行政院國家科學委員會專題研究計畫 成果報告

下一代行動網路多媒體資訊服務的服務品質保證之研究
研究成果報告(精簡版)

計 畫 類 別 ：個別型
計 畫 編 號 ：NSC 96-2416-H-004-019-
執 行 期 間 ：96 年 08 月 01 日至 97 年 07 月 31 日
執 行 單 位 ：國立政治大學企業管理學系

計 畫 主 持 人 ：郭更生

計畫參與人員：碩士班研究生-兼任助理人員：古昌平
碩士班研究生-兼任助理人員：羅嘉彥
碩士班研究生-兼任助理人員：黃世民
碩士班研究生-兼任助理人員：游詩怡

處 理 方 式 ：本計畫可公開查詢

中 華 民 國 97 年 12 月 17 日

行政院國家科學委員會專題研究計畫成果報告

下一代行動網路多媒體資訊服務的服務品質保證之研究

Research on QoS Guarantee of Multimedia-based Information Services on Next-Generation Mobile Networks

計畫編號：NSC 96-2416-H-004-019

執行期限：96 年 8 月 1 日至 97 年 7 月 31 日

主持人：郭更生　執行機構及單位名稱：國立政治大學

一、中文摘要

服務品質保證是下一代行動網路多媒體資訊服務的核心問題，不能容忍一定的資料丟失，對資料的傳輸延遲和延時抖動有嚴格的要求。同時，行動網路的傳輸鏈路由於無線介質的特性以及行動終端的頻繁移動而表現出極大的不穩定性，這些因素對行動網路多媒體資訊服務的服務品質保證提出了很高的要求，也是電信運營商持續不斷地擴大行動用戶資訊服務需求的關鍵。服務品質問題的解決不是頻寬建設單一方面的事情，有了寬頻網路基礎設施，僅僅具備了多媒體資訊服務服務品質問題的前提條件，還需要行動網路具備多層次、高效率的服務品質保證技術。下一代行動網路將是整合多種寬頻無線和有線接入技術、統一提供整合資訊服務的開放融合的網路架構，多種現存和新興的有線或寬頻無線接入技術將共存於下一代行動網路，例如 WLAN、WPAN、WMAN 及 2G/3G/4G 行動通信系統等。因此，對下一代行動網路中多媒體資訊服務服務品質保證機制與相關機制進行研究，有著極為重要的必要性。

關鍵詞：多媒體資訊服務、服務品質保證、下一代行動網路。

Abstract

The multimedia information service is the critical topic of next-generation mobile networks; the QoS guarantee is the central issue of the multimedia information service in next-generation mobile networks. The main purpose of this research project is to explore the relationship between the QoS guarantee of multimedia information service and the optimal mechanism of resource allocation. The optimal mechanism can be designed for achieving the required QoS guarantee of multimedia information services on the next-generation mobile networks.

Keywords: QoS guarantee, multimedia information service, resource allocation mechanism, next-generation mobile networks.

二、研究成果

■已出版

Thomas M. Chen, Geng-Sheng (G.S.) Kuo, Zheng-Ping Li, and Guomei Zhu, "Intrusion Detection in Wireless Mesh Networks," in *Security in Wireless Mesh Networks*, edited by Yan Zhang, Jun Zheng, and Honglin Hu, Auerbach Publications, Feb. 2008.

李爭平與郭更生, "基於 802.11 無線網狀網的准動態通道分配演算法," accepted by 电子与信息学报. 此期刊屬於 EI.

李爭平與郭更生, 2008, 4, "無線 Mesh 網路中基於自相似流的通道分配演算法," 高技术通讯, Apr. 2008. 此期刊屬於 EI.

Xuelian Long and Geng-Sheng (G.S.) Kuo, 2008, 5, "A Novel Dynamic Fuzzy Analysis Hierarchy Model Enabling Context-aware Service Selection in IMS for Future Next-Generation Networks," *Proc. of 2008-Spring IEEE Vehicular Technology Conference (VTC 2008-Spring)*, in Marina Bay, Singapore, on May 11 – 14, 2008. 此論文集屬於 EI.

Zhongbin Qin and Geng-Sheng (G.S.) Kuo, 2008, 1, "Performance Optimization for

Uplink Transmission in IEEE 802.16e BWA Networks," *Proc. of 2008 IEEE Consumer Communications & Networking Conference (CCNC 2008)*, in Las Vegas, Nevada, U.S.A., on January 10-12, 2008. 此論文集屬於 EI.

Xing-Jian Zhu and Geng-Sheng (G.S.) Kuo, Invited Paper, 2008, 1, "A Cross-Layer Routing Scheme for Multi-channel Multi-hop Wireless Mesh Networks," *Proc. of the Second Workshop on Broadband Wireless Access (BWA 2008)*, in Las Vegas, Nevada, U.S.A., on January 10-12, 2008. 此論文集屬於 EI.

Lifeng Le and Geng-Sheng (G.S.) Kuo, 2008, 1, "A Novel P2P Approach to S-CSCF Assignment in IMS," *Proc. of 2008 IEEE Consumer Communications & Networking Conference (CCNC 2008)*, in Las Vegas, Nevada, U.S.A., on January 10-12, 2008. 此論文集屬於 EI.

Zhongbin Qin and Geng-Sheng (G.S.) Kuo, Invited Paper, 2007, 9, "Cross-Layer Design for QoS-Oriented Resource Allocation with Fairness Provision in IEEE 802.16 OFDMA Networks," *Proc. of the First Workshop on Broadband Wireless Access (BWA 2007)*, in Cardiff, Wales, UK, on Sep. 13, 2007. 此論文集屬於 EI.

■已投稿

Tian Wu and Geng-Sheng (G.S.) Kuo, "Seamless Integration of IMS and Peer-to-Peer Technologies for Unified Service Provisioning Platform in Next-Generation Networks," submitted to *IEEE Communications Magazine*, under first revision.

Yahui Hu and Geng-Sheng (G.S.) Kuo, "Space-frequency Subchannel Allocation and Adaptive Modulation in SDMA MIMO OFDM Beamforming Systems with Limited Feedback," submitted to *IEEE Transactions on Communications*.

Jie Zhang and Geng-Sheng (G.S.) Kuo, "An IMS-based Novel Service Discovery Architecture for Next-Generation Networks," submitted to *IEEE Transactions on Networking*.

三、結論

本研究的研究成果已在國際上有多篇論文發表，已有三篇論文投稿國際傑出期刊，現正在審查中，等待最後結果。另進行兩篇論文撰寫，將投稿國際傑出期刊。就個人自己評估，已達到計畫申請書的預期水準。現將一篇自認結果很好、正在審查中的期刊論文附上，供參考。

An IMS-based Novel Service Discovery Architecture for Next-Generation Networks

Journal:	*IEEE/ACM Transactions on Networking*
Manuscript ID:	TNET-00390-2008
Manuscript Type:	Original Article
Date Submitted by the Author:	19-Oct-2008
Complete List of Authors:	Kuo, Geng-Sheng (G.S.); National Chengchi University, National Chengchi University
Keywords:	Next-Generation Networks (NGNs), IP Multimedia Subsystem (IMS), performance optimization, service discovery architecture (SDA)

An IMS-based Novel Service Discovery Architecture for Next-Generation Networks

Jie Zhang and Geng-Sheng (G.S.) Kuo

***Abstract*—IP Multimedia Subsystem (IMS) standardized by the Third Generation Partnership Project (3GPP) realizes the convergence of fixed and mobile networks. In virtue of providing various types of multimedia services among different kinds of access networks, IMS is considered as the core network for Next-Generation Networks (NGNs). For detecting the requested services quickly and efficiently, it is necessary for IMS to adopt a valid service discovery function. In this paper, we propose an IMS-based novel service discovery architecture (SDA) to provide a universal service discovery function independent of any specific network access technology. It is supported by the existing IMS-related entities and functionalities, and is well integrated in IMS-based networks. We analyze the performance of the SDA on considering the average update interval of service announcement (SA), the mean interval between two adjacent service requests, and the network condition variability. All the performance analyses are conducted by using mathematical modeling, theoretical analyses, and numerical simulations. Furthermore, we find an optimal value for the average update interval of SA, which enables the performance to reach the trade-off. And, the value is correspondingly stationary, and can be set and modified by the network operator.**

***Index Terms*—Next-Generation Networks (NGNs), IP Multimedia Subsystem (IMS), performance optimization, service discovery architecture (SDA).**

I. Introduction

IP Multimedia Subsystem (IMS), converging the fixed and mobile networks together, is specified to provide various multimedia services based on diverse network access technologies [1] – [2]. IMS is migrating to All-IP-based Next-Generation Networks (NGNs), and is considered to be the core network for future NGNs [3] – [5]. For end-users, one of the most important requirements is getting their desired services easily, quickly, and successfully. Accordingly, as services become more numerous, various and complicated, it is more urgent and of great significance for IMS-based networks to provide an adapted and valid service discovery (SD) function.

Manuscript received October 19, 2008. This work of the second author was supported by the National Science Council of Taiwan under Grants NSC 96-2416-H-004-019 and NSC 97-2410-H-004-116-MY3.

Jie Zhang is with the Telecommunications Engineering School, Beijing University of Posts and Telecommunications, 100876 Beijing, China.

Geng-Sheng (G.S.) Kuo is with National Chengchi University, Taipei, 116 Taiwan (e-mail: gskuo@ieee.org).

However, up to now, there has been no relevant specification proposed, defined or described by the Third Generation Partnership Project (3GPP) yet. Moreover, the studies of SD in all these years have been mainly emphasized on some specific environments, such as mobile ad-hoc networks [6] – [8], web services frameworks and grid systems [9] – [10], and peer-to-peer overlay networks [11] – [12]. Besides, very few papers [13] – [15] were focused on theoretical analyses.

To the best of our knowledge, there is only one paper that proposed a SD system coined location-based SD system [16] using the components of IMS to realize the SD function, but no analytical model describing SD architecture (SDA) based on IMS has been published. Comparing with the concepts that the services using heterogeneous network technologies are available and treated as the IMS services by mapping methods and middleware mechanisms [17] – [18], in this paper, we propose an IMS-based novel SDA to provide a universal SD (USD) function covering the whole and global IMS-based network without adding any new functional entities. In our SDA, Home Subscriber Servers (HSSs) are used as the directories, User Equipments (UEs) are considered as the agents of customers, Application Servers (ASs) are recognized as the service providers, and Subscription Locator Function (SLF) is used for searching HSSs. Here, we model HSS as a queuing system to evaluate the cache size for service announcements (SAs). The success rate of SD and the relevant traffic load generated by SD are also studied. Considering the impact of the average update interval of SA, the performance optimizations are made.

The main contributions of this paper are concreted as follows. First, because our SD function only works in the IP Multimedia Core Network (IM CN) Subsystem, our new SDA can coexist with any existing SD technology which is the solution for certain specific access network. Secondly, the new SDA is based on the existing entities and functionalities of IMS, and integrated well in the IMS-based network. There is no any new entity needed, therefore its cost is much lower. More important, our SDA provides a global user approach independent of any specific network access; it is a universal solution of SD. Furthermore, we find an optimal value for the average update interval of SA, which makes the performance reach the trade-off, and can be assigned and controlled by the network operator.

The rest of this paper is organized as follows. Section II proposes our IMS-based novel SDA. In Section III, analytical

models and theoretical analyses are described. The performance analyses and optimizations are conducted in Section IV. Finally, conclusions are made.

II. IMS-Based Novel Service Discovery Architecture

Generally, there are two (decentralized SD) or three (centralized SD) components in SDA: a service requestor, a service provider, and sometimes a directory. The directory is a cache used for aggregating information from different service providers. According to the characteristics of IMS, we adopt the centralized SDA based on SIP.

As the directory, HSS is used. It saves service information coming from ASs and communicates with UEs to complete SD procedure. The service information includes location information (having the accessible addresses and ports), security information (having both authentication and authorization information), service profile information, and the S-CSCF (Serving-Call Session Control Function) allocated to the service. As sending service information to HSS, AS is recognized as the service information provider. Also, UE is considered as the service requestor. It finds desired service information via sending Service Requests (SrvRqts) to and receiving Service Replies (SrvRplys) from HSS. Considering the whole IMS-based network, i.e., multiple HSSs, SLF is used for searching neighbor HSSs in order to get more service information for UE.

Here, we introduce a concept named HSS Domain (HSSD). In general, there is more than one HSS in the whole IMS-based network, because there are too many UEs and ASs that need to communicate with HSS and one HSS does not have the capability to deal with all. A HSSD is used to express a domain in which a single logical HSS (LHSS) is in charge of some UEs and ASs. The whole IMS-based network usually comprises a number of HSSDs. A single LHSS is a logical entity which consists of N physical HSSs. These physical HSSs are redundantly collocated in order to avoid single point of failure and to improve reliability. For the sake of searching services in neighbor HSSDs, we also use a logical SLF (LSLF) which includes some physical SLFs redundantly collocated. The following discussions are based on both LHSS and LSLF.

We consider the storage states of service information in LHSS. The service information consists of the SA and its lifetime. When a new service appears, the relevant AS sends its service information by a Service Registration message to the LHSS in the same HSSD, and periodically updates its service information to let LHSS know the service variation in time. When a SA is stored in LHSS, it will be in one of the four states: 1) keeping alive in its lifetime, 2) being discarded when its lifetime is expired, 3) being updated in its lifetime due to the service change, or 4) being re-registered by sending a Re-Registration message before its lifetime is expired. So, a SA keeps valid by being updated or Re-Registered in its lifetime. These conditions are depicted in Fig. 1.

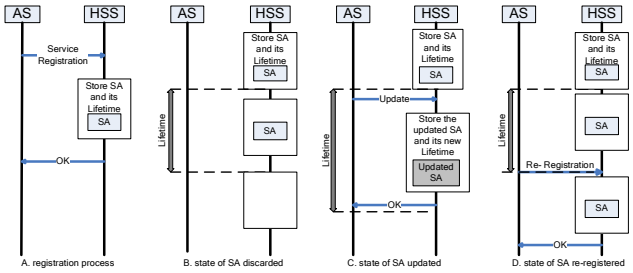


Fig. 1. States of SA in HSS.

The proposed novel SDA is illustrated in Fig. 2. The procedure of SD in UE's own HSSD is described as steps 1-2-3-A-B-4-5-6 and that in a neighbor HSSD is depicted as steps 1-2-3-A-a-b-A'-B'-4-5-6 (including the process of getting a new LHSS address from LSLF). The principle of searching service information is as follows. First, UE looks for its desired service in its own HSSD. If there is no suitable service in its HSSD, it is possible that UE goes on with further searches in its neighbor HSSDs. Whether searching in the neighbor HSSDs and the times of searching are decided by UE depending on the relevant customer's requirements. The searching order and scope of the neighbor HSSDs are determined by LSLF according to the network topology of LHSSs.

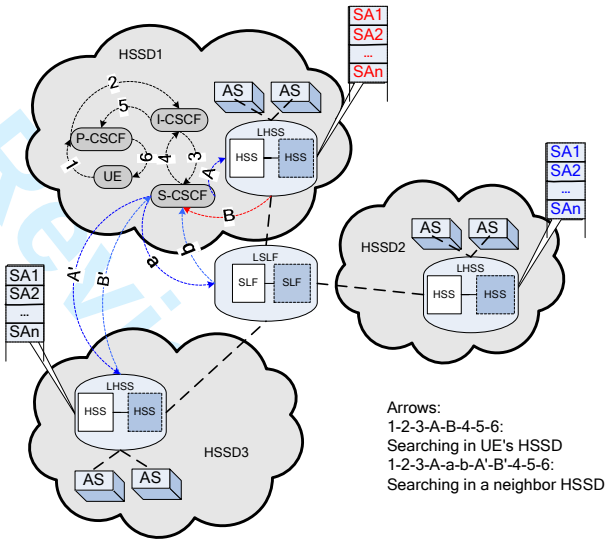


Fig. 2. The proposed IMS-based SDA.

The details of SD procedure are expatiated as follows. Here, we only study the steps starting from UE, but those from the customer to UE are not included. Also, the details between UE and Proxy-CSCF (P-CSCF) in the IP connectivity access network are not considered. First, UE searches its desired service information in its own HSSD. A SrvRqt is sent by UE then is forwarded to S-CSCF via P-CSCF and Interrogation-CSCF (I-CSCF). According to the address information of LHSS in its caches, S-CSCF finds and contacts the LHSS in the same HSSD. After receiving the SrvRqt, LHSS searches the matching service information in its database and replies with a SrvRply in case of finding the

suitable service information. Then, the service information is sent back from LHSS to UE in the SrvRply message. After receiving the SrvRply, if there is more than one service fitting the SrvRqt, UE will select the best one. We set an expiration time to each SrvRqt. If there is no SrvRply back to UE during this time interval, UE will reckon that there is no matching service in its HSSD. In this condition, UE may start one or several new searches in its neighbor HSSDs.

The steps of searching in a neighbor HSSD are described as follows. UE sends the same SrvRqt to LSLF through P-CSCF, I-CSCF, and S-CSCF, but set a new expiration time. After receiving the SrvRqt, LSLF chooses the nearest neighbor HSSD based on the network topology and responds to S-CSCF with the corresponding LHSS address information. Then, S-CSCF contacts the new LHSS. After searching in the new LHSS, a new SrvRply is replied to UE (having suitable service) or no response (having no matching service). If there is no matching service in the nearest neighbor HSSD, i.e., UE can not find suitable service the second time, UE may ask LSLF again in the same manner as in its nearest neighbor HSSD to search service in another neighbor HSSD. The process may repeat several times. The times of search and the expiration time of each search are determined by UE according to relevant service requirements including the service priority, the real-time demand, and so on.

Obviously, our SDA is solely based on IMS and only uses the existing entities and functionalities in the IM CN Subsystem. The advantage is that our SDA is well integrated with IMS and can realize the USD function among different access networks based on IMS without adding any new entities or functionalities.

III. Mathematical Modeling and Theoretical Analyses

In this section, we analyze the performance of our SDA through mathematical modeling and theoretical analyses. We model LHSS as an M/G/c/c queuing system, and evaluate the performance in terms of cache size in LHSS for SAs, success rate of SD, and the relevant traffic load caused by SD procedure. The impacts of the average update interval of SA, the mean interval between two adjacent SrvRqts, and the variation of network conditions are all considered.

A. Cache Size in LHSS for SAs

As described in Section II, in each HSSD, SAs coming from ASs are stored in the database of LHSS and discarded when there is no update or re-registration within their lifetimes. That is, each LHSS stores and deletes SAs according to their lifetimes.

First, we study the state in one HSSD and analyze the cache size for SAs in one LHSS. We assume the arrival of SAs is a Poisson process, the service time of each SA is its lifetime (an SA updated or re-registered is considered as a new SA), and the cache capacity of LHSS for SAs equals to the number of available services in the HSSD. Then, we model LHSS as an M/G/c/c queuing system [10]. The model is depicted in Fig. 3.

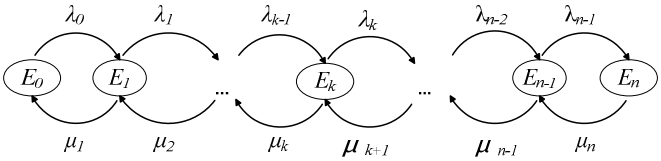


Fig. 3. State transition diagram of LHSS caches.

The parameters in Fig. 3 are defined as follows.

n: the cache capacity of LHSS for SAs.

E_k: the state of k SAs stored in LHSS.

λ_i: the arrival rate of SAs when in the state of E_i. Here, we assume the average arrival rate of SAs is λ, and set $\lambda_i=\lambda$, $i=0, 1, \ldots, n-1$.

μ_i: the service rate of SAs when in the state of E_i. We set $\mu_i=i\mu$, $i=1, 2, \ldots, n$. μ is defined as the mean service rate of SAs, and it equals to the mean discarded rate of SAs, i.e., $\mu=1/T_L$. Here, T_L is the average lifetime of SAs.

Let $H_{AS\text{-}HSS}$ be the mean hop distance between AS and LHSS, and P_f be the average one-hop loss rate in the network. Then, the success rate of an SA transmitted from AS to LHSS is derived as

$$P_{AS-HSS} = \left(1-P_f\right)^{H_{AS-HSS}}. \quad (1)$$

Let N_{AS} be the total number of ASs providing services and assume each AS can only provide one service and send one SA message simultaneously. I_{SA} is denoted as the average update interval of SA including its registration, update, and re-registration. Then, we have,

$$\lambda = \frac{N_{AS} P_{AS-HSS}}{I_{SA}} \quad (2)$$

We denote the traffic intensity as $\rho=\lambda/\mu$. According to the queuing theory [19], the probability of k SAs being stored in LHSS is derived as,

$$P_k = \frac{\frac{\rho^k}{k!}}{1+\rho+\frac{\rho^2}{2!}+\ldots+\frac{\rho^n}{n!}} \quad for\ k = 0,1,\ldots,n \quad (3)$$

Let the average size in bytes of an SA be b_{SA}, and set the cache capacity of LHSS for SAs n as the total number of ASs providing services N_{AS} in the same HSSD. Then, the cache size in LHSS for SAs is computed as

$$C_{HSS-SA} = b_{SA}\sum_{k=0}^{n} kP_k = b_{SA}\sum_{k=0}^{N_{AS}} k\frac{\left[\frac{N_{AS}}{I_{SA}}\left(1-P_f\right)^{H_{AS-HSS}}T_L\right]^k}{k!\sum_{j=0}^{N_{AS}}\frac{\left[\frac{N_{AS}}{I_{SA}}\left(1-P_f\right)^{H_{AS-HSS}}T_L\right]^j}{j!}} \quad (4)$$

As mentioned above, we only studied the cache size for SAs in one LHSS covering one HSSD, and the condition in each HSSD is equal in terms of modeling. So, it is not necessary to analyze the cache usage in each LHSS.

B. Success Rate of SD

First of all, the success rate of SD in UE's own HSSD is considered. Under this precondition, all the following requirements should be satisfied: 1) LHSS receives SrvRqt sent by UE, 2) LHSS finds the suitable service information in its database, and 3) UE gets SrvRply responded by LHSS.

We consider the process of sending SrvRqt from UE to LHSS. Let $H_{UE\text{-}HSS}$ denote the mean hop distance from UE to LHSS. Then, the success rate in this part is

$$P_{SrvRqt} = \left(1-P_f\right)^{H_{UE-HSS}}. \quad (5)$$

We assume the success rate of receiving SrvRply from LHSS to UE is equal to P_{SrvRqt}, namely, $P_{SrvRply}=P_{SrvRqt}$. The success rate of searching suitable service information in LHSS means the probability of suitable SAs cached in LHSS successfully. We denote the mean interval between two adjacent SrvRqts as I_{SrvRqt} and set

$$m = \left\lfloor \frac{I_{SrvRqt}}{I_{SA}} \right\rfloor. \quad (6)$$

Then, $I_{SrvRqt}=mI_{SA}+x$. Here, m=0, 1, 2, …; $x=mod(I_{SrvRqt}, I_{SA})$, and $x\in[0, I_{SA})$.

When its lifetime expires, an SA will be discarded. That means only the latest several transmissions (named valid transmissions) of an SA influence the success rate of storing SA in LHSS.

We set $m' = \left\lfloor \frac{T_L}{I_{SA}} \right\rfloor$, then have $T_L=m'I_{SA}+x'$. Here, m'=0, 1, 2, …; $x'=mod(T_L, I_{SA})$, and $x'\in[0, I_{SA})$.

When $x\le x'$, the number of valid transmissions of an SA is $n_1=m'+1$, otherwise, it is $n_2=m'$. The possibilities in the two conditions are $P(n_1)=x'/I_{SA}$ and $P(n_2)=1-P(n_1)$. Then, the success rate of finding a suitable SA cached in LHSS is expressed as

$$P_{SA} = \begin{cases} P_{SA_1} = 1-\left(1-P_{AS-HSS}\right)^{n_1}, & if\ x \le x' \\ P_{SA_2} = 1-\left(1-P_{AS-HSS}\right)^{n_2}, & if\ x > x' \end{cases} \quad (7)$$

$$\begin{cases} P\left(x \le x'\right) = P\left(n_1\right) = \dfrac{\mathrm{mod}(T_L, I_{SA})}{I_{SA}} \\ P\left(x > x'\right) = P\left(n_2\right) = 1-\dfrac{\mathrm{mod}(T_L, I_{SA})}{I_{SA}} \end{cases} \quad (8)$$

Assume there are s services in this HSSD matching SrvRqt, then the success rate of SD in UE's own HSSD is expressed as

$$P_{SD} = P_{SrvRqt}P_{SrvRply}\left(1-\left(1-P_{SA}\right)^s\right) \quad (9)$$

The average success rate of SD in UE's own HSSD is considered as the expected value of P_{SD}, that is,

$$\begin{aligned} E\left(P_{SD}\right) &= P\left(n_1\right)P_{SrvRqt}P_{SrvRply}\left(1-\left(1-P_{SA_1}\right)^s\right) \\ &\quad + P\left(n_2\right)P_{SrvRqt}P_{SrvRply}\left(1-\left(1-P_{SA_2}\right)^s\right) \\ &= \frac{\mathrm{mod}\left(T_L, I_{SA}\right)}{I_{SA}}\left(\left(1-P_f\right)^{H_{UE-HSS}}\right)^2 \\ &\quad \left(1-\left(1-\left(1-\left(1-P_f\right)^{H_{AS-HSS}}\right)^{\left(\left\lfloor\frac{T_L}{I_{SA}}\right\rfloor+1\right)}\right)\right)^s \\ &\quad + \left(1-\frac{\mathrm{mod}\left(T_L, I_{SA}\right)}{I_{SA}}\right)\left(\left(1-P_f\right)^{H_{UE-HSS}}\right)^2 \\ &\quad \left(1-\left(1-\left(1-\left(1-P_f\right)^{H_{AS-HSS}}\right)^{\left\lfloor\frac{T_L}{I_{SA}}\right\rfloor}\right)\right)^s \end{aligned} \quad (10)$$

If it receives SrvRply including the information of more than one suitable service, UE will choose the best one according to its specific requirements. When there is no suitable service in UE's HSSD, S-CSCF may communicate with LSLF and continues to search in neighbor HSSDs. Here, we only analyze the procedure of search in UE's own HSSD because in a neighbor HSSD, the procedure is the same ignoring the communications between S-CSCF and LSLF.

C. Traffic Load Generated by SD Procedure

Assume each UE only looks for its desired services in its own HSSD, and in this precondition we analyze the traffic load generated by SD procedure in one HSSD.

Assume there are q SrvRqts and w SAs in a random period of time T. Using the similar analytical method in Subsection B, we derive

$$q=\begin{cases} q_1=\left\lfloor \dfrac{T}{I_{SrvRqt}} \right\rfloor+1, & \mathrm{P}(q_1)=\dfrac{\mathrm{mod}(T,I_{SrvRqt})}{I_{SrvRqt}} \\ q_2=\left\lfloor \dfrac{T}{I_{SrvRqt}} \right\rfloor, & \mathrm{P}(q_2)=1-\dfrac{\mathrm{mod}(T,I_{SrvRqt})}{I_{SrvRqt}} \end{cases} \tag{11}$$

$$w=\begin{cases} w_1=N_{AS}\left(\left\lfloor \dfrac{T}{I_{SA}} \right\rfloor+1\right), & \mathrm{P}(w_1)=\dfrac{\mathrm{mod}(T,I_{SA})}{I_{SA}} \\ w_2=N_{AS}\left\lfloor \dfrac{T}{I_{SA}} \right\rfloor, & \mathrm{P}(w_2)=1-\dfrac{\mathrm{mod}(T,I_{SA})}{I_{SA}} \end{cases} \tag{12}$$

The number of SrvRplys is determined by the success rates of SrvRqts and SAs, and all the SrvRplys constitute n-time Bernoulli trial. Assume there are r SrvRplys in T, using binomial distribution, we have

$$\begin{aligned} r&=\sum_{k=0}^{q} kC_q^k\left(P_{SA}P_{SrvRqt}\right)^k\left(1-P_{SA}P_{SrvRqt}\right)^{q-k} \\ &=\sum_{k=0}^{q} k\frac{q!}{k!(q-k)!}\left(P_{SA}P_{SrvRqt}\right)^k\left(1-P_{SA}P_{SrvRqt}\right)^{q-k} \end{aligned} \tag{13}$$

The traffic load caused by SAs, SrvRqts, and SrvRplys in T are denoted as L_{SA}, L_{SrvRqt}, and $L_{SrvRply}$, respectively, then the total traffic load in UE's HSSD can be expressed as

$$L_{HSSD}=L_{SA}+L_{SrvRqt}+L_{SrvRply}\,. \tag{14}$$

Let m_{SA}, m_{SrvRqt}, and $m_{SrvRply}$ be the number of mean total messages caused by one SA, SrvRqt, and SrvRply, respectively. We derive

$$m_{SA}=H_{AS-HSS}\left(1-P_f\right)^{H_{AS-HSS}}+\sum_{i=1}^{H_{AS-HSS}} iP_f\left(1-P_f\right)^{i-1} \tag{15}$$

$$m_{SrvRqt}=H_{UE-HSS}\left(1-P_f\right)^{H_{UE-HSS}}+\sum_{j=1}^{H_{UE-HSS}} jP_f\left(1-P_f\right)^{j-1} \tag{16}$$

$$m_{SrvRply}=m_{SrvRqt} \tag{17}$$

Using b_{SrvRqt}, $b_{SrvRply}$, and b_{SA} express the average size of SrvRqt, SrvRply, and SA in bytes respectively, then we can further derive

$$L_{SA}=wm_{SA}b_{SA} \tag{18}$$

$$L_{SrvRqt}=qm_{SrvRqt}b_{SrvRqt} \tag{19}$$

$$L_{SrvRply}=rm_{SrvRply}b_{SrvRply} \tag{20}$$

We consider the whole IMS-based network including more than one HSSD. Regard the SrvRqts of searching in neighbor HSSDs as new SrvRqts, assume there are H HSSDs in all, and ignore the communications between S-CSCF and LSLF. The total traffic load in the whole IMS-based network can be calculated approximately as

$$L_{total}=HL_{HSSD} \tag{21}$$

The average traffic load caused by SD procedure in the whole IMS-based network is considered as the expected value of L_{total}, that is,

$$\begin{aligned} E(L_{total})=&P(q_1)P(w_1)P(n_1)L_{total}(q_1,w_1,n_1)\\ &+P(q_1)P(w_2)P(n_1)L_{total}(q_1,w_2,n_1)\\ &+P(q_1)P(w_1)P(n_2)L_{total}(q_1,w_1,n_2)\\ &+P(q_2)P(w_1)P(n_1)L_{total}(q_2,w_1,n_1)\\ &+P(q_1)P(w_2)P(n_2)L_{total}(q_1,w_2,n_2)\\ &+P(q_2)P(w_2)P(n_1)L_{total}(q_2,w_2,n_1)\\ &+P(q_2)P(w_1)P(n_2)L_{total}(q_2,w_1,n_2)\\ &+P(q_2)P(w_2)P(n_2)L_{total}(q_2,w_2,n_2) \end{aligned} \tag{22}$$

We set

$$\begin{aligned} &E(L_{total1})=P(q_1)P(w_1)P(n_1)L_{total}(q_1,w_1,n_1)\\ &=H\frac{\mathrm{mod}(T,I_{SrvRqt})}{I_{SrvRqt}}\frac{\mathrm{mod}(T,I_{SA})}{I_{SA}}\frac{\mathrm{mod}(T_L,I_{SA})}{I_{SA}}\\ &\left(b_{SrvRqt}\left(\left\lfloor\frac{T}{I_{SrvRqt}}\right\rfloor+1\right)\left(H_{UE-HSS}\left(1-P_f\right)^{H_{UE-HSS}}+\sum_{j=1}^{H_{UE-HSS}}jP_f\left(1-P_f\right)^{j-1}\right)\right.\\ &\left.+b_{SrvRply}\left(H_{UE-HSS}\left(1-P_f\right)^{H_{UE-HSS}}+\sum_{j=1}^{H_{UE-HSS}}jP_f\left(1-P_f\right)^{j-1}\right)\right)\\ &\sum_{k=0}^{\left(\left\lfloor\frac{T}{I_{SrvRqt}}\right\rfloor+1\right)}kC^k_{\left(\left\lfloor\frac{T}{I_{SrvRqt}}\right\rfloor+1\right)}\left(\left(\left(1-P_f\right)^{H_{UE-HSS}}\left(1-\left(1-\left(1-P_f\right)^{H_{AS-HSS}}\right)^{\left(\left\lfloor\frac{T_L}{I_{SA}}\right\rfloor+1\right)}\right)\right)^k\right.\\ &\left.\left(1-\left(1-P_f\right)^{H_{UE-HSS}}\left(1-\left(1-\left(1-P_f\right)^{H_{AS-HSS}}\right)^{\left(\left\lfloor\frac{T_L}{I_{SA}}\right\rfloor+1\right)}\right)\right)^{\left(\left\lfloor\frac{T}{I_{SrvRqt}}\right\rfloor+1\right)-k}\right)\\ &+b_{SA}N_{AS}\left(\left\lfloor\frac{T}{I_{SA}}\right\rfloor+1\right)\left(H_{AS-HSS}\left(1-P_f\right)^{H_{AS-HSS}}+\sum_{i=1}^{H_{AS-HSS}}iP_f\left(1-P_f\right)^{i-1}\right) \end{aligned} \tag{23}$$

$$\begin{aligned} &E(L_{total2})=P(q_1)P(w_2)P(n_1)L_{total}(q_1,w_2,n_1)\\ &=H\frac{\mathrm{mod}(T,I_{SrvRqt})}{I_{SrvRqt}}\left(1-\frac{\mathrm{mod}(T,I_{SA})}{I_{SA}}\right)\frac{\mathrm{mod}(T_L,I_{SA})}{I_{SA}}\\ &\left(b_{SrvRqt}\left(\left\lfloor\frac{T}{I_{SrvRqt}}\right\rfloor+1\right)\left(H_{UE-HSS}\left(1-P_f\right)^{H_{UE-HSS}}+\sum_{j=1}^{H_{UE-HSS}}jP_f\left(1-P_f\right)^{j-1}\right)\right.\\ &\left.+b_{SrvRply}\left(H_{UE-HSS}\left(1-P_f\right)^{H_{UE-HSS}}+\sum_{j=1}^{H_{UE-HSS}}jP_f\left(1-P_f\right)^{j-1}\right)\right)\\ &\sum_{k=0}^{\left(\left\lfloor\frac{T}{I_{SrvRqt}}\right\rfloor+1\right)}kC^k_{\left(\left\lfloor\frac{T}{I_{SrvRqt}}\right\rfloor+1\right)}\left(\left(\left(1-P_f\right)^{H_{UE-HSS}}\left(1-\left(1-\left(1-P_f\right)^{H_{AS-HSS}}\right)^{\left(\left\lfloor\frac{T_L}{I_{SA}}\right\rfloor+1\right)}\right)\right)^k\right.\\ &\left.\left(1-\left(1-P_f\right)^{H_{UE-HSS}}\left(1-\left(1-\left(1-P_f\right)^{H_{AS-HSS}}\right)^{\left(\left\lfloor\frac{T_L}{I_{SA}}\right\rfloor+1\right)}\right)\right)^{\left(\left\lfloor\frac{T}{I_{SrvRqt}}\right\rfloor+1\right)-k}\right)\\ &+b_{SA}N_{AS}\left\lfloor\frac{T}{I_{SA}}\right\rfloor\left(H_{AS-HSS}\left(1-P_f\right)^{H_{AS-HSS}}+\sum_{i=1}^{H_{AS-HSS}}iP_f\left(1-P_f\right)^{i-1}\right) \end{aligned} \tag{24}$$

$$E(L_{total3}) = P(q_1)P(w_1)P(n_2)L_{total}(q_1, w_1, n_2)$$
$$= H\frac{\mathrm{mod}(T, I_{SrvRqt})}{I_{SrvRqt}}\frac{\mathrm{mod}(T, I_{SA})}{I_{SA}}\left(1-\frac{\mathrm{mod}(T_L, I_{SA})}{I_{SA}}\right)$$
$$\left(b_{SrvRqt}\left(\left\lfloor\frac{T}{I_{SrvRqt}}\right\rfloor+1\right)\left(H_{UE-HSS}(1-P_f)^{H_{UE-HSS}}+\sum_{j=1}^{H_{UE-HSS}}jP_f(1-P_f)^{j-1}\right)\right.$$
$$+b_{SrvRply}\left(H_{UE-HSS}(1-P_f)^{H_{UE-HSS}}+\sum_{j=1}^{H_{UE-HSS}}jP_f(1-P_f)^{j-1}\right)$$
$$\sum_{k=0}^{\left(\left\lfloor\frac{T}{I_{SrvRqt}}\right\rfloor+1\right)}kC^k_{\left(\left\lfloor\frac{T}{I_{SrvRqt}}\right\rfloor+1\right)}\left(\left((1-P_f)^{H_{UE-HSS}}\left(1-\left(1-(1-P_f)^{H_{AS-HSS}}\right)^{\left\lfloor\frac{T_L}{I_{SA}}\right\rfloor}\right)\right)^k\right.$$
$$\left.\left(1-(1-P_f)^{H_{UE-HSS}}\left(1-\left(1-(1-P_f)^{H_{AS-HSS}}\right)^{\left\lfloor\frac{T_L}{I_{SA}}\right\rfloor}\right)\right)^{\left(\left\lfloor\frac{T}{I_{SrvRqt}}\right\rfloor+1\right)-k}\right)$$
$$\left.+b_{SA}N_{AS}\left(\left\lfloor\frac{T}{I_{SA}}\right\rfloor+1\right)\left(H_{AS-HSS}(1-P_f)^{H_{AS-HSS}}+\sum_{i=1}^{H_{AS-HSS}}iP_f(1-P_f)^{i-1}\right)\right)$$
(25)

$$E(L_{total4}) = P(q_2)P(w_1)P(n_1)L_{total}(q_2, w_1, n_1)$$
$$= H\left(1-\frac{\mathrm{mod}(T, I_{SrvRqt})}{I_{SrvRqt}}\right)\frac{\mathrm{mod}(T, I_{SA})}{I_{SA}}\frac{\mathrm{mod}(T_L, I_{SA})}{I_{SA}}$$
$$\left(b_{SrvRqt}\left\lfloor\frac{T}{I_{SrvRqt}}\right\rfloor\left(H_{UE-HSS}(1-P_f)^{H_{UE-HSS}}+\sum_{j=1}^{H_{UE-HSS}}jP_f(1-P_f)^{j-1}\right)\right.$$
$$+b_{SrvRply}\left(H_{UE-HSS}(1-P_f)^{H_{UE-HSS}}+\sum_{j=1}^{H_{UE-HSS}}jP_f(1-P_f)^{j-1}\right)$$
$$\sum_{k=0}^{\left\lfloor\frac{T}{I_{SrvRqt}}\right\rfloor}kC^k_{\left\lfloor\frac{T}{I_{SrvRqt}}\right\rfloor}\left(\left((1-P_f)^{H_{UE-HSS}}\left(1-\left(1-(1-P_f)^{H_{AS-HSS}}\right)^{\left(\left\lfloor\frac{T_L}{I_{SA}}\right\rfloor+1\right)}\right)\right)^k\right.$$
$$\left.\left(1-(1-P_f)^{H_{UE-HSS}}\left(1-\left(1-(1-P_f)^{H_{AS-HSS}}\right)^{\left(\left\lfloor\frac{T_L}{I_{SA}}\right\rfloor+1\right)}\right)\right)^{\left\lfloor\frac{T}{I_{SrvRqt}}\right\rfloor-k}\right)$$
$$\left.+b_{SA}N_{AS}\left(\left\lfloor\frac{T}{I_{SA}}\right\rfloor+1\right)\left(H_{AS-HSS}(1-P_f)^{H_{AS-HSS}}+\sum_{i=1}^{H_{AS-HSS}}iP_f(1-P_f)^{i-1}\right)\right)$$
(26)

$$E(L_{total5}) = P(q_1)P(w_2)P(n_2)L_{total}(q_1, w_2, n_2)$$
$$= H\frac{\mathrm{mod}(T, I_{SrvRqt})}{I_{SrvRqt}}\left(1-\frac{\mathrm{mod}(T, I_{SA})}{I_{SA}}\right)\left(1-\frac{\mathrm{mod}(T_L, I_{SA})}{I_{SA}}\right)$$
$$\left(b_{SrvRqt}\left(\left\lfloor\frac{T}{I_{SrvRqt}}\right\rfloor+1\right)\left(H_{UE-HSS}(1-P_f)^{H_{UE-HSS}}+\sum_{j=1}^{H_{UE-HSS}}jP_f(1-P_f)^{j-1}\right)\right.$$
$$+b_{SrvRply}\left(H_{UE-HSS}(1-P_f)^{H_{UE-HSS}}+\sum_{j=1}^{H_{UE-HSS}}jP_f(1-P_f)^{j-1}\right)$$
$$\sum_{k=0}^{\left(\left\lfloor\frac{T}{I_{SrvRqt}}\right\rfloor+1\right)}kC^k_{\left(\left\lfloor\frac{T}{I_{SrvRqt}}\right\rfloor+1\right)}\left(\left((1-P_f)^{H_{UE-HSS}}\left(1-\left(1-(1-P_f)^{H_{AS-HSS}}\right)^{\left\lfloor\frac{T_L}{I_{SA}}\right\rfloor}\right)\right)^k\right.$$
$$\left.\left(1-(1-P_f)^{H_{UE-HSS}}\left(1-\left(1-(1-P_f)^{H_{AS-HSS}}\right)^{\left\lfloor\frac{T_L}{I_{SA}}\right\rfloor}\right)\right)^{\left(\left\lfloor\frac{T}{I_{SrvRqt}}\right\rfloor+1\right)-k}\right)$$
$$\left.+b_{SA}N_{AS}\left\lfloor\frac{T}{I_{SA}}\right\rfloor\left(H_{AS-HSS}(1-P_f)^{H_{AS-HSS}}+\sum_{i=1}^{H_{AS-HSS}}iP_f(1-P_f)^{i-1}\right)\right)$$
(27)

$$E(L_{total6}) = P(q_2)P(w_2)P(n_1)L_{total}(q_2, w_2, n_1)$$
$$= H\left(1-\frac{\mathrm{mod}(T, I_{SrvRqt})}{I_{SrvRqt}}\right)\left(1-\frac{\mathrm{mod}(T, I_{SA})}{I_{SA}}\right)\frac{\mathrm{mod}(T_L, I_{SA})}{I_{SA}}$$
$$\left(b_{SrvRqt}\left\lfloor\frac{T}{I_{SrvRqt}}\right\rfloor\left(H_{UE-HSS}(1-P_f)^{H_{UE-HSS}}+\sum_{j=1}^{H_{UE-HSS}}jP_f(1-P_f)^{j-1}\right)\right.$$
$$+b_{SrvRply}\left(H_{UE-HSS}(1-P_f)^{H_{UE-HSS}}+\sum_{j=1}^{H_{UE-HSS}}jP_f(1-P_f)^{j-1}\right)$$
$$\sum_{k=0}^{\left\lfloor\frac{T}{I_{SrvRqt}}\right\rfloor}kC^k_{\left\lfloor\frac{T}{I_{SrvRqt}}\right\rfloor}\left(\left((1-P_f)^{H_{UE-HSS}}\left(1-\left(1-(1-P_f)^{H_{AS-HSS}}\right)^{\left(\left\lfloor\frac{T_L}{I_{SA}}\right\rfloor+1\right)}\right)\right)^k\right.$$
$$\left.\left(1-(1-P_f)^{H_{UE-HSS}}\left(1-\left(1-(1-P_f)^{H_{AS-HSS}}\right)^{\left(\left\lfloor\frac{T_L}{I_{SA}}\right\rfloor+1\right)}\right)\right)^{\left\lfloor\frac{T}{I_{SrvRqt}}\right\rfloor-k}\right)$$
$$\left.+b_{SA}N_{AS}\left\lfloor\frac{T}{I_{SA}}\right\rfloor\left(H_{AS-HSS}(1-P_f)^{H_{AS-HSS}}+\sum_{i=1}^{H_{AS-HSS}}iP_f(1-P_f)^{i-1}\right)\right)$$
(28)

$$E(L_{total7}) = P(q_2)P(w_1)P(n_2)L_{total}(q_2, w_1, n_2)$$
$$= H\left(1-\frac{\mathrm{mod}(T, I_{SrvRqt})}{I_{SrvRqt}}\right)\frac{\mathrm{mod}(T, I_{SA})}{I_{SA}}\left(1-\frac{\mathrm{mod}(T_L, I_{SA})}{I_{SA}}\right)$$
$$\left(b_{SrvRqt}\left\lfloor\frac{T}{I_{SrvRqt}}\right\rfloor\left(H_{UE-HSS}(1-P_f)^{H_{UE-HSS}}+\sum_{j=1}^{H_{UE-HSS}}jP_f(1-P_f)^{j-1}\right)\right.$$
$$+b_{SrvRply}\left(H_{UE-HSS}(1-P_f)^{H_{UE-HSS}}+\sum_{j=1}^{H_{UE-HSS}}jP_f(1-P_f)^{j-1}\right)$$
$$\sum_{k=0}^{\left\lfloor\frac{T}{I_{SrvRqt}}\right\rfloor}kC^k_{\left\lfloor\frac{T}{I_{SrvRqt}}\right\rfloor}\left(\left((1-P_f)^{H_{UE-HSS}}\left(1-\left(1-(1-P_f)^{H_{AS-HSS}}\right)^{\left\lfloor\frac{T_L}{I_{SA}}\right\rfloor}\right)\right)^k\right.$$
$$\left.\left(1-(1-P_f)^{H_{UE-HSS}}\left(1-\left(1-(1-P_f)^{H_{AS-HSS}}\right)^{\left\lfloor\frac{T_L}{I_{SA}}\right\rfloor}\right)\right)^{\left\lfloor\frac{T}{I_{SrvRqt}}\right\rfloor-k}\right)$$
$$\left.+b_{SA}N_{AS}\left(\left\lfloor\frac{T}{I_{SA}}\right\rfloor+1\right)\left(H_{AS-HSS}(1-P_f)^{H_{AS-HSS}}+\sum_{i=1}^{H_{AS-HSS}}iP_f(1-P_f)^{i-1}\right)\right)$$
(29)

$$E(L_{total8}) = P(q_2)P(w_2)P(n_2)L_{total}(q_2, w_2, n_2)$$
$$= H\left(1-\frac{\mathrm{mod}(T, I_{SrvRqt})}{I_{SrvRqt}}\right)\left(1-\frac{\mathrm{mod}(T, I_{SA})}{I_{SA}}\right)\left(1-\frac{\mathrm{mod}(T_L, I_{SA})}{I_{SA}}\right)$$
$$\left(b_{SrvRqt}\left\lfloor\frac{T}{I_{SrvRqt}}\right\rfloor\left(H_{UE-HSS}(1-P_f)^{H_{UE-HSS}}+\sum_{j=1}^{H_{UE-HSS}}jP_f(1-P_f)^{j-1}\right)\right.$$
$$+b_{SrvRply}\left(H_{UE-HSS}(1-P_f)^{H_{UE-HSS}}+\sum_{j=1}^{H_{UE-HSS}}jP_f(1-P_f)^{j-1}\right)$$
$$\sum_{k=0}^{\left\lfloor\frac{T}{I_{SrvRqt}}\right\rfloor}kC^k_{\left\lfloor\frac{T}{I_{SrvRqt}}\right\rfloor}\left(\left((1-P_f)^{H_{UE-HSS}}\left(1-\left(1-(1-P_f)^{H_{AS-HSS}}\right)^{\left\lfloor\frac{T_L}{I_{SA}}\right\rfloor}\right)\right)^k\right.$$
$$\left.\left(1-(1-P_f)^{H_{UE-HSS}}\left(1-\left(1-(1-P_f)^{H_{AS-HSS}}\right)^{\left\lfloor\frac{T_L}{I_{SA}}\right\rfloor}\right)\right)^{\left\lfloor\frac{T}{I_{SrvRqt}}\right\rfloor-k}\right)$$
$$\left.+b_{SA}N_{AS}\left\lfloor\frac{T}{I_{SA}}\right\rfloor\left(H_{AS-HSS}(1-P_f)^{H_{AS-HSS}}+\sum_{i=1}^{H_{AS-HSS}}iP_f(1-P_f)^{i-1}\right)\right)$$
(30)

Then, we have,

$$E(L_{total}) = E(L_{total1}) + E(L_{total2}) + E(L_{total3}) + E(L_{total4}) + E(L_{total5}) + E(L_{total6}) + E(L_{total7}) + E(L_{total8}) \quad (31)$$

IV. Performance Analyses and Optimizations

In this section, we evaluate the performance of the proposed novel SDA by numerical simulations first. All the simulations are based on the analytical models and theoretical analyses deduced in Section III. The input parameters and respective values are documented in Table I.

TABLE I.
Parameter Setting.

Parameter	Description	Value
H	The total number of HSSDs in the whole IMS-based network.	10
N_{AS}	The total number of ASs providing services in each HSSD.	20
T	The total simulation time.	1800 seconds
T_L	The average lifetime of SAs.	150 seconds
$H_{AS\text{-}HSS}$	The average hop distance between AS and LHSS.	2
$H_{UE\text{-}HSS}$	The average hop distance between UE and LHSS.	10
b_{SA}	The mean size of SA.	200 bytes
b_{SrvRqt}	The mean size of SrvRqt.	50 bytes
$b_{SrvRply}$	The mean size of SrvRply.	100 bytes
I_{SA}	The average update interval of SA.	[30, 150] seconds
I_{SrvRqt}	The average interval between two adjacent SrvRqts.	[350, 500] seconds

Also, according to the performance analyses, we propose a performance optimization policy based on choosing an optimal value for the average update interval of SA. The policy makes all the performances get the trade-off.

A. Cache Size in LHSS for SAs

Through the analytical expression derived for the cache size in one LHSS for SAs, $C_{HSS\text{-}SA}$ in (4), with the average update interval of SA I_{SA}, the numerical simulation results are presented in Fig. 4. It is evidently showed that the cache size in one LHSS for SAs, $C_{HSS\text{-}SA}$, decreases while the average update interval of SA I_{SA} increases. That means reducing the update frequency of SA can save the cache in LHSS for SAs. On the other hand, the variation of network condition P_f (we set P_f changes from 0.01 to 0.05) also influences the cache size. But, comparing with I_{SA}, the influence of P_f to $C_{HSS\text{-}SA}$ is negligible.

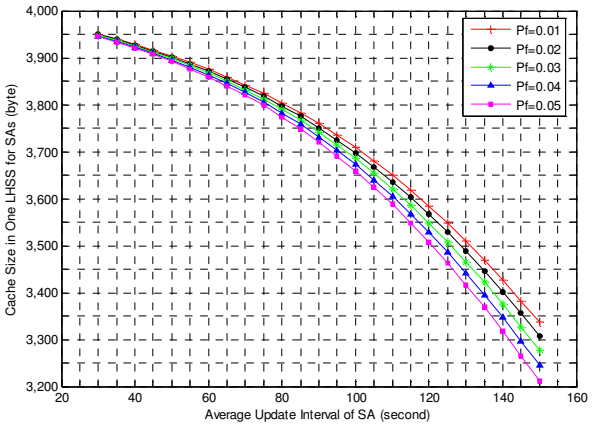
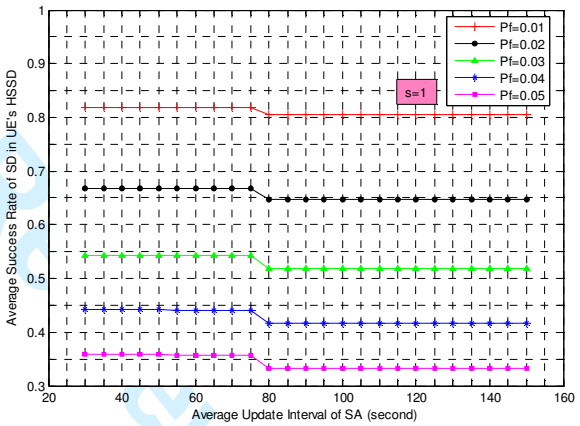


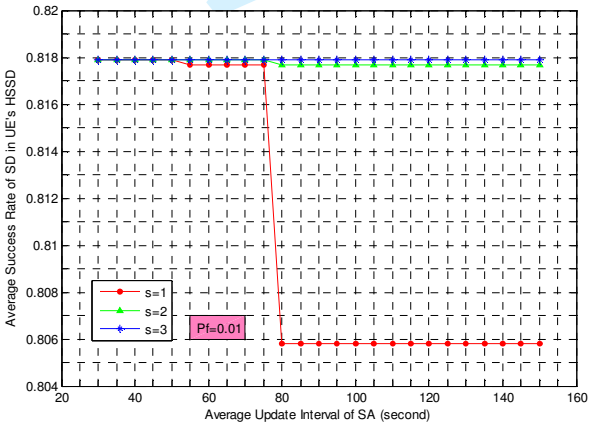
Fig. 4. Cache size in LHSS for SAs.

B. Success Rate of SD

According to the analyses in Subsection B of Section III, the procedure of SD in a neighbor HSSD is the same as that in UE's own HSSD. So, we only consider the average success rate of SD in UE's own HSSD here. The simulations are shown in Figs. 5(a) and 5(b).



(a)



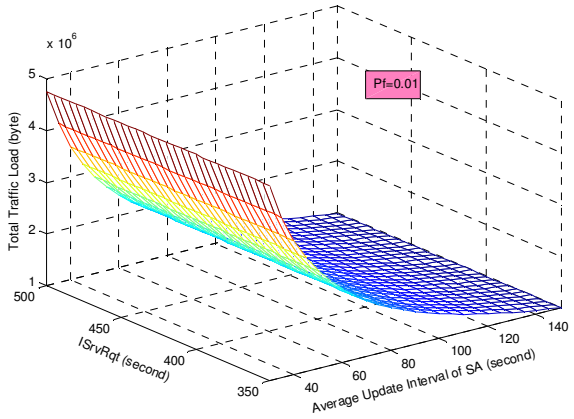
(b)

Fig. 5. Success rate of SD.

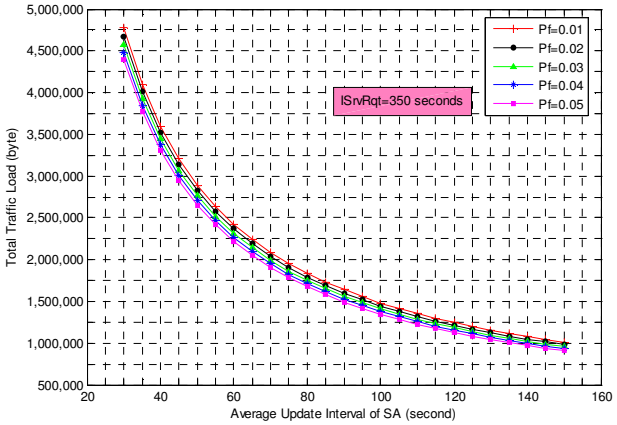
From the figures, we note that the average success rate of SD $E(P_{SD})$ presents the ladder-like decline as the average update interval of SA I_{SA} is getting longer. The turning point is at the point of $I_{SA}=75$ seconds, that is $I_{SA}=T_L/2$ (here, T_L is 150 seconds). When I_{SA} is longer than $T_L/2$, the success rate of SD has a slight reduction. In addition, Fig. 5(a) shows that $E(P_{SD})$ descends markedly when the network condition becomes bad, i.e., many signaling messages get lost when P_f increases. In order to ensure higher success rate of SD, P_f should better not be higher than 0.02. Also, comparing with the influence of P_f, the influence of I_{SA} to $E(P_{SD})$ can be ignored. Take the condition of $S=1$ for example. If the average number of matching services for each UE is one, when I_{SA} increases, the success rate of SD reduces just from 0.818 to 0.806. When the average number of matching services for each UE is more than one ($s\geq 2$), the reduction is almost void (see Fig. 5(b), $P_f=0.01$).

C. Traffic Load Generated by SD Procedure

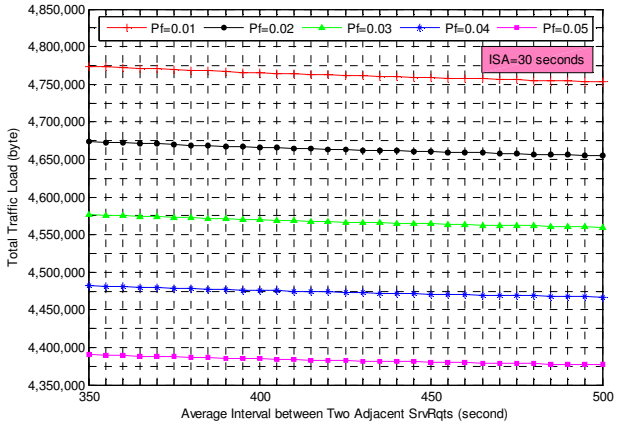
Considering the influence of the average update interval of SA I_{SA}, the mean interval between two adjacent SrvRqts I_{SrvRqt}, and the network condition P_f, Fig. 6 shows the variation trend of the average traffic load generated by SD procedure in the whole IMS-based network $E(L_{total})$.



(a)



(b)



(c)

Fig. 6. Traffic load generated by SD procedure.

It is observed that the traffic load generated by SD procedure increases when the update frequency of SA improves or/and the SrvRqts come thick and fast, but the traffic load decreases when the network condition is worse. That means prolonging the update interval of SA or/and the interval between SrvRqts, or/and worse network condition can reduce the traffic load. Fig. 6(a) ($P_f=0.01$) shows that the traffic load decreases sharply when I_{SA} is getting longer, but I_{SrvRqt} is ignorable comparing to I_{SA}. On the other hand, the curves of $E(L_{total})$ metrics for each different P_f are very close to one another (see Fig. 6(b) ($I_{SrvRqt}=350$ seconds)). That means the variety of the network condition only has a small impact comparing to I_{SA}. Fig. 6(c) ($I_{SA}=30$ seconds) presents that $E(L_{total})$ cuts down as the I_{SrvRqt} increases, but the impact is slight comparing to P_f. So, the importance of the influence factors to the traffic load is $I_{SA}>P_f>I_{SrvRqt}$.

D. Performance Optimization Policy

From the above performance analyses, we draw the following conclusions. 1) Prolonging the update interval of SA I_{SA} saves cache size in LHSS for SAs evidently. 2) There is a turning point at $I_{SA}=T_L/2$ (here, 30 seconds$\leq I_{SA}\leq 150$ seconds, $T_L=150$ seconds). When I_{SA} is longer than $T_L/2$, the average success rate of SD presents ladder-like reduction. We can get higher success rate by setting $I_{SA}\leq T_L/2$. 3) Prolonging I_{SA} makes the average traffic load generated by SD procedure in the whole IMS-based network decrease much. On balance, we can choose an appropriate I_{SA} to get the best trade-off among the performances, i.e., we can set I_{SA} slightly less than $T_L/2$ to satisfy the requirements of higher success rate, less cache space, and lower traffic load simultaneously.

All the above conclusions are based on the parameter setting in Table I. To any arbitrary IMS-based network, now we consider how to find a suitable I_{SA} to get the best trade-off among all the performances we have analyzed. From Figs. 4, 5, and 6(b), we know the cache size in LHSS for SAs $C_{HSS\text{-}SA}$, the success rate of SD $E(P_{SD})$, and the traffic load caused by SD

procedure $E(L_{total})$ decrease when the I_{SA} increases. In order to get good performance, both $C_{HSS\text{-}SA}$ and $E(L_{total})$ are the lower the better, but $E(P_{SD})$ is the higher the better. According to (4), (10), and (31), $C_{HSS\text{-}SA}$, $E(P_{SD})$, and $E(L_{total})$ are related to I_{SA}. We use nonlinear optimization theory [20] to get the optimal value of I_{SA} for the sake of optimizing the whole IMS-based network performances.

First of all, we use unitary method to deal with (4), (10), and (31), because $C_{HSS\text{-}SA}$, $E(P_{SD})$, and $E(L_{total})$ have different dimensions and units which can not be added directly. After the unitary disposal, we have

$$\left(C_{HSS-SA}\right)_{unit} = \frac{C_{HSS-SA} - \min\left(C_{HSS-SA}\right)}{\max\left(C_{HSS-SA}\right) - \min\left(C_{HSS-SA}\right)} = \frac{C_{HSS-SA} - C_{HSS-SA}\left(\max\left(I_{SA}\right)\right)}{C_{HSS-SA}\left(\min\left(I_{SA}\right)\right) - C_{HSS-SA}\left(\max\left(I_{SA}\right)\right)} \quad (32)$$

$$\left(E\left(P_{SD}\right)\right)_{unit} = \frac{E\left(P_{SD}\right) - \min\left(E\left(P_{SD}\right)\right)}{\max\left(E\left(P_{SD}\right)\right) - \min\left(E\left(P_{SD}\right)\right)} = \frac{E\left(P_{SD}\right) - \left(E\left(P_{SD}\right)\left(\max\left(I_{SA}\right)\right)\right)}{\left(E\left(P_{SD}\right)\left(\min\left(I_{SA}\right)\right)\right) - \left(E\left(P_{SD}\right)\left(\max\left(I_{SA}\right)\right)\right)} \quad (33)$$

$$\left(E\left(L_{total}\right)\right)_{unit} = \frac{E\left(L_{total}\right) - \min\left(E\left(L_{total}\right)\right)}{\max\left(E\left(L_{total}\right)\right) - \min\left(E\left(L_{total}\right)\right)} = \frac{E\left(L_{total}\right) - \left(E\left(L_{total}\right)\left(\max\left(I_{SA}\right)\right)\right)}{\left(E\left(L_{total}\right)\left(\min\left(I_{SA}\right)\right)\right) - \left(E\left(L_{total}\right)\left(\max\left(I_{SA}\right)\right)\right)} \quad (34)$$

Then, we try to find the optimal value of I_{SA} to get the best trade-off among $C_{HSS\text{-}SA}$, $E(P_{SD})$, and $E(L_{total})$. Based on the nonlinear optimization theory, we set

$$\begin{aligned} Trade-off\left(I_{SA}\right) = {} & Weight_1 * \left(C_{HSS-SA}\right)_{unit} \\ & - Weight_2 * \left(E\left(P_{SD}\right)\right)_{unit} \\ & + Weight_3 * \left(E\left(L_{total}\right)\right)_{unit} \end{aligned} \quad (35)$$

Here, $Weight_1$, $Weight_2$, and $Weight_3$ mean the weights of $C_{HSS\text{-}SA}$, $E(P_{SD})$, and $E(L_{total})$, respectively. Considering their physical meanings, $Weight_1$, $Weight_2$, and $Weight_3$ are all positive quantities. As mentioned above, both $C_{HSS\text{-}SA}$ and $E(L_{total})$ are the lower the better, but $E(P_{SD})$ is the higher the better. So, in (35), we put the positive signs in front of both $Weight_1$ and $Weight_3$, but the negative sign in front of $Weight_2$. As the whole IMS-based network operator, it can set different weights of $C_{HSS\text{-}SA}$, $E(P_{SD})$, and $E(L_{total})$ to satisfy the specific network conditions and requirements. Basically, the setting of weights needs to consider the essential characteristics and special demands of the certain given IMS-based network. For example, if an IMS-based network has limited bandwidths, and saving bandwidth is the most important aspect comparing with reducing the cost of adding caches in LHSSs for SAs and improving the success rate of SD, the relevant network operator can realize the trade-off among the performances based on the network requirements by adding the weight of $E(L_{total})$. Of course, according to different IMS-based networks or the variations of the same network condition, the network operator can adjust the weights easily and freely. From (35), it is evident that using the expression of $Trade\text{-}off(I_{SA})$, we can find the optimal value of I_{SA} to get the best trade-off among the performances according to the requirements of a certain given IMS-based network. Draw the curves of $Trade\text{-}off(I_{SA})$ and find the corresponding value of I_{SA} which makes $Trade\text{-}off(I_{SA})$ minimum, and is the optimal value we need.

To be a certain IMS-based network, there are some parameters relatively determinate, such as b_{SrvRqt}, $b_{SrvRply}$, b_{SA}, $H_{AS\text{-}HSS}$, $H_{UE\text{-}HSS}$, N_{AS}, and H. For the values of these parameters, we still use the setting presented in Table I. Also, we still set the total simulation time T to be *1800* seconds. In the following, we discuss the performance optimization policy concretely for choosing the optimal value of I_{SA}.

1) The Influence of the Average Number of Services Matching Each SrvRqt in One HSSD

Setting I_{SrvRqt}=*350* seconds, P_f=*0.01*, $Weight_1$=$Weight_2$=$Weight_3$=*1*, T_L=*150* seconds, and the range of I_{SA} is from *30* to *150* seconds, when the average number of services matching each SrvRqt s changes from *1* to *3*, the curves of $Trade\text{-}off(I_{SA})$ are presented in Fig. 7. It is evident that the values of s do not influence the selection for the optimal value of I_{SA}. Also, at the point of I_{SA}=$T_L/2$=*75* seconds, we can get the best trade-off among the performances. Here, T_L=$max(I_{SA})$.

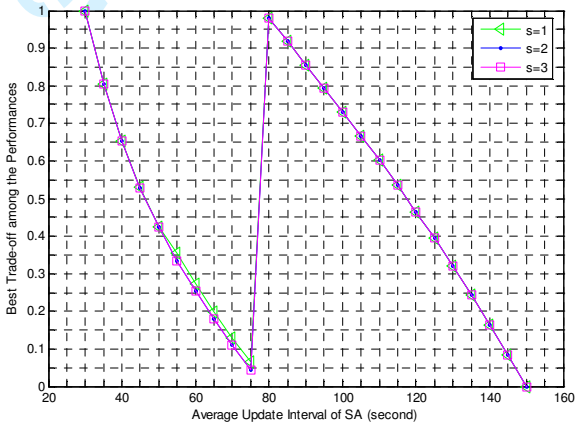
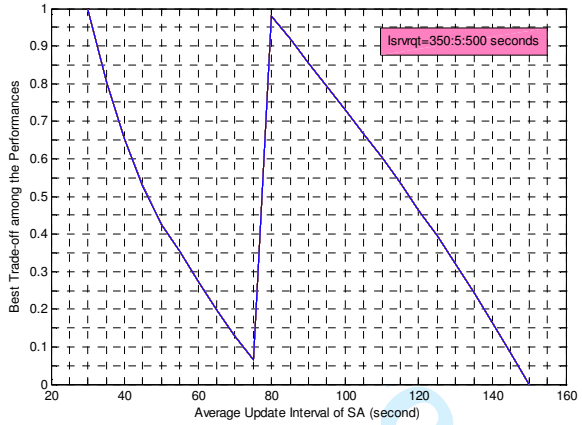


Fig. 7. Best trade-off for different values of s.

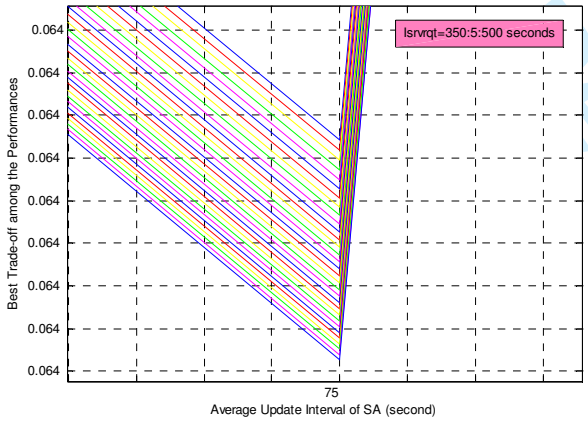
2) The Influence of the Mean Interval between Two Adjacent SrvRqts

We set the average number of matching services for each UE s=*1*, P_f=*0.01*, $Weight_1$=$Weight_2$=$Weight_3$=*1* and T_L=*150* seconds. The value of I_{SrvRqt} changes from *350* to *500* seconds and I_{SA} ranges from *30* to *150* seconds. From Fig. 8, we see the value of I_{SrvRqt} does not impact the choice for the optimal value

of I_{SA}. Also, it is presented that at the point of $I_{SA}=T_L/2=75$ seconds, the best trade-off among the performances can be got. Fig. 8(b) is the amplificatory figure of Fig. 8(a) at the point of $I_{SA}=T_L/2=75$ seconds. Here, $T_L=max(I_{SA})$.



(a)



(b)

Fig. 8. Best trade-off for different average intervals between two adjacent SrvRqts.

3) The Influence of Network Condition

Fig. 9 shows the influence of network condition variation to the selection of the optimal value of I_{SA}. Here, $s=1$, $I_{SrvRqt}=350$ seconds, $Weight_1=Weight_2=Weight_3=1$, and $T_L=150$ seconds. Also, set the range of I_{SA} is *30* to *150* seconds. Considering different network conditions (one-hop loss rate in the network P_f changes from 0.01 to 0.05), in order to get the best trade-off among the performances, we only need to set $I_{SA}=T_L/2=75$ seconds. Here, T_L also equals to $max(I_{SA})$. Obviously, the variety of the network condition does not influence the optimal value of I_{SA}.

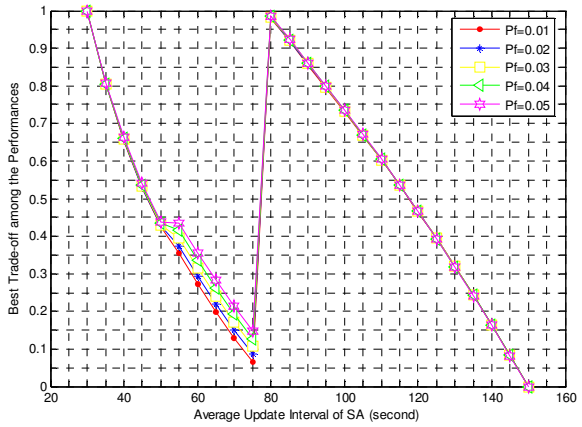


Fig. 9. Best trade-off for different network conditions.

4) The Influence of Setting Different Weights of C_{HSS-SA}, $E(P_{SD})$ and $E(L_{total})$

$Weight_1$, $Weight_2$ and $Weight_3$ reflect the importance of C_{HSS-SA}, $E(P_{SD})$, and $E(L_{total})$ respectively to a certain given IMS-based network. In order to consider all the conditions for the influence of weights, we set the rate among $Weight_1$, $Weight_2$, and $Weight_3$ from *1*:*1*:*1* to *10*:*10*:*10*. Also, we set $s=1$, $I_{SrvRqt}=350$ seconds, $P_f=0.01$, and $T_L=150$ seconds. The value of I_{SA} ranges from *30* to *150* seconds. Apparently, $T_L=max(I_{SA})$. The best trade-off among the performance curves are shown in Fig. 10. The optimal value of I_{SA} is also at the point of $I_{SA}=T_L/2=75$ seconds. That means the weights do not impact the selection for the optimal value of I_{SA}.

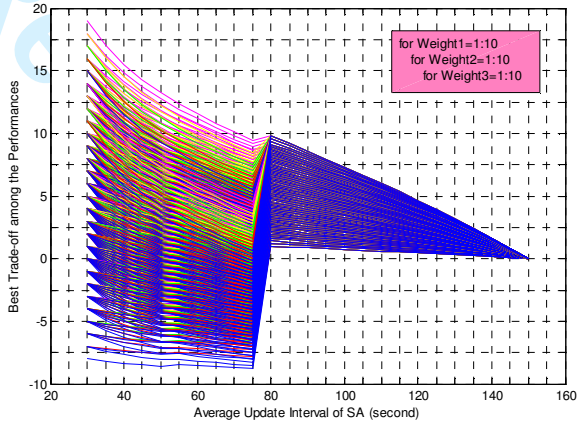


Fig. 10. Best trade-off for different weights of the performances.

5) The Influence of Average Lifetime of SAs

We set $s=1$, $I_{SrvRqt}=350$ seconds, $Weight_1=Weight_2=Weight_3=1$ and $P_f=0.01$. The I_{SA} is ranged from *30* to *150* seconds. From Fig. 11, giving different values to T_L, the setting of T_L influences the optimal value of I_{SA} which can make the performance get the best trade-off.

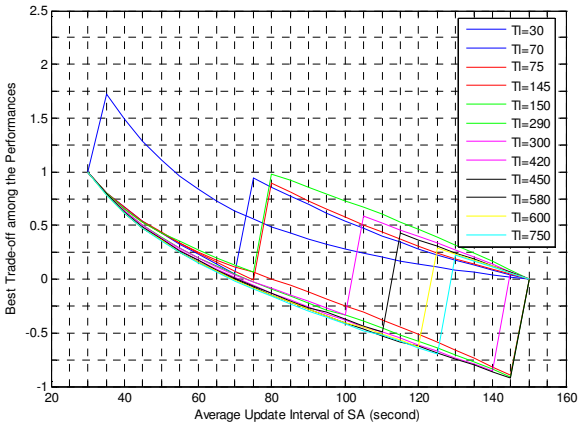


Fig. 11. Best trade-off for different average update intervals of SA.

When $T_L<(1/2)I_{SA}$, the optimal value of I_{SA} is at the point of $I_{SA}=max\ (I_{SA})$. For example, set $T_L=30$ or 70 seconds, the optimal value of I_{SA} is at the point of $I_{SA}=150$ seconds (I_{SA} ranges from 30 to 150 seconds).

When $(1/2)I_{SA}\le T_L<I_{SA}$, the optimal value of I_{SA} is at the point of $I_{SA}=T_L$. For example, set $T_L=75$ or 145 seconds, the optimal value of I_{SA} is at the point of $I_{SA}=75$ or 145 seconds (I_{SA} ranges from 30 to 150 seconds).

When $I_{SA}\le T_L<2I_{SA}$, the optimal value of I_{SA} is at the point of $I_{SA}=(1/2)T_L$. For example, set $T_L=150$ or 290 seconds, the optimal value of I_{SA} is at the point of $I_{SA}=75$ or 145 seconds (I_{SA} ranges from 30 to 150 seconds).

When $2I_{SA}\le T_L<3I_{SA}$, the optimal value of I_{SA} is at the point of $I_{SA}=(1/3)T_L$. For example, set $T_L=300$ or 420 seconds, the optimal value of I_{SA} is at the point of $I_{SA}=100$ or 140 seconds (I_{SA} ranges from 30 to 150 seconds).

When $3I_{SA}\le T_L<4I_{SA}$, the optimal value of I_{SA} is at the point of $I_{SA}=(1/4)T_L$. For example, set $T_L=450$ or 580 seconds, the optimal value of I_{SA} is at the point of $I_{SA}=112$ or 145 seconds (I_{SA} ranges from 30 to 150 seconds).

When $4I_{SA}\le T_L<5I_{SA}$, the optimal value of I_{SA} is at the point of $I_{SA}=(1/5)T_L$. For example, set $T_L=600$ seconds, the optimal value of I_{SA} is at the point of $I_{SA}=120$ seconds (I_{SA} ranges from 30 to 150 seconds).

When $5I_{SA}\le T_L<6I_{SA}$, the optimal value of I_{SA} is at the point of $I_{SA}=(1/6)T_L$. For example, set $T_L=750$ seconds, the optimal value of I_{SA} is at the point of $I_{SA}=125$ seconds (I_{SA} ranges from 30 to 150 seconds).

Based on the principle of mathematical induction, we conclude that when $nI_{SA}\le T_L<(n+1)I_{SA}$, the optimal value of I_{SA} is at the point of $I_{SA}=(1/(n+1))T_L$. Here, n can be any arbitrary integer.

We use (36) to express the rule of selecting the optimum I_{SA},

$$(I_{SA})_{optimum}=\begin{cases}\max(I_{SA}), & T_L<\frac{1}{2}\max(I_{SA})£\quad\neg\\ T_L, & \frac{1}{2}\max(I_{SA})\le T_L<\max(I_{SA})£\quad\neg\\ \frac{1}{n+1}T_L, & n\max(I_{SA})\le T_L<(n+1)\max(I_{SA}),\end{cases}\tag{36}$$

To a certain given IMS-based network, the selection of optimum I_{SA} is just related to the value of T_L. However, T_L and I_{SA} are correspondingly fixed and all can be set by ASs, so the value of the optimum I_{SA} is easy to be found. We coin the point of the optimum I_{SA} as the stationary point of I_{SA}. It can be controlled by the network operator to realize the whole IMS-based network optimization and make all the network performances trade-off. Comparing optimizing the network by controlling the average interval between two adjacent SrvRqts (determined by the users) and the network condition (real-time changing), it is a very easy and useful policy for the network operator to control and adjust the network according to some certain parameters and requirements of the network to get the optimal network performance.

V. Conclusion

This paper proposed an SDA based on IMS which can provide an USD function for different types of services. And the services depend on different network access technologies. The proposed SDA configures a flexible, fully integrated SD mechanism, which surpasses traditional SD techniques and can satisfy the requirements imposed by the new network architectural trends for future NGNs. It can facilitate common use of services in different systems.

In view of being supported by IMS entities and functionalities, the SDA can be integrated well in IMS-based networks. As the directories, HSSs are considered to use, but they need not to broadcast their presences to others. Also, SLF is used for searching neighbor HSSs to find desired services. By mathematical modeling, theoretical analyses, and numerical simulations, we do the performance evaluations and optimizations which are mainly focused on the usage of cache size in HSS for SAs, the success rate of SD, and the traffic load caused by SD procedure. All of the analyses are based on the variation of the mean update interval of SA, the average interval between two adjacent SrvRqts, and the network conditions. More important, we find a special point, an optimal value of the average update interval of SA, which can make all the performances get trade-off. Moreover, the value is relatively stationary and controllable to a specific IMS-based network, and can be evaluated and set by the network operator.

REFERENCES

[1] 3GPP TS 23.228, "IP Multimedia Subsystem (IMS); Stage 2," Rel. 8, V8.4.0, March 2008.

[2] G. Camarillo and M.A. Garcia Martin, *The 3G IP Multimedia Subsystem (IMS): Merging the Internet and the Cellular Words*, Second Edition, Wiley, 2006.

[3] 3GPP TS 22.228, "Service Requirement for the Internet Protocol (IP) Multimedia Core Network Subsystem; Stage 1," Rel. 8, V8.4.0, March 2008.

[4] ITU-T FG NGN 9, "Definition of the Base for Access Independent IMS Core," June 2004.

[5] ITU-T Recommendation Y.2021, "IMS for Next Generation Networks," September 2006.

[6] C. I. Katsigiannis, D.A. Kateros, E. A. Koutsoloukas, N. L. D. Tselikas, and I. S. Venieris, "Architecture for Reliable Service Discovery and Delivery in MANETs Based on Power Management Employing SLP Extensions," *IEEE Wireless Communications*, vol. 13, pp. 90–95, October 2006.

[7] Y. Chen and Z. K. Mi, "A Novel Service Discovery Mechanism in MANET Using Auto-configured SDA," *Proc. of IEEE WiCom 2007*, pp. 1660–1663, September 2007.

[8] J. C. Liang, J. C. Chen, and T. Zhang, "Mobile Service Discovery Protocol (MSDP) for Mobile Ad-Hoc Networks," *Proc. of IEEE ISADS 2007*, pp. 352-362, March 2007.

[9] H. Song, D. Cheng, A. Messer, and S. Kalasapur, "Web Service Discovery Using General-Purpose Search Engines," *Proc. of IEEE ICWS 2007*, pp. 265-271, July 2007.

[10] Y. C. Tao, H. Jin, X. H. Shi, and L. Qi, "GNSD: A Novel Service Discovery Mechanism for Grid Environment," *Proc. of IEEE NWeSP 2006*, pp.17-26, September 2006.

[11] C. Lee, A. Helal, N. Desai, V. Verma, and B. Arslan, "Konark: A System and Protocols for Device Independent, Peer-to-Peer Discovery and Delivery of Mobile Services," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 33, pp. 682-696, November 2003.

[12] W. Louati and D. Zeghlache, "SPSD: A Scalable P2P based Service Discovery Architecture," *Proc. of IEEE WCNC 2007*, pp. 2586-2591, March 2007.

[13] T. Wu and G. S. Kuo, "An Analytical Model for Service Discovery Architectures in Next-Generation Networks," *Proc. of IEEE GLOBECOM 2006*, pp. 1–5, November 2006.

[14] D. Charkraborty, et al., "Queuing Theoretic Model for Service Discovery in Ad-hoc Networks," *Proc. of CNDS*, January 2004.

[15] T. Wu and G. S. Kuo, "An Analytical Model for Centralized Service Discovery Architecture in Wireless Networks," *Proc. of IEEE VTC-2006 Fall*, pp. 1–5, September 2006.

[16] G. S. Kuo, T. Wu, X. Zhang, and G. Li, "Location-based Service Discovery System for Next-Generation IMS in Beyond-3G Converged Networks," *Proc. of WWRF 15*, December 2005.

[17] A. D. Jun, et al., "An IMS-Based Service Platform for the Next-Generation Wireless Networks," *IEEE Communications Magazine*, vol. 44, no. 9, pp. 88–95, September 2006.

[18] R. Levenshtayn and I. Fikouras, "Mobile Services Interworking for IMS and XML WebServices," *IEEE Communications Magazine*, vol. 44, pp. 80–87, September 2006.

[19] A. O. Allen, *Probability, Statistics, and Queuing Theory: with Computer Science Applications, Second Edition*, Academic Press, New York, 1997.

[20] A. Ruszczynski, *Nonlinear Optimization*, Princeton University Press, 2006.