

Service Initiation Procedure with On-demand UE Registration for Scalable IMS Services

Y. Kitatsuji, Y. Noishiki, M. Itou and H. Yokota

KDDI R&D Laboratories, Inc.
2-1-15, Ohara, Fujimino, Saitama, 356-8502, Japan,
{kitaji, yujin, mn-itou, yokota}@kddilabs.jp

ABSTRACT

The IMS-based service management is a key for emerging ALL-IP-based mobile network services. The session initiation protocol (SIP) employed for the IMS supposes that a user equipment (UE) is constantly registered with a call session control function (CSCF). This obliges the mobile network operators (MNOs) to have a large number of CSCF nodes to maintain the SIP transaction when a large number of UEs are maintained.

In this paper, we propose a modification to the session initiation procedure of IMS to include an on-demand UE registration. This allows the UE not to be registered until beginning their services, and which results in the MNO having a smaller IMS (maintaining less CSCF nodes than the standard service provisioning). We discuss the requirements to fit the modification to 3GPP standards. The evaluation with packet-based simulation experiments reveals that the proposal can be superior for the workload of CSCF nodes to the standard procedure and reduce the required nodes up to 40% from the case of the standard procedure.

Keywords: IMS, SIP, CSCF, and 3GPP

1 INTRODUCTION

The third generation partnership project (3GPP) [1] has specified the IP multimedia subsystem (IMS) that enables user equipments (UEs) to setup, modify, and teardown multimedia sessions in managed networks [2]. The UEs and the IMS employ session initiation protocol (SIP) [3] as their signaling control. The IMS is a key for most emerging mobile network operators (MNO) to migrate to all IP-based network from the conventional networks.

SIP is a client/server application protocol, where client applications send SIP request messages to maintain multimedia sessions. Server applications reply with one or more responses for each SIP request. RFC3261 [3] specifies an intermediary node called a SIP proxy which helps to process and route SIP messages from/to SIP endpoints (UEs and service-specific servers). A SIP transaction represents a request and its corresponding response exchanged between two adjacent SIP nodes (SIP endpoints and proxies).

The IMS architecture enables the MNOs to accommodate a large number of users with a SIP proxy farm. The IMS has a mechanism to balance the workload of the SIP transactions over a number of SIP proxies. Besides load-balancing

[4] [5], reducing the workload [6] is one of important tasks that the MNOs pursue. The success of this task can make the SIP proxy farm smaller and probably result in reducing the operational cost.

The workload of the UE registrations to SIP proxies is a relatively large portion of the entire workload in the SIP proxy farm. The IMS services maintained by the SIP is on the basis of a preliminary registration of the end-user terminal to the SIP proxy. The UE registration enables IMS to control sessions of interactive applications with two or more UEs. Furthermore, IMS architecture specifies that the UEs to be registered with session call control function (CSCF) itself periodically. This feature imposes a large workload of SIP transactions, as described in Section 2.2.

This paper introduces and evaluates the modification to the session initiation procedure. Although the modification can remove the periodic UE registration, IMS can still maintain interactive applications. A key to the modification is to include UE registrations in the session initiation procedure. This adds to the IMS an interaction with the mobile core network which is a transport network managing the network resources, the UE attachment, and mobility. The evaluation shows that the modification has no major impact on the UE and service initiation, and allows a reduction in the number of proxies developed on the SIP proxy farm.

The rest of the paper is organized as follows. Section 2 illustrates the IMS architecture, and describes scalable issues with the burden of the periodic UE registration. Section 3 clarifies the difference of our approach from the previous studies. Section 4 describes the modification and the requirement to change the current 3GPP standard. Section 5 analyzes the impact on the workload of the SIP proxies in the IMS and reveals that some proxies reduce their workload although the workload of other proxies increases. Section 5 demonstrates that the required SIP proxies are largely reduced. Section 6 summarizes the paper and concludes with future tasks.

2 WORKLOAD OF SESSION MANAGEMENT IN IMS

2.1 IMS Architecture

All IP-based mobile networks have two function sets: the IMS as the service control stratum, and the mobile core network as the network transport stratum. Figure 1 illustrates the simple architecture of IMS which is composed of call session control function (CSCF) maintaining the service control for

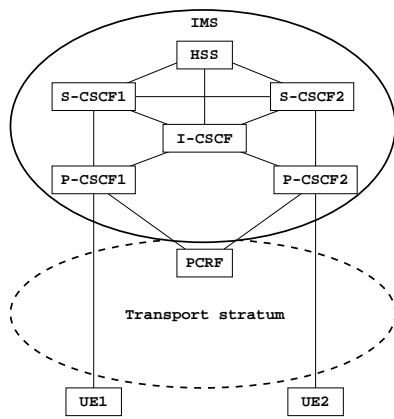


Figure 1: Architecture of IMS.

UEs, and a home subscriber server (HSS) maintaining user subscription information.

There are three types of CSCF: Proxy-CSCF (P-CSCF), Serving-CSCF (S-CSCF) and Interrogating-CSCF (I-CSCF). The P-CSCF routes the SIP requests between the UE and the other S/I-CSCF and establishes the security association with the UE. The P-CSCF is also an entry point between the IMS and the mobile core network. The I-CSCF assigns an S-CSCF node in receiving the first SIP registration request from the P-CSCF, and routes the second registration request between the P-CSCF and the S-CSCF. The I-CSCF is also a gateway from the other network operators. The S-CSCF performs the session control services for the UE, e.g., UE authorization, forwarding the SIP request to the other MNO, routing the SIP request to the application servers, and notifying the UE to change the registration.

The policy and charging rules function (PCRF) is a gateway from the IMS to the mobile core network. The PCRF receives, from the P-CSCF, requests for the preparation of the network resource for the user data coming from/going to the UE.

2.2 Load Balancing in IMS

Figure 2 shows the UE registration procedure. The UE is registered with the IMS twice during the first registration. Through the first request, I-CSCF assigns an S-CSCF node with which the UE is registered in the latter request, and an algorithm for the integrity protection is shared between the UE and the P-CSCF for establishing a security association. The second request allows the UE to be authorized and registered to the assigned S-CSCF. The UE and P-CSCF apply the integrity protection, e.g., IPsec [7], from this request.

The registration procedure is conducted periodically (e.g., every 30 minutes). In the periodic registration the registration procedure is conducted once because the UE and P-CSCF continue to maintain the security association.

In order for the IMS to maintain a large number of UEs, each of P-CSCF, I-CSCF and S-CSCF is composed of multiple nodes and the IMS has the CSCF node assignment mechanisms among them in the registration procedure. A P-CSCF is discovered, e.g., via procedures with DHCP and DNS-based

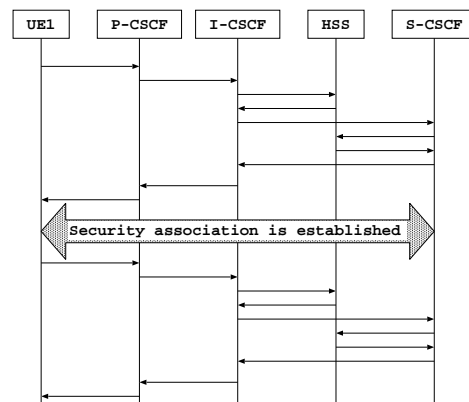


Figure 2: User registration procedure.

load-balancing, as described in Section 4.2, when the UE is attached to the mobile core network. The I-CSCF discovery is also on the bases of the DNS-based load balancing at the selected P-CSCF. The I-CSCF queries the HSS to find the S-CSCF every time for the UE, and this enables the I-CSCF nodes in the first and second registrations to be different. The S-CSCF assignment is performed with capability information for the operator preference, the user subscription, the availability of S-CSCF, topological information (e.g., based on P-CSCF's IP address), and the other information. Once the UE is registered with the S-CSCF, the P-CSCF memorizes the S-CSCF for the UE. Therefore, the P-CSCF routes the request for the other procedure, e.g., service initiation, to the S-CSCF directly.

The IMS supposes that all the UEs are registered to the S-CSCF and P-CSCF because the SIP is employed in the IMS. Note that this consumes a large amount of computing and memory resources of SIP nodes accommodating a large number of UEs in order to maintain the periodic registration procedure. E.g., P-CSCF and S-CSCF nodes accommodating 500,000 UEs handle about 280 registration procedures per second on average for the UE registering every 30 minutes. In the case of VoIP services provided by the IMS the call arrival rate reaches 170 calls per second during on-peak periods as described in the next subsection.

2.3 Treatment of Busy Hour Call Attempts

The telecom data book in Japan [8] says that the annual average rate of call arrivals from a single user terminal is about 1.5 times per day. The call arrival rate increases to 5 and 20 times during on-peak periods in a day and a year, respectively, and results in 43 and 173 calls per second, respectively.

Most MNOs have a call acceptance rate of 20% through 70% to limit arrival calls during on-peak periods. This limit allows the MNOs to reduce the maintained CSCF nodes to some level from the maximum level of all calls attempted by users. In the case of 50 % of the call acceptance rate, the MNO handles the sessions in total, including the UE registration and service initiation of up to 366 procedures per second during on-peak periods. Although the UE registration

procedures create generally lighter workload for CSCF nodes compared to the service initiation one, the periodic UE registration still represents a large share of the workload in the CSCF nodes

In this paper, we modify the session initiation procedure to include the UE registration and to reduce CSCF nodes involved in the session initiation procedure. This modification reduces the entire workload of CSCF nodes.

3 RELATED WORK

The performance and scalability of a SIP infrastructure have been the subject of several studies. Existing studies are grouped into the performance (call throughput) improvement of SIP proxies, and the workload controls [4] [5]. Furthermore, studies of performance improvement are grouped into the entire SIP proxy farm and a single SIP proxy. Our study is related to the former study.

The studies evaluating the impact of SIP state management, transport protocol and authentication on the workload [9] [10] are categorized into the performance improvement for the SIP proxy farm. Their findings show that the SIP proxy performance varies greatly depending on the SIP proxy configuration, and that authentication has the greatest impact across the various configurations. Dacosta *et al* also improved the entire performance of SIP proxies in the node configuration where SIP proxies were distributed whereas the database servers were centralized [11]. They analyzed the relationship between the latency of message delivery and call throughput, and indicated that the request batch mechanism could reduce the required bandwidth at the database.

Although our study is grouped into this, none of the previous studies has taken our approach. The performance improvement of the SIP proxy farm is achieved by mitigating the bottleneck in handling a large number of signaling call flows, or reducing the overhead of signaling call flows. Our study reduces the number of sessions maintained by the SIP proxies in the IMS by modifying the session initiation procedure to include the on-demand registration.

Our proposal modifies a small part of the session initiation procedure, of the mobile core network, and of the UE implementation. However, this does not include another signaling interface, nor does it bring load balancers or load monitors in the IMS and the mobile core network. We presume that these features are of great importance for the MNOs that migrate their infrastructure to include our modification.

Most studies of performance improvement for a single SIP proxy are basically combined with our proposal. The influence of parsing, string processing, memory allocation, and thread overhead on overall capacity was evaluated and optimized by Coltes *et al*. [12]. Furthermore, Janak proposed parsing the limited portion in the message and the assignment of the different parsed portion to each of the SIP proxies improved call throughput in the single SIP proxy [6]. Since our proposal changes the signaling call flow of the session initiation, combining their proposal with ours requires an analysis

for the optimization. However, we still presume that optimization for the performance improvement is viable.

4 SESSION INITIATION WITH ON-DEMAND REGISTRATION

The IMS includes interactions between itself and the mobile core network for network resource preparation and charging procedures. Our proposal adds another interaction initiating UE registration through the mobile core network. Although this increases the additional interactions, this can dramatically reduce the maintained inactive session.

The following subsections explain the modifications to session initiation and the requirements.

4.1 Modification to Session Initiation Procedure

Figure 3 shows the 3GPP-standardized session initiation procedure for the VoIP application. Session termination after the call is not included in this procedure. The thick arrows are the messages relayed via P-CSCF1, S-CSCF1, S-CSCF2, and P-CSCF2. Although IMS can provide various services with different session managements, this paper employs signaling call flows for the VoIP application. This is because the signaling call flow is relatively complex, and all types of CSCF are included.

Through the message exchange from the INVITE to the second 200 OK messages, IMS enables UE1 to meet UE2, and to determine the employed CODEC and then later queries network resource authorization for the media traffic beginning between the UEs. Next, UE2 notifies UE1 that UE2 begins to ring through the messages from 180 Ringing to the third 200 OK. Last, UE2 notifies UE1 that the user of UE2 has received the call, and P-CSCF1/2 request the gate open in order to establish the communication path for the media traffic by opening the packet filter in the gateways of the mobile core network.

Figure 4 shows the proposed modification to relay INVITE message. The feature of the modification is that UE registration is included in the INVITE request, and that a single S-CSCF is involved in the session between the UEs. Hereafter the modification is termed the on-demand-registration session initiation procedure (OSIP).

The followings are the requirements for the modification.

- **UE registration:** A caller UE (UE1) is first registered with the IMS before beginning IMS-based services (A in Figure 4).
- **S-CSCF assignment by HSS:** I-CSCF notifies S-CSCF1 as the candidate S-CSCF for the following UE2 registration, when the HSS is queried S-CSCF for the UE2 (B in Figure 4). HSS replies to S-CSCF1 with which UE1 is registered to I-CSCF in registering UE2 (D in Figure 4).

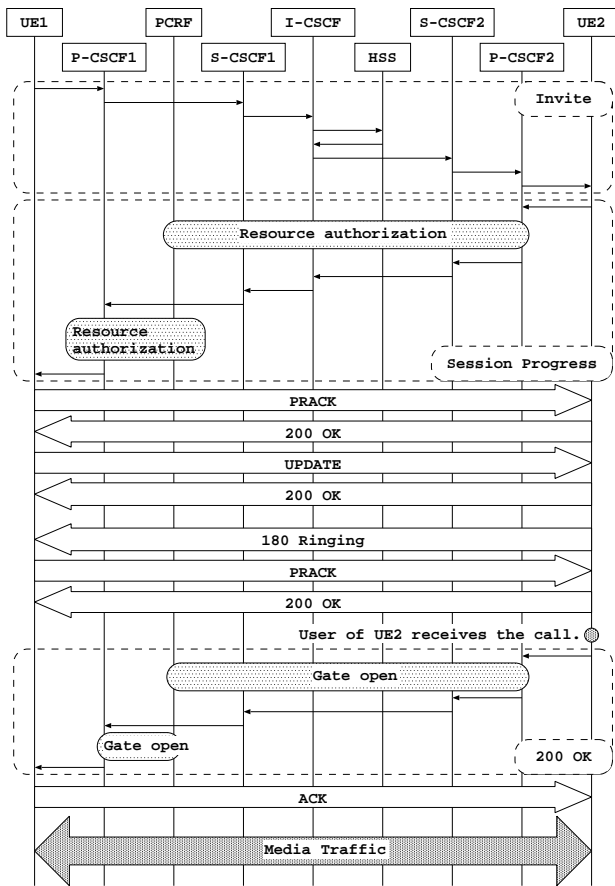


Figure 3: IMS service session initiation procedure.

- **Registration request:** the I-CSCF requests that the mobile core network (PCRF) call a callee (UE2) to be registered with IMS when the I-CSCF does not find that UE2 registered with any S-CSCF (C1 – C3 in Figure 4).
- **Network-initiated UE registration:** UE2 registers with the IMS from the transport stratum request (not specified in Figure 4);
- **IMS-initiated UE deregistration:** this is specified in the 3GPP architecture, that is, the S-CSCF initiates deregistration of the UE from the IMS (not specified in Figure 4).

The following subsections describe how to modify the standards to realize the corresponding requirements above.

4.2 Registration Before Service Initiation

In the standardized procedure, the UE obtains P-CSCF hostname statically or in the DHCP procedure when the UE is attached to the mobile core networks. DNS name resolution from the hostname enables the workload to be balanced over multiple P-CSCF nodes.

In the OSIP, the UE is not registered with the IMS when the UE is attached to the network. Instead, it begins the registration just before the UE begins the VoIP application. This may

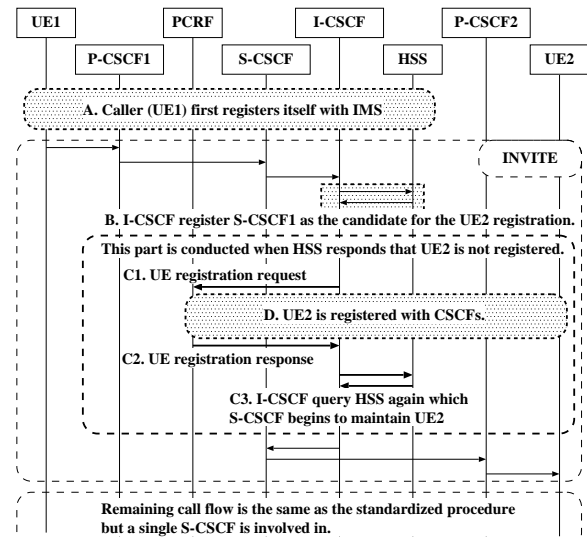


Figure 4: Proposed modification to the IMS service session initiation procedure.

impede the service initiation procedure taking longer time. The influence is evaluated in Section 5.

4.3 Registration Request from I-CSCF to PCRF

The I-CSCF requests the PCRF to let UE2 be registered with the IMS (S-CSCF node) when UE2 is not registered. The I-CSCF often encounters this case when the S-CSCF requests that the I-CSCF finds the other S-CSCF with which UE2 is registered. This is the main difference from the standard procedure. If the I-CSCF finds that UE2 has already been registered with an S-CSCF node, remaining half of the INVITE procedure progresses.

The identities employed in the IMS and mobile core network are usually different in the 3GPP standard. The IMS employs the public user identity [13], e.g., the universal resource indicator (URI) printed on a business card, whereas the transport stratum employs the international mobile subscriber identifier (IMSI) [14]. Because these identities are stored in the HSS, the HSS is required to respond with the IMSI to I-CSCF when UE2 is not registered in the S-CSCF.

When the PCRF responds that UE2 has been registered with the IMS, the I-CSCF queries which S-CSCF begins to maintain UE2 to the HSS again. This is because the assignment of S-CSCF node for UE2 may differ from S-CSCF1.

4.4 Network-Initiated UE Registration

This requires a communication interface between the mobile core network and the UE. In the 3GPP architecture [1], there are interfaces between the PCRF and the access gateway (Serving-GW), between the Serving-GW and the mobility management entity (MME), and between the MME and the UE. Furthermore, the standard has the service request procedure conducted over these interfaces, which makes the

UE recover the wireless segment resource released in the idle mode.

The modification includes the additional framework carrying the service/application specific parameter in the service request procedure. The UE is also modified that it examines which service/application is requested to initiate as well as preparing the wireless segment resource when the UE receives a service request from the network.

UE2 may begin the registration procedure just before the request from the network-initiated UE registration arrives. In this case, the UE waits for the response for the already-begun procedure and responds by reporting success to the mobile core network.

4.5 S-CSCF Assignment by HSS

The proposed modification enables the HSS to give the I-CSCF the S-CSCF node (S-CSCF1) with which UE1 is registered, as the candidate of UE2 registration. For this, I-CSCF in B notify S-CSCF1, and the HSS memorizes it for following UE2 registration. The HSS has a timer to release memorized S-CSCF candidate for UE2 to meet the failure of the UE2 registration.

HSS can respond S-CSCF2 if UE2 had already been registered with S-CSCF2 and the registration has been remained by the following call attempt. In this case, HSS does not memorize S-CSCF1.

4.6 IMS-Initiated UE deregistration

The IMS-initiated UE deregistration is specified in the 3GPP architecture. A key is that the timer for each UE is maintained by the S-CSCF. It is recommended that the timer is much shorter than the standard re-registration timer, although it depends on the operational policy. The duration of the timeout also influences the workload of the S-CSCF and P-CSCF nodes accommodating a large number of UEs.

4.7 Effectiveness of Modification

The main difference between the standard and OSIP is when the UE registration is performed. The OSIP includes two registrations (for caller and callee UEs) every call in the worst case. On the other hand, the UE following the standard procedure has periodic registrations.

A rough estimate is that the entire workload in IMS caused by the OSIP is lower than that from the standard procedure during off-peak periods. This is because the number of the attempted registrations for the OSIP is less than that for the standard procedure. However, on-peak periods have the opposite results.

For example, a UE attempts the registration 48 times a day in the standard procedure in the case of a registration interval of 30 minutes. If the periodic registration has a shorter (longer) period in an MNO, they see a much lower (higher) workload in the CSCF nodes during off-peak periods when OSIP is adopted. When the UE employing the OSIP makes

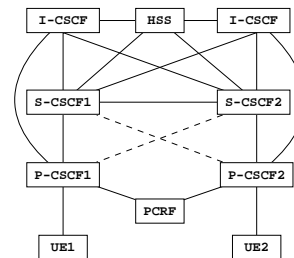


Figure 5: CSCF node composition in the experiments.

a phone call 10 times a day, it performs 20 registrations at maximum in total.

The OSIP can also reduce the workload by limiting the single S-CSCF node involved in the session initiation procedure. However, the workload is increased when I-CSCF is much more involved in the registration request at the same time. The evaluation in the next section verifies the effectiveness of the OSIP, although there is a tradeoff in the workload between the S-CSCF and I-CSCF.

5 EVALUATION

The effectiveness of the modified session initiation procedure is verified with the packet-based simulator. First, we reveal the impact of the modification on the P/S/I-CSCF nodes with regard to two aspects: one is how much longer the modification makes the procedure; and the other is the queuing delay of the requests stacked in the CSCF nodes as the CSCF node workload. Second, the required CSCF nodes are estimated for the standard and OSIP to demonstrate effectiveness.

5.1 Model of Experiments

The two-set CSCF/UE model as specified in Figure 5 was employed in the experiments. UE1 and UE2 represent the sets of UEs accommodated to P-CSCF1 and P-CSCF2, respectively. The experiment had a scenario where each set of UEs made calls to the other set of UEs. The standard session initiation procedure took two S-CSCF nodes (the solid lines between S-CSCF and P-CSCF), whereas the OSIP took one of S-CSCF nodes with which the caller UE was registered (the solid and dashed lines between S-CSCF and P-CSCF).

Table 1 shows the employed parameters for the time to process a message in the nodes, the propagation delay of links, and the other elapsed time in the signaling call flow. The processing time in receiving a single signaling message, and the propagation delay of the message delivery were basically united to 200 microseconds. The short propagation delays represent that the CSCF and HSS database servers are placed together as the centralized structure of the IMS. We supposed that the HSS had a short processing time to represent that the HSS had a sufficient processing performance (responded quickly).

The PCRF had a relatively long processing time (uniform random numbers) because the request is sent to the transport

Table 1: Parameters employed in experiments

Parameters		Duration
Process time (microsecond)	UE	200
	P/S/I-CSCF	200
	HSS	10
Propagation delay (microsecond)	UE/P-CSCF	5,000
	other links	200
Others (second)	PCRF	0.005-0.015
	Ringling	1-5
	Call duration	120

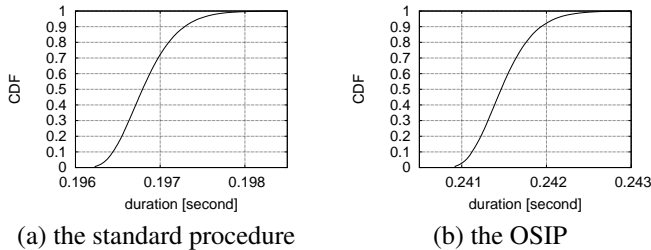


Figure 6: Cumulative Distribution Function of durations for the session initiation procedure.

stratum including multiple nodes and the relatively long propagation delay in the wireless segments. This includes the interactions with the PCRF including requesting the UE to begin the SIP registration, preparing the network resource, and opening the packet filters for the media traffic. The duration of ringing was uniform random numbers between 1 and 5 seconds. The call duration followed the exponential distribution with 120 seconds as the average [8].

5.2 Duration of the Session Initiation Procedure

Figure 6 shows the cumulative distribution functions of the durations for (a) the standard and (b) the OSIP session initiation procedures, respectively. To represent a on-peak period, the offered calls were 20 times larger than the one-year average call arrival rate (1.5 calls per user a day). The call arrivals followed the Poisson process for 100,000 UEs, which was about one-tenth of a single SIP proxy capacity of the current product [15].

The figures show that the durations have about 45-millisecond difference. The result does not include the PCRF interaction estimated as 5 milliseconds. Therefore, the actual difference is about 50 milliseconds. This difference comes from eight additional messages sent between the UE and P-CSCF (four messages for each of UE1 and UE2 registrations as specified in Figure 2), and durations for message deliveries.

The difference can be larger if the interaction in the mobile core network is larger. However, the difference is still relatively small compared to the duration of ringing until the

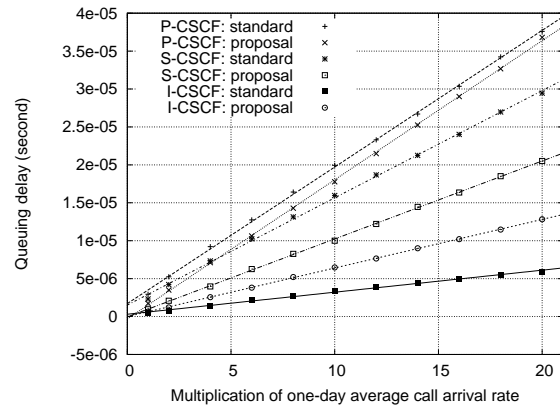


Figure 7: Mean of queuing delays.

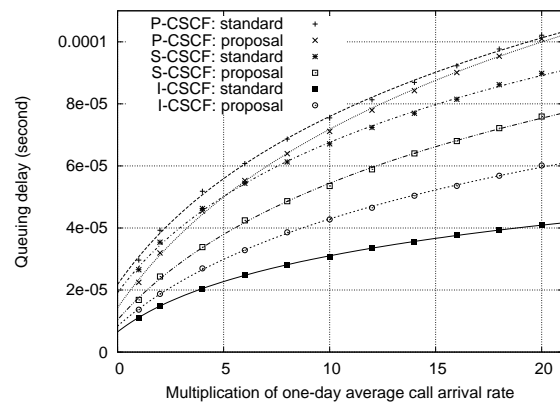


Figure 8: Standard deviation of queuing delays.

callee receives the call. Furthermore, the INVITE message takes longer in the mobile network when the callee UE is paged because the network resource in the wireless segment is idle. Comparing these network features, the increased duration brought by the OSIP is sufficiently small.

Comparing two distributions, the results indicate that there is no significant impact (workload increased) on the CSCF nodes. This is assumed from the fact that each of distributions has almost the same difference (2 milliseconds) between the minimum and the maximum durations. This indication is also verified in the following subsections.

5.3 Workload of Session Initiation Procedures

Figures 7 and 8 show the mean and standard deviation of the queuing delay, respectively, for each of P/S/I-CSCF nodes.

The queuing delays were collected from the latter half of a two-hour simulation during which the offered calls continued to arrive. The queuing delay represents what a SIP message sees when it arrives at a CSCF node. The x-axis presents the offered calls labeled with the multiplication of the one-year average call arrival rate (1.5 calls per user a day). Table 2 shows the call arrival rates and average numbers of active calls. The points in the figure are the obtained results from the

Table 2: Offered calls in experiments

Multiplication	2	4	6	8	10
Arrival rate	1.74	3.47	5.21	6.94	8.68
Active calls	208	416	624	833	1042

simulations, whereas the lines are approximations from the method of least squares (the employed function is described in Section 5.4).

The results show that the OSIP has a shorter queuing delay than the standard procedure for the P-CSCF and S-CSCF, whereas the I-CSCF has the opposite results. For the mean of the queuing delay, the difference in the queuing delay for P-CSCF decreases as the call arrival rate increases, whereas that for S/I-CSCF increases. For the standard deviation of the queuing delay, the difference in the queuing delay of P-CSCF almost disappears at 20 times larger than the one-year average. In the both results the increase of the call arrival results in the increase of the UE registration, and therefore, P-CSCF workload of the OSIP increases.

Furthermore, we verified these tendencies for the larger number of UEs accommodated in the CSCF nodes from 200,000, 300,000 and 400,000. We also found that the increasing and decreasing ratio of the queuing delay was almost the same over these numbers of UEs. The results probably imply that full capacity case, that is, a million UEs accommodated in a CSCF node, also has almost the same result.

In the following subsection, we demonstrate, with the approximated queuing delay in Figures 7 and 8, that the OSIP can reduce the required nodes when aiming at a certain level of queuing delay or lower.

5.4 Estimation of Required CSCF Nodes

In the network operation of the IMS, the limitations of the call acceptance rate are defined for each CSCF node in general. Although there are several derivations of the limitation, e.g., CPU utilization, signaling response time, or signaling throughput, we employ the signaling response time based on the queuing delay. In this case, the call acceptance rate derived from the queuing delay boundary: $\mu(x) + 3\sigma(x)$ is one of the candidates, where $\mu(x)$ and $\sigma(x)$ denote the mean and standard deviation of the approximated queuing delay, which are obtained from the multiplication of one-year average, x . This section discusses with this queuing delay boundary.

The approximations in the previous section were obtained from the linear and logarithm-based functions:

$$\mu(x) = ax + b \text{ and } \sigma(x) = c \log(x + d) + e, \quad (1)$$

for the mean and standard deviation of queuing delays, respectively. Table 3 shows the parameters obtained from the method of least squares.

From these approximations, the required number of CSCF nodes with the lower target queuing delay is estimated. For example, the following derivation estimates the number of P-CSCF nodes with the standard session initiation procedure,

Table 3: Offered calls in experiments

CSCF node	Mean		Stdev		
	a	b	c	d	e
P-CSCF standard	1.77	2.12	49.8	5.08×10^6	-59.1
proposal	1.86	-0.71	57.4	5.79×10^6	-86.5
S-CSCF standard	1.37	2.14	40.9	4.41×10^6	-41.5
proposal	1.03	-1.31	41.9	5.39×10^6	-60.3
I-CSCF standard	2.69	6.31	17.6	3.31×10^6	-14.6
proposal	6.40	7.51	33.7	5.60×10^6	-49.7

unit: 10^{-6}

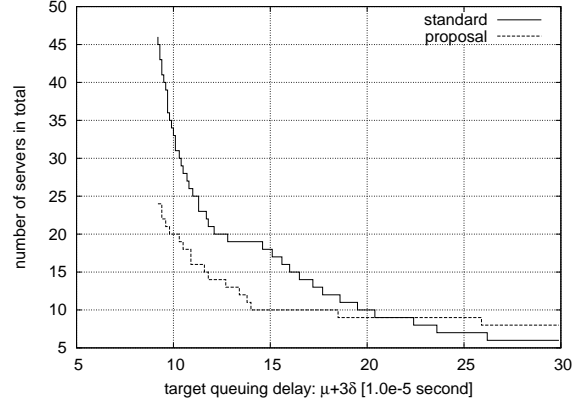


Figure 9: Estimated number of total CSCF nodes.

which accept 20 times more calls than the one-year average at maximum, and which satisfy 1.5×10^{-5} for the mean of queuing delay or lower,

$$n_{(\text{mean}, 1.5 \times 10^{-5})} \geq \lceil 20 / \mu_{p, \text{std}}^{-1}(1.5 \times 10^{-5}) \rceil = 3 \quad (2)$$

This estimation is based on the fact that the arrival calls can be equally balanced over the multiple P-CSCF nodes and that the call arrival rate can be divided by the number of nodes.

Based on this calculation, the estimated minimum number of CSCF nodes is shown in Figure 9. The results were computed by searching the minimum number of CSCF nodes, each of which satisfies the target queuing delay. There are multiple combinations of the P/S/I-CSCF nodes for the target queuing delay because the target queuing can be composed of various combinations of the mean and standard deviation. The figure shows the minimum number from the various estimation, which is derived from min-max of the numbers of nodes n_{mean} and n_{stdev} ,

$$\sum_{i \in \{\text{P,S,I-CSCF}\}} \min_{0 \leq q_m \leq q_t} (\max(n_{(\text{mean}, q_m)}^{(i)}, n_{(\text{stdev}, \{q_t - q_m\}/3)}^{(i)})), \quad (3)$$

where q_t denotes the target queuing delay.

The results indicate that the OSIP can reduce the required nodes comparing to the standard procedure. When the target queuing delay is 1.4×10^{-5} , the node reduction reaches to more

than 40 % (to 11 from 19). P/S/I-CSCF nodes are 5, 3 and 2 for numbers from 10, 8 and 1, respectively, for accepting 20 times more call arrivals than the one-year average.

This outcome is for specific conditions: the registration period is limited to 30 minutes; the message process times are equal among CSCF nodes and among the message types; and the offered call is one-tenth of the state-of-the-art node capacity. However, the feature of the OSIP that reduces the workload of P/S-CSCF nodes which generally have a larger workload than I-CSCF nodes allows the OSIP to decrease the workload at some level. Thus, this enables the MNOs that maintain a large number of UEs with multiple CSCF nodes, to reduce the number of nodes by adopting the OSIP.

6 CONCLUDING REMARKS

This paper introduced a novel approach to include the UE registration in the IMS session initiation procedure. We presented the UE registration and the session initiation procedures in the IMS and discussed the fact that the periodical UE registration consumed a large amount of computing and memory resource. Our approach lets UEs going to have an active session, e.g., a VoIP session, be registered with the CSCF nodes, and reduce the workload of the entire CSCF nodes.

This effectiveness was shown with simulation experiments and its analysis. The experiments showed the modification to the session initiation procedure taking a longer time for completion of the procedure. However, this extension was still small compared to paging the callee UE in the mobile network, and ringing by the callee user receiving the call. The distribution of the procedure completion time during on-peak periods indicated that the modification provided no major impact on the session initiation procedure.

The queuing delays in each of the P/S/I-CSCF nodes maintaining a large number of users were analyzed. The results revealed that the workload of the P/S-CSCF nodes decreased, while that of I-CSCF nodes increased. This is because the modification includes two UE registrations, where the I-CSCF is involved to assign an S-CSCF node. Although such a trade-off was observed, the demonstration comparing the procedures showed that the number of required CSCF nodes could be reduced. This reduction was as high as 40 %.

The applicability of the proposed modification to the other session control procedure, and the influence of the difference of the processing time for each of SIP messages remain as future work.

REFERENCES

- [1] General Packet Radio Service (GPRS) Enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) Access, Technical Specification Group Services and System Aspects, TS 23.401, Rel. 8, Ver. 8.5.0, (2009).
- [2] IP Multimedia Subsystem (IMS); Stage 2, 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects, TS 23.228, Rel. 8, Ver. 8.8.0, (2009)
- [3] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, E. Schooler, SIP: Session Initiation Protocol, IETF RFC-3261, (2002).
- [4] C. Shen, H. Schulzrinne, and E. Nahum, Session Initiation Protocol (SIP) Server Overload Control: Design and Evaluation, Proceedings of Conference on Principles, Systems and Applications of IP Telecommunications, pp. 149–173 (2008).
- [5] K. Singh and H. Schulzrinne, Failover and load sharing in SIP telephony, International Symposium on Performance Evaluation of Computer and Telecommunication Systems(SPECTS) (2005).
- [6] J. Janak. SIP Proxy Server Effectiveness, Masters Thesis, Department of Computer Science, Czech Technical University (2003)
- [7] S. Frankel, R. Glenn and S. Kelly, The AES-CBC Cipher Algorithm and Its Use with IPsec, IETF RFC-3602, (2003).
- [8] Telecom Data Book 2008, Telecommunications Carriers Association in Japan, <http://www.tca.or.jp/databook/index.html>
- [9] S. Salsano, L. Veltri, and D. Papalilo, SIP security issues: the SIP Authentication Procedure and its Processing Load, Network, IEEE, Vol. 16, No. 6, pp. 38–44 (2002).
- [10] E. M. Nahum, J. Tracey, and C. P. Wright, Evaluating SIP Server Performance, Proceedings of the 2007 ACM SIGMETRICS international conference on Measurement and modeling of computer systems, pp. 349–350 (2007).
- [11] I. Dacosta, V. Balasubramaniyan, M. Ahamad and P. Traynor, Improving Authentication Performance of Distributed SIP Proxies, Proceedings of Conference on Principles, Systems and Applications of IP Telecommunications (IPTComm'09), (2009).
- [12] M. Cortes, J. R. Ensor, and J. O. Esteban, On SIP Performance, Bell Labs Technical Journal, Vol. 9, No. 3, pp. 155–172 (2004).
- [13] Service Requirements for the IP Multimedia Core Network Subsystem, 3GPP Technical Specification Group Services and System Aspects, TS 22.228, Rel. 8, Ver. 8.5.0, (2009)
- [14] The International Identification plan for Mobile Terminals and Mobile Users, ITU-T, E.212 (1998)
- [15] H. Jiang, A. Iyengar, E. Nahum, W. Segmuller, A. Tantawi, and C. P. Wright, Load Balancing for SIP Server Clusters, Proceedings of the IEEE INFOCOM 2009, pp. 2286–2294 (2009).