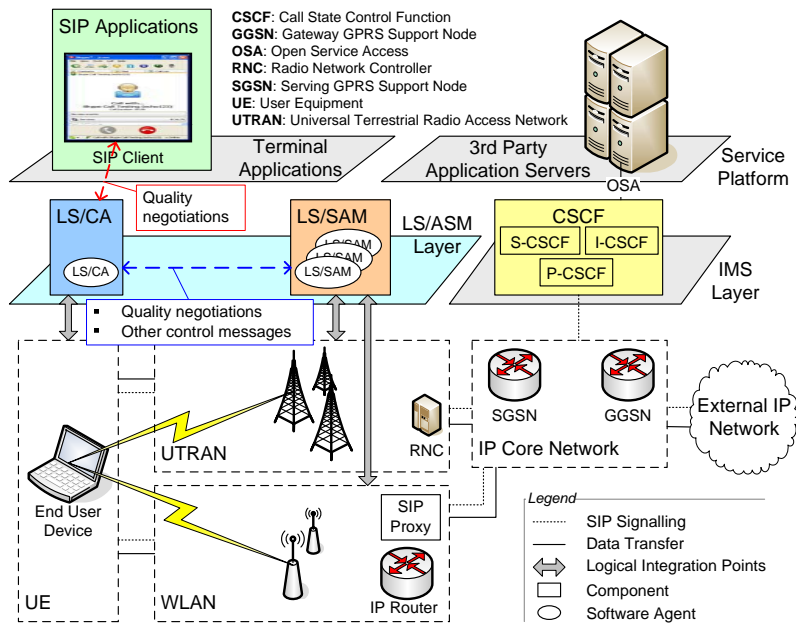


FIGURE 5. Deployment model of the LS/ASAM Suite when integrating with an IMS/SIP-based architecture.

user devices and in the access network. A set of preliminary experimental results are provided, obtained from the comparison of the discussed access strategies.

5.1. The ASAM Simulator

The ASAM simulator is an instrument built in Java for validating the ASAM concepts, in particular, how different autonomic access strategies deployed into LS/CA and LS/SAM modules should perform in real network access scenarios.

In the real world people use their portable devices to request services in accordance with changes in location, activity, and other requirements. Using radio communication they are able to connect to a network operator offering a heterogeneous infrastructure of different access node types (e.g., WLAN access nodes, GPRS/UMTS antennas, etc.) In the ASAM simulator, both user devices and access network components are modeled using software agents which. Agents that simulate a user device can make use of the LS/CA where specific service access strategies are pre-loaded. In the same way, an agent representing an access network component can make use of the LS/SAM capabilities. The interaction between a device and an access node is then mapped through an exchange of FIPA-compliant messages.

Input Parameters. Within the ASAM Simulator time is discrete and the simulations are based on the *quasi-static* condition. For this reason, input parameters related with the time are:

- Start time of the experiment.
- Duration of the experiment.
- Duration of a time step (e.g., 1 minute).

Furthermore, other parameters are required to describe the scenario:

- Locations represented in the experiment (e.g., train station, street, offices, etc.).
- Types of available network interfaces (e.g., UMTS, EDGE, etc.).
- Set of network services that are simulated in this experiment (e.g., phone call, VOD, email, etc.).
- Set of access nodes.
- Set of end user devices.

For each access node (e.g., WLAN access point, UMTS cell, etc.) the following input parameters are required:

- Type of network technology represented by this access node (e.g., UMTS cell).
- Nominal bandwidth of an access node measured in Bit/s.
- Maximum number of concurrent connections.
- Maximum bandwidth deployable on a single connection.
- Version of the LS/SAM the access node makes use (e.g., none, *LS/SAM-BN²*).

Finally, for each user device to be simulated, the following input parameters are necessary:

- Location of the device at the beginning of the simulation (e.g., street).
- Set of network adapters installed in device (e.g., only GPRS).
- Amount of bandwidth that can be used at maximum given the network technology (e.g., 11Mbit/s for WLAN).
- Set of service descriptors denoting the services that are available to the user who operates the device (e.g., a normal mobile phone can perform only calls).
- Version of the LS/CA the user device makes use (e.g., none, *LS/CA-APF²*).
- A set of input parameters used to define a mathematical model which describes the behaviour of the end-user while using the device. This is defined in terms of movements among locations, usage rate and duration of services while being located in a given place. In particular, the following matrices must be provided:
 - The average time before an user changes her location, moving from one environment to another one.

²The suffix (APF in this case) determines the type of autonomic access strategies the component implements.

- The average time before an user issues a service demand while being located in a given space.
- The duration of a started service while being located in a given space.

The implemented mathematical model is based on the *Markovian property* that the probability of the occurrence of an event does not depend on the history of previous events. Based on this property, events like the starting of a service or the movement between locations are simulated with an occurrence rate equal to the inverse of the λ parameter of a *negative exponential* distribution. Furthermore, the duration of of a started service is simulated using the *Erlang-k* distribution. The expected average and standard deviation of the service duration are used to define the distribution.

In the ASAM simulator, each device has a user event generator that implements this mathematical model. The generated *user events* represent movements or service initiations with a stochastic duration. When an event occurs, the action is simulated (e.g., start a VOIP call in a road for 2 minutes). Each time an event is consumed, a new one is immediately generated. The generator also terminates elapsed services.

Output Variables. The ASAM simulator provides a set of output parameters that measure the performance of the LS/CA and LS/SAM strategies. The following list of output parameters includes only the subset of those used in Section 5.3:

- M_{ur} : The average used bandwidth of an access node in relation to its nominal bandwidth. High M_{ur} values encounter a high average utilization of the access nodes which means that the infrastructure is more efficiently utilized.
- M_{sr} : The satisfaction rate of a demand is an indicator for the service quality that a user receives. Currently, this variable considers only the amount of bandwidth consumed versus the amount of requested.
- M_{fc} : The accumulated time span during which an end user device receives the bandwidth it requests and thus can deliver full service quality to the user. Values are normalized in the range [0..1].
- $M_{d.vho}$: The average occurrence rate of vertical handoffs in a time step when triggered by a user device.
- $M_{n.vho}$: The average occurrence rate of vertical handoffs in a time step when triggered by an access node.

5.2. Simulation Setup

This section presents preliminary laboratory experiments conducted to validate the ASAM concepts through the use of the ASAM simulator. The presented simulation evaluates what might happen in a normal working day during which a large number of people arrive at a train station before dispersing to their places of work where they spend most of their day.

Figure 6 illustrates the simulated access network topology where different access nodes (UMTS/GPRS cells and WLAN access points) cover different locations

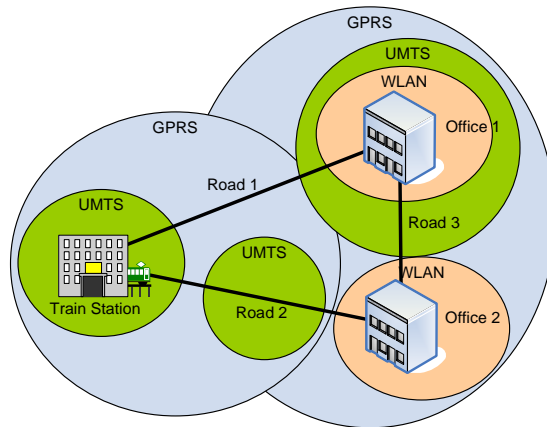


FIGURE 6. Access network topology used in the presented simulation.

TABLE 1. Access nodes properties.

Type	Nominal Bandwidth	Max Bandwidth/Device	Max. Connections
WLAN	2000 Kbit/s	2000 Kbit/s	120
UMTS	1500 Kbit/s	300 Kbit/s	6
GPRS	400 Kbit/s	50 Kbit/s	10

TABLE 2. Average movement rate exhibit by the users.

From \ To	Train Station	Road 1	Road 2	Road 3	Office 1	Office 2
Train Station	-	5 mins	5 mins	-	-	-
Road 1	10 mins	-	-	-	5 mins	-
Road 2	10 mins	-	-	-	-	5 mins
Road 3	-	-	-	-	10 mins	10 mins
Office 1	-	5 hours	-	4 hours	-	-
Office 2	-	-	5 hours	4 hours	-	-

(a train station, three roads and two offices). The access nodes exhibit the characteristics presented in Table 1. The reported values are similar to the characteristics offered by typical network components deployed in most access networks.

In this simulation, 40 users, starting from the train station, move around this scenario with their devices consuming network services. Their devices are able to handle communication with all the available network technologies (WLAN, UMTS and GPRS).

Table 2 describes the average movement rates exhibited by the users. It is important to notice that these rates are unidirectional, that is, the frequency of

TABLE 3. Occurrence rate and duration of services started in specific locations.

	eMail 80Kbit/s	VOIP 128Kbit/s	Internet 240Kbit/s	VOD 1Mbit/s
Train Station	40 mins 5 mins \pm 1	2 hours 2 mins \pm 1	3 mins 5 mins \pm 1	-
Road 1	1 hour 5 mins \pm 1	2 hours 2 mins \pm 1		
Road 2	1 hour 5 mins \pm 1	2 hours 2 mins \pm 1		
Road 3	1 hour 5 mins \pm 1	2 hours 2 mins \pm 1		
Office 1	20 mins 5 mins \pm 1	1 hour 10 mins \pm 1	1 hour 15 mins \pm 2	4 hours 30 mins \pm 5
Office 2	20 mins 5 mins \pm 1	1 hour 10 mins \pm 1	1 hour 15 mins \pm 2	4 hours 30 mins \pm 5

moving from one location to another is not necessarily the same of the reverse direction between the same locations.

The information about how services are used is described in Table 3. In particular, the table presents the average occurrence rate and the duration of a service, given the user location. Moreover, the service duration is described in terms of the average time and its standard deviation. Additionally, the bandwidth consumed by each type of service is defined in the table column headers. Values in Table 2 and Table 3 do not represent a specific case but we believe these quantities are sufficient to analyze how the selected access strategies behave.

The described scenario was simulated for 24 hours with a time step of 1 minute. Given the non-deterministic property of the experiments, 10 simulation runs of the same scenario have been performed.

Network access strategies. In this experiment we consider two network access strategies: one implemented by the LS/CA and one implemented by the LS/SAM.

The access strategy implemented in the LS/CA equipped user devices, named *Adapter Priority Function* (APF), assigns a dynamic priority function at each adapter of the user device. This function takes as input measured and expected parameters values that are: battery power, time since last handover, used bandwidth, end-to-end delay, adapter statistics, adapter cost and creates a weighted linear combination of a set of sub-fuctions built on the listed parameters. If the adapter with the highest function value is different then the current one, an handover is triggered. We named an LS/CA that implements the APF access strategy as *LS/CA-APF*.

TABLE 4. Comparison of the results obtained simulating four different network access configurations.

	Without LS/ASAM	Only LS/CA- APF	Only LS/SAM-BN	With LS/ASAM
M_{ur}	0.2821 ± 0.0062	0.2897 ± 0.0265	0.4048 ± 0.0066	0.4199 ± 0.005
M_{sr}	0.5433 ± 0.0070	0.5839 ± 0.0637	0.7382 ± 0.0137	0.7621 ± 0.0138
M_{fc}	0.1153 ± 0.0088	0.1540 ± 0.0752	0.2216 ± 0.0156	0.2583 ± 0.0260
$M_{n.vho}$	0	0	0.1685 ± 0.0212	0.1478 ± 0.0064
$M_{d.vho}$	0.0169 ± 0.0008	0.0066 ± 0.0006	0.0095 ± 0.0010	0.0015 ± 0.0001

The access strategy implemented in the LS/SAM equipped access nodes, named *Balance* (BN), tries to keep high the quality of the services users require balancing the load among the access nodes available in the user device's neighborhood. Whenever an established connection obtains less bandwidth than the requested one, the access node using a Contract-Net protocol asks to other nodes how much bandwidth they could offer to that connection. The candidate access node should be able to satisfy the requested bandwidth and minimizes the gap between the bandwidth demand and the bandwidth offered. If there are no access nodes that offer more than the requested bandwidth, the connection is assigned to that access node with the highest bandwidth offered. If no proposals are better than what the current access node offers, no handover is performed. We name an LS/SAM that implements the BN access strategy as *LS/SAM-BN*.

If both the LS/CA and the LS/SAM are deployed in the network, that is, the whole LS/ASAM system is in use, a mechanism to avoid conflicts between provider and user strategies is adopted.

In order to understand the benefits provided by LS/ASAM, the scenario where LS/ASAM is not present was also simulated. In this case, access nodes do not exhibit any access logic and the devices select the preferred access node based on the highest nominal bandwidth a network technology provides (e.g., WLAN, UMTS, GPRS).

5.3. Results

Table 4 presents the results obtained simulating four different network access configurations: the case without the LS/ASAM system, the case with only LS/CA-APF components, the case with only the LS/SAM-BNs, and the last case where the whole LS/ASAM system is deployed.

The results show that if only LS/CA components are deployed in the network (third column), they are able to improve all the evaluated metrics when compared with the case where no LS/ASAM components are in place. For example, LS/CAs generate 33% more time where users receive the requested bandwidth even with a lower number of vertical handovers.

Results are even better if we compare the case without LS/ASAM to the case where only LS/SAMs are deployed. In this case, for example, the user satisfactory

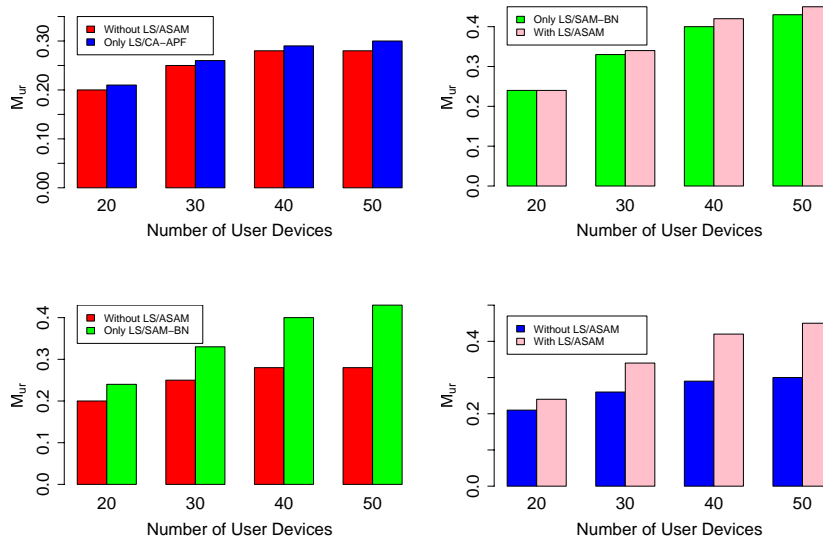


FIGURE 7. The benefits provided by LS/ASAM (or some of its components) against an increasing number of network users.

metric, M_{fc} , improves by 100% and the usage rate metric, M_{ur} , improves by 42%. From this we can conclude that an operator may be able to mitigate the need for extensions to network infrastructure in lieu of deploying some software intelligence into existing infrastructure.

Moreover, we can observe from these results that adding intelligence in the access network brings about greater benefits than adding intelligence to user devices. This is because the network has a broader and real knowledge of the current infrastructure status than a user device that also bases its decisions on estimated values.

Furthermore, we can notice that the combined use of LS/CA and LS/SAM components (the whole LS/ASAM system) generates yet greater benefits than those one generated by one of the two components in isolated use. We believe that further improvements of the evaluated metrics will be obtained with as yet to be reported work regarding the simulation of collaborative strategies between LS/CA and LS/SAM as presented in Section 3.3.

Finally, Figure 7 shows the benefits provided by LS/ASAM (or some of its components) with the increasing of the network users when analysing the M_{ur} metric, that is, usage of the network. Considering for example the histogram in the bottom-right of the figure, where a system with LS/ASAM deployed is compared to one without, it is notable that as the volume of users increases, so does the gain achieved with LS/ASAM.

To measure the real significance of the obtained results the Wilcoxon Paired Rank Sum Test was applied. This test stated with a confidence level higher than

95% that the improvements generated by the LS/ASAM system are statistically significant.

6. Discussion and Conclusions

The LS/ASAM Suite is a distributed and resilient system that exhibits high adaptivity to its network environment. This has been achieved by properly combining multi-agent systems concepts and technology with powerful resource allocation algorithms and reasoning strategies.

The central idea is that loosely-coupled distributed management functions and control methods can be well-modeled and implemented by making use of automated, goal-driven and proactive software entities. These lightweight components are able to operate on resource-scarce devices and support asynchronous communication with intermittent network connections. Moreover, according to the results of proactive monitoring information received from the environment within which they are embedded, the LS/ASAM components directly assist with autonomous management of network resources. They are able to configure themselves and dynamically optimize their operations according to the way their environment changes and in-line with operator and client user policies. They thus assist with the speed-up and automation of simple, tedious and repetitive service management tasks currently performed most commonly by human operators. The ultimate result of this is potentially substantial cost savings to the operator. In particular, by hiding low-level networking aspects that, especially in converged network scenarios, can continuously change due to end users mobility, the LS/ASAM middleware provides transparent service access in heterogeneous networks and becomes an essential complement to (bearer unaware) service delivery platforms.

However, to achieve the potential of autonomous management systems in today's networks is not a straightforward task. Migrating intelligence and complex management functions toward the edge of the network reduces the degree of manual intervention needed, but increases somehow the complexity of the management system itself. The network has indeed to be adaptable, but at the same time stable and controllable. Therefore, populating the networking environment with autonomous software components requires some additional configuration and monitoring capabilities. In this sense, middleware technologies for highly dynamic and heterogeneous networks must become able to monitor and control the middleware itself by integrating with traditional, relatively static infrastructures often populated by legacy solutions and adapting to different operating systems and connection technologies. This is a challenging task that still requires additional investigation.

The system described in this paper has been implemented as a fully-functional prototype with an accompanying scenario simulator for experimental evaluation. The results presented in this paper are rather preliminary in that we cannot yet report on the fully mediated solution, but they are nevertheless extremely encouraging. In particular, the demonstrable performance improvement with the

complete LS/ASAM suite in operation, as shown in Figure 7, is a quite significant result.

Our ongoing and future work includes more refined and extensive characterization of LS/ASAM performance, especially on the network side, when adopting different user and operators policies, network allocation strategies and algorithms. While the LS/CA has been already successfully deployed in a variety of real-world scenarios, the adoption of the LS/SAM requires some additional work given the wide assortment of existing and upcoming service and network management architectures. In particular, by simulating and analyzing the LS/ASAM Suite performance in a variety of networking scenarios and consequently refining the behavior of the various system components, we expect to better characterize and consequently improve performance and scalability. In addition, by means of selected testbed demonstrations and experiments we are assessing the feasibility and complexity of integrating LS/ASAM entities in specific service delivery frameworks including IMS.

Acknowledgment

Many thanks to the colleagues at Whitestein Technologies who contributed significantly toward this work, in particular Thomas Lozza, Martin Stangel, Oliver Hoeffleur, Oliver Carl and the LS/CA team.

References

- [1] Strassner, J.: Autonomic Networking - Theory and Practice. In: Proceedings 9th IFIP/IEEE International Symposium on Integrated Network Management, Nice, France (May 2005)
- [2] Jennings, N.R.: Agent-Oriented Software Engineering. Lecture Notes in Computer Science **1647** (1999) 1–7
- [3] Ferrus, R., Gelonch, A., Sallent, O., Perez-Romero, J.: Vertical Handover Support in Coordinated Heterogeneous Radio Access Networks. In: Proceedings 14th IST Mobile and Wireless Communications Summit, Dresden, Germany (June 2005)
- [4] Perkins, C.: RFC 3220: IP Mobility Support for IPv4 (January 2002)
- [5] Kent, S., Atkinson, R.: RFC 2401: Security Architecture for the Internet Protocol (November 1998)
- [6] Yokoo, M., Hirayama, K.: Algorithms for Distributed Constraint Satisfaction: A Review. *Autonomous Agents and Multi-Agent Systems* **3**(2) (2000) 185–207
- [7] Foundation for Intelligent Physical Agents: FIPA Iterated Contract Net Interaction Protocol Specification (2001)
- [8] Hofbauer, J., Sandholm, W.H.: On the global convergence of stochastic fictitious play. *Econometrica* (70) (2002) 2265–94
- [9] Shamma, J.S., Arslan, G.: Unified convergence proofs of continuous-time fictitious play. *IEEE Trans. on Automatic Control* **49**(7) (2004) 1137–42

- [10] Ahmavaara, K., Haverinen, H., Pichna, R.: Integration of wireless LAN and 3G wireless - Interworking Architecture between 3GPP and WLAN systems. *IEEE Communications Magazine* **11**(41) (2003) 74–81
- [11] Wang, X.G., Min, G., Mellor, J.E., Al-Begain, K., Guan, L.: An adaptive QoS framework for integrated cellular and WLAN networks. *Computer Networks* **47**(2) (2005)
- [12] Song, W., Jiang, H., Zhuang, W., Shen, X.: Resource management for QoS support in cellular/WLAN interworking. *IEEE Network* **19**(5) (2005)
- [13] 3GPP: TS 23.107 v7.4.0: Quality of Service (QoS) concept and architecture. (June 2006)
- [14] Cuevas, M.: Admission control and resource reservation for session-based applications in next generation networks. *BT Technology Journal* **23**(02) (2005)
- [15] Kolbehdari, M., Lizotte, D., Shires, G., Trevor, S.: Session Initiation Protocol (SIP) Evolution in Converged Communications. *Intel Technology Journal* **10**(01) (2006)
- [16] ABIresearch: IP Multimedia Subsystem Industry Survey Results (2005)

Monique Calisti, Roberto Ghizzioli, Dominic Greenwood
Whitestein Technologies AG
Pestalozzistrasse 24
CH-8032, Zurich,
Switzerland
e-mail: {mca,rgh,dgr}@whitestein.com