Hierarchical Security Architecture for Next Generation Mobile Networks

Fazirulhisyam Hashim[1], M. Rubaiyat Kibria[1], Damien Magoni[2] and Abbas Jamalipour[1]

[1]The University of Sydney, NSW 2006, Australia
Email: {fhisyam,rkibria,abbas}@ee.usyd.edu.au

[2]Université Louis Pasteur, 67412 Illkirch Cedex, France
Email: magoni@dpt-info.u-strasbg.fr

***Abstract*— The integration of existing and emerging technologies in Next Generation Mobile Networks (NGMN) exposes the interworked infrastructure to malicious security threats arising from individual networks and heightens the possibility of their migration across network boundaries. Owing to their autonomous characteristics, the proprietary security solutions of legacy networks cannot be extended to address such multi-faceted security threats affecting NGMN functionality. In this paper, we propose a generic hierarchical security architecture that identifies and eliminates/isolates the dominant security threats in NGMN. While the architecture utilizes an anomaly-based security detection mechanism, elimination/isolation is carried out through a cooperative approach between the node under attack and its corresponding higher tier nodes. Performance evaluation indicates that the architecture is capable of isolating threats such as denial-of-service (DoS) and worm attacks in a timely manner.**

I. Introduction

The explosive growth in mobile computing has inspired the rapid emergence of new technologies with diverse data rates and quality of service (QoS) requirements. The homogeneity of these technologies and the existing legacy networks is likely to disappear in the foreseeable future due to the introduction of Next Generation Mobile Networks (NGMN). With the aim to offer "always good connectivity" *anytime anywhere*, NGMN has emerged as a promising platform to offer inter-connectivity among disparate networks through a common framework. This framework not only provides seamless mobility across network boundaries, but also offers guaranteed QoS.

Unfortunately, such inter-connectivity has exposed NGMN (including the roaming user) to security threats stemming from individual legacy networks. In addition to these threats, NGMN is also affected by the migration of security threats across the network boundaries via infected terminals. While access-specific terminals usually incorporate network-specific security measures for homogeneous service access, no such approach is available in a heterogeneous environment like NGMN. In that sense, security has become a major research issue, perhaps the greatest obstacle to the growth and reliability of NGMN.

This work is partly supported by the the Australian Research Council. The corresponding author: [1]Fazirulhisyam Hashim {fhisyam@ee.usyd.edu.au}, is sponsored by the Malaysian Ministry of Higher Education.

In this paper we propose a generic security architecture design with a potential hostile heterogeneous environment in mind. To the best of our knowledge, no such work exists in current literature. The proposed system is structured in a hierarchical manner in order to reduce the burden of the attacked node (e.g., mobile terminal (MT), base station (BS) etc.) and to provide a fair load distribution among the NGMN network entities. This architecture considers the detection of two dominant security threats namely denial-of-service (DoS) and worm attack due to their high severity level and their potential for leading to other types of security attacks. Since in a multi-tier system (as in the proposed NGMN architecture [1]) any layer can come under malicious attacks, the proposed security architecture employs a distributive mechanism with cooperation between the node under attack and its corresponding upper-tier nodes. The architecture comprises of three security elements namely Detection Unit (DU), Decision Maker Unit (DMU) and Security Database Unit (SDU), which are responsible for detection and containment of the security threats, and database storage of the corresponding solutions (in the form of an algorithm or a software patch). Every node in the NGMN hierarchy is capable of carrying the DU and the DMU responsibilities, while the SDU is stored at the Mobility Anchor Point (MAP) (the highest node in a domain, responsible for mobility, resource and QoS management). Once a suspicious threat is detected by a network entity (representing DU) at tier n (the tier of the attacked entity), it triggers an alarm message and sends it to the tier $n+1$ network entity (the corresponding upper-tier node: DMU) to halt the attack from propagating to other peer networks. The tier $n+1$ entity then retrieves a specific solution for the attack from the SDU located at the MAP. In that sense, the proposed security architecture is lightweight and feasible, and can be applied to attacks at all tiers, subject to the availability of a higher tier node to perform the DMU functionality.

The remainder of this paper is structured as follows. The following section gives a brief overview of the interworked NGMN architecture and the security issues arising from NGMN interconnectivity. The proposed hierarchical security architecture is presented in Section III while the performance evaluation is discussed in Section IV, followed by some

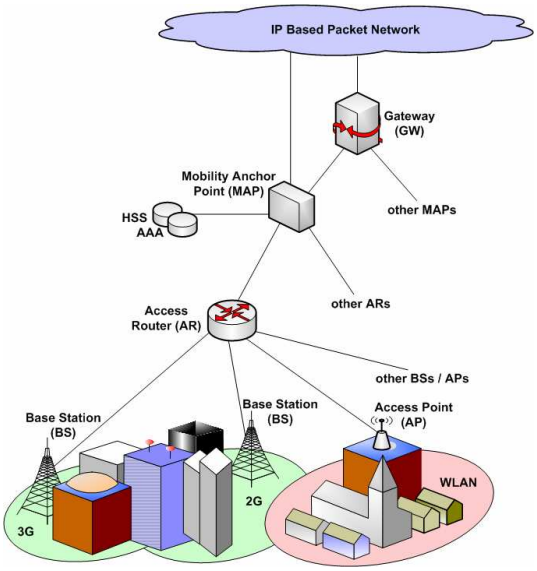


Fig. 1. NGMN hierarchical architecture.

concluding remarks.

II. Overview of NGMN Architecture

Fig. 1 shows a distributed and hierarchical NGMN architecture [1] inter-connecting disparate access technologies such as wired, cellular, Wireless Local Area Networks (WLAN), Worldwide Interoperability for Microwave Access (WiMAX), Wireless Mesh Networks (WMN) and emerging access technologies, through an IP based packet network. Besides providing a platform for the inter-connection of different radio access networks, the hierarchical structure also improves roaming capabilities of mobile users by localizing the handoff and thus reducing the signaling overhead. Interested readers may refer to [1] for a detailed description of the architecture.

Based on the NGMN architecture, it becomes evident that the success of NGMN design relies heavily on its ability to provide seamless mobility across network boundaries and service reachability *anytime anywhere* with guaranteed QoS. Any discontinuity stemming from unpredictable network dynamics (e.g., link failure, congestion and so on) and security threats from malicious nodes, can be termed as a violation of the NGMN service profile (or service level agreement (SLA)). While a mountain of literature is available for combating disruptions caused by network dynamics [2][3][4], security measures investigated so far has largely been confined to individual legacy networks and their access-specific security threats (e.g., denial-of-service (DoS), distributed DoS (DDoS), worms, viruses, and so on). Based on our literature review, we found that each network is susceptible to different kind of security threats. As an example, while most of the wired (e.g., Ethernet) and wireless networks (e.g., WLAN, WiMAX and WMN) are vulnerable to worm attack, man-in-the-middle (MITM), eavesdropping, hijack, replay and so on, the aforementioned threats are not considered as dominant threats in cellular network. With relatively good authentication and access control mechanism, it is very difficult for a malicious attacker to infiltrate the cellular system. The only practical way to breach the security profile of cellular network is via the internal network infrastructure (through rogue software) as recently reported in [5]. In addition, the vulnerability of cellular network to DoS and DDoS have been reported in [6][7]. With the advent of all-IP features the situation will become further compounded and is likely to make these attacks along with worm and virus as the dominant threats.

As per the above discussion, considering the homogeneity of current radio access technology it is reasonable for each network to implement proprietary solutions in combating their specific dominant security threats. Unfortunately, such an individualistic approach cannot mitigate the propagation of security threats across network boundaries via roaming terminals. As such, an open security architecture is required that can adequately identify and subsequently eliminate/isolate a malicious node from inflicting further damage to the NGMN. Since in NGMN each network entity can become the subject of an attack which eventually affects the whole radio network (e.g., AP or BS is under attack), a more cooperative approach is necessary between the node under attack and the corresponding higher tier nodes. The following section presents a detailed description of this cooperative mechanism.

III. The Proposed Security Architecture

A. Key Components of the Proposed Architecture

A multi-tier security architecture, as shown in Fig. 2, is proposed to provide distributed security functions, as well as correlation with the proposed hierarchical NGMN architecture. It enables the formulation of a framework where any node in the NGMN hierarchy, when under attack, either through the network infrastructure (i.e., BS/AP, AR and MAP) or directly (e.g., Personal Area Network (PAN) using Bluetooth) can be supported by a higher tier node in mitigating the threats. The architecture consists of three main security elements namely the Detection Unit (DU), Decision Maker Unit (DMU) and Security Database Unit (SDU). The DU is responsible for detecting any anomaly which is based on predefined threshold values. This implies that when the threshold is exceeded for a particular class of traffic, the traffic is defined as a suspicious flow. The DU then generates a trigger message to update the information (e.g., information regarding previous attacks, types of attacks etc.) in the upper-tier DMU. Upon receiving the trigger message, the DMU determines the area that is affected by the attacks (i.e., the location area) and updates its information. Using a client-server method, the DMU informs the MAP of the trigger event and requests for solutions related to the particular attack. This information is stored in the repository (i.e., SDU) located at the MAP, and usually available in the form of an algorithm/program/software patch, that the terminal can deploy to prevent harmful attacks from causing damages. In reply to this request, the MAP responds with the required solution for the requesting DMU. The MAP also transmits alert messages to other network elements within its domain that are most likely to be affected by the attacks.

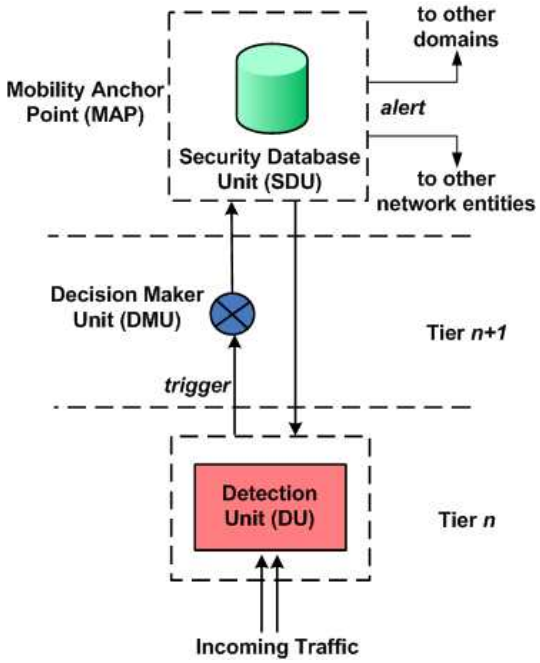


Fig. 2. General layout of NGMN security architecture

This mechanism improves the security level significantly as malicious attacks especially worms and DoS can be detected and contained at an early stage before they are propagated to other elements. A bounce diagram of the signaling process required to trigger/detect and mitigate an attack (for a MT) is shown in Fig. 3. Although the bounce diagram depicts signaling for MT under attack, it can be easily extended to other nodes in the NGMN hierarchy.

B. Working Principle of the Security Elements

The working principle of individual elements in the security architecture is described below:

1) Detection Unit (DU): The DU analyzes the incoming traffic and generates a trigger message when it exceeds the predefined threshold. In hierarchical NGMN, it is assumed that every node (i.e., MT, BS, AP and AR) is capable of handling the detection operation. Although for network entities such as BS/AP and AR, the processing capability is not a major issue, the inclusion of DU in the MT is based on the assumption that with technological advancement more powerful terminals will become available in the future. Therefore, when a MT is under attack (subject to threshold crossing which is described in the next section), the DU in the terminal generates a trigger message, attaches the attacker's IP address, and sends it to the corresponding BS (i.e., DMU in this case) which determines the type of attack. In an identical fashion, if the BS is attacked, then the trigger message is generated by the BS itself and sent to the corresponding AR (which now functions as the DMU). Note that the extension of this mechanism is limited to the hierarchical structure only whereas the core network security is separately administrated and is not considered for service access purposes.

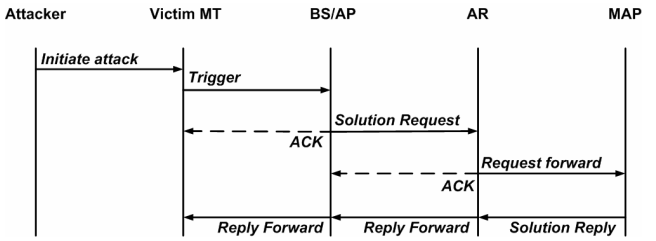


Fig. 3. Bounce diagram of the signaling transmission during the detection process

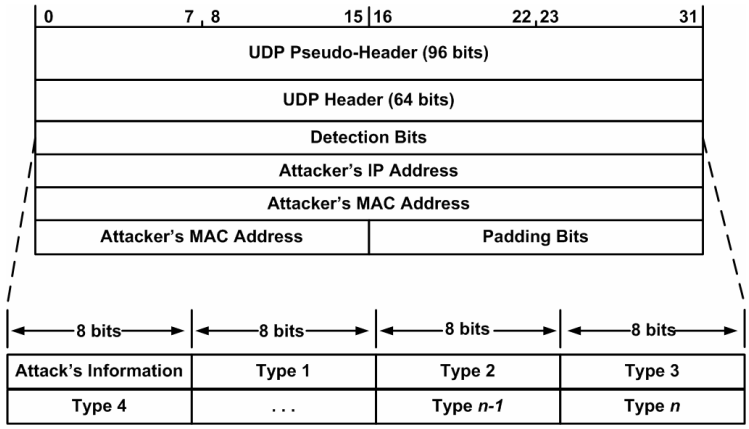
2) Decision Maker Unit (DMU): The main function of the DMU is to provide additional security measurements in terms of malicious attack containment. This unit is designed to halt the spread of worm and DoS attacks in a network domain by isolating the infected terminals and preventing them from contacting other domains. It does so by maintaining lists of suspicious IP addresses and confirmed attacker's IP addresses namely in the form of Attacker List and Attacker Blacklist respectively. When the DMU receives a trigger message from its lower-tier element (the affected node), it carries out a look-up operation in the lists to ascertain earlier instance of the attacker. For a new attack, it appends the attacker's IP address to the Attacker List. The DMU transfers the suspicious IP address into the Attacker Blacklist only if the same IP address causes message triggers several times (exceeding a predefined threshold value). Additionally a time-out period is also implemented in both the lists to prevent them from going beyond limit. The DMU also possesses a list of worm signatures [8] that have either already been detected in the domain or have been defined for the system.

3) Security Database Unit (SDU): The SDU is responsible for maintaining a database of security solutions as well as a list of worm signatures and Attacker Blacklist. Since reconfigurable MT is considered for global service access, the solutions can be regarded as access-specific. The SDU is placed at the MAP in conformance to the other repositories i.e., HSS and AAA.

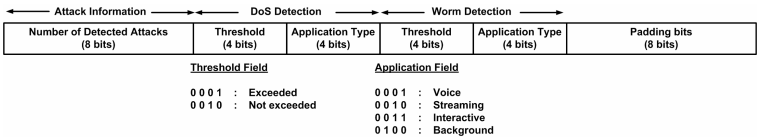
C. Anomaly Detection

Two main security threats i.e., worm and DoS are considered in this paper due to their severity and their potential for leading to other security attacks.

1) DoS Detection: DoS attacks have gained enormous attention in recent years because of the progressive sophistication and organization of the attackers. In general, there are two types of DoS attacks namely the direct attack and the reflection attack. Both of these attacks generate enormous number of User Datagram Protocol (UDP) packets, Transport Control Protocol (TCP) packets, Internet Control Message Protocol (ICMP) packets, and even data packets, and consequently send them to a large number of targeted machines. In a typical DoS attack where an attacker sends a huge amount of DoS packets (f_{dos}) to a victim node directly, the attack can be easily



(a) Packet format of trigger message.



(b) Detection bits format.

Fig. 4. General packet format of UDP-based trigger message.

detected when the number of packets surpasses the predefined threshold value (TH_{dos}). In that sense, for this scenario an attack is qualified as DoS when $f_{dos} > TH_{dos}$. Note that the threshold values are operator-defined and can be customized as per the QoS requirements. Nevertheless, since NGMN architecture enables the interconnection among different access networks, it is crucial to highlight the possibility of attack via the network infrastructure (e.g., BS/AP, AR and MAP). For example, consider a malicious MT initiating a DoS attack through its upper tier node (e.g., AP). In order to find the weakest point in the network, it scans the targeted network through the AP. In this case, the AP functions as the victim node and subsequently carries out the detection process.

2) Worm Detection: Accurate detection and quick defense are always difficult tasks for unprecedented worm attacks. For the purpose of simplicity, we exploit some common characteristics of worms in order to mitigate the attacks. Many worms utilize random scanning to search for vulnerable machines before infecting them. In addition, worms generate a substantial volume of identical or similar traffic targeting different destinations. By monitoring both intrinsic characteristics which we refer to in this paper as scanning-based detection [9] and payload-based detection [10], we can decrease the possibilities of the false positives and false negatives. In scanning-based detection, any packet destined to inactive and unused addresses is considered as a suspicious traffic flow from worms. The monitoring technique (i.e., DMU) in the security architecture collects these addresses and monitors them on the individual networks. Conversely, in payload-based detection mechanism, each suspicious traffic flow is examined by inspecting the payload of the suspicious packet. For each packet in the traffic flow, addressed to the same IP destination, the DU examines the existence of similar substring and bit sequence (referred to as worm traffic) within the payload of successive packets.

Suppose that F represents the total traffic flow through the DU and f_i is the scanning traffic destined for inactive addresses where $f_i \in F$. The suspicious flow f_s is then identified if $\sum f_i > TH_{scan}$, where TH_{scan} represents the threshold value of the scanning detection. The suspicious f_s is then examined for the same and repetitive substring $f_w \in f_s$, where the occurrence frequency of f_w is increased by one whenever the same substring appears in f_s. As flows with high occurrence frequencies of similar substrings are considered to be part of worm payloads, a trigger message is generated if $\sum f_w > TH_{string}$ which indicates the threshold crossing of the occurrence frequencies (for similar substrings). Note that, as before, the threshold values are operator-defined.

D. Signaling Packet Format

Fig. 4(a) depicts the packet format of the signaling transmission during the detection process. In order to simplify and reduce the signaling overhead, we utilize the data field of the 32-bits UDP packet. The choice of UDP is influenced

by its widespread use as a connectionless protocol in both wired and wireless networks which translates into reduced data transmission time - a key requirement of our proposed architecture. As can be seen from Fig. 4(a), the proposed frame structure is divided into multiple 8-bit fields where the first field represents the types of attacks while the subsequent fields carry distinct information about the individual attack. Additional 32 bits and 48 bits are reserved for the attacker's IP and MAC address respectively. Since we utilize the data field of UDP packet, the trigger message is therefore expandable and can be utilized to detect a number of security threats. However, in this paper we only consider two types of security attacks (i.e., worm and DOS), denoted as the Type 1 and Type 2 fields in Fig. 4(a). Both DoS detection and worm detection use a total of eight bits where four bits each are reserved for the Threshold field and the Application Type field. In principle, the Threshold field is used to determine the status of the attacked terminal (instead of the actual value), thus indicating whether the threshold value is exceeded or not while the Application Type field is used for QoS purposes. Note that the QoS is guaranteed in the NGMN. Therefore the Application Type field enables the attacked domain and the upper-tier network elements to initiate necessary measures (e.g., reduce transmission rate, reduce data coding rate etc.) so as to improve the degraded QoS resulting from security attacks. The details of the frame format, together with the detection bits for both types of attacks, are shown in Fig. 4(b).

IV. Performance Evaluation

In this section, we demonstrate the effectiveness of the proposed security architecture against the DoS and worm attacks. In order to evaluate and validate the proposed architecture, we have conducted extensive simulations using the NS2 simulator [11]. It is important to highlight that the effectiveness of our architecture is influenced by the implemented detection mechanism. Since we implement the threshold-based anomaly detection for the attacks, the results may vary depending on the selected threshold value. Moreover, since the detection mechanism relies on a counter exceeding a threshold before it issues an alert, higher value of the DoS attacking rates and larger packet size lead to faster and easier detection of the attacks. Note that due to the inherent characteristics of the NS2 simulator, a maximum achievable throughput of 700kbps at a constant bit rate (CBR) of 1Mbps is considered.

A. Simulation Model

In the DoS attack, we model a direct attack from a MT to the target AP. In this attack, the attacker generates and sends a large number of TCP SYN packets (packet size set to 512 bytes) to the target node (i.e., AP) to initiate a connection/session. This creates enormous number of half-open connections at the AP which saturate the connection table and consequently drops any connection attempt from legitimate nodes. On the other hand, worm attack is simulated as worm propagation between two MTs located at different access networks. Therefore, the worm generated packets need to propagate through the corresponding AP at the target network. Since every node in the NGMN possess the DU capabilities, we assume that the detection process will be executed at the respective AP. The simulation runtime is 250 seconds while attacks are initiated at t=100 seconds, for both DoS and worm simulations.

B. Simulation Results

Fig. 5 and Fig. 6 show the throughput of the victim node as a function of time, with and without the security architecture for DoS attack respectively. Note that the victim's throughput in both figures represents the incoming packets from legitimate node(s) (i.e., non-attacker) only. From Fig. 5, it can be seen that the throughput of the AP is heavily affected by the large amount of incoming DoS packets. However, when the security architecture is implemented into the network (refer to Fig. 6), the AP is able to detect a rapid increase of suspicious packets from the malicious MT, thus indicating a deviation from a normal traffic pattern. Given that in DoS attack, the main intention of the attacker is to halt the operation of the victim node by congesting its buffer, it is crucial for the AP to detect the attack in a short period of time. With the security architecture, once the AP detects any abnormality from the incoming traffic, it subsequently sends a trigger message to its corresponding upper tier AR and waits for the solution. It is evident from Fig. 6 that the proposed architecture is effective in detecting and defeating the DoS attack.

On the other hand, Fig. 7 and Fig. 8 illustrate the impact of the worm attack on the victim node's throughput, with and without the security architecture respectively. Similar to DoS simulation, the victim's throughput represents incoming packets from legitimate node(s) (i.e., non-attacker) only. As stated earlier, since every incoming packet from the infected MT has to travel through the AP, we consider the respective AP as the victim node (i.e., in which the DU functionality is executed). Similar to DoS attack, the throughput of the AP is also affected by the worm attack (refer to Fig. 7). However, in comparison to the rapid DoS traffic, during the initial stage of worm attack (i.e., scanning port), the infected MT scans for victims by sending scan packets at constant time interval which describes the less stormy effect of the worm attack. Since the worm attack can be detected at the early stage through the scanning port activities and through the matching of similar packet substring [8], the process of worm detection and containment can be accomplished at the scanning port stage. Therefore, as shown in Fig. 8, the worm propagation is restricted before it can send the actual worm generated packets.

V. Conclusions

In this paper, we proposed a generic security architecture for NGMN which considers anomaly-based detection technique for two dominant security threats (i.e., DoS and worm), which have been identified from our extensive literature survey. The proposed architecture facilitates cooperation between NGMN network entities in carrying out the detection and mitigation

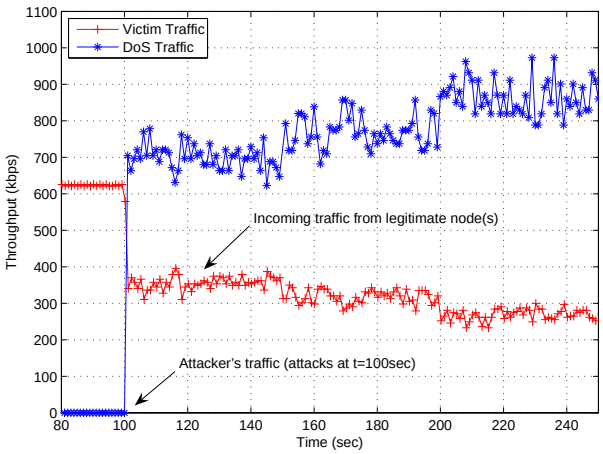


Fig. 5. Throughput at AP for DoS attack - without NGMN security architecture.

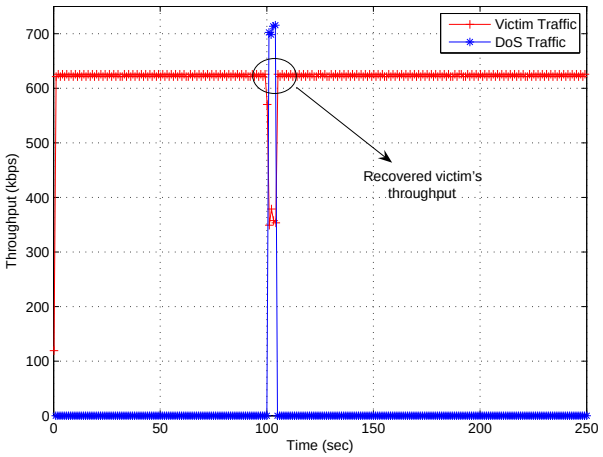


Fig. 6. Throughput at AP for DoS attack - with NGMN security architecture.

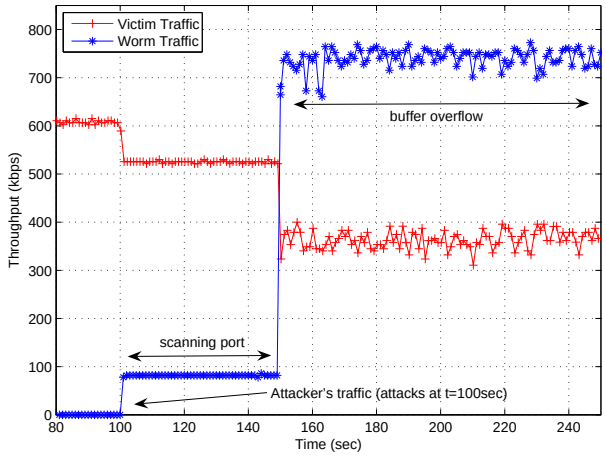


Fig. 7. Throughput at AP for worm attack - without NGMN security architecture.

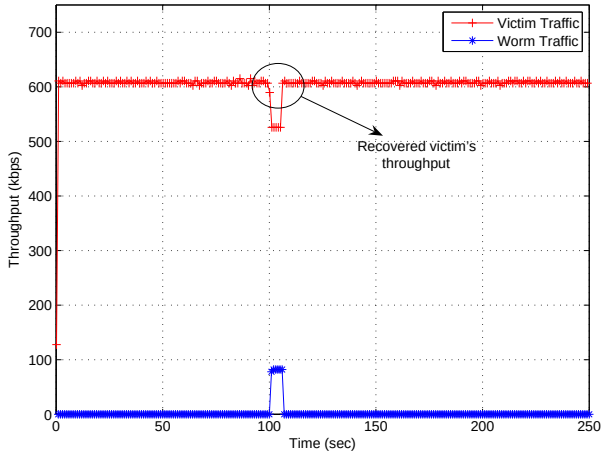


Fig. 8. Throughput at AP for worm attack - with NGMN security architecture.

of security threats. The preliminary simulation results show that the architecture is effective and capable of detecting and isolating the threats within an acceptable period of time. Given the promising results of the proposed architecture, we intend to extend the simulation to provide more comprehensive results in various situations and further adjust the proposed architecture and its parameters.

REFERENCES

[1] M. Rubaiyat Kibria and Abbas Jamalipour, "On designing issues of the next generation mobile network," IEEE Network, vol. 21, no. 1, pp. 6-13, Jan 2007.

[2] A. Calvagna, A.L. Corte, and S. Sicari, "Mobility and quality of service across heterogeneous wireless networks," Computer Networks: The International Journal of Computer and Telecommunications Networking, vol. 47, no. 2, pp. 203-217, Feb 2005.

[3] Q. Zhang, C. Guo, Z. Guo, and W. Zhu, "Efficient mobility management for vertical handoff between WWAN and WLAN," IEEE Communications Magazine, vol. 41, no. 11, pp. 102-108, Nov 2003.

[4] C. Shigang and K. Nahrstedt, "An overview of quality of service routing for next-generation high-speed networks: problems and solutions," IEEE Network, vol. 12, no. 6, pp. 64-79, Dec 1998.

[5] V. Prevelakis and D. Spinellis, "The Athens affair," IEEE Spectrum, vol. 44, no. 7, pp. 26-33, Jul 2007.

[6] S. Grech, P. Eronen, "Implications of unlicensed mobile access (UMA) for GSM security," in Proc. First International Conference on Security and Privacy for Emerging Areas in Communications Networks (SecureComm), pp. 3-12, Athens, Greece, Sep 2005.

[7] P. Traynor, W. Enck, P. McDaniel, T.L. Porta , "Mitigating attacks on open functionality in SMS-capable cellular networks," in Proc. International Conference on Mobile Computing and Networking (MobiCom), pp. 182-193, Los Angeles, USA, Sep 2006.

[8] K. Simkhada, T. Taleb, W. Yuji, A. Jamalipour, N. Kato, and Y. Nemoto, "An efficient signature-based approach for automatic detection of Internet worms over large-scale networks," in Proc. IEEE International Conference on Communications (ICC), vol. 5, pp. 2364-2369, Istanbul, Turkey, Jun 2006.

[9] S. Chen and S. Ranka, "Detecting Internet worms at early stage," IEEE Journal on Selected Areas in Communications, vol. 23, no. 10, pp. 2003-2012. Oct 2005.

[10] P. Akritidis, K. Anagnostakis, and E.P. Markatos, "Efficient content-based detection of zero-day worms," in Proc. IEEE International Conference on Communications (ICC), vol. 2, pp. 837-843, Seoul, Korea, May 2005.

[11] The Network Simulator - ns-2, available at http://www.isi.edu/nsnam/ns/, retrieved on 17/08/2007.