Policy Based SIP Signaling Management in IMS

Bao Li (bbli@alcatel-lucent.com)
Daping Wang (dapingwang@alcatel-lucent.com)
Shufang Zhang (elsiez@alcatel-lucent.com)

Alcatel-Lucent
1-6/F Sec. D Hi-Tech Software Park
Qingdao, 266101, China

Abstract

Manageability, the theme for IMS to compete with existing heterogeneous Voice over IP (VoIP) applications, runs over every network element in NGN for security and quality of experience (QoE). Further decomposition of IMS Core Network (CN) has resulted in a rapid increase of signaling traffic. At the same time, the ever rising of applications over signaling, the needs of message validation and screening, and request prioritizations are all making SIP signaling management a required capability for IMS commercialization. This paper analyzes the current problems in IMS signaling handling by setting up different application scenarios. In light of efficient and consistent services delivery, a policy based SIP signaling management mechanism is proposed, in order to meet the versatile requirements from both operation and applications perspectives. Example scenarios and deployment suggestions are also provided.

1. Challenges to IMS Signaling Stratum

Ever since the Signaling System 7 (SS7) introduced the out-of-band signaling, people have tended to think differently of the signaling and bearer network. Although the use of User-to-User Information (UUI), like Short Message Service (SMS), has demonstrated a kind of reviving of the in-bind signaling in mobile network, the limitation of message size mitigates its impact to the overall divine signaling network. When IMS takes the dominant role as the Next Generation Net-work (NGN) core for both wireless and wireline, however, a careful study of IMS signaling stratum is necessary.

The increase of signaling traffic in IMS originates from various IMS applications as well as IMS complex interfaces. A vivid characteristic of IMS is its flexibility in orchestrating various applications, be it traditional telecom based like customized applications for mobile network enhanced logic (CAMEL) or internet-originated like Google Map. Such flexibility has inevitably introduced more signaling information exchange and processing, especially when most application information is exchanged in EXtended Markup Language (XML) format. XML, though rich in its applications, clear as it is self-explanatory, is infamous of its size [1]. Compared with legacy ISUP messages, the SIP methods/headers will support more applications either coupled or de-coupled with the signaling path. The priority for each method/header is dynamic and flexible as they are associated with specific applications. How to assure the timely processing of the most needed request, especially during system congestion, is a key issue to be resolved for IMS deployment.

1.1 More Applications over Signaling in IMS

With the introduction of SMS, the signaling network is no longer dedicated to signaling itself because the real contents have been directly passed from user to user. In IMS, traffics of the SIP-based Instant Messaging (IM) and Presence applications are combined with the session control messages in the signaling stratum. In order to distinguish them from normal applications, we call them Application over Signaling (AoS), i.e., application contents are bundled with the signaling protocol in the same path.

Presence and IM, which are two typical IMS applications, are tightly coupled with the SIP signaling. Presence utilizes SIP NOTIFY method, while IM relies on the SIP INVITE or MESSAGE method to transfer the service information. The contents of these two are transferred via the SIP signaling path, in XML format. Compared with typical session related SIP messages, Presence messages are relatively large (can be up to 1M bytes).

The burgeoning of AoS challenges the IMS signaling network engineering. First, unlike the well-known Erlang traffic model, Presence and IM traffics have more unpredictable factors. For Presence, the number of...
federated presentities [2] per watcher will greatly impact the message size as well as the sending rate. Furthermore, user behavior will vary widely as we could expect a fast-moving user to generate more location update information than a busy newspaper editor at the desk [3]. Second, besides the need to safeguard the normal signaling traffic, it is necessary to provide a kind of Quality of Service (QoS) for these AoS, which should be considered as one indispensable part of Service Level Agreement (SLA).

1.2 The Need of Signaling Screening

Traditional QoS control in IMS network only handles bearer stratum resources. When everything goes in IP, however, there is a need to scrutinize signaling network too. Compared with bearer network, we always treat signaling network as a safe, reliable, and fast one. This is true in SS7 when user information can hardly be passed to the signaling network. Such restriction gets loosened when SIP rules over both user-to-network interface (UNI) and network-to-network interface (NNI). Since SIP is text-based, there is left enough space for hacks to abuse the signaling network resource for their real contents exchange. Similar to the local policy that Application Function (AF) exerts to authorize the proper bandwidth request [4], SIP signaling should be screened to authorize subscriber’s signaling request based on operator’s regulation.

Unlike the SS7 network, IMS has no explicit definition of UNI and inherits the nature of SIP to tolerate extensions. P-CSCF can perform screening on sensitive headers related to UE access, charging; but it won’t prevent IMS client from utilizing other SIP headers, tokens, and parameters for its proprietary solution. SIP protocol is lack of over-length signaling control, while Message Transfer Protocol (MTP) in SS7 has explicit requirement of 272-octet message length limit. Potential unauthorized signaling usage will increase call setup delay, and hence brings in more burdens on IMS signaling processing.

1.3 Overload Control Needs

An IMS system is said to be overloaded when it is offered more traffic than its designed capacity. In the NGN control plane, signaling traffic is the major concern of the IMS system congestion. Overload control, as a key performance index, is targeted to maximize the successful call setup rate and reduce the network resource on calls that will ultimately fail. The distributed architecture of IMS is increasing the signaling exchange among IMS network elements, and the various usages of IMS signaling, such as multimedia session, AoS, QoS, etc., are bringing in more uncertainties to the overall system signaling traffic model. Growing signaling complexity triggers more consideration on the efficiency and flexibility of the IMS system overload control.

With the evolution of UMTS CN, highly distributed architecture is complicating the signaling exchange in terms of the number of network elements and the interfaces between them, which is shown in Figure 1.

![Figure 1: Network Elements and Interfaces Increment in 3GPP](image)

Taking a multimedia session setup procedure with pre-condition and QoS operations as example, it requires 123 different kinds of signaling messages exchanged between network elements [5], including both SIP and Diameter messages. This is a dramatic increase compared with circuit-switched networks.

For a simple UE to UE call establishment within the same switch/network, without considering charging information report, there are a significant difference of the number of signaling messages used in UMTS R99 network, R4 network, and IMS network respectively (Figure 2).

![Figure 2: Number of Signaling Message for Session Setup](image)

Considering its numerous signaling entities and interfaces, the signaling traffic of IMS network could basically be grouped into three categories, each addressed by different characteristics.

- **Session related signaling**, which is responsible for establishing, updating, and terminating a session in session-based multimedia service. It is based on SIP protocol, e.g., INVITE, UPDATE, BYE, etc.
- **Non-session related signaling**, which is to authenticate subscriber and deliver non-session-
based applications. E.g. SUBSCRIBE, NOTIFY, PUBLISH for presence and availability, MESSAGE for IM, REGISTER for registration, re-registration, and de-registration, etc.

- **Supporting signaling**, which provides the critical charging report, QoS guaranty, and necessary database query functionalities. This category of signaling is primarily based on Diameter protocol, e.g. Rf/Ro interface for offline/online charging information report, Cx/Dx/Sh/Dh interface for HSS/SLF query, Rx/Gx interface for QoS control, etc.

Take into account of the complexity of the IMS signaling, special cares must be put to make sure the signaling traffic is properly routed and the right decision is made for overload control. Simply rejecting all traffic from a user performing new request attempt is a straightforward way utilized by the general overload control mechanism. Considering the heavy signaling traffic inherited from the IMS architecture and the different categories of signaling messages, such excess signaling traffic blocking might not be an ideal one.

According to different IMS application traffic models, e.g. for presence service, based on certain value of the registration per day, status changes per day, popularity factor, additions per day, and removals per day, each presentity can generate ~12 SIP messages per busy hour [6], shaping the signaling traffic appropriately based on the system policy and user preference should help manage the IMS signaling traffic in a more effective manner. For example, different user might have the same set of signaling messages with distinct priority for their specific needs:

- Under overload, system may not terminate an ongoing session, but could reject a new registration request;
- System would still accept session release request, but may not offer the QoS operations;
- System has to guarantee the correct charging information reporting, but might have the presence availability status update decreased in frequency.

### 1.4 Signaling Management Alignment among IMS Entities

As there are many entities involved in the IMS signaling path, alignment among them with a proper signaling handling is necessary to assure the end to end signaling transfer without impacting service, especially for the AoS. For example, presence AS may send SIP message encapsulated with large XML data chunks while PCSCF could reject the request due to its own handing limitation of message size. Such limitation will hence block presence service being successfully delivered even though AS is capable to handle it. It is suggested to keep the signaling management information into a centralized place accessible by IMS CN entities, to simplify the alignment work.

### 2. IMS Signaling Management

#### 2.1 State-of-Art of SIP Signaling Management

Two response codes are defined in RFC 3261 [7] -- 503 Service Unavailable and 513 Message Too Large. There is no detailed requirement of 513 response code usage.

A few internet drafts have been issued to address the problems raised by the 503 response code. As John Rosenberg summarized in “draft-rosenberg-sipping-overload-reqs-02” [8], the 503 response code defined in RFC 3261 will cause load amplification, underutilization, ambiguous usage and on/off retry-after problems. This is why Internet-draft draft-hilt-sipping-overload-01 [9] proposes several new overload control mechanisms and defines a new “Load” header and the 520 response code.

At the same time, RFC 4412 [10] defines 2 new SIP headers, “Resource-Priority” and “Accept-Resource-Priority”, to allow a request originator to indicate to a SIP element that it wishes the request to invoke resource prioritization, which can be utilized for emergency services support. RFC 4411 [11] also defines the reason header to indicate the network preemption event or access preemption event.

#### 2.2 Signaling Management Infrastructure

In order to resolve the challenge discussed in previous sections, we are proposing a signaling management infrastructure as in figure 3. Similar to policy Resource Admission and Control (RAC) [12] on bearer stratum, IMS signaling management adopts policy based approach to satisfy the dynamic, flexible and versatile signaling management requirements.

All the policies regarding to signaling management are stored in a centralized database that are accessed by Policy Engine (PE). There are two types policy engine - network based PE (NBPE) as the default one for all subscribers and subscriber based PE (SBPE) that enables individual preference setting.

PE is the core element in the signaling management, which is in charge of generating the policy information from the database, collecting the signaling usage status report from those enforcement elements, and resolving the conflict to align among different signaling handling policies. Here we define five policy types for signaling management.
1. Message validation, which should be applied on the UNI (P-CSCF) to minimize the impact of invalid messages to the IMS CN. Such kind of policy information may include invalid SIP message handling, the allowed maximum length associated with message type, treatment applied when received message length exceeded, such as returning the 513 response code, or discarding less important SIP headers, etc.

2. Signaling screening, which is to be applied on the UNI (P-CSCF) and NNI boundaries (IBCF). It includes the signaling screening criteria, screening treatment, and schedule, etc. Such signaling screening policy drives enforcement entities to screen the unauthorized SIP signaling.

3. Request prioritization, which is to be applied to the entities directly involved with request handling and service interaction, i.e., S-CSCF and AS. Request prioritization policy may include request priority list (request type, target URI, etc.), lifetime, and resource reservation mechanism. Therefore, IMS CN entities can be utilized for the most needed services, especially when the system is under or close to be overloaded.

4. Overload control, which is preferred to all the IMS signaling entities. It takes care of the overload related treatment, from overload threshold with different congestion level to proper traffic reduction mechanism selection upon receiving of the notification from those overloaded nodes, etc.

5. Regulatory service support, which helps to provide the equivalent regulatory service in IMS as PSTN does. Emergency call is one typical regulatory service that defines the rule of service detection, and resource reservation mechanism.

Policy enforcement may be implemented across all IMS CN entities. Based on network operator’s preference, policy enforcement entity should execute some or all of the signaling management functionalities. These entities communicate with corresponding PEs to retrieve policy information, install, enforce them, and provide result report per requests from the policy engine.

To satisfy the needs from individual subscribers, SBPE is introduced to define individual screening, message validation and request prioritization policy, along with network based ones.

3. Example Scenarios
In this section, we will cover two scenarios for the usage of signaling management

3.1 Request Prioritization
We select IMS based SMS voting as an example because of its popularity in China. For example, one entertainment program that is held in 2007 has set a record -- high SMS traffic of 2 million voting SMS for favorite singer during a 15-minute interval.

There are two key entities in the IMS based SMS voting -- S-CSCF to contact the SMS GW AS and SMS GW AS to interwork with legacy SMS center. MESSAGE method is used for IMS users’ voting, which has higher priority than other requests during the voting window. To assure the timely response to MESSAGE request, NBPE assigns a high priority on MESSAGE to the S-CSCF and SMS GW AS.
- Life time: voting window
- Top priority request: MESSAGE with request URI as the SMS GW AS PSI;

With the signaling management policy enforced, SMS voting request will take the highest priority among requests, putting aside the regulatory services. During the voting window, S-CSCF will first handle SMS voting request, and may delay other requests processing. This will, to the best capacity of S-CSCF/AS, guarantee the timely processing the SMS voting request according to the request prioritization policy.

Figure 4: Request Prioritization to support SMS Voting

3.2 Signaling Screening

It’s expected that some users could explore the usage of X-SMS header to convey the SMS content in SIP INVITE request. PE can request enforcement entity (e.g., P-CSCF) to probe the SIP message. Based on the report analysis, i.e., the finding of proprietary parameter, PE will modify network based and subscriber based policy (if needed). It may store the information into the databases, and install the new policy into policy enforcement entities finally. The SBPE may also be invoked to generate/modify the subscriber specific policy based on the newly generated network-based policy.

In order to apply proper policies to X-SMS header handling, we can put subscribers into 3 categories:
- Golden subscriber: allowed to use the X-SMS header for SMS.
- Silver subscriber: X-SMS header usage is not allowed; request is not rejected but X-SMS header will be removed;
- Ordinary subscriber: X-SMS header usage is not allowed and requests with such header will be rejected, following network-based policy.

Figure 5 showed the request and report on X-SMS detection, and the follow up policy generation. Figure 6 showed the policy enforcement result.

Figure 5: Policy Dynamic Generation based on Probe Result

4. Deployment Suggestions

As discussed before, policy based signaling management should be applied to meet the versatile signaling management requirements. Here we list some deployment suggestions with the hope to efficiently manage signaling network.

1. It’s suggested to carefully define network-based signaling management policy, the default one and the only policy eligible to be executed under overload condition
2. Centralized policy database is helpful to align the signaling management between CSCF and AS, to assure no service impact introduced with considering singling handling
3. It is suggested to deploy the related signaling management capability on the edge entities in order to minimize impact to CN, such as signaling screening is preferred in P-CSCF (for UNI) and IBCF (for NNI) and message validation is preferred in P-CSCF (for UNI)
4. It is not suggested to deploy over-length control between IMS core network entities. Connection oriented protocol, such as TCP/SCTP, is relatively
suitable for large message transfer, and it is expected that TCP/SCTP connections between limited CN entities will only consume a limited number of socket resources.

5. Regulatory service will be deployed based on a network-based policy with no subscriber exception.

6. Considering the performance impact, request prioritization should be deployed on the S-CSCF and AS that are directly involved with the service invocation and session setup.

7. Special consideration is needed on SIP signaling traffic engineering with the newly defined management mechanism, which may consume additional system resource to perform the needed signaling deep inspections.

5. Summary

With the maturity and deployment of IMS, more challenges are revealed regarding to the SIP signaling usage. More applications are tight-coupled with the SIP signaling, especially using XML for application data encapsulation. This will change the signaling traffic model and hence dramatically increases the number of SIP signaling messages together with their sizes. Lack of signaling screening will put the system under great risk considering client intelligence and holes in SIP extensibility. Although there are a few draft RFCs to address the weakness of overload control with the SIP protocol, the complexity of IMS solution has put more requirements on it. Both operator and subscriber require a flexible, dynamic and versatile signaling management mechanism, which is beyond the SIP protocol’s capability.

In this paper, a policy based signaling management infrastructure is proposed to effectively manage the IMS SIP signaling and efficiently utilize precise IMS CN entities. PE is responsible to generate and modify the policy information via policy database and signaling usage report from policy enforcement entities. NBPE provides the default signaling management in the network, while SBPE served to provide subscriber specific signaling management policy. Using request and report mechanism, PE can understand the signaling network usage and modify the policy according to the real time status of the network.

Signaling management may consume more system resource when it performs deep inspection on SIP signaling compared with normal session setup, but it will provide more manageability on the signaling stratum. Therefore, special consideration on signaling traffic engineering is needed to balance these two factors. Quantified analysis for signaling traffic with certain IMS traffic model will be the focus of future study.

6. Acknowledgements

We would like to thank Igor Faynberg to lead and encourage us to participate the ICIN 2007 conference. Many thanks to Huilan Lu for her extensive discussion of this idea, and detailed review and contribution.

7. References:


[5] 3GPP TS 24.228: “Signaling flows for the IP multimedia call control based on Session Initiation Protocol (SIP) and Session Description Protocol (SDP); Stage 3”


[8] Internet Draft: “draft-rosenberg-sipping-overload-reqs-02”

[9] Internet Draft: “draft-hilt-sipping-overload-01”


