Caching in I-CSCF of UMTS IP Multimedia Subsystem

Yi-Bing Lin, *Fellow, IEEE*, and Meng-Hsun Tsai, *Student Member, IEEE*

***Abstract*—The IP multimedia core network subsystem (IMS) provides multimedia services for Universal Mobile Telecommunications System (UMTS). In IMS, any incoming call will first arrive at the interrogating call session control function (I-CSCF). The I-CSCF queries the home subscriber server (HSS) to identify the serving CSCF (S-CSCF) of the called mobile user. The S-CSCF then sets up the call to the called mobile user. This paper investigates the performance of the IMS incoming call setup. We also propose cache schemes with fault tolerance to speed up the incoming-call-setup process. Our study indicates that the I-CSCF cache can significantly reduce the incoming-call-setup delay, and checkpointing can effectively enhance the availability of I-CSCF.**

***Index Terms*—Call session control function (CSCF), general packet radio service (GPRS), home subscriber server (HSS), IP multimedia core network subsystem (IMS), registration, universal mobile telecommunications system (UMTS).**

I. Introduction

UNIVERSAL Mobile Telecommunications System (UMTS) is one of the major standards for the third-generation (3G) mobile telecommunications. In UMTS, the IP multimedia core network subsystem (IMS) provides multimedia services by utilizing the Session Initiation Protocol (SIP) [4]. Fig. 1 illustrates a simplified UMTS network architecture (the reader is referred to [1], [3], [5], and [6] for the detailed descriptions). This architecture consists of a radio access network, the general packet radio service (GPRS) core network and the IMS network. The GPRS core network connects to the IMS network through gateway GPRS support nodes (GGSNs). The home subscriber server (HSS) is the master database containing all user-related subscription information. Both the GPRS and the IMS networks access the HSS for mobility management and session management. A mobile user utilizes a mobile station (MS) or user equipment (UE) to access IMS services. To provide a data session for the UE, a connection between the UE and the GGSN is established. This connection is specified by a Packet Data Protocol (PDP) context. The PDP context must be activated before a UE can access the IMS network. The IMS user data traffic is transported through the media gateways (MGWs). The IMS signaling is carried out by proxy call session control function (P-CSCF), interrogating CSCF (I-CSCF), and serving CSCF (S-CSCF). The I-CSCF determines how to route incoming calls to the S-CSCF and then to the destination UEs. That is, the I-CSCF is the contact point for the IMS network of the destination UE, which may be used to hide the configuration, capacity, and topology of the IMS network from the outside world. When a UE attaches to the GPRS/IMS network and performs PDP context activation, a P-CSCF is assigned to the UE. The P-CSCF contains limited address translation functions to forward the requests to the I-CSCF. Authorization for bearer resources in the network (where the UE visits) is performed by the P-CSCF. By exercising the IMS registration, to be described in Section II, an S-CSCF is assigned to serve the UE. This S-CSCF supports the signaling interactions with the UE for call setup and supplementary-services control (e.g., service request and authentication). This paper investigates the performance of the IMS incoming call setup. Specifically, we propose cache schemes with fault tolerance to speed up the incoming-call-setup process.

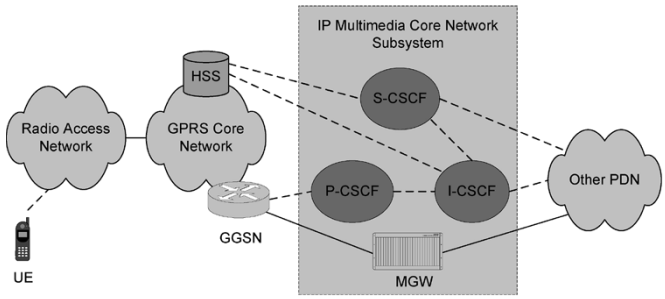

Fig. 1. UMTS network architecture.

II. IMS Registration and Call Setup

This section describes the registration and the incoming-call-setup procedures for UMTS IMS. We first elaborate on the basic scheme proposed in 3G Partnership Project (3GPP) 23.228 [3]. Then, we propose a cache scheme and two checkpoint schemes that speed up the incoming-call-setup process.

A. The Basic Scheme

Suppose that a UE already obtained the IP connectivity through the PDP context activation, and has performed at least one IMS registration. The UE may issue reregistration due to,

Manuscript received October 6, 2003; revised September 25, 2004; accepted September 25, 2004. The editor coordinating the review of this paper and approving it for publication is Y. Fang. This paper was sponsored in part by National Science Council (NSC) Excellence project NSC93-2752-E-0090005-PAE, by NSC 93-2213-E-009-100, by National Telecommunication Development Program (NTP) Voice over IP (VoIP) Project under Grant NSC 92-2219-E-009-032, by IIS/ACADEMIA Sinica, and by the Industrial Technology Research Institute (ITRI)/National Chiao Tung University (NCTU) Joint Research Center.

The authors are with the Department of Computer Science and Information Engineering, National Chiao Tung University, Hsinchu 30050, Taiwan, R.O.C. (e-mail: liny@csie.nctu.edu.tw; tsaimh@csie.nctu.edu.tw).

Digital Object Identifier 10.1109/TWC.2005.858332

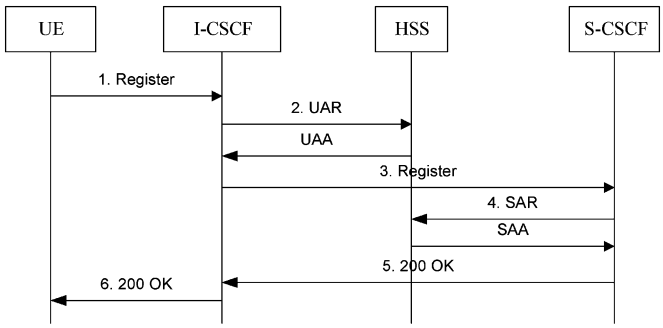

Fig. 2. Registration procedure for the basic scheme (B-RP).

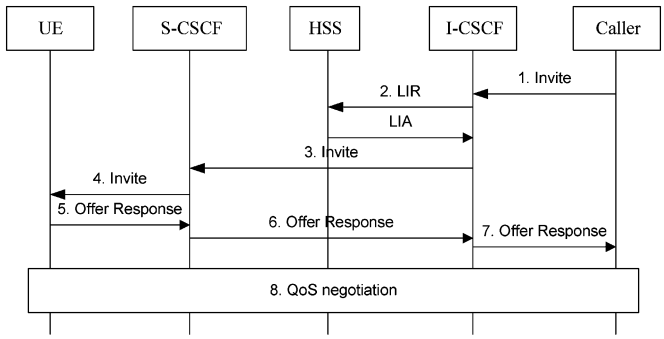

Fig. 3. Incoming call setup for the basic scheme (B-ICS).

for example, movement among different service areas. Fig. 2 illustrates the registration message flow for the basic scheme [called basic registration procedure (B-RP)] defined in 3GPP [2], [3], which includes the following steps.

Step 1) The UE issues the Register message to the I-CSCF through the P-CSCF (not shown in Fig. 2).

Step 2) The I-CSCF exchanges the User-Authorization-Request (UAR) and User-Authorization-Answer message pair with the HSS to obtain the S-CSCF name for the UE.

Step 3) By using a name–address resolution mechanism, I-CSCF identifies the S-CSCF address and sends the Register message to the S-CSCF.

Step 4) Through the Server-Assignment-Request (SAR) and Server-Assignment-Answer (SAA) message-pair exchange between the S-CSCF and the HSS, the S-CSCF obtains the user profile of the UE from the HSS. The user profile will be used in call setup.

Steps 5) and 6) The 200 OK message is sent from the S-CSCF to the I-CSCF, and then from the I-CSCF to the UE, which indicates that the registration is complete.

The IMS incoming call setup defined in 3GPP [2], [3] referred to as the basic incoming call setup (B-ICS), is illustrated in Fig. 3 with the following steps.

Step 1) The caller sends the Invite message to the I-CSCF. The initial media description offered in the Session Description Protocol (SDP) is contained in this message.

Step 2) The I-CSCF exchanges the Location-Info-Request (LIR) and Location-Info-Answer (LIA) message pair with the HSS to obtain the S-CSCF name for the destination UE.

Steps 3) and 4) The I-CSCF forwards the Invite message to the S-CSCF. Based on the user profile of the destination UE, the S-CSCF invokes whatever service logic is appropriate for this session setup attempt. Then, it sends the Invite message to the UE (through the P-CSCF of the IMS network where the UE resides).

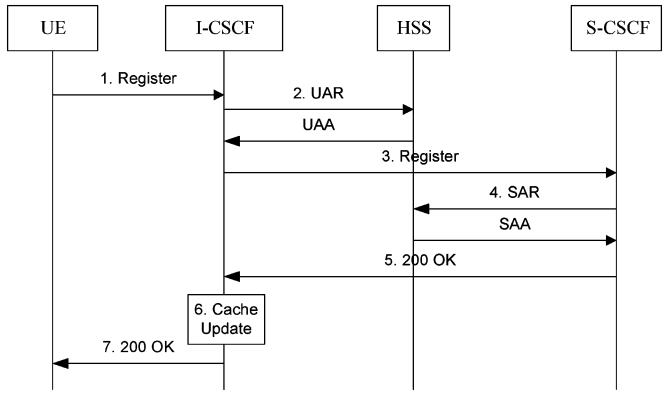

Fig. 4. Registration with cache update for the C (C1 and C2) schemes (C-RP).

Steps 5–7) The UE responds with an answer to the offered SDP. This Offer Response message is passed along the established session path back to the caller.

Step 8) The quality of service (QoS) for this call is negotiated between the originating network (of the caller) and the terminating network (of the UE), and the details are omitted.

B. The Cache Schemes

This paper utilizes a cache at the I-CSCF to speed up the incoming-call-setup process. We first describe a basic cache scheme. To enhance availability and reliability, we then consider two checkpoint schemes that immediately recover the I-CSCF cache after an I-CSCF crash (failure).

1) The Basic Cache Scheme (The C Scheme): Fig. 4 illustrates the registration message flow for the C scheme (called C-RP). C-RP is the same as B-RP except that when the 200 OK is sent from the S-CSCF to the I-CSCF, the (UE, S-CSCF) mapping (called the S-CSCF record) is saved in the cache [step 6) in Fig. 4]. The incoming call setup for the C scheme (C-ICS) is illustrated in Fig. 5. In C-ICS, the LIR and LIA message pair exchanged [step 2) of B-ICS in Fig. 3] is replaced by a cache retrieval [step 2), Fig. 5] to obtain the S-CSCF address. If an I-CSCF failure occurs and the whole cache content is lost, then the S-CSCF records are gradually rebuilt through the IMS registration procedure. If an incoming call arrives earlier than the registration, then B-ICS must be executed to set up the call.

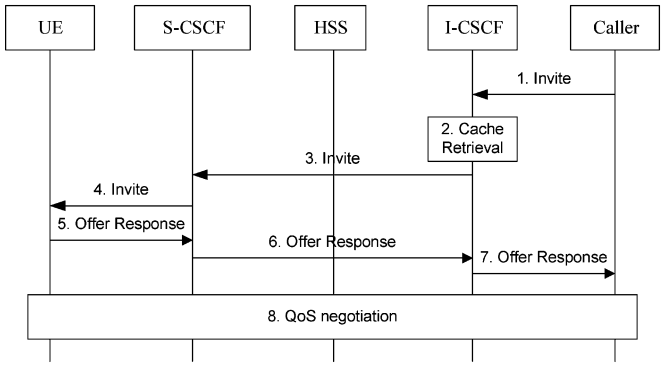

Fig. 5. Incoming call setup with cache retrieval for the C (C1 and C2) schemes (C-ICS).

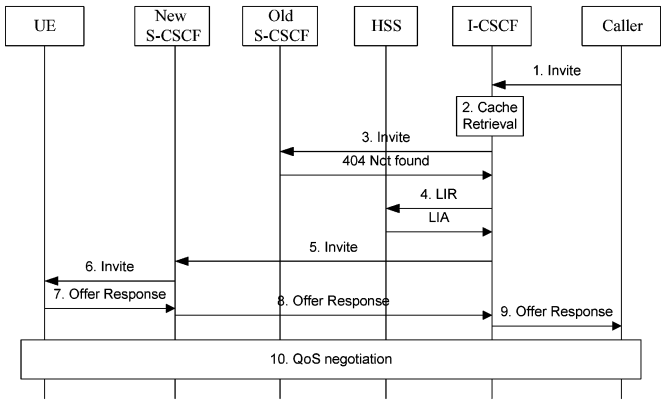

Fig. 6. First incoming call setup after I-CSCF failure: cache miss for the checkpoint-1 scheme.

2) The Checkpoint Scheme 1 (The C1 Scheme): To immediately recover the I-CSCF cache after a failure, we may save the content of the cache (only for the modified records) into a backup storage. In the C1 scheme, we periodically save the cache into the backup. When an I-CSCF failure occurs, the cache content is restored from the backup. Therefore, in the normal operation, the registration procedure and the incoming-call-setup procedure for the C1 scheme is the same as that for the C scheme. After a failure, the incoming-call-setup procedure is the same as C-ICS (see Fig. 5) if the S-CSCF is up to date (called a cache hit). Note that between a failure and the previous checkpoint, the S-CSCF record of an MS may be modified. In this case, the S-CSCF may be obsolete when an incoming call arrives (called a cache miss), and the call-setup message flow is illustrated in Fig. 6. The first three steps of this message flow are the same as C-ICS. Since the UE already moves from the old S-CSCF to the new S-CSCF, at the end of step 3), the old S-CSCF replies the 404 Not Found message to the I-CSCF. The I-CSCF then retrieves the new S-CSCF information from the HSS and sets up the call following steps 2)–8) of B-ICS in Fig. 3.

3) The Checkpoint Scheme 2 (The C2 Scheme): It is clear that after an I-CSCF failure, the call-setup cost for the C1 scheme is very expensive if a cache miss occurs. We resolve this issue by introducing the C2 scheme. Like the C1 scheme, this scheme periodically checkpoints the cache content into the backup. Furthermore, an S-CSCF record in the backup is invalidated if the corresponding record in the cache is modified. The C2 registration procedure is the same as C-RP except for step 6) in Fig. 4. In this step, we check if the S-CSCF record at the backup has been invalidated since the last checkpoint. If so, no extra action is taken. If not, the record in the backup is invalidated. Therefore, if multiple registrations for the same UE occur between two checkpoints, the S-CSCF record in the backup is only invalidated for the first registration. After an I-CSCF failure, the C2 scheme knows exactly which S-CSCF records are invalid. For the first incoming call after the failure, if the S-CSCF record is valid, then the call-setup procedure follows C-ICS in Fig. 5. On the other hand, if the S-CSCF record is invalid, the procedure (see Fig. 7) follows B-ICS in Fig. 3. Features of the B, C, C1, and C2 schemes are summarized in Tables I and II.

In the remainder of this paper, we will analyze the registration and incoming call setup by using analytic and simulation models. Input parameters used in these models are listed in Table III.

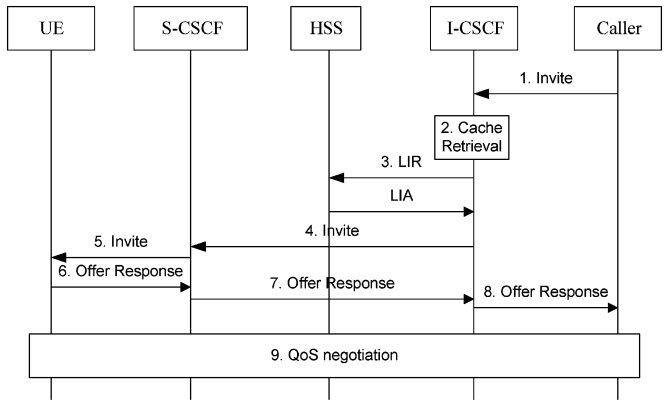

Fig. 7. First incoming call setup after I-CSCF failure: cache miss for the checkpoint-2 scheme.

TABLE I
CACHING AND CHECKPOINTING OPERATIONS

Scheme	Periodic Checkpointing	Backup Record Invalidation	Cache Restoration after I-CSCF Failure
Basic	no	no	no
Cache	no	no	no
Checkpoint 1	yes	no	yes
Checkpoint 2	yes	yes	yes

TABLE II
IMS REGISTRATION AND CALL SETUP

Scheme	Registration	Normal Incoming Call Setup	First Incoming Call Setup after Failure
Basic	B-RP	B-ICS	B-ICS
Cache	C-RP (B-RP + cache update)	C-ICS (B-ICS without HSS query)	B-ICS
Checkpoint 1	C-RP	C-ICS	Cache hit: C-ICS; Cache miss: B-ICS plus extra access to S-CSCF
Checkpoint 2	C-RP + possible backup record invalidation	C-ICS	Cache hit: C-ICS; Cache miss: B-ICS

TABLE III
NOTATION (INPUT PARAMETERS)

μ	the checkpoint frequency
λ	the registration arrival rate
V	the variance of the inter-registration intervals
γ	the incoming call arrival rate
$1/\delta$	the mean transmission delay between two IMS nodes
$V1$	the variance of the transmission delay between two IMS nodes

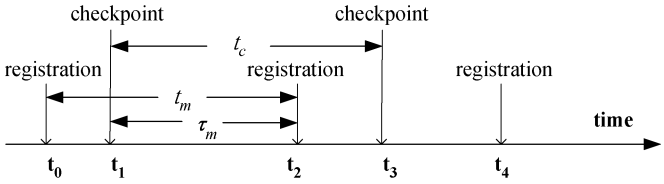

Fig. 8. Timing diagram for registration and checkpointing.

III. OVERHEAD OF CHECKPOINTING

This section investigates the costs for the C1 and the C2 schemes. Fig. 8 illustrates the timing diagram for the registration and checkpointing activities of a UE. At t_0, t_2, and t_4, the UE issues registration requests either because it attaches to the network, or it moves from one service area to another service area. The inter-registration intervals $t_2 - t_0$, $t_4 - t_2$, etc., are represented by a random variable t_m. In this figure, periodic checkpoints are performed at t_1 and t_3, where the checkpointing interval is represented by a random variable t_c. The interval τ_m between a checkpoint and the next registration (e.g., $t_2 - t_1$) is called the excess life of the inter-registration interval. At a checkpoint, only the modified S-CSCF records are saved into the backup. Let $p_{\rm u}$ be the probability that the S-CSCF record for the UE is saved at a checkpoint, then

$$p_{\rm u} = \Pr[t_c > \tau_m].$$

It is clear that the checkpoint cost increases as $p_{\rm u}$ increases. Two types of checkpoint intervals are often considered. Fixed checkpointing performs checkpoints with fixed period $1/\mu$. In exponential checkpointing, the inter-checkpointing interval has an exponential distribution with mean $1/\mu$. Assume that the inter-registration intervals t_m have an exponential distribution with mean $1/\lambda$ (i.e., the registration stream forms a Poisson process). For an arbitrary time interval T, the number X of registrations occurring in this period has a Poisson distribution. That is

$$\Pr[X = x, T = t] = \left[\frac{(\lambda t)^x}{x!}\right] {\rm e}^{-\lambda t} \tag{1}$$

and

$$\Pr[t_c > \tau_m] = 1 - \Pr[X = 0, T = t_c]. \tag{2}$$

From (2), the probability $p_{\rm u}$ for fixed checkpointing is expressed as

$$p_{\rm u} = 1 - {\rm e}^{-\frac{\lambda}{\mu}} \quad \text{(fixed checkpointing)}. \tag{3}$$

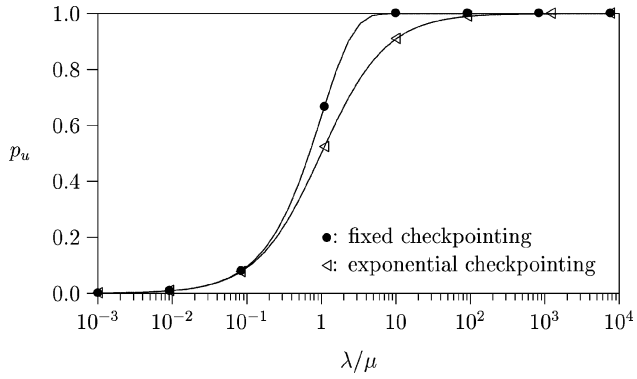

Fig. 9. Comparing fixed and exponential checkpointing (Poisson registration stream with the rate λ).

Similarly, the probability $p_{\rm u}$ for exponential checkpointing is expressed as

$$\begin{aligned} p_{\rm u} &= 1 - \int_{t_c=0}^{\infty} {\rm e}^{-\lambda t_c} \mu {\rm e}^{-\mu t_c} dt_c \\ &= \frac{\lambda}{\lambda + \mu} \text{ (exponential checkpointing)}. \end{aligned} \tag{4}$$

Fig. 9 plots $p_{\rm u}$ for fixed and exponential checkpointing approaches based on (3) and (4). The figure indicates that $p_{\rm u}$ for fixed checkpointing is larger than that for exponential checkpointing (i.e., exponential checkpointing yields better performance than fixed checkpointing). In the remainder of this paper, we only consider exponential checkpointing. General conclusions drawn from this paper also apply to fixed checkpointing. Consider the inter-registration interval random variable t_m with mean $1/\lambda$, density function $f(\cdot)$, and Laplace transform $f^*(s)$. The excess life τ_m has a distribution function $R(\cdot)$, density function $r(\cdot)$, and Laplace transform $r^*(s)$.

Since exponential checkpointing is a Poisson process, t_1 in Fig. 8 is a random observer of the t_m intervals. From the excess-life theorem [7]

$$r^*(s) = \left(\frac{\lambda}{s}\right)[1 - f^*(s)]. \tag{5}$$

Based on (5), we derive $p_{\rm u}$ as

$$\begin{aligned} p_{\rm u} &= \Pr[t_c > \tau_m] \\ &= \int_{t_c=0}^{\infty} \int_{\tau_m=0}^{t_c} r(\tau_m) \mu {\rm e}^{-\mu t_c} d\tau_m dt_c \\ &= \int_{t_c=0}^{\infty} R(t_c) \mu {\rm e}^{-\mu t_c} dt_c \\ &= \left.\frac{\mu r^*(s)}{s}\right|_{s=\mu} \\ &= \left(\frac{\lambda}{\mu}\right)[1 - f^*(\mu)]. \end{aligned} \tag{6}$$

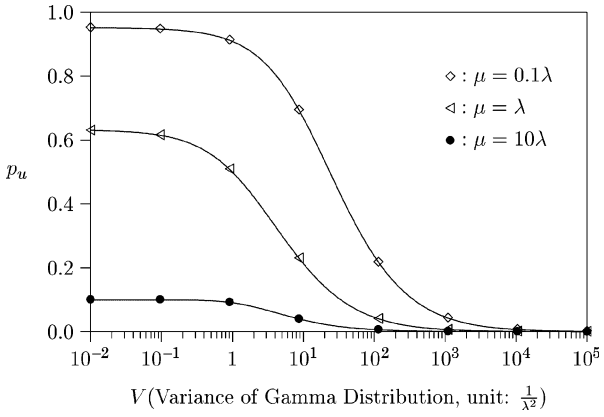

Fig. 10. Effects of the variance of the inter-registration intervals on $p_{\rm u}$.

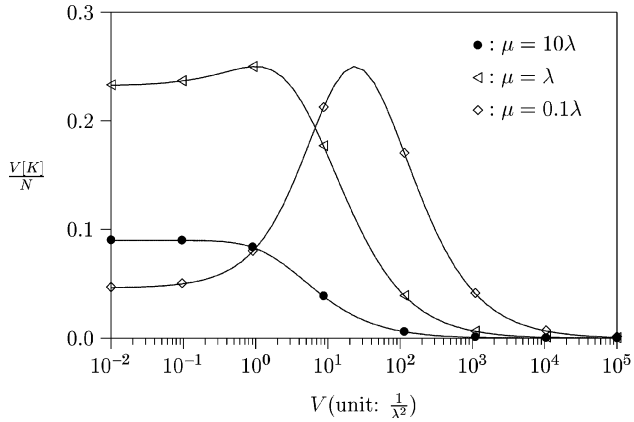

Fig. 11. Variance of the K distribution.

Assume that t_m is a Gamma random variable with mean $1/\lambda$, variance V, and Laplace transform

$$f^*(s) = \left(\frac{1}{V\lambda s + 1}\right)^{\frac{1}{V\lambda^2}}. \tag{7}$$

Then, (6) is rewritten as

$$p_{\rm u} = \left(\frac{\lambda}{\mu}\right)\left[1 - \left(\frac{1}{V\lambda\mu + 1}\right)^{\frac{1}{V\lambda^2}}\right] \quad (\text{for Gamma-distributed } t_m). \tag{8}$$

When t_m is exponentially distributed, $V = 1/\lambda^2$, and (8) is rewritten as $p_{\rm u} = \lambda/(\lambda + \mu)$, which is the same as (4). Fig. 10 plots $p_{\rm u}$ for Gamma inter-registration intervals with different variance values. The figure indicates that $p_{\rm u}$ decreases as the variance V increases. This phenomenon is explained as follows. When the registration behavior becomes more irregular, we will observe more checkpoint intervals with many registrations and more checkpoint intervals without any registration. In other words, smaller $p_{\rm u}$ is observed. Therefore, the checkpointing performance is better when the registration activity becomes more irregular (i.e., V is larger).

Suppose that N UEs have registered to the IMS network, then there are N S-CSCF records in the I-CSCF cache. Let $\Pr[K = k]$ be the probability that k records are modified between two checkpoints. Then

$$\Pr[K = k] = \binom{N}{k} p_{\rm u}^k (1 - p_{\rm u})^{N-k} \tag{9}$$

is a binomial probability mass function, and the random variable K has mean $E[K] = Np_{\rm u}$ and variance $V[K] = Np_{\rm u}(1 - p_{\rm u})$. A small $V[K]$ value implies that the S-CSCF record-saving overheads for the checkpoint intervals are roughly the same (which provides stable, i.e., better performance for the checkpointing system). The $E[K]$ curves have the same shapes as that for the $p_{\rm u}$ curves (see Fig. 10). That is, the S-CSCF record-saving cost at a checkpoint decreases as the variance of t_m increases. The $V[K]$ curves are illustrated in Fig. 11. For small μ values, (e.g., $\mu \le \lambda$), the variance of K increases and then decreases as the variance of t_m increases. For large μ (i.e., $\mu \ge 10\lambda$), the variance of K decreases as the variance of t_m increases.

IV. Costs for Incoming Call Setup

This section investigates the incoming-call-setup costs. We first study the normal incoming call setup. Then, we analyze the first incoming call setup after I-CSCF failure. We note that the checkpoint action is a background process, and the cost for retrieving the cache [e.g., step 6) in Fig. 4] and the cost for saving an S-CSCF record into the backup is negligible as compared with the communications cost between I-CSCF and HSS (the I-CSCF-cache operation is typically 1000 times faster than the inter-I-CSCF and HSS communications). Therefore, we will ignore the I-CSCF-cache-operation costs in the incoming-call-setup study.

A. Normal Incoming Call Setup

Let $t_{\rm H}$ be the round-trip transmission delay between the I-CSCF and the HSS, $t_{\rm S}$ be the round-trip delay between the I-CSCF and the S-CSCF, and $t_{\rm M}$ be the round-trip delay between the S-CSCF and the UE. Let T_x be the incoming-call-setup delay from the I-CSCF to the UE (without QoS negotiation) for the "x" scheme (where $x \in \{B, C, C1, C2\}$).

Consider the B scheme. In Fig. 3, $t_{\rm H}$ is the delay for step 2); $t_{\rm S}$ is the delay for steps 3) and 6); and $t_{\rm M}$ is the delay for steps 4) and 5). We have

$$T_{\rm B} = t_{\rm H} + t_{\rm S} + t_{\rm M}.$$

In Fig. 5, $t_{\rm S}$ is the delay for steps 3) and 6); and $t_{\rm M}$ is the delay for steps 4) and 5). We have

$$T_{\rm C} = T_{\rm C1} = T_{\rm C2} = t_{\rm S} + t_{\rm M}.$$

Suppose that $t_{\rm H}$, $t_{\rm S}$, and $t_{\rm M}$ have the same distribution with density function $f_{\rm d}(\cdot)$ and mean $1/\delta$. It is clear that

$$E[T_{\rm C}] = E[T_{\rm C1}] = E[T_{\rm C2}] = \frac{2E[T_{\rm B}]}{3}.$$

In other words, the cache/checkpoint schemes can save 33% of the incoming-call-setup overhead (between the I-CSCF and the UE). Furthermore, a timeout timer is typically maintained

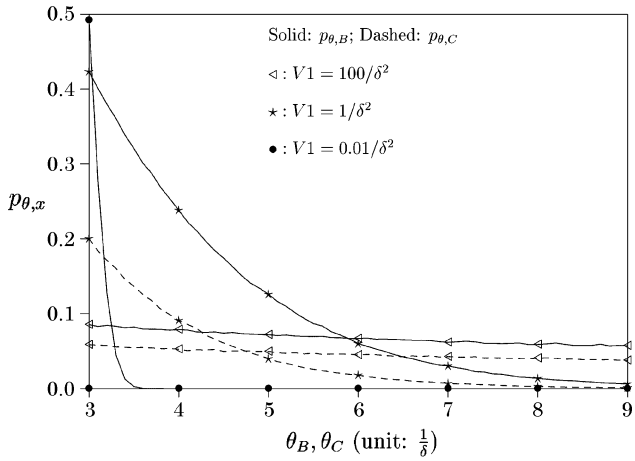

Fig. 12. Comparing the timeout thresholds for the B and C (C1 and C2) schemes ($V1$: the variance of the gamma transmission delay).

in the I-CSCF. For a call setup, if the I-CSCF does not receive the Offer Response message within a timeout period, the call is considered lost in the I-CSCF (i.e., the call is aborted). If the timeout period is set too short, then many normal incoming call setups may be misleadingly terminated due to timeouts. If the timeout threshold is set too long, then many incomplete call setups will not be detected early. Let θ_x be the timeout threshold for the "x" scheme (where $x \in \{B, C, C1, C2\}$), and

$$p_{\theta,x} = \Pr[T_x > \theta_x] \tag{10}$$

is the probability that a call setup is misleadingly aborted because its transmission delay is longer than the timeout period. Suppose that $t_{\rm M}$, $t_{\rm S}$, and $t_{\rm H}$ have the same Gamma density function $f_{\rm d}(\cdot)$ with mean $1/\delta$. We utilize the simulation approach to compute $p_{\theta,x}$. Specifically, we repeatedly generate the sum of two gamma random numbers (for $T_{\rm C}$) and the sum of three gamma random numbers (for $T_{\rm B}$). Then, we derive $p_{\theta,{\rm B}}$ and $p_{\theta,{\rm C}}$ by using (10). Fig. 12 compares $p_{\theta,{\rm B}}$ with $p_{\theta,{\rm C}}$ $(= p_{\theta,{\rm C1}} = p_{\theta,{\rm C2}})$. The figure shows the intuitive results that when the variance $V1$ of the transmission delay is small, the performance of the timeout mechanism is better (i.e., $p_{\theta,x}$ is small). The figure also indicates that to ensure the same $p_{\theta,x}$ values, the timeout period for the B scheme is much longer than the C (C1 and C2) schemes. For example, when $V1 = 100/\delta^2$, to ensure $p_{\theta,{\rm B}} = p_{\theta,{\rm C}} = 5.8\%$, $\theta_{\rm B} = 9/\delta$ and $\theta_{\rm C} = 3/\delta$ (i.e., the timeout threshold for the B scheme must be set three times as large as that for the C scheme).

B. First Incoming Call Setup After Failure

For the "x" scheme (where $x \in \{B, C, C1, C2\}$), let T_x^* be the round-trip transmission delay of the first incoming call setup after an I-CSCF failure. The delay T_x^* is derived as follows. For the B scheme, the first incoming call setup is not affected by the I-CSCF failure. That is

$$T_{\rm B}^* = T_{\rm B} = t_{\rm H} + t_{\rm S} + t_{\rm M}. \tag{11}$$

For the C scheme, the cache in the I-CSCF is cleared after the failure. If the first event after the failure is an incoming call,

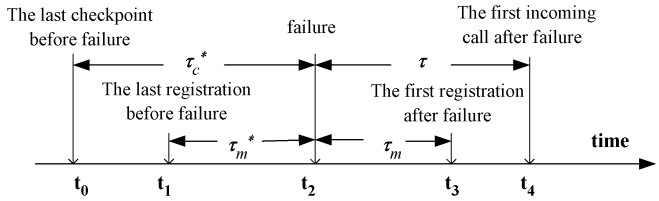

Fig. 13. Timing diagram before and after an I-CSCF failure.

then $T_{\rm C}^* = T_{\rm B}^*$. If the first event is a registration event, then the S-CSCF record is restored in the cache before the first incoming call arrives. When the incoming call arrives, $T_{\rm C}^* = T_{\rm C}$. Let α be the probability that the first event after I-CSCF failure is a registration. Then

$$T_{\rm C}^* = \alpha T_{\rm C} + (1-\alpha) T_{\rm B} = T_{\rm C} + (1-\alpha) t_{\rm H}. \tag{12}$$

The probability α is derived as follows. Consider the timing diagram in Fig. 13. Suppose that an I-CSCF failure occurs at time t_2. For a UE, the first registration after the failure occurs at t_3 and the first incoming call after the failure occurs at t_4. Let $\tau = t_4 - t_2$ and $\tau_m = t_3 - t_2$. Since the failure-occurring time t_2 is a random observer of the inter-call arrival times and the inter-registration times, τ is the excess life of an inter-call arrival time. Also, τ_m is the excess life of an inter-registration time. Suppose that the call arrivals to the UE are a Poisson process with the rate γ. Then, from the excess-life theorem [7], τ has an exponential distribution with mean $1/\gamma$. Similarly, τ_m has the density function $r(\cdot)$ and Laplace transform $r^*(s)$, as expressed in (5). Therefore, similar to the derivation for (6)

$$\alpha = \Pr[\tau > \tau_m] = \left(\frac{\lambda}{\gamma}\right)[1 - f^*(\gamma)]. \tag{13}$$

For the C1 scheme, two cases are considered when the first incoming call after the failure occurs.

- Case 1) The S-CSCF record restored from the backup is invalid (with probability β) and the first event after the failure is an incoming call (with probability $1-\alpha$). In this case (see Fig. 6), $T_{\rm C1}^* = T_{\rm B} + t_{\rm S}$.
- Case 2) The S-CSCF record restored from the backup is valid (with probability $1-\beta$), or the restored record is invalid (with probability β) and the first event after the failure is an IMS registration (with probability α). In this case, $T_{\rm C1}^* = T_{\rm C}$.

Based on the above two cases, we have

$$T_{\rm C1}^* = T_{\rm C} + (1-\alpha)\beta(t_{\rm H} + t_{\rm S}). \tag{14}$$

Probability β is derived as follows. In Fig. 13, the last checkpoint before the I-CSCF failure occurs at time t_0. The last registration before the failure occurs at time t_1. From the reverse excess-life theorem [7], $\tau_c^* = t_2 - t_0$ has an exponential distribution with mean $1/\mu$, and $\tau_m^* = t_2 - t_1$ has the density function $r(\cdot)$ and Laplace transform $r^*(s)$. Similar to the derivation for (6)

$$\beta = \Pr\left[\tau_c^* > \tau_m^*\right] = \left(\frac{\lambda}{\mu}\right)[1 - f^*(\mu)] = p_{\rm u}. \tag{15}$$

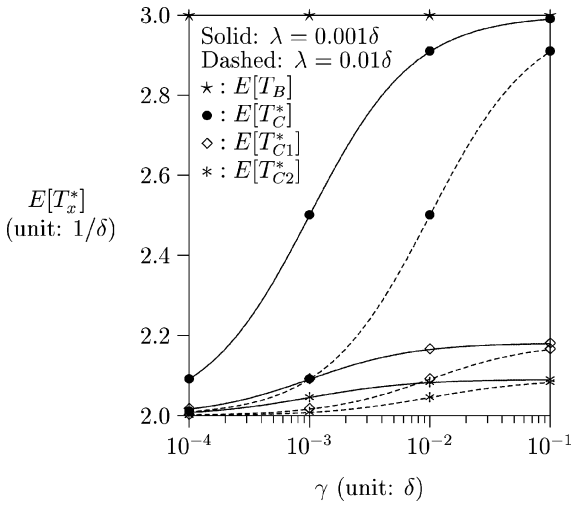

Fig. 14. Transmission delays for the first incoming call after an I-CSCF failure ($\mu = 10\lambda$, $V = 1/\lambda^2$).

For the C2 scheme (see the message flow in Fig. 7), we have

$$T_{\mathrm{C2}}^* = T_{\mathrm{C}} + (1-\alpha)\beta t_{\mathrm{H}}. \tag{16}$$

If $f^*(s)$ is a Gamma Laplace transform as expressed in (7), then from (11), (12), (14), and (16), we have

$$E\left[T_{\mathrm{B}}^*\right] = \frac{3}{\delta}$$

$$E[T_{\mathrm{C}}^*] = \frac{3}{\delta} - \left(\frac{\lambda}{\gamma\delta}\right)\left[1 - \left(\frac{1}{V\lambda\gamma+1}\right)^{\frac{1}{V\lambda^2}}\right]$$

$$E\left[T_{\mathrm{C1}}^*\right] = \frac{2}{\delta} + \left(\frac{2\lambda^2}{\mu\gamma\delta}\right)\left[1 - \left(\frac{1}{V\lambda\mu+1}\right)^{\frac{1}{V\lambda^2}}\right] \times \left[\frac{\gamma}{\lambda} - 1 + \left(\frac{1}{V\lambda\gamma+1}\right)^{\frac{1}{V\lambda^2}}\right]$$

$$E\left[T_{\mathrm{C2}}^*\right] = \frac{2}{\delta} + \left(\frac{\lambda^2}{\mu\gamma\delta}\right)\left[1 - \left(\frac{1}{V\lambda\mu+1}\right)^{\frac{1}{V\lambda^2}}\right] \times \left[\frac{\gamma}{\lambda} - 1 + \left(\frac{1}{V\lambda\gamma+1}\right)^{\frac{1}{V\lambda^2}}\right].$$

Fig. 14 plots the $E[T_x^*]$ curves where $\mu = 10\lambda$ and $V = 1/\lambda^2$. Similar results are observed for different μ and V values, which will not be presented in this paper. The figure indicates that, for C, C1, and C2, $E[T_x^*]$ increases as γ (the incoming-call arrival rate) increases. In this scenario, $E[T_{\mathrm{C1}}^*]$ is limited to $2.2/\delta$ (less than 10% extra overhead for the normal incoming call setup of the C1 scheme), and $E[T_{\mathrm{C2}}^*]$ is limited to $2.1/\delta$ (less than 5% extra overhead for the normal incoming call setup of the C2 scheme). For the C scheme, $E[T_{\mathrm{C}}^*]$ approaches $E[T_{\mathrm{B}}^*]$ as γ increases. The figure also indicates that as the registration rate λ increases, the call-setup time decreases for C, C1, and C2. This phenomenon is due to the fact that α increases as λ increases, and it is more likely that the S-CSCF record is valid when the first incoming call arrives.

V. Conclusion

The IMS network provides multimedia services for UMTS. This paper investigated the performance of the IMS incoming call setup, and proposed cache schemes with fault tolerance to speed up the incoming-call-setup process. Our study indicated that by utilizing the I-CSCF cache, the average incoming-call-setup time can be effectively reduced, and smaller I-CSCF timeout threshold can be set to support early detection of incomplete call setups. To enhance fault tolerance, the I-CSCF cache is periodically checkpointed into a backup storage. When an I-CSCF failure occurs, the I-CSCF cache content can be restored from the backup storage. Since the checkpointing process is conducted in the background, this activity does not affect the incoming-call-setup delays. As a final remark, if both the I-CSCF and the HSS fail, the S-CSCF records can only be recovered from the backup. In this case, our checkpoint schemes can significantly enhance the availability and fault tolerance of the IMS network.

References

[1] 3GPP. 3rd Generation Partnership Project, *Technical Specification Group Services and Systems Aspects; General Packet Radio Service (GPRS); Service Description; Stage 2. Technical Specification 3G TS 23.060 version 3.6.0 (2001-01)*, 2000.

[2] ——. *Technical Specification Core Network; IP Multimedia Subsystem Cx and Dx Interfaces; Signaling Flows and Message Contents (Release 5).Technical Specification 3G TS 29.228 version 5.4.0 (2003-06)*, 2003.

[3] ——. *Technical Specification Group Services and Systems Aspects; IP Multimedia Subsystem Stage 2. Technical Specification 3G TS 23.228 version 6.2.0 (2003-06)*, 2003.

[4] IETF, *SIP: Session Initiation Protocol*, 2002. IETF RFC 3261.

[5] Y.-B. Lin and I. Chlamtac, *Wireless and Mobile Network Architectures*. New York: Wiley, 2001.

[6] Y.-B. Lin, Y.-R. Huang, A.-C. Pang, and I. Chlamtac, "All-IP approach for UMTS third generation mobile networks," *IEEE Network*, vol. 5, no. 16, pp. 8–19, Jan. 2002.

[7] S. M. Ross, *Introduction to Probability Models*. New York: Academic, 1985.

Yi-Bing Lin (M'95–SM'95–F'03) is the Chair Professor of Computer Science and Information Engineering, National Chiao Tung University (NCTU), Hsinchu, Taiwan, R.O.C. His current research interests include wireless communications and mobile computing. He has published over 180 journal articles and more than 200 conference papers. He is the author of the book *Wireless and Mobile Network Architecture* (coauthored with Imrich Chlamtac; published by John Wiley & Sons).

Dr. Lin is a Fellow of the Association for Computing Machinery (ACM).

Meng-Hsun Tsai (S'04) received the B.S. and M.S. degrees from National Chiao Tung University (NCTU), Hsinchu, Taiwan, R.O.C., in 2002 and 2004, respectively. He is currently working toward the Ph.D. degree at NCTU.

His current research interests include design and analysis of personal communications services networks, mobile computing, and performance modeling.