Joint Conference of 1996 International Computer Symposium
December 19~21, Kaohsiung, Taiwan, R.O.C.

Performance Comparison on Multi-Tier Database Systems in Multi-Tier Personal Communications Services Systems

Ruoh-Wen Tzeng#, Jyhi-Kong Wey*, and Wei-Pang Yang#
Institute of Computer and Information Science, National Chiao Tung University, Hsinchu
* Network Planning Laboratory, Chunghwa Telecom. Labs.
Taiwan, ROC
E-mail: jkwey@ms6.hinet.net Tel: +886-2-3445307 Fax: +886-2-3445327

Abstract

Two possible multi-tier database system architectures, Centralized HLR Architecture (CHA) and Decentralized HLR Architecture (DHA), for multi-tier PCS systems are proposed. We also propose mobility management strategies for the two proposed architectures. Under some assumptions, performance metrics are derived on the query response time, the update response time, and the traffic volume produced in the two architectures. Finally, the analysis results of CHA and DHA are made, and the pros and cons of the two proposed architectures are figured out. From the analysis results, it is found that the DHA performs better than the CHA except the update response time.

1. Introduction

Recently, Personal Communications Services (PCS) has become a hot topic. The development of PCS is driven by the evolution of Intelligent Network (IN) [4], the wireless access technologies [1, 13], and the portable (mobile station, MS) technologies [16]. The system integrated with these technologies to provide PCS is called PCS system. Roughly PCS systems can be classified into two classes: one is high-tier PCS systems including AMPS (Advanced Mobile Phone System) and GSM (Global System for Mobile Communications). The other is low-tier PCS systems including PACS (Personal Access Communications Systems) [19] and PHS (Personal Handyphone System).

High-tier PCS systems and low-tier PCS systems have different characteristics [6, 7, 15, 17]. In high-tier PCS systems, high speed mobility is supported. The coverage area (cell) of a base station is large. The call connection charge is expensive. However, the voice quality is poorer than low-tier PCS systems. On the other hand, the usage charge of low-tier PCS systems is cheaper and the voice quality is higher. The microcell coverage area is small so that more base stations are needed. However, it just supports low speed mobility.

Both high-tier PCS systems and low-tier PCS systems have advantages and disadvantages. When a user is roaming whereabouts, he would need different tiers services [8]. To satisfy the need of different tiers services, multi-tier PCS systems have been proposed [6]. A multi-tier PCS system integrates different single-tier PCS systems to provide subscribers access to multiple tiers services. The multi-tier PCS system causes the emergence of multi-tier database system architectures to maintain the users' profiles and location information.

The research issues of this paper are focused on the performance study on multi-tier database systems for PCS. We propose two possible multi-tier database system architectures— Centralized HLR Architecture (CHA) and Decentralized HLR Architecture (DHA) [17]. We also propose mobility management strategies which deal with mobility-related operations [9, 11, 14] such as registration, deregistration, call origination, and call termination for the two proposed architectures. Performance metrics are also derived on the query response time, the update response time, and the traffic volume produced in these two architectures.

The proposed CHA and DHA architectures, and the mobility management strategies for the two architectures are introduced in Section 2. The mathematical analysis for the two architectures is derived, and the analysis results are summarized and discussed in Section 3 and Section 4. Finally, conclusions are made in Section 5.

2. Network architectures and mobility management strategies

The basic and traditional architecture of databases for a single-tier PCS system is referred to two-tier database system in this paper. Conventional PCS systems (e.g., GSM) belong to this case. It consists of

one HLR (Home Location Register) and several VLRs (Visitor location Register) and forms a star-like connection. The HLR in the single-tier PCS system is a centralized database which stores all the information including authentication data, user profiles, call records, and charging information, etc., for all subscribers. As opposed to HLR, the VLR stores just a subset data of HLR. Each VLR mainly caches the portion of data which is related to the location information of mobile subscribers who currently visit the registration areas belonged to this VLR.

The multi-tier PCS systems are the integration of multiple single-tier PCS systems. Therefore, the integration of multiple two-tier database systems forms the multi-tier database systems. In this section, the two analyzed network architectures are described.

2.1. Centralized HLR architecture

The CHA consists of a centralized HLR (CHLR), and several single-tier PCS systems. Each HLR of the single-tier PCS systems are connected to the centralized database CHLR as depicted in Fig. 1. Namely, the CHA is a three-tier database system. The CHLR stores all the information about the subscribers of the multi-tier PCS system. Besides, each single-tier PCS system stores the information about the mobile users who are currently within the system.

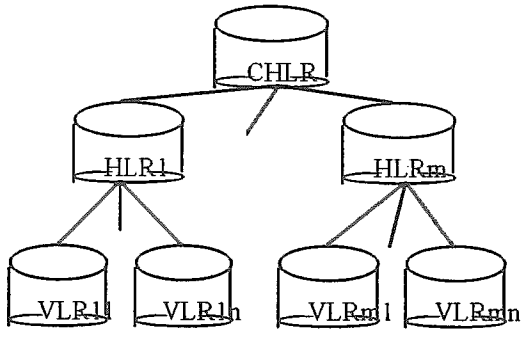

Fig. 1. Centralized HLR Architecture (CHA).

2.2. Decentralized HLR architecture

For a multi-tier PCS system, there is another possible architecture of databases to maintain users profiles and location information as depicted in Fig. 2. Several single-tier PCS systems that are connected through the linking of their HLRs form the DHA architecture. As opposed to CHA, DHA is a decentralized and two-tier database system. In this architecture, each single-tier PCS system stores the information about the mobile users who are currently within it.

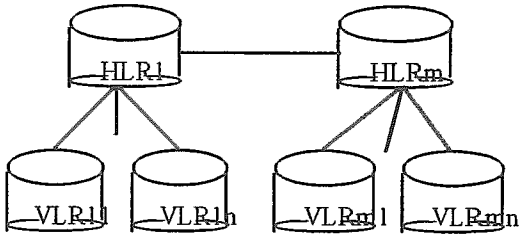

Fig. 2. Decentralized HLR Architecture (DHA).

2.3. Mobility-related operations

The primary and the most important mobility-related operations considered in this paper include registration, deregistration, call origination and call termination. These operations in the single-tier PCS systems are referred to [4, 9, 11]. Not losing generality, we extend GSM operations as an example to illustrate the call control processes (i.e., call origination and call termination) [5, 12].

2.3.1. Registration and deregistration

The scenarios for inter-VLR roaming in CHA and DHA are the same as that in single-tier PCS systems. The multi-tier PCS system introduces inter-HLR roaming, i.e., the mobile user enters a new service area which is belonged to a different HLR. Fig. 3 shows the registration and deregistration processes performed when the inter-HLR roaming happens for CHA architecture. As Fig. 3 shows, the registration request must be forwarded to the CHLR to inform the CHLR that the current visited HLR has changed. In the meanwhile, the CHLR performs the deregistration process to remove the obsolete records in the old HLR and old VLR.

Fig. 4 depicts the scenario of inter-HLR roaming in the DHA. The new HLR must issue an update request to the old HLR so that the old HLR can have an entry points to the new HLR, and then the user profile migrates from the old HLR to the new HLR. In the meanwhile, the old HLR performs deregistration process to inform the old VLR to remove the obsolete record.

With such registration and deregistration processes, a pointer chain is created if a mobile user roams across several HLRs. For analysis simplicity, we regard the cost of forwarding the pointer chain as the cost of forwarding along just one step of the chain.

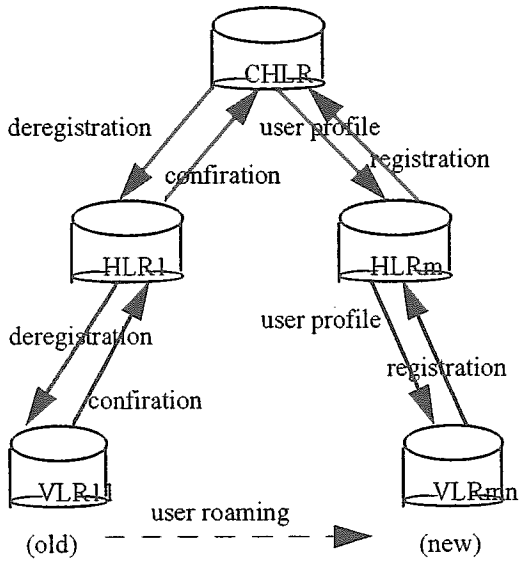

Fig. 3. Inter-HLR roaming in CHA.

2.3.2. Call origination and call termination

The call origination processes in the CHA and DHA are the same as that of single-tier PCS systems except that in CHA the process may involve the CHLR when the authentication process is performed at CHLR or when the user profile is not entirely cached in HLRs. In these cases, the query request will be forwarded to the CHLR to get solved.

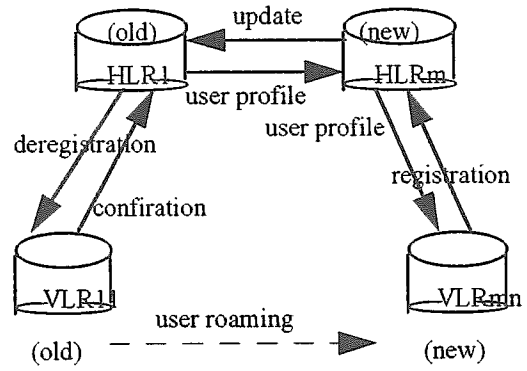

Fig. 4. Inter-HLR roaming in DHA.

The call termination processes in the CHA and DHA are also the same as that in the single-tier PCS system except the case that the called party is not within his home HLR. The process is performed as Fig. 5 in CHA. First, the GMSC (Gateway Mobile Switching Center) interrogates the home HLR (in this case, HLR_h) of the called MSISDN (Mobile Station ISDN number) for roaming number. But there is no entry for that called user. The HLR_h then queries the CHLR which records the address of the HLR that the called user currently visits. The CHLR then issues a query to the visited HLR to get the MSRN. Upon the CHLR receiving MSRN (Mobile Station Roaming Number), it sends the result back to the HLR_h and then to the GMSC. The rest steps are the same as that of the call termination process in the single-tier PCS system.

In the DHA, we assume that a pointer chain is maintained for each user who has roamed out of his home HLR. Fig. 6 shows that when the system wants to deliver the call to the user, firstly, the GMSC interrogates the user's home HLR to get the MSRN. The home HLR has an entry points to other HLR; therefore, the query is forwarded through the pointer chain until the currently visited HLR is reached. Fig. 6 only shows the case when the length of the pointer chain is one. The remaining steps are to query the visited VLR for roaming number, set up trunks, and page the called party to complete the call delivery.

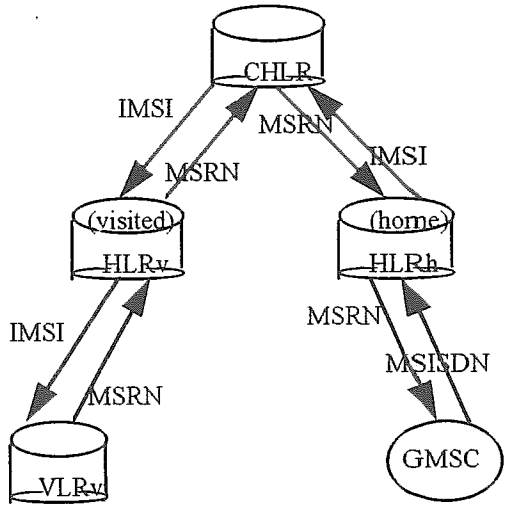

Fig. 5. Call termination in CHA if the called user is not within his home HLR.

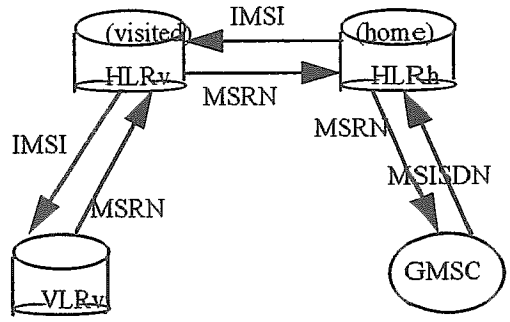

Fig. 6. Call termination in DHA if the called user is not within his home HLR.

3. Performance comparison

3.1. Performance metrics

We assume that the *M/G/1* queuing model under FCFS discipline is applied to each database in the networks [2, 17, 19, 20]. We assume that the service time distribution is gamma distribution [3]. The population of mobile users are uniformly distributed and the database systems are homogeneous and symmetric. This means that each database at the same tier serves the equivalent number of subscribers and has equivalent processing power. We also assume that each subscriber has the same behavior pattern. The evaluated performance metrics in our study include:

(A) *The response time of a query from a mobile station*: the time elapsed from issuing a query request to getting the query result by the VLR which is queried by an MS.

(B) *The response time of a query from PSTN*: the time elapsed from issuing a query to getting the query result by an HLR when the call termination happens.

(C) *The update response time*: the time elapsed from issuing an update request to getting the acknowledgment by a VLR.

(D) *Expected total number of messages exchanged in the network*: the expected total number of messages exchanged per second between databases in the network.

3.2. Common tunable parameters

Common tunable parameters for our analysis are listed in Table 1. Specific parameters for CHA and DHA and the detailed derivation process are referred to [17]. The assumption values of some parameters are referred to [10, 18, 20]. For all notations appeared in the following, the subscript '*C*' denotes that the notation is used for analysis of the CHA. Similarly, the subscript '*D*' means that the notation is for DHA. As for subscripts '*c*', '*h*', and '*v*', they stand for CHLR, HLR, and VLR respectively.

From the description of mobility-related operations in Section 2, we can derive the mean service time by considering the ratios of query and update arrival rates at each database of different tier (i.e., VLR, HLR, and CHLR). With mean service time and total traffic arrival rate at each database, the waiting time at each database can be easily computed. The response time of each kind of request can be derived based on the database activities and the scenarios of mobility-related operations described in Section 2. We can easily evaluate the expected number of messages exchanged between databases per second by observing the message flow in each mobility-related operation. Finally, the total number of messages exchanged can be derived. However, for the limitation of space all of these derivations can be found in [17].

Table 1 Common tunable parameters for CHA and DHA

Notation	*Value*	*Description*
q_1	$0 < q_1 \le 0.5$	probability of inter-VLR roaming
q_2	$0 < q_2 \le \min(0.1, q_1)$	probability of inter-HLR roaming
p_1	$0 \le p_1 \le 1$	probability of a query resolved at VLR
p_2	$0 \le p_2 \le (1 - p_1)$	probability of a query resolved at HLR
T_{tx}	8.75 ms	transmission time for each message(ms)
R	0.55/0.45	mobile termination rate over origination rate
r	$1 - q_2$	probability of user data stored in the home HLR
N_h	$2 \le N_h \le 6$	number of HLRs
N_v	$2 \le N_v \le 5$	number of VLRs
λ_v^q	$0 \le \lambda_v^q \le 0.05$	local query arrival rate at VLR (per ms)
λ_v^u	$0 \le \lambda_v^u \le 0.05$	local update arrival rate at VLR (per ms)

3.3. Special tunable parameters

The impacts of four pairs of special tunable parameters on the performance metrics are interesting to us. They are cache probabilities (p_1, p_2), roaming rates (q_1, q_2), traffic rates (λ_v^q, λ_v^u), and the number of HLRs and VLRs (N_h, N_v). By tuning them, the impacts of different cache strategies, different degree of user mobility, and subscriber behavior pattern on the performance metrics can be observed. We are also able to know which architecture can serve more subscribers, bear higher request rate, and have better extendibility.

4. Analysis results

Four performance metrics with four predetermined condition sets are compared, i.e.,

$\Omega_1 = \{(\alpha, q_1, q_2, N_h, N_v, \lambda_v^q, \lambda_v^u) = (1, 0.5, 0.1, 4, 5, 0.05, 0.05)\}$

$\Omega_2 = \{(\alpha, p_1, p_2, N_h, N_v, \lambda_v^q, \lambda_v^u) = (1, 0, 1, 4, 5, 0.05, 0.05)\}$

$\Omega_3 = \{(\alpha, p_1, p_2, q_1, q_2, N_h, N_v) = (1, 0, 1, 0.5, 0.1, 4, 5)\}$

$\Omega_4 = \{(\alpha, p_1, p_2, q_1, q_2, \lambda_v^q, \lambda_v^u) = (1, 0, 1, 0.5, 0.1, 0.05, 0.05)\}$

where α denotes the shape parameter of gamma density.

4.1. Response time of a query from an MS

Intuitively, the query path in the CHA is longer than that in the DHA. It is expected that the DHA has better performance on the response time of a query from a mobile station. However, the query path (or the message flow) of the query from a mobile station depends on the cache probabilities (i.e., cache strategies).

The effects of cache probabilities are illustrated in Fig. 7. In Fig. 7, the response time of a query from a mobile station in the CHA is always longer than that in the DHA except the case $p_1 + p_2 = 1$. With $p_1 + p_2 = 1$ the response time in the two architectures are the same. In this case, the query is not issued from the HLR to the CHLR because all of the information can be retrieved from the HLR (i.e., the query messages do not go through the CHLR). Thus, $p_1 + p_2 = 1$ is the best case for cache strategy to reduce the response time of a query from a mobile station in the CHA.

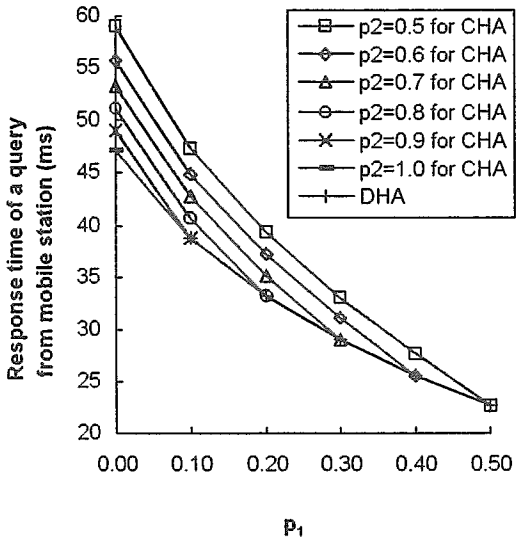

Fig. 7. The impact of different cache probabilities with Ω_1 on the response time of a query from MS.

4.2. Response time of a query from PSTN

It is also expected intuitively that the response time of a query from PSTN in the CHA is longer than that in the DHA because the query path in the CHA is longer than that in the DHA.

In Fig. 8, the response time of a query from PSTN in the CHA is longer with the smaller p_2 at a fixed value of p_1. Strictly speaking, the response time of a query from PSTN in the CHA is not related to p_2 directly, referred to Fig. 8, but with smaller p_2 the traffic arrival rate at the CHLR increases. Therefore, the difference between each line of Fig. 8 with different p_2 values is very small. The differences of the response time of a query from PSTN in the CHA and the DHA increase when p_1 increases. This accounts for that caching data in the VLR improves the response time of a query from PSTN for the DHA more than that for the CHA.

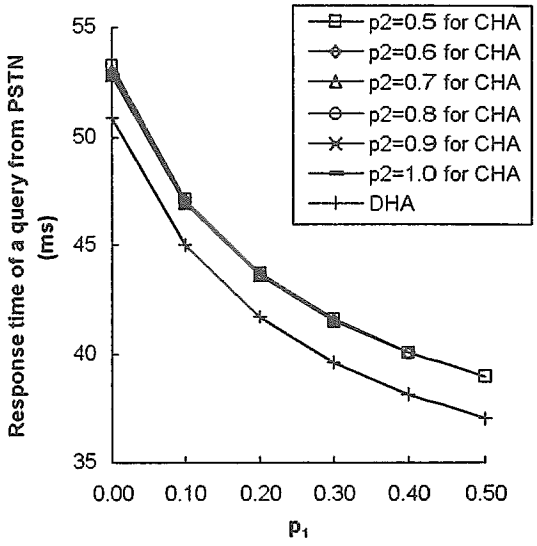

Fig. 8. Response time of a query from PSTN with Ω_1 and different cache probabilities.

4.3. Update response time

When the sum of p_1 and p_2 is small, the waiting time at CHLR in the CHA will be longer than the waiting time at HLR in the DHA. This is because the small value of $p_1 + p_2$ causes high traffic arrival rate at CHLR in the CHA, and increases the waiting time at the CHLR. Therefore, small value of $p_1 + p_2$ causes update response time in the CHA longer than that in the DHA. The situation is not depicted because it happens only when the sum of p_1 and p_2 is very small (i.e., below 0.4). It is not deserved to pay attention to such a bad cache policy because it causes high traffic arrival rates at HLRs and CHLR and long query response time (Figs. 7- 8). Therefore, we only observe the analysis results with $p_1 + p_2 \geq 0.5$ in the CHA.

Since CHA is with the best performance in the query response time when $p_1 + p_2 = 1$; therefore we compare the update response time in the CHA ($p_1 + p_2 = 1$) with that in the DHA as depicted in Fig. 9. It is found that the difference is getting smaller with the increase of p_1. The update response time in the DHA is always longer than that in the CHA because the traffic arrival rate at HLR is higher than that at CHLR, and waiting time at HLR is longer than that in CHLR. Refer to Figs. 3-4, this analysis result can be validated.

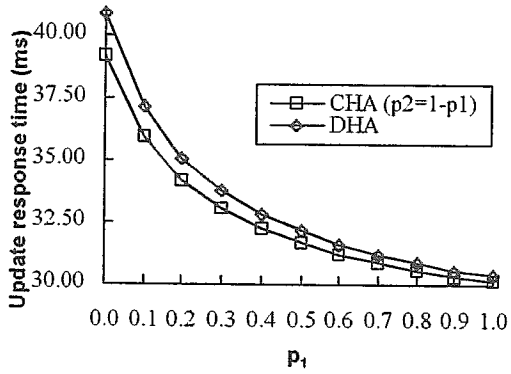

Fig. 9. The impact of various cache probabilities with Ω_1 and ($p_2 = 1 - p_1$) on the update response time.

With large value of $p_1 + p_2$, the update response time is longer in the DHA as shown above. The difference is small but gets larger when roaming rates increase as illustrated in Fig. 10 because of the impact of roaming rates on HLRs.

The impact of traffic rates on the update response time is depicted in Fig. 11. It shows that the differences increase with the increase of query and update rates. Based on Fig. 11, when query and update rates at VLR increase, the difference of update response time gets larger. This is because as the query and update rates at VLR increase, the impact on the traffic arrival rate at HLR is more than that at the CHLR.

4.4. Total number of messages

The comparison of the expected total number of messages exchanged per second in the CHA and the DHA with different traffic rates is shown in Fig. 12. The message increasing slope of CHA is sharper than that of DHA. This accounts for the traffic rate impact on the number of messages exchanged between CHLR and HLRs in the CHA is more than that between HLR pair in the DHA. Because the number of messages exchanged between each HLR and VLR in CHA is the same as that in DHA.

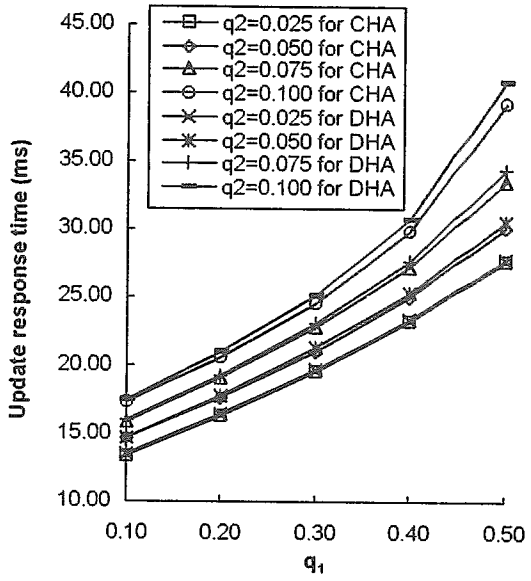

Fig. 10. The impact of roaming rates with Ω_2 on the update response time.

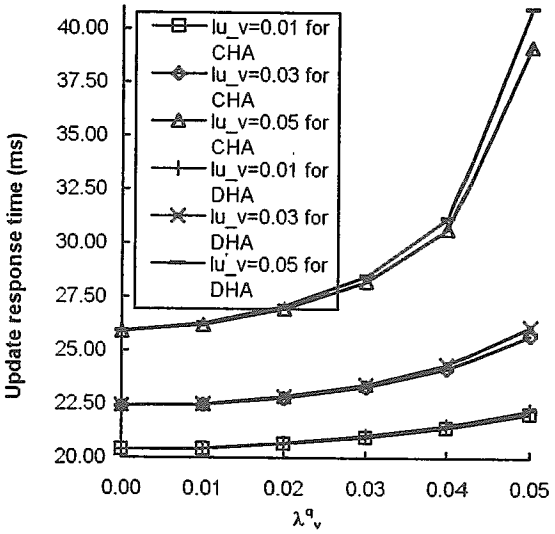

Fig. 11. The impact of various traffic rates with Ω_3 on the update response time. (lu_v denotes λ_v^u)

The expected number of messages exchanged between each HLR and VLR is not affected by the change of the number of HLRs and VLRs. It is found that the number of messages exchanged between CHLR and each HLR is not affected by N_h. While in the DHA, the number of messages exchanged between

each HLR pair decreases with the increase of the number of HLRs as shown in Fig. 13. This means that when the number of HLRs increases, the messages exchanged between HLRs in the DHA are more distributed, while in the CHA, the CHLR may become the bottleneck. The impact of the number of HLRs and VLRs on the total number of messages exchanged in the CHA is larger than that in the DHA as shown in Fig. 14.

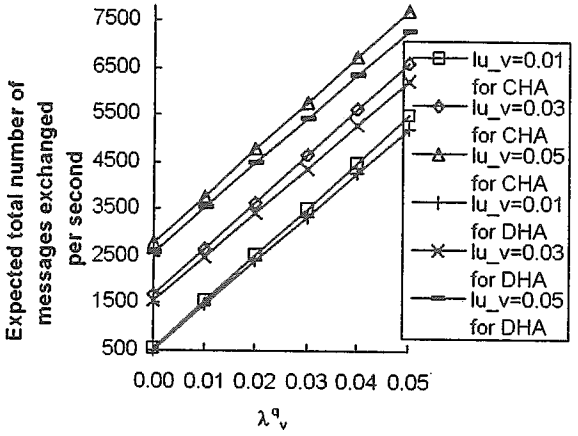

Fig. 12. The expected total number of messages exchanged per second with Ω_3 and various traffic rates. (lu_v denotes λ^u_v)

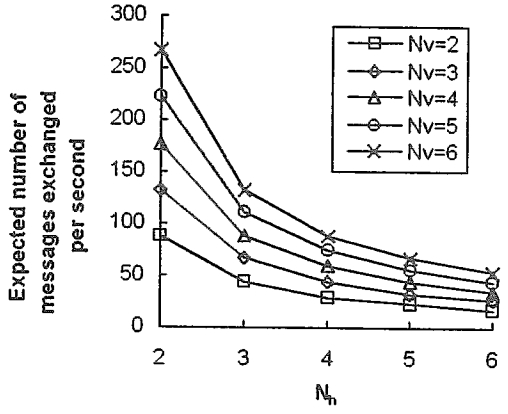

Fig. 13. The impact of the number of HLRs and VLRs with Ω_4 on the expected number of messages exchanged per second between each HLR pair in the DHA.

5. Conclusions

We have proposed two possible multi-tier database system architectures, the CHA and the DHA, and the mobility management strategies for them. The analysis model has been proposed and the performance metrics have been derived.

It is found that the DHA performs better than the CHA in the query response time and the number of messages exchanged in the network but not the update response time. However, by adopting more powerful HLRs compared to the CHLR in the two architectures, the DHA may have shorter update response time. Besides, we find that the CHA is not an efficient architecture; every update or query request has global effect because request messages must go through high tier databases. Therefore, decentralized database systems are better than centralized database systems with these evidences.

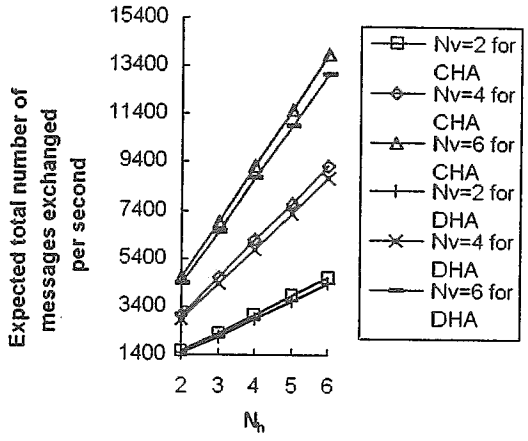

Fig. 14. The impact of different number of HLRs and VLRs with Ω_4 on the expected total number of messages exchanged per second.

The cost of forwarding the pointer chain when location tracking is performed in the DHA is ignored in our analysis. We also ignore the individual characteristics of each single-tier system, i.e., we just analyze homogeneous systems and not take the priorities of different tier services into consideration. In the future, we may take the cost of forwarding the pointer chain into consideration, design different forwarding pointer strategies for the DHA, and make comparisons of them. We may also concern the priorities of different tier services and model a heterogeneous system. More critical tunable parameters and more performance metrics are needed to be observed when such heterogeneous systems are analyzed.

References

[1] D.C. Cox, "Wireless personal communications: what is it?" *IEEE Personal Commun. Mag.*, pp. 20-35, April 1995.

[2] J.N. Daigle, *Queuing Theory for Telecommunications*, Addison-Wesley, 1992.

[3] N.A.J. Hasting and J.B. Peacock, *Statistical Distributions*, Wiley, 1975.

[4] B. Jabbari, "Intelligent network concepts in mobile communications," *IEEE Commun. Mag.*, pp. 64-69, February 1992.

[5] Y.-B. Lin, "No wires attached: reaching out with GSM," *IEEE Potentials*, October/ November, 1995.

[6] Y.-B. Lin, L.F. Chang, A.R. Noerpel, and K. Park, "Performance modeling of multi-tier PCS system," *Int'l Journal of Wireless Information Networks*, vol. 3, no. 2, pp. 67-78, 1996.

[7] Y.-B. Lin, "A comparison study of the two-tier and the single-tier personal communications services systems," *ACM/Baltzer Mobile Networks and Applications*, 29-38, 1996.

[8] Y.-B. Lin and I. Chlamtac, "Heterogeneous personal communications services," *IEEE Commun. Mag.*, 1996.

[9] Y.-B. Lin, "Mobility management for cellular telephony networks," To appear in *IEEE Parallel & Distributed Technology Mag.*

[10] M. Listanti and S. Salsano, "Impact of signaling traffic for mobility management in an IN based PCS environment," *Proceeding of ICUPC.*, pp. 107-111, Japan 1995.

[11] S. Mohan and R. Jain, "Two user location strategies for personal communications services," *IEEE Personal Commun. Mag.*, pp. 42-50, First Quarter 1994.

[12] M. Mouly and M.B. Pautet, *The GSM System for Mobile Communications*, M. Mouly, 49 rue Louise Bruneau, Palaiseau, France, 1992.

[13] J.E. Padgett, C.G. Gunther, and T. Hattori, "Overview of wireless personal communications," *IEEE Commun. Mag.*, January, 1995.

[14] K.I. Park and Y.-B. Lin, "Registration methods for multi-tier pesonal communications services," To appear in *IEEE Trans. on Veh. Technol.*

[15] J.F. Rizzo and N.R. Sollenberger, "Multitier wireless access," *IEEE Personal Commun. Mag.*, June 1995, pp. 18-30.

[16] J.E. Russell, "Universal personal communications: emergence of a paradigm shift in the communications industry," *Int'l Journal of Wireless Information Networks*, vol. 1, no. 3, 1994.

[17] R.W. Tzeng, "A performance study on multi-tier database systems for personal communications services," Master thesis, Dept. of Computer and Information Science, National Chiao Tung University, Hsinchu, Taiwan, R.O.C., June 1996.

[18] M. Vudali, R. Wank, and S. Tekinay, "BHME: a new capacity benchmark in PCS switching," *Proceeding of IEEE Globecom.*, pp. 2294-2296, Singapore, 1995.

[19] J.K. Wey, W.P. Yang, and Y.-B. Lin, "Mobility traffic analysis for PACS using various subscriber profiles," *Proceeding of IEEE PIMRC*, pp. 133-137, Taiwan, 1996.

[20] J.K. Wey, W.P. Yang, and L.F. Sun, "Traffic impacts of international roaming with transparent signaling connection," *Proceeding of 2nd Int'l Conference on Mobile Computing*, pp. 114-122, Taiwan, 1996.