An analysis of mobile signalling in converged networks

Oredope, A.; Exarchakos, G.; Menkovski, V.; Liotta, A.

Published in:
Proceedings of the 9th IEEE Malaysia International Conference on Communications (MICC 2009, Kuala Lumpur, Malaysia, December 15-17, 2009)

DOI:
10.1109/MICC.2009.5431419

Published: 01/01/2009

Citation for published version (APA):
An Analysis of Mobile Signalling in Converged Networks

Adetola Oredope, George Exarchakos, Vlado Menkovski and Antonio Liotta

Abstract—Converged networks are viewed as a multi-access platform on which fixed and mobile communications can be easily merged into a unified system thus enabling the deployment of rich and personalized services. In this paper, we provide the outcomes of our research into the challenges introduced by terminal mobility in converged networks. This was achieved by carrying out performance based evaluations on various mobile network properties at the access, core and application levels of converged networks and then analysing how they affect the platform individually and as a whole. We also provide the details of our state-of-the-art experimental tests, and analyses that were carried out on emulated converged network test beds, novel architectures, recommendations and extensions which were specifically developed to improve the performance of mobile applications and services in converged networks building on key technologies such as Mobile IPv6, Session Initiation Protocol (SIP) and Peer-to-Peer (P2P) Networks.

Keywords—Converged Networks, MIPv6, SIP, P2P, Mobility

I. INTRODUCTION

The phrases “Triple Play” and “Quad Play” have been made quite popular due to the recent trends and developments within the telecommunications and Internet industry, in order to describe the capabilities and advantages of converged networks. However, converged networks can be described based on two broad categories, namely services and infrastructure. In terms of services, converged networks are perceived as a single platform for seamlessly providing voice, video, and data services or an integration of these three services (i.e. Triple Play) to end-users [1]. In terms of infrastructure, converged networks can be described as a single multi-access platform providing mobile, fixed, broadband, and television services (i.e. Quad Play) seamless to end-users. In this paper, we focus on the infrastructural aspect of converged network especially in terms of its mobile properties.

However, before the advent of converged networks, mobile networks gained access to the fixed networks (e.g. the Internet) by enabling the mobile devices to act like modems, thereby using analogue signalling and circuits which have various limitations such limited amount of bandwidth and services. On the other hand, fixed networks were restricted in the deployment of mobile network services due to the limitations in mobile properties such as handovers and a wide range of air interfaces (i.e. WLAN, Bluetooth, GPRS, UMTS) available in mobile networks [2]. Therefore, merging both fixed and mobile networks onto a single platform will definitely raise quite a number of issues since these two technologies were developed independently and each containing respective key properties, which must be retained in order for them to maintain their individual functionalities and advantages. The effects of these issues on the performance of converged networks can be studied based on a broad range of properties such as the service architectures, security, signalling and internetworking properties [3].

In this paper we provide the details of our research into the performance of converged networks with emphasis on the investigation into the effects of the signalling generated due to the merger of mobile and fixed networks onto converged networks. These effects were studied, understood and then analysed at the access, core and application levels of converged networks. Based on the experimental tests, and analyses carried out on real test beds and under production network environments, recommendations were provided, which will further improve converged networks and also support the deployments of future applications using both fixed and mobile properties. The paper is organised as follows: Section II provides a high level introduction into the concept of signalling in converged networks describing the types of signalling at the access, core and application levels. Our experimental testbeds, test and analysis are provided in Section III, showing how the performance of the converged network are affected by key properties such as Mobile IPv6. The experimental test results are discussed in Section IV. Finally in Section V we provided our recommendations and conclusions.

II. BACKGROUND TECHNOLOGIES

Terminal Mobility [4] in converged networks allows devices and end-users ubiquitous access to next generation services over a pure IP multi-access platform at anytime and from anywhere. The challenges introduced by terminal mobility have only been partially addressed, mainly because the integration of mobility and converged network properties poses more stringent requirements that have not yet been fully addressed [5][6]. In this paper we are mainly interested in the signalling effects on the integration of mobility and converged network performance at the access, core and application levels of the platform [7]. This is because signalling in converged networks facilitates the exchange of information between entities on the network in order to create, maintain, manage and terminate communications. The main aim of signalling is...
to negotiate parameters that will allow the entities to communicate effectively. This is achieved by using a wide range of protocols. However, a large amount of signalling could be highly detrimental to the performance of the network as it uses up vital resources that are to be used for actual communication. On the other hand, a limited amount of signalling could also affect the quality of communication due to the fact that limited amounts of information would be exchanged between the entities. The aim of any signalling protocol should be to communicate as much information as possible using minimal data exchange. Converged networks, at each of the access, core and applications levels, use different protocols for signalling which varies in quantity and quality. In order to provide a proper understanding on the concepts of signalling at each level we will use various protocols and services as explained below.

A. Access Level Signalling

At the access level, we study the signalling effects due to the integration of mobility and security because in fixed networks, secured communications over public infrastructures i.e. Virtual Private Networks (VPNs) [8] generates quite a lot of signalling due the authentication, authorisation and accounting procedures involved. Although VPNs has since reached a matured level in fixed networks, their deployments in mobile networks i.e. Mobile VPNs (MVPNs) is still in its infancy. Therefore the integration of MVPNs in converged networks is bound to raise some concern [9]. Although converged networks already have some mechanisms in place to enhance the deployment of mobile services, we were particularly interested in determining if the performance of MVPNs will be improved due to these mechanisms, or if the performance of converged networks will be affected as a whole due to the deployment of MVPNs. Furthermore, significant attention has been devoted to individual studies of mobility at the access network level in converged networks with little regard to the integration of mobility and security. However, little attention has been given to MVPNs based entirely on the latest freely available software, which is also one of the motivations of our research.

B. Core Level Signalling

The core level is made up of a back-end infrastructure which enables the converged architecture. We are particularly interested in the integration of mobility and the core signalling protocol due to the fact that most of the signalling is generated at the core level of the converged networks [10]. We are particularly interested in the IP Multimedia Subsystem (IMS) [3,11,12], which is the Third Generation Partnership Project (3GPP) standardised core network for the all-IP convergence of fixed and mobile networks. The core signalling protocol in IMS is the Session Initiation Protocol (SIP) [13], a lightweight standardised protocol for creating, modifying and terminating multimedia session over the Internet. SIP is a mature protocol in fixed networks since it is used by a wide range of next generation services such as Voice over IP (VoIP) and multimedia conferencing. Although SIP has also been deployed in mobile networks, a different SIP profile has to be developed for mobile networks due to its varying mobile properties and also as an improvement to the performance of the current SIP profile in mobile networks. These two SIP profiles (i.e. fixed and mobile) are used in converged networks in order to support its fixed and mobile services. In order to allow the interworking of both the fixed and mobile SIP profiles, the IMS uses a gateway known as the Back-to-Back User Agent (B2BUA) [14]. These gateways introduce some limitations in terms of end-to-end communications, Quality of Service (QoS), charging, signalling interworking, media adaptation and protocol translation. These are major drawbacks, which lead to a dramatic reduction in the overall system performance as we will show in Section III.

C. Application Level Signalling

At the Application level, there is a motivation to study the signalling effects on the integration of mobility and distributed applications in converged networks. Distributed applications are known to generate quite a lot of signalling traffic especially when deployed in serverless environments like in Peer-to-peer (P2P) applications which are known to be highly scalable and redundant when fully stabilised [15-17]. This is to investigate the extent of these P2P services scalability and redundancy properties in converged networks especially when integrated with mobility properties in highly volatile environments, where the peers are constantly joining and leaving, thus not allowing the network to stabilise. Although most of the efforts on distributed applications over converged networks have been directed towards the development of the core decentralized SIP protocol i.e. Peer-to-peer SIP (P2P-SIP) [18], there has been little or no publicly available efforts towards development of tools to study, test, and analyze the properties, effects, and limitations of large-scale P2P deployments in converged networks, to see whether theoretical efficiencies can be maintained in a practical environment. Another motivation for this study at the application level is the fact that there was an urgent need to build specific frameworks for the assessment of P2P over converged networks. This also motivated the research to illustrate the potential and applicability of a suitable testing environment and develop a test to carry out early trials of P2P applications and services over converged networks.

III. EXPERIMENTAL TESTS AND RESULTS

A. Experimental Test Bed

The test bed in an emulated converged network which has been implemented according to 3GPP specifications for both the protocol and interfaces. All the nodes run on the Linux 2.6 kernel and the user environment of Fedora Core 5- Linux. The kernel supported various protocols such as IPv6 and IPSec out-of-the-box, which makes the deployment of next generation services possible. In order to emulate the application level, we developed a P2P-SIP application which was deployed over a high performance computing infrastructure made up of 12 Intel Quadcore high end nodes with each node having 2GB of RAM, 250GB of disk space, and an external attached RAID storage of 6TB using Network
File System. A high-end infrastructure was used because it enabled the test bed to run applications that could span to over a million nodes. Parallel communication between individual nodes was based on the Sun Grid Engine (SGE), which allowed the test bed to provide a platform for distributed and parallel processing.

At the core level, an Optical Crossconnect (OXC) was integrated into the test bed. This is because the OXC could easily be partitioned to a certain amount of virtual independent switches, and the virtual switches can be configured for any network topology required. The mode of transmission of the OXC was based on Dense Wavelength Division Multiplexing (DWDM), with transmission rates of 1Gbps, 10Gbps and 40Gbps in order to accommodate large amounts of signalling that are generated at the core level. Furthermore, the Anritsu MD1230B traffic analyser [172] was used to generate SIP traffic with similar patterns to that of a real core level in a converged network.

At the access level, a Foundry FastIron Edge X424HF Layer-2/Layer-3 edge switch was deployed. The edge switch interfaces the application and core levels, both supporting a wide range of speeds at 10Mbps, 100Mbps, 1Gbps and 10Gbps, with additional support multiple Virtual LANs (VLANs), tagging queue in queue. (Q-in-Q). In order to provide mobility at the access level, a pool of wireless routers were connected to the edge switch via an impairment node in order to emulate mobile environments such as Wireless LAN (WLAN), General Packet Radio Service (GPRS) and Universal Mobile Telecommunications System (UMTS). In order to validate the test bed, over 120 tests were carried out ranging from interconnectivity tests to inter process communications. To also show that the test bed supported various implementations, P2P-SIP applications were also deployed.

B. Test and Results

1) Access Level: Mobile Virtual Private Network (MVPNs)

The aim of these tests is to validate and assess the effects of signalling at the access level by investigating the functionality, stability and handover properties of MVPNs in a multi-access network including WLAN 802.11b, GPRS and UMTS. The sample of results presented herein is focused on the performance under vertical handover conditions. On the testbed, MIPv6 was enabled on both the Home Agents (HA) and the Corresponding Nodes (CN) with routes to a web server in our controlled network. The test starts by downloading a file to the MN – at first located in the home network – from the web server. During the transfer, a handover is forced. To obtain this result, the transition between Home and Foreign Networks was performed by setting different Service Set Identifiers or SSID, a token used to identify an 802.11 WLAN network, and forcing the MN to change from one SSID to another. This procedure is known as hard handover since the IP point of attachment is terminated before the MN is moved to a new point of attachment (break-before-make). By analysing the TCPDUMP pcap file it was possible to identify the key MIPv6 packet exchanges during handover, determining their precise position in the file-transfer timeline. In order to emulate GPRS and UMTS functionalities we used an impairment node to limit both bandwidth and Round Trip Times as follows: For GPRS, we used an Upstream Bandwidth = 10kbps; Downstream Bandwidth = 40kbps; Round Trip Time = 700msec. For UMTS, we used Upstream/Downstream Bandwidth = 128kbps; Round Trip Time = 500msec.

The study highlights, some similarity between the distribution of overheads of pure WLAN/WLAN, WLAN/GPRS and WLAN/UMTS handovers, as shown in the percentage values of the overheads in Figure 2. Comparing these WLAN/UMTS results to the pure WLAN environment, the average figures are extremely close (less than 1% difference) as shown in keys 1 and 2 in Figure 2 but the difference in average handover times increases as IPSec overheads are added. On the other hand, the addition of IPSec overheads makes the difference in performance between the UMTS and GPRS environments quite significant. In the WLAN/UMTS IPSec mode, the distribution of overheads is similar to those of the WLAN/GPRS mode with differences below 5%. Vertical handover in GPRS is on average 158.1%
slower in RO and 200.5% slower in BT, but the overall average time in UMTS is much better than the GPRS results. This places vertical UMTS handover between vertical GPRS handover and horizontal WLAN handover. This was expected, since the MIPv6 control packets incurred by the MN benefit from the UMTS conditions. However, the measurements help quantifying those differences. On average next-TCP-packet accounted for 49% (UMTS), 45% (GPRS) and 61.5% (WLAN) of the total handover time respectively. The overall handover time in UMTS in tunnel mode required 84.4% less time in average than GPRS to complete. From the results it is obvious that due to level of amount of signalling generated the time for the handover process is further increased, which may not be acceptable in some applications or services in converged networks.

2) Core Level: Session Initiation Protocol (SIP) in Mobile Domains

This aim of these tests were to investigate the scalability and redundancy properties of the Back-to-Back User Agent (B2BUA) which was described earlier in Section II; as it is subjected variation of inter-domain (mobile calls to Internet generated calls) loads. Under this scenario, a call rate 900 calls per second was injected into the system; which is the call rate in which the system became saturated as observed from earlier performance based test which are not provided in this paper but can be found in [14]. This experiment aim to study how the B2BUA will recover from issues like congestions when the originating traffic is from an external domain like the Internet in which it has no control over.

In order to properly analyse the behaviour of the consumption of the local resources, the results obtained from the tests are broken down into two different graphs as shown in Figure 3. The graphs show the relationship between the CPU usage and the memory usage. From the graph, it is evident that as the number of calls increase, the per cent CPU utilization also increases. This is because the B2BUA needs to process the messages generated as it is generated. Analyzing the CPU utilization with respect to the memory utilization as, the graphs show that the memory consumption increases based on a logarithmic scale which can be as a result of the fixed sizes of the memory.

The graph also shows that the consumption of the CPU increases faster than the consumption of the memory usage, before the CPU performance starts to degrade. This is to show that at higher call rates, more of the SIP messages are actually processed and rewritten, thus generating more signalling in the system. As shown in the results, when the calls are at the peak, the CPU is congested and at this level, the memory usage becomes constant discarding all other incoming messages, which results into a time-out. Because most the messages are discarded due to the time-out, the CPU consumption is reduced even when the calls were increased to 1000 calls per second.

Based on the above results and discovered patterns, it is evident that most of the limitations to signalling at the core of converged networks, is related to delay introduced into the system due to the processing time of the CPU re-writing the SIP messages. As the number of SIP messages that transverse the B2BUA increases, the performance of the systems degrades. In order to reduce the amount of time used in processing the calls and continually re-writing the SIP messages, results from the analysis suggest that an additional component is introduced into the B2BUA. This is further discussed in Section IV.

3) Application Level: Peer Applications in Converged Networks

The aim of these tests were to study and understand the level of the impact of Peer-to-Peer (P2P) applications when deployed over converged networks. P2P applications were selected because they are known to generate quite a lot signalling even in fixed networks. The signalling generated are used to maintain the P2P overlay, while the overlay is used for routing and peer lookups at the application level. In order to study the impact of P2P application on converged networks, we subjected the overlay to the worst case scenario in terms of overlay maintenances, that is, subjecting the overlay to continuous peer join and leaves; never allowing the system to converge (stabilise). In order to emulate this type of environment, we used P2P-SIP overlays. Once the overlay was fully converged, varying churn rates were introduced into the system. The peers in the overlay were constantly forced to leave and join the overlay over a fixed period of 300 seconds. For example, for a churn rate of 40%, 20% of the peers were continually forced to leave the overlay whilst another 20% were forced to join at an incremental interval of 30 seconds until a total time of 300 seconds was reached. The time specified for each scenario was fixed at 300 seconds throughout all the experiments. All experiments for each scenario were carried out 10 times to minimise any effects of errors.

Figure 4 shows a consistent pattern for all the churn rate values used. The maximum amount of bandwidth consumed tended to peak at a particular point and then decrease rapidly. Thus, at higher churn rates, the bandwidth consumed by the system is lower. Although logically, more signalling packets are expected to be generated with higher numbers of peers, but from the graph it is evident that at a 10% churn rate for 20 nodes, causes 2122 signalling packets, which in turn,

![Figure 3: B2BUA Scalability and Redundancy Tests](image-url)
consumed a bandwidth of 40.42 Kbits/s as compared to 200 peers, which generated 4093 packets using a bandwidth of 78.78 Kbits/s.

![Figure 4: The Impact of P2P Signalling on Bandwidth](image)

Even though the amount of packets generated, and bandwidth consumed in 200 peers was far higher than that for the 10 peers, the graph also shows that the relative bandwidth was in fact dropping. This could be due to a variety of factors, of which one is based on the finger tables, which shows that with a higher number of peers in the overlay, only a small fraction of the ring is affected by any join or leave updates, which in turn leads to a reduced level of signalling. Another factor could be based on the key space or identifier chord circle in which each space in the ring allocated a fixed key. When a node leaves the ring and the same node or a new node returns, it returns to the same space in the identifier circle or key space previously vacated. In this situation, the peer’s neighbours may not have noticed that the peer had previously left the overlay for a while and therefore, no extra overhead signalling is generated to make up for this disappearance. Paradoxically perhaps, but this graph again confirms that the amount of bandwidth generated is actually reduced as churn rate increases. Based on the results and graphs from the experimental tests, it is evident that under the configuration of P2P-SIP, the protocol appears to scale well in mobile and volatile environments. There was a higher amount of signalling generated as compared to a centralised system. From the results, a network operator or service provider can calculate the cost of these signalling overheads and compare them with the cost of running and maintaining centralised resources for conventional SIP. The network operator or service provider can also take advantage of P2P-SIP systems and provide a Web 2.0 array of services in which different peers develop new services and even host these services. Although, the P2P-SIP architecture is still not fully matured. Various issues still need to be addressed, such as security (i.e. the trust-worthiness of peers). However, these issues were not the focus of this research as we more interested in how the amount of signalling generated affects the performance of the system.

IV. DISCUSSION

Based on the experimental tests in the last section, a testing-out approach has been used to investigate the effects of terminal mobility on the integration of mobile and fixed networks over a multi-access, all-IP network platform i.e. mobile converged network. The key properties of converged network analysed in this work include signalling, handovers, service provisioning, inter-domain networking, standardised interfaces and network architectures. As shown in our tests, most of these converged network properties work well for stationary nodes on fixed networks, but are not well suited to the requirements of mobile components, which are constrained in terms of processing power, memory, and battery life. In order to further analyse the signalling effects, the experimental tests carried out were broken into three broad categories i.e. access, core and application levels. These are the architectural levels defined in mobile converged networks. Each level uses a different protocol for signalling. The protocols differ from each other in terms of algorithm, design and implementation.

At the access level, the MVPN test-bed helped to assess the level of maturity of some of the access level signalling protocols MIPv6, and IPSec in relation to secure and mobile networking over multi-access networks. It also helped in the analysis of the amount of signalling that was generated at the access level. This is of special interest in mobile network management where the administrative overheads for traffic exchange between operators are always considerable. Based on the results from the experimental tests, it became evident that various shortcomings in the signalling were performance related. The most significant shortcoming arose from the fact that MIPv6 expected handover between networks to trigger with layer-3 awareness of change. This has significant (negative) impact on time-sensitive applications, which suggests that there is scope for improvement by looking at ways to trigger handover prior to the interruption of layer-3 communication.

At the core level, a performance based evaluation of the developed test-bed provided a platform for the assessment of the signalling interworking performance parameters which is one of the many limitations suffered by the core network due to the ongoing convergence of the Internet and the cellular worlds. A full assessment was also performed on these shortcomings such as loss and delay of packets due to the process of rewriting SIP messages and its effects on the end-to-end services. The experimental study represented a first step towards solving the open issues that still hamper the realization of terminal mobility in converged IP networks. The results provide a proven foundation for future research in
studying the other “bottlenecks” between the ordinary Internet and the cellular context. The interconnection of these two networks also requires the tackling issues such as end-to-end quality of service management, and charging. The major obstacle today is however at the business model level. The cellular IP world originated from the conventional network operator’s domain, whose business models are completely different from those of the Internet.

At the application level, various experiments to study the performance of P2P-SIP in a highly volatile environment were also carried out, since SIP is used for session signalling in converged networks and provides access to both fixed and mobile end-users. The results of the experiments showed that in the presence of mobile and volatile factors, P2P-SIP tended to scale quite well even though it generated more signalling than conventional SIP. The overall bandwidths per node also tended towards a download slope allowing the overlay to scale quite well.

V. CONCLUSION AND RECOMMENDATIONS

Although the main focus of this paper was around the effects of signalling on the performance mobile converged networks; this is just one of the many challenges faced by this next generation platform. A summary is given below of some recommendations for future development in the area. Although most of these recommendations were beyond the scope of this paper, due to the fact that they are mainly based on protocol design and extensions, it is believed that they will provide a tested foundation for future works on mobile converged networks. The recommendations are as follows.

• Time-sensitive applications require fast-handover. It is imperative to develop alternatives to layer-3 based trigger mechanisms. There have been various efforts in this area such as,
  • IPsec moves part of the computation load away from the network into the terminal. It will be interesting to assess to which extent this can be sustained by thin mobile terminals. This is based on the fact that running the current MVPN implementation on a thin mobile terminal can be quite resource demanding. Terminals are also expected to run other client applications and SIP at the same time. More work needs to be done in producing a lightweight (MVPN) protocol for mobile terminals.
  • Additional IPsec/MIPv6 network analysis tools are needed because at present most tools only support native IPv4, except for a few such as Ethereal and Click Router Project that have support for IPv6.
  • The development of the database, which is a distributed and redundant database that is used to store the mappings between 3GPP SIP profiles and non-3GPP SIP profiles that successfully transverse the B2BUA. Most of the design and implementation approaches to achieve this has already been provided Chapter 4.

• The P2P-SIP architecture is still not fully matured and various issues still need to be considered such as security (which peers can be trusted?) etc. However, these are issues outside the scope of this paper as we have focussed specifically on understanding its behaviour under stressful conditions.

• The development a single platform to accommodate tests and analysis of access, core and application protocols cumulatively

VI. REFERENCES