Global Mobility Management by Replicated Databases in Personal Communication Networks*

Kin K. Leung and Yonatan Levy
AT&T Laboratories
101 Crawfords Corner Road
Holmdel, NJ 07733

{kkleung, ylevy}@att.com

October 1, 1996

April 18, 1997 (revised)

ABSTRACT

This paper explores the use of replicated databases for management of customer data (e.g., mobility data, call routing logic) in global, intelligent and wireless networks. We propose and analyze two, full and partial, data replication schemes - which are compatible with industry protocol standards - and compare them with the traditional, centralized database scheme. By identifying a set of key teletraffic and mobility parameters, we develop a modeling framework based on queueing models, and apply it to assess the relative performance and merits of these schemes.

Our results reveal that the full replication scheme outperforms the centralized one over a range of parameters. Furthermore, if customers update some of their data frequently (such as location data for highly mobile customers in wireless networks) and each call launches multiple queries into the databases, the partial replication scheme offers further performance improvement.

1. Introduction

To provide advanced telecommunication services, such as Personal Communication and Software Defined Network (SDN) services, today's networks are equipped with many databases of customer information for call routing and signaling purposes. To meet growing customer demands for additional advanced features and customization, network databases will be increasingly used in the future. In particular, although customers are becoming mobile, they still expect to be able to access their services anywhere, anytime, and in any medium. Thus, mobility management has become a major component of the future network design.

Mobility management can be divided into three different aspects: *terminal*, *personal* and *service mobility*. Terminal mobility is a network capability that enables mobile terminals to access the network even although they are moved from place to place. In contrast, personal mobility allows customers to access the network, independent of location and type of terminal that a customer happens to use at a given time. Service mobility enables customers to access the same types of network services, regardless of their current locations and terminals in use, as if the customers were at home or office.

To support service, personal, and terminal mobility in the *Personal Communication Network (PCN)* environment, the network has to store, maintain and retrieve customers' information for signaling purposes. This information includes the type and characteristics of the terminal in use, the services subscribed to, the location data and the call routing logic. Similarly, the provisioning of many advanced services such as the *Universal Personal Telecommunications (UPT)* services in future Advanced Intelligent Networks (AIN) also relies strongly upon the efficient use of databases of customer information such as call routing logic, call forwarding data, authentication information, etc. Thus, a key challenge for the management of PCN is to develop an efficient database architecture so that customer data can be readily available for signaling functions such as call setup and routing. Such architecture is needed especially for providing advanced services to global travelers.

The current approach to supporting terminal mobility requires a Home Location Register (HLR) and a Visitor Location Register (VLR). This HLR-VLR architecture [GPM92] has been established as an industry standard in the Global System for Mobile Telecommunication (GSM) for Europe [R93] and the IS-41 recommendations for North America [E91]. The routing and other signaling functions of each call initiated from or destined for a mobile customer requires the use of the location information stored in the databases. Although the protocol and the associated architecture for supporting terminal mobility have been standardized, the HLR-VLR architecture may cause potential performance problems for the signaling links and database systems due to the anticipated high terminal mobility and traffic demands in future wireless networks (for example, see [MA92], [LMW92] and [LWB92]).

The notion of personal and service mobility (especially in the context of global networks) is relatively new, and the protocols and architecture required to support them are yet to be standardized. We propose here signaling-network architectures based on centralized and replicated databases to support personal and service mobility for customers traveling throughout the world.

The architecture with a centralized database represents a traditional approach to managing customer data in such global networks. In this design, all call setup and other signaling functions access the centralized database in the signaling network. Evidently, the centralized design may not be efficient, especially for signaling functions across multiple international networks half way around the world, as

---

* This paper was presented in part at IEEE ICUPC'96 and International Teletraffic Congress 1997.

needed and anticipated for supporting global mobility in the future. In contrast, the replicated database design uses replicated databases for management of customer data (e.g., mobility data, call routing logic). This paper explores and analyzes two, *full* and *partial*, data replication schemes - which are compatible with IS-41, GSM, PCS and UPT protocol standards - and compares them with the centralized database design.

Replicating customer records at multiple places in communication networks for easy access have been suggested previously. For example, [BT92], [BTV92] and [JLL94] proposed the replication of location information for tracking customer locations in wireless networks, and [BJ92] suggested a similar data replication in signaling networks. In addition, the HLR-VLR architecture for terminal mobility can be viewed in a way as a form of partial data replication because parts but *not* all of customer data stored on the HLR and VLR are identical. We find in Section 5 that when compared with the full replication scheme, the partial data replication scheme yields better delay performance only for certain parameter settings. More importantly, the focus of this paper lies on the full and partial replication of HLR data on another database, while keeping the HLR-VLR architecture unchanged. As a result, the proposed data replication schemes in conjunction with the IS-41 and GSM standards can support terminal mobility efficiently.

The focus of our analysis of the centralized and replicated database architectures is on a framework for performance modeling that can quantify the query and update response time associated with the different designs. Clearly, there are many business and operational issues not covered here that would influence architectural decisions, and remain open for study and discussion. Many of the assumptions we make in order to illustrate the modeling approach can be easily relaxed or modified to study other scenarios. Hence, although our analysis and numerical examples provide insight into the relative benefits of each scheme, the main contribution of our performance analysis is in identifying the key variables and in developing the modeling framework.

The rest of this paper is organized as follows. We first examine in Section 2 a centralized database design to support global mobility across multiple, international networks. Due to the unique characteristics of global mobility, we then propose in Section 3 a signaling network with replicated databases for global mobility management. Section 4 analyzes the performance of the centralized and fully replicated database designs in terms of query and update response times. Due to the similarities in the derivation of results, the analysis for the partial replication scheme is presented in Appendix B. Based on some typical model parameters, we present several numerical examples in Section 5 to compare the performance of the database designs. In Section 6, we further characterize the differences among the schemes in terms of facility utilization. Finally, Section 7 presents our conclusions.

2. A Centralized Database Design for Global Mobility Management

Let us first consider the customer location information associated with personal mobility services such as UPT. (Since all signaling data under consideration is specified on a per customer basis, our results and conclusions can easily be extended to cover service mobility and other customer data such as call routing logic and terminal characteristics.) These services allow a user to initiate and receive calls by using a

unique personal telephone number (referred to as *Personal Telecommunication Number, PTN*) on any fixed or mobile terminal, irrespective of its geographical location throughout the world. Calls destined for a PTN can be routed to the customer's home, office, car, or answering service according to the routing logic specified by the customer. Such personal mobility is the defining attribute of the UPT service [FW92]. In order for the network to support the service, it has to rely on the extensive use of databases for call routing and other signaling functions. It is worth noting that each of the originating and receiving ends of a UPT call can be a wireless or wired terminal.

Our focus here is to consider such mobility services across multiple, international networks. To make our ideas concrete, we assume in this paper that global customers are based in the U.S. and possibly traveling in a foreign country. A centralized design, where the customer location information is stored in a *centralized database* physically located in the U.S, is shown in Figure 1. In order to make this design feasible, it requires that customers traveling abroad register from the visited country so that their home network is informed of their current locations (e.g., in terms of POTS numbers in wired networks, or network identity and temporary numbers in wireless networks). In addition, every time a customer changes location, the location data for the customer in the database is updated. Such database updates can be initiated: a) explicitly by customers connected to a wired network, as similar to the current AT&T 500 services, or b) automatically by the wireless network at which the customers are located. When a call is destined for a customer located either in the U.S. or abroad, the signaling network queries the centralized database for the location information for call setup and other signaling functions.

Figure 1: Centralized Database Architecture

The centralized database architecture can be generalized into a *distributed database* design. In this distributed design, customer records are partitioned and stored (but not replicated) in multiple databases at different physical locations. The specific database where the record of a customer is placed is referred to as the *home database*. An extra requirement for the distributed design is that the signaling network or switches have the knowledge of which database among many contains the signaling data for a given customer. If a single service provider is involved, we expect that the centralized and distributed designs would yield very similar query and update response times in normal situations, if the databases have identical design and traffic loads.

The centralized database design can also be used to provide advanced, global services in AIN. For example, AT&T International 800 Services currently allow calls originated in foreign countries to query a database physically located in the U.S. for call routing and signaling information. The centralized design can be adopted to support future advanced services such as Global Virtual Network Services (GVNS), by which, for example, employees of a multi-national company can place calls to each other as if they were located in the same company location. The database for these AIN services is updated when the call routing logic, authentication information, features subscribed, or other customer data are changed. Besides different reasons for database updates and rates of queries and updates, there is no substantive difference between network databases for supporting the AIN services and those for mobility management. Hence, in

the sequel, we focus on database activities for mobility management.

Given that signaling networks in the U.S. are likely to provide a very high degree of connectivity, querying the distributed databases across the U.S. does not cause unacceptable delays nor incur high costs for call setup and other signaling functions. As far as personal and service mobility supported within the U.S. is concerned, the use of distributed databases such as the architecture discussed in [MJ94] will continue to be a viable approach. If the high connectivity is maintained in the future, there will not be much need to replicate customer records at multiple places for reducing query delays and associated costs. However, the scenario can be entirely different for signaling functions across multiple international networks half way around the world, as needed and anticipated for supporting global mobility in the future. In that situation, depending on the specific mobility patterns and other traffic parameters, replicating customer records at databases located in foreign countries may shorten query and update delays. This is the subject of our discussion in the following section.

3. A Replicated Database Design for Global Mobility Management

A signaling network based on replicated databases is presented in Figure 2. Depending on the anticipated traffic load and other engineering and economic considerations, the world outside of the U.S. is divided into one or more regions, each of which covers a group of countries. An identical pair of *regional databases (RDB)* is installed in each region and connected to the U.S. signaling network, e.g., via two additional STPs. Functionally, the pair of regional databases can be viewed as one database, and the purpose of installing an identical pair is mainly to enhance system availability. This regional database serves as one of multiple regional databases for the global carrier, and can provide HLR functionality, but it will not replace the VLRs of wireless networks in the region. While we discuss this design mainly in the context of U.S.-based customers traveling to other countries, it also enables a global carrier to serve more effectively global customers whose home country is in one of the regions.

Figure 2: Replicated Database Architecture for Global Networks

The replicated database can be accessed for call setup and other signaling purposes as follows. Calls initiated in the U.S. will query the *home database (HDB)*. On the other hand, for calls originated from a visited country, queries are first launched to the RDB. If a call is destined for a customer visiting the region, the RDB has the customer record, and is capable of handling the associated queries. Otherwise, when the RDB cannot find the needed records, the queries will then be forwarded over the signaling network to the HDB for further processing. (The method to process updates will be discussed later.)

There are two possible approaches to the registration of a customer upon arrival at a foreign region:

1. Setup by registration messages - If all customer data, location information and other signaling data are included in the registration message, the RDB naturally can set up the customer record for future call processing. Since the message is subsequently sent to and processed as an update by the HDB in the U.S., the new customer data is also posted in the corresponding record on the HDB.

2. Setup by record download - If the registration message does not contain all customer data, the RDB cannot create a customer record solely based on the message. In this case, the RDB forwards the message to the HDB so that the associated customer record can be updated according to the new information (such as the new location data) in the message. After updating, the HDB downloads a copy of the entire, new customer record to the RDB.

Clearly, the appropriate choice of the setup methods depends on the amount of data in customer records and whether the data can be supplied by the customers while they are located overseas. As explained later, the methods of setting up replicated records on the RDB will affect the overall performance of the replicated database design.

3.1 Full and Partial Data Replication Schemes

The replicated database design can be further divided into two, *full* and *partial*, data replication schemes. In the full replication scheme, each customer visiting a foreign region has two complete, replicated copies of customer data: one in the HDB and the other in the RDB. In contrast, in the partial replication scheme, only the RDB keeps a complete set of customer data, whereas the HDB maintains (replicates) only a subset of customer data that is *infrequently* updated. For example, the HDB does not keep the location data for highly mobile, wireless customers visiting a foreign region. The idea of the partial data replication scheme is to reduce the overhead of updating certain customer data that is frequently changed. It is likely to be advantageous to maintain all customer data, including the most frequently updated data such as location information, at the RDB because updates are most likely generated by the customers themselves located in the foreign country.

During the registration process of a customer upon arriving at a foreign country, the customer supplies the RDB (which in turns forwards new data as an update to the HDB) with customer information, location data, and other signaling data. As pointed out, a copy of the customer record can be set up (replicated) on the RDB either by processing the data supplied by the customer or by record downloads from the HDB. In either replication scheme, the data record in the RDB will be removed when the customer leaves the region.

3.2 Potential Advantages and Implementation Issues

There are several potential advantages and benefits that data replication can bring about, and at the same time some potential drawbacks and implementation issues that can be identified. In this section, we explore several significant issues, recognizing that others may exist.

1. Improved delay performance - Call setup time and response times for other signaling functions can be reduced. This is the focus of this paper and is discussed in the next section.
2. Potential savings and cost of facilities - Due to the local availability of needed customer data, transmission cost for signaling messages for calls originated from the visited country or region, and possibly signaling facilities can be saved. The savings increase with: a) the proportion and distance of customers in foreign regions, b) the number and size of queries per call, and c) the fraction of calls originated from foreign regions. However, these savings and the reduced load on the HDB have to be

considered together with the cost of additional transmission and database facilities in the foreign region.

3. Compatibility with existing wireless standards - The replicated database design is compatible with industry standards such as the IS-41 [E91] and the GSM [R93] specifications for wireless networks. This is so because by adding the HLR functionalities to the RDB in a region, it can be utilized as the HLR by the VLR in the wireless network of the visited country.
4. Overhead due to access to RDB - The following strategies can be taken to minimize the overhead associated with the initial screening of queries. First, customers of global mobility services can use a distinct numbering plan such that the STPs in the region can simply forward queries associated with other services directly to the signaling network in the U.S. Thus, the RDB is accessed only for calls associated with the global services. Second, only when a customer is visiting a foreign country, does his or her record exist in the RDB, which can then be structured efficiently to speed up the access time.
5. Data consistency - Another potential issue is how to maintain the consistency between the replicated copies of customer data stored in the HDB and RDB. A low-cost data consistency control protocol such as the *Primary-Writer Protocol (PWP)* [L97] can be applied to maintain weak data consistency. When the PWP is used, the RDB and the HDB can be designated as the primary site and secondary site, respectively, to update records of customers visiting a foreign country. Updates for customers located in the U.S. are processed only by the HDB and there is no need to apply the PWP because these customers do not have replicated records on the RDB.

4. A Modeling Framework for Replicated Databases

A modeling framework based on queueing models for the replicated database design is presented in Figure 3. Since our main objective is to compare the system performance among the centralized design and the full and partial replicated database designs, only four components of the signaling network are considered in the framework: a) the home database (HDB) in the U.S., b) a regional database (RDB) in a foreign region, c) the bi-directional signaling link connecting the HDB and RDB, and d) the signal propagation delay. The framework remains applicable when different queueing models are used to model the components as needed. For example, depending on the database machines in use and their operation details, queueing models of different levels of complexity can be adopted in the framework.

As indicated in the figure, queries (marked as *H-queries*) and updates generated in the U.S. are forwarded to the HDB for processing. In the full data replication scheme, the HDB always contains data of all customers, regardless of whether they are located in the U.S. or in a foreign region. Thus, after handling the H-queries, the HDB sends responses to the switches in the U.S. In the partial replication scheme, however, a fraction of H-queries associated with calls destined for the overseas customers do not find the frequently updated data on the HDB. As a result, the HDB forwards these queries over the signaling link to the RDB, which processes the forwarded H-queries and sends the responses via the signaling link back to switches in the U.S.

In both full and partial replication schemes, queries (marked as *F-queries*) and updates generated in a foreign region first access the RDB. If an F-query finds the needed customer record (i.e., the customer is visiting the region), the RDB can process the query and send the response directly to switches overseas. However, if the needed record is not found, the RDB sends the failed F-query over the signaling link to the HDB. In turn, the HDB processes the failed F-queries, and sends the responses via the signaling link back to switches overseas.

As discussed in Section 3, there are two possible approaches to the registration of a customer upon arrival a foreign region. In either approach, registration messages including new location information and other signaling data are treated as updates, which are first processed by the RDB and then by the HDB, as shown in Figure 3. If replicated data on the RDB is setup by downloading, a copy of the entire, new customer record is downloaded to the RDB after the associated update is processed by the HDB.

Since our main objective here is to identify the key variables and gain a general understanding of the relative performance of the replicated and centralized database architectures for global networks, we make the following modeling assumptions (most of which can be easily relaxed to allow more complexity):

1. Each of the HDB and RDB has a single processor responsible for query and update processing, which is modeled as a single-server queue where queries and updates are processed on a first-come-first-served (FCFS) basis.
2. For the method of record downloads to set up replicated records on the RDB, since the downloading mainly involves direct memory access (DMA) or disk access, the additional processing (CPU) time for record downloads on the HDB and RDB is assumed to be negligible.
3. The signaling link in each direction is modeled as a single-server queue. Query and update messages and downloaded records are transmitted on a FCFS basis.
4. The signaling propagation delay is modeled by an infinite-server queue with mean service time equal to the mean propagation delay.
5. As an approximation, the databases and the signaling links are treated as independent M/G/1 queues. (Note that if the service discipline is the *processor sharing* policy, this assumption is not needed because the resultant open queueing network model has a product-form distribution. The same comment also applies to the FCFS service discipline, if all processing and transmission times have exponential distributions with identical averages. Further discussion on the validity of this approximation approach is given in Appendix A.)
6. Each call is assumed to require access to the record of the associated customer N_q times (i.e., each call generates N_q database queries) and acknowledgements are not considered. The analysis can be easily adapted to accommodate other scenarios; for example, the number of queries per call can be a random variable whose expected value is N_q.

The following additional assumptions are pertinent only to the partial replication scheme:

7. Only one of the N_q queries for each call requires access to the frequently updated data.

8. For simplicity, only the frequently changed data (such as location information) will be updated. Therefore, updating of infrequently updated data happens only at registration time. (This assumption can be easily relaxed by having two types of updates.)

4.1 Key Performance Parameters

In order to model the performance of the centralized and replicated database schemes, we define the following key parameters. To avoid cumbersome notation, we carry out the analysis for a single foreign region and then explain how it can be generalized.

ϕ : proportion of the global-service customers visiting a foreign region (i.e., a fraction ϕ of all calls should be terminated in a foreign region and similarly, ϕ of all updates are first generated by the customers overseas),

α_h : proportion of calls originated from a foreign region, given that the calls are destined for customers currently located in the U.S.,

α_f : proportion of calls originated from a foreign region, given that the calls are destined for customers currently located in the foreign region,

R_q : ratio of query to update rate (which is the average number of record queries per record update),

R_g : ratio of query to registration rate (which is the average number of record queries per registration in a foreign region),

N_q : number of database queries for each call,

X_q and $\overline{X_q^2}$: average and second moment of query processing time,

X_u and $\overline{X_u^2}$: average and second moment of update processing time,

ρ : loading of database (i.e., server occupancy),

L_q : average query message length in bytes,

L_u : average update message length in bytes,

L_d : average customer record length in bytes for record download,

Y_q and $\overline{Y_q^2}$: average and second moment of transmission time for a query message or response on the signaling link,

Y_u and $\overline{Y_u^2}$: average and second moment of transmission time for an update message on the signaling link,

Y_d and $\overline{Y_d^2}$: average and second moment of transmission time for a downloaded record on the signaling link,

C : signaling link speed,

τ : signal propagation delay.

Before the performance analysis, let us provide a brief qualitative insight on how some of these parameters can affect the system performance. The benefits of the replicated database design strongly depend on the parameter ϕ, which is the fraction of customers roaming in a foreign region. It is clear that if $\phi = 0$, the RDB does not contain any customer records and all queries eventually need to be processed by the HDB. Thus, the replicated database design does not offer any performance advantages over the

centralized design. Actually, query response time in the replicated design is longer than that in the centralized design because of the double dips to the RDB followed by the HDB. On the other hand, if ϕ approaches 1, all queries generated in the foreign region would be satisfied by the RDB. In this case, the query response time is reduced due to the local availability of the needed customer data on the RDB. In general, with all other parameters held fixed, the higher ϕ is, the better the replicated schemes can be expected to perform.

The performance gain of the replicated database design also depends strongly on the parameters α_h and α_f. Generally, the performance gain for the replicated design can be expected to increase with α_f and with $1 - \alpha_h$, i.e., with the fraction of intra-region calls.

If the method of record download is used to set up replicated copies of customer records on the RDB, the frequency of downloads and the volume of customer data would play a major role in the performance of the replicated database design. In particular, the signaling link from the HDB to the RDB can be saturated due to frequent record download or large record size.

Another factor that can degrade the performance of the replicated design is the query-to-update ratio. If the ratio is high (i.e., the customer information is seldom changed), the overhead in updating the replicated records will be minimal. Otherwise (e.g., for a mobile customer moving from location to location), the update overhead can be significant. In this case, the partial data replication scheme can perform better than the full replication scheme because the former scheme can reduce the update overhead by replicating only infrequently updated data on the HDB.

4.2 Performance of the Replicated Database Design

Since the main goal of this performance study is to analyze the respective merits of the replicated and centralized database designs for global mobility, a traffic load scenario for fair performance comparison is needed. For this purpose, we choose to define the common traffic scenario as the traffic rates that correspond to a fixed database (server) occupancy of ρ_c for the centralized database design. However, the arrival rates of queries and updates from the U.S. and the foreign region in both designs can vary according to the system and mobility parameters listed above.

Given the database load, the processing times and the record query-to-update ratio, the arrival rates of queries and updates at the centralized database are

$$\lambda_q = \frac{\rho_c}{X_q + \dfrac{X_u}{R_q}}, \qquad \lambda_u = \frac{\lambda_q}{R_q}. \tag{1}$$

The corresponding call load (which is not required for this analysis) is $\lambda_c = \lambda_q / N_q$. Given the arrival rates λ_q and λ_u in (1), the query and update rates from the U.S. and the foreign region are changed according to the traffic parameters ϕ, α_h and α_f, as discussed in the following. In turn, the traffic loads on the RDB, the HDB and the signaling link in the replicated design are affected.

In the replicated design, there are four basic types of calls, which are discussed below and summarized in Table 1:

1. A fraction $\alpha_f \phi$ are foreign calls destined for customers with records in the RDB resulting in successful F-queries.
2. A fraction $(1 - \alpha_f)\phi$ of calls are originated in the U.S. and destined for customers in the foreign region. These result in H-queries, and under the partial replication scheme, $1/N_q$ of those are forwarded to the RDB for further processing.
3. A fraction $\alpha_h(1 - \phi)$ are foreign calls that are destined for customers at the U.S. and result in failed F-queries and double dips to the RDB and then the HDB.
4. Finally, the remainder of the calls, $(1 - \alpha_h)\ (1 - \phi)$, result in successful H-queries and do not use the signaling link and the RDB.

Origination	Termination	
	Domestic	Foreign
Domestic	$(1-\alpha_h)(1-\phi)$	$(1-\alpha_f)\phi$
Foreign	$\alpha_h(1-\phi)$	$\alpha_f\phi$

Table 1. Fractions of Four Call Types

We now proceed to derive the expressions for the performance of the fully replicated (FR) database design. A similar analysis for the partial replication (PR) scheme is given in Appendix B, and the analysis of the centralized design (C) is given in the next section. In order to make the comparison easier, we use the same variable notation for all three designs and label equations for the same performance measure by using the same number with a suffix, PR, FR or C, to denote the corresponding design.

Now, let us consider the average response times on the RDB. Given that the query arrival rate is λ_q, a fraction $\alpha_h(1-\phi) + \alpha_f\phi$ of the queries correspond to queries generated in the foreign region (i.e., the F-queries in Figure 3). According to the replicated database design, these queries are first processed by the RDB. In addition, among the updates at the rate λ_u, a fraction ϕ of them is generated in the foreign region. Combining the query and update load, the server occupancy of the RDB is thus

$$\rho_r = \left[\alpha_h(1-\phi) + \alpha_f\phi \right]\lambda_q X_q + \phi\lambda_u X_u. \qquad \text{(2-FR)}$$

By Assumptions 1 and 5, the average response times (i.e., waiting time plus processing time) for queries and updates on the RDB are given by [K75]

$$T_{rq} = W_r + X_q \quad \text{and} \quad T_{ru} = W_r + X_u \qquad (3)$$

respectively, where

$$W_r = \frac{\left[\ \alpha_h(1-\phi) + \alpha_f\phi\ \right]\lambda_q\overline{X_q^2} + \phi\lambda_u\overline{X_u^2}}{2(1-\rho_r)} \qquad \text{(4-FR)}$$

is the average waiting time.

If a query finds the needed customer record on the RDB, T_{rq} in (3) is the average response time for the query. However, if the customer is not visiting the foreign region, the failed query is forwarded via the signaling link to the HDB for re-processing. Furthermore, after processing by the RDB, all updates generated in the foreign region are sent to update the customer record on the HDB at the rate of $\phi\lambda_u$. Thus, the traffic load on the signaling link from the RDB to the HDB (referred to as the *inward* direction) is

$$\beta_{in} = \alpha_h(1-\phi)\lambda_q Y_q + \phi\lambda_u Y_u \qquad \text{(5-FR)}$$

where Y_q and Y_u denote the average transmission times for query and update messages, which are given by $Y_q = 8L_q/C$ and $Y_u = 8L_u/C$, respectively.

If $\beta_{in} < 1$, the link in the inward direction is not saturated, and can operate properly at steady state. In this case, by Assumptions 3 and 5, we obtain the average response times (i.e., waiting plus transmission time) for query and update messages on the signaling link in the inward direction as

$$R_{in}^q = W_{in} + Y_q \quad \text{and} \quad R_{in}^u = W_{in} + Y_u \qquad (6)$$

respectively, where

$$W_{in} = \frac{\alpha_h(1-\phi)\lambda_q\overline{Y_q^2} + \phi\lambda_u\overline{Y_u^2}}{2(1-\beta_{in})}\ . \qquad \text{(7-FR)}$$

To consider the delay on the HDB, we note that the factor $\alpha_h(1-\phi)\lambda_q$ in (5-FR) is the arrival rate of the failed queries at the HDB. Combining this with the queries generated in the U.S., the total query arrival rate at the HDB is $(1-\alpha_f\phi)\lambda_q$. On the other hand, since all updates regardless of where they are generated are processed by the HDB, the update total arrival rate is simply λ_u. As a result, the server occupancy of the HDB is

$$\rho_h = (1-\alpha_f\phi)\lambda_q X_q + \lambda_u X_u. \qquad \text{(8-FR)}$$

By Assumptions 1 and 5, the average waiting time for queries and updates on the HDB is given by

$$W_h = \frac{(1-\alpha_f\phi)\lambda_q\overline{X_q^2} + \ + \lambda_u\overline{X_u^2}}{2(1-\rho_h)}. \qquad \text{(9-FR)}$$

Before considering the query response time, let us discuss the response time for query responses on the signaling link in the direction from the HDB to the RDB (referred to as the *outward* direction). Note that the traffic load on the link in the outward direction includes query responses and downloaded records, if record download is used as the method for setting up replicated records on the RDB. By the definition of the read-registration ratio R_g and the fact that the fraction of customers located in the foreign region is ϕ,

the rate of record downloading to the RDB is $\phi\lambda_q/R_g$. As for query responses, their rate is equal to that of the failed queries from the RDB. Thus, the arrival rate of query responses is given by $\alpha_h(1-\phi)\lambda_q$. Combining the query response messages and possibly download records, the link occupancy in the outward direction is

$$\beta_{out} = \alpha_h(1-\phi)\lambda_q Y_q + \frac{\phi\lambda_q}{R_g} Y_d \qquad \text{(10-FR)}$$

where Y_q and Y_d are the average transmission time for a query response and a downloaded record and Y_d is given by $8L_d/C$. Note that in case the method of record download is not used, the second term in (10-FR) does not exist. If $\beta_{out} < 1$, the link can operate stably at steady state, and, by Assumptions 3 and 5, the average response times for query (response) messages and downloaded records on the signaling link in the outward direction are

$$R^q_{out} = W_{out} + Y_q \quad \text{and} \quad R^u_{out} = W_{out} + Y_u \qquad (11)$$

respectively, where

$$W_{out} = \frac{\alpha_h(1-\phi)\lambda_q \overline{Y_q^2} + + \frac{\phi\lambda_q}{R_g}\overline{Y_d^2}}{2(1-\beta_{out})} . \qquad \text{(12-FR)}$$

Based on the delays at the databases and the signaling link in inward and outward directions in (3), (6), (9-FR) and (11), we finally obtain the average query response time for the replicated database design as

$$T_q = \left[\alpha_f\phi + \alpha_h(1-\phi)\right]T_{rq} + \alpha_h(1-\phi)\left[R^q_{in} + 2\tau + R^q_{out}\right] + (1-\alpha_f\phi)(W_h + X_q). \qquad \text{(13-FR)}$$

Note that the factor $[\alpha_f\phi + \alpha_h(1-\phi)]$ in (13-FR), which is the sum of two entries in the bottom row of Table 1, represents the probability that a query is generated in the foreign region and thus first incurs the delay on the RDB T_{rq}. A fraction $\alpha_h(1-\phi)$ of these queries cannot find the needed customer records on the RDB, and are thus forwarded to the HDB for further processing. As a result, these queries experience the signaling link delay in both directions and the propagation delay, as given in the second term. Finally, since the proportion of queries that can be successfully handled by the RDB (i.e., those F-queries associated with calls destined for the roaming customers) is $\alpha_f\phi$, all other $(1-\alpha f\phi)$ queries have to be processing by the HDB eventually. This delay is captured by the last term in (13-FR).

The update response time is defined as the time interval from the arrival of an update (at the RDB or HDB) until the updated record becomes available for query processing on the RDB or HDB. Using the corresponding probabilities for various types of queries as the weighting factors, we can obtain the average update response time as follows:

$$T_u = \phi T_{ru} + (1-\alpha_f)\phi\left[R^u_{in} + \tau\right] + (1-\alpha_f\phi)\left[W_h + X_u\right]. \qquad \text{(14-FR)}$$

As suggested in Section 3.2, the PWP can be used to maintain data consistency between the regional and home databases. Since the PWP is a weak concurrency control protocol, obsolete customer data may exist on the HDB for short periods. As a result, in addition to query and update response times as the key performance measures, it is also important to consider the probability of accessing such obsolete data, referred to as the *call misroute* probability. An analysis of the call misroute probability for the PWP has been presented in [L97]. Although the analysis there does not consider the link delay as part of the vulnerable period, the analysis approach is readily applicable to the setting of the replicated database for global mobility. For brevity, the analysis results of the misroute probability are omitted here and interested readers are referred to the reference for details.

In order to generalize the analysis to several regions, the parameters ϕ, α_f, and α_h, which are scalars in the above, can be indexed by region. Other parameters, such as the signal propagation delay, can also depend on the region. The analysis is then carried out in a similar manner.

4.3 Performance of the Centralized Database Design

The performance model for the centralized database design is a special case of the replicated design model in Figure 3. First, since the centralized architecture does not have the RDB, all queries and updates are forwarded to the centralized (home) database for processing. Second, all queries and updates generated in a foreign region, and associated responses, are transmitted over the signaling link. Lastly, no record download is needed and parameters such as R_g, L_d and Y_d are no longer needed. As pointed out earlier, the traffic load for a fair performance comparison is to fix the server occupancy of the centralized database to ρ_c. Thus, the query and update arrival rates for the centralized database design are given in (1).

Recall that the rate of updates generated overseas is equal to $\phi\lambda_u$. Similarly, the arrival rate of queries from the foreign region (i.e., the F-queries) is $[\alpha_h(1-\phi)+\alpha_f\phi]\lambda_q$. As a result, the traffic occupancy on the signaling link in the inward direction is

$$\beta_{in} = \Big[\ \alpha_h(1-\phi)\ +\ \alpha_f\phi\ \Big]\lambda_q Y_q\ +\ \phi\lambda_u Y_u \tag{5-C}$$

where Y_q and Y_u are the average transmission time of query and update messages, respectively. If β_{in} is less than 1, the signaling link in the inward direction is not saturated, and the average response times for query and update messages on the signaling link in the inward direction are

$$R_{in}^q\ =\ W_{in}\ +\ Y_q \quad \text{and} \quad R_{in}^u\ =\ W_{in}\ +\ Y_u \tag{6-C}$$

respectively, where

$$W_{in}\ =\ \frac{\Big[\ \alpha_h(1-\phi)\ +\ \alpha_f\phi\ \Big]\lambda_q \overline{Y_q^2}\ +\ \phi\lambda_u \overline{Y_u^2}}{2(1\ -\ \beta_{in})}\ . \tag{7-C}$$

Given that the server occupancy of the centralized database has been specified to be ρ_c, the average waiting time for queries and updates on the database is

$$W_c = \frac{\lambda_q \overline{X_q^2} + \lambda_u \overline{X_u^2}}{2(1 - \rho_c)}. \qquad (9\text{-C})$$

In contrast, the signaling link in the outward direction carries only responses of the queries generated overseas because record download is not needed for the centralized database design. Thus, the link occupancy in the outward direction is simply

$$\beta_{out} = \left[\alpha_h(1-\phi) + \alpha_f \phi \right] \lambda_q Y_q \ , \qquad (10\text{-C})$$

and the average response time for the query responses on the link is given by

$$R_{out}^q = \frac{\left[\alpha_h(1-\phi) + \alpha_f \phi \right] \lambda_q \overline{Y_q^2}}{2(1 - \beta_{out})} + Y_q. \qquad (12\text{-C})$$

Finally, considering all delay components together, the average query response time for the centralized database design is

$$T_q = \left[\alpha_h(1-\phi) + \alpha_f \phi \right] \left[R_{in}^q + R_{out}^q + 2\tau \right] + W_c + X_q \qquad (13\text{-C})$$

where the first bracketed factor on the right-hand side is the proportion of queries generated in the foreign region and only these queries incur the link delay and the propagation delay in both the inward and outward directions. Similarly, the average update response time for the centralized database design is

$$T_u = \phi \left[R_{in}^u + \tau \right] + W_c + X_u \qquad (14\text{-C})$$

where the weighting factor ϕ is needed because the link and propagation delay in the inward direction are incurred only by those updates generated for customers who are currently visiting the foreign region.

5. Numerical Performance Results

As explained earlier (Section 4.3), to enable a fair performance comparison between the centralized and replicated database designs, the server occupancy ρ_c for the centralized design is fixed to be 90% (which is typically considered to be a heavy load), while the arrival rates of queries and updates generated in the foreign country and in the U.S. can vary according to other traffic and mobility parameters. For simplicity, we assume $\alpha_f = \alpha_h$ in this numerical study. Furthermore, assume that the processing times for queries and updates are exponentially distributed with averages $X_q = 5$ msec and $X_u = 10$ msec, respectively. Unless stated otherwise, the average length of query and update messages (and query responses) is 50 bytes and the average record length for download is 2K bytes. The transmission speed *C* of the signaling link is assumed to be 64K bits/sec in each direction, which is the current CCITT standardized rate for international signaling links. The transmission times for query and update messages and downloaded records on the link are assumed to be exponentially distributed with averages $Y_q = Y_u = 6.25$ msec and $Y_d = 253$ msec, as obtained from the message lengths and the link speed. In addition, the signal propagation delay is 50 msec.

We remark that the above database processing times and message lengths are based on our observations of typical parameters on AT&T Network Control Point (NCP) database and signaling networks for existing advanced telecommunication services.

We first present results comparing the centralized and full replication designs, which are independent of the value of N_q. The average query response times as a function of the fraction of calls originated from the foreign region (i.e., α_f and α_h as they are assumed to be equal) for both the fully replicated and centralized database designs are depicted in Figure 4. In this numerical example, we assume that the RDB is capable of setting up replicated records based on the information contained in the registration messages. The parameter R_q is chosen to be 10. Thus, on average, a customer has 10 database queries between two consecutive updates (e.g., due to change of location) for the customer. Note that $R_q = 10$ is a few times higher than the corresponding values considered by [MJ94] for cellular networks. Since we also consider global services provided by AIN where subscribers tend to be less mobile than those in wireless networks, we consider $R_q = 10$ is relatively low, representing active and AIN or not highly mobile wireless customers. Results in the figure reveal that, as long as some portion of the calls are originated from the foreign country, the fully replicated database always yields an average query response time lower than that of the centralized database design. As one would intuitively expect, the query response time for the replicated database decreases as the parameter ϕ increases. It is also noteworthy that when the majority of calls are originated in the foreign country, an excessive number of query messages can saturate the signaling link in the inward direction (from the RDB to the HDB), thus causing an unacceptable delay increase for the centralized design which would require additional signaling capacity.

Let us continue to use the same traffic parameters except that the record download is now used as the method to set up replicated records on the RDB. Figure 5 shows the average response times for both database designs. In contrast to Figure 4, results in Figure 5 clearly reveal that the fully replicated design does not always yield query delays lower than those of the centralized design when records are downloaded. Specifically, depending on the combination of the values of R_g and ϕ, the fully replicated design can perform better or worse than the centralized design. When the replicated design yields a high query response time, it is due to the high delay in transmitting the downloaded records on the signaling link in the outward direction. This potential performance problem can be avoided if customers are allowed to upload their signaling data from smart cards to the RDB at registration overseas and the new data is forwarded to the HDB, if applicable. Generally, we observe that when the query-to-registration ratio R_g is sufficiently large (i.e., many calls for each foreign visit), the fully replicated design can outperform the centralized design in term of the average query response time.

We now turn to comparing all three schemes considering different values of N_q. In the following, we assume an average message length of 100 bytes and replicated records set up by registration messages. The average query response times as a function of the fraction of calls originated from the foreign country for the centralized, and full and partially replicated database designs are depicted in Figures 6 to 8 for $\phi = 0.1$, 0.5 and 0.9, respectively. The parameter R_q is set to 1, so, on the average, a customer receives one

database query between two consecutive updates to the customer record (e.g., due to change of location). We remark that $R_q = 1$ is very low, representing highly mobile wireless customers and/or small cells.

First, we note that the query response time of the centralized and full data replication designs are independent of the number of queries per call, N_q. This is due to our analysis method which holds ρ_c fixed, so the query and update rates do not vary with N_q. On the other hand, the response time is lowered when N_q increases for the partial replication scheme. This is so because when N_q increases, more queries can be successfully processed by the HDB, thus reducing the traffic load for the RDB and the link in the outward direction. Second, when the majority of calls are originated in the foreign country, an excessive number of query messages can saturate the signaling link in the inward direction, thus causing a large delay increase for the centralized design.

More importantly, when ϕ is relatively small (e.g., 0.1 to 0.2), the partial replication scheme always yields an average query response time lower than that of the centralized and fully replicated database designs. However, as ϕ increases, the partial replication scheme does not perform as well as the full replication design at times. As revealed in Figure 8, for high ϕ and small N_q, when most of the calls are initiated in the U.S., their queries are likely to have double dips to the HDB first and then the RDB, thus prolonging the delay in obtaining the needed signaling data.

Figure 9 shows the comparison of the average update response time among the three designs. We observe that the partial replication design generally yields lower update response time. This is mainly due to the fact that the frequently updated data is stored only on the RDB and updates are assumed to apply to that data only. In contrast, in the full replication scheme, both the RDB and HDB have to process each update generated overseas, thus prolonging the response time. In comparison, the update response time for the centralized database design can be very high due to the saturation of signaling link when a large fraction of calls are originated overseas or most of customers are located overseas.

Numerical results for high values of R_q have been obtained, and, in general, the full replication scheme performs better than the partial replication scheme when the customers are less mobile (i.e., $R_q >> 1$).

The above numerical results demonstrate the value of the modeling framework by illustrating potential advantages of the replicated schemes over the centralized one and providing insight into the range of parameters for which this advantage is significant. For network engineering and planning, the modeling framework can be useful for identifying the actual tradeoffs for specific traffic parameters under consideration. We make an initial step in this direction in the next section.

6. Facility Utilization and Deployment

The analysis so far compares the different designs for a traffic load corresponding to a fixed occupancy level at the centralized database. In this section, we examine more closely the database utilization for all three designs and use it to describe how a gradual deployment of the replicated architecture could occur as the traffic volume grows.

Given an offered call rate of λ_o, the corresponding query and update loads are given by

$$\lambda_q = N_q \lambda_o \qquad \lambda_u = \frac{\lambda_q}{R_q}. \tag{15}$$

By comparing the occupancy of the centralized database, $\rho_c = \lambda_q X_q + \lambda_u X_u$, with the occupancy of the HDB in the replicated designs in (8-FR) and (8-PR), we get

$$\rho_h^{FR} = \rho_c - \alpha_f \phi \lambda_q X_q \tag{16-FR}$$

$$\rho_h^{PR} = \rho_h^{FR} - (1 - \frac{R_q}{R_g}) \phi \lambda_u X_u \ . \tag{16-PR}$$

As registrations are treated as one type of updates, we have $R_g \geq R_q$. Thus by (16-FR) and (16-PR), we establish that with $\phi > 0$

$$\rho_h^{PR} \leq \rho_h^{FR} < \rho_c \ . \tag{17}$$

Hence, as expected, the load on the HDB in both replicated designs will be less than the load on the centralized database required to support the same overall call volume. However, the replicated designs require additional database(s) overseas. So next we compare the total processing capacity required by the replicated designs with that of the centralized design, which is ρ_c. By combining the RDB occupancy in (2) with (16), we obtain

$$\rho^{FR} = \rho_c + \alpha_h (1 - \phi) \lambda_q X_q + \phi \lambda_u X_u \tag{18-FR}$$

$$\rho^{PR} = \rho_c + \left[\alpha_h (1 - \phi) + (1 - \alpha_f) \phi / N_q \right] \lambda_q X_q + \phi \frac{R_q}{R_g} \lambda_u X_u \ . \tag{18-PR}$$

Hence, the total processing capacity required by the replicated design is higher due to the need to perform double dips and replicated updates. Similar results can be easily obtained for the signaling link utilization in both directions, showing that for most scenarios the replicated designs would require less signaling capacity. All these results can be used in an economic model that would incorporate the total costs of the designs with the benefits. In this context, it is important to realize that the RDB can also function as the HLR for a regional wireless network.

A reasonable approach to the development of a global network based on the replicated design would be to employ a gradual deployment as the traffic grows. The initial network is based on a centralized database. Then, as either the occupancy of the response time approach critical values, instead of adding another central database, a regional database is deployed in a distant region that generate the most inter-regional traffic. As the demand continues to grow, regions can be further divided, and more regional databases can be added.

7. Conclusion

Two, full and partial, replicated database and the centralized database designs for supporting mobile services in global, intelligent and wireless networks are presented. The key performance parameters are

identified and queueing models are used to study and compare the system performance of the database designs. Our performance results indicate that the choice of the database architecture for global services depends on the specific traffic and mobility parameters such as the proportion of customers visiting a foreign region, the fraction of calls generated overseas, the query-to-update ratio and the query-to-registration ratio. However, it is also found that for a range of traffic parameters, the replicated database design can yield query and update response times lower than those of the centralized design. Thus, to support the same amount of traffic load, the replicated design can require fewer transmission facilities for the signaling messages. The concepts presented in this paper can be further developed into a set of network planning guidelines that will help identify the best design alternative for mobility management based on the characteristics and parameters of a given application.

8. Acknowledgment

Thanks are due to P. Mastoris of AT&T for pointing out the applicability of the replicated database design to the Global Virtual Network Services (GVNS).

REFERENCES

[BT92] B.R. Badrinath and T. Imielinski, "Replication and Mobility," *IEEE Proc. of the 2nd Workshop on Management of Replicated Data,* Monterey, California, Nov. 1992, pp.9-12.

[BTV92] B.R. Badrinath, T. Imielinski and A. Virmani, "Locating Strategies for Personal Communication Networks," *IEEE Globecom'92 Workshop on Networking of Personal Communications Applications,* Orlando, Florida, Dec. 1992.

[BJ92] A.B. Bondi and V.Y. Jin, "Performance Analysis of a Minimally Replicated Distributed Database for Universal Personal Telecommunications Services," *Proc. of 8th ITC Specialist Seminar on Universal Personal Telecommunication,* Genova, Italy, Oct. 1992, pp.131-140.

[E91] EIA/TIA IS-41.1, "Cellular Radiotelecommunications Intersystem Operations," July 1991.

[FW92] R.L. Franks and P.E. Wirth, "UPT Traffic Issues - An Agenda for the 90's," *Proc. of 8th ITC Specialist Seminar on Universal Personal Telecommunication,* Genova, Italy, Oct. 1992, pp.107-115.

[GPM92] D.J. Goodman, G.P. Pollini and K.S. Meier-Hellstern, "Network Control for Wireless Communications," *IEEE Communication Magazine,* Dec. 1992, pp.116-124.

[JLL94] R. Jain, Y. Lin, C. Lo and S. Mohan, "A Caching Strategy to Reduce Network Impacts of PCS," *IEEE J. Sel. Areas in Commun.,* Vol.12, No.8, Oct. 1994, pp.1434-1444.

[K75] L. Kleinrock, *Queueing Systems, Volume I: Theory* John Wiley & Sons, New York (1975).

[K79] Paul J. Kuehn, "Approximate Analysis of General Queuing Networks by Decomposition," *IEEE Trans. on Commun.,* Vol.COM-27, No.1, Jan. 1979, pp.113-126.

[L97] K.K. Leung, "An Update Algorithm for Replicated Signaling Databases in Wireless and Advanced Intelligent Networks," *IEEE Trans. on Computers,* Vol.46, No.3, March 1997, pp.362-367.

[LL96] K.K. Leung and Y. Levy, "Use of Centralized and Replicated Databases for Global Mobility Management in Personal Communication Networks," presented at ICUPC'96, Boston, October 1996.

[LMW92] C.N. Lo, S. Mohan and R.S. Wolff, "An Estimate of Network Database Transaction Volume to Support Voice and Data Personal Communications Services," *Proc. of 8th ITC Specialist Seminar on Universal Personal Telecommunication,* Genova, Italy, Oct. 1992, pp.293-311.

[LWB92] C.N. Lo, R.S. Wolff and R.C. Bernhardt, "An Estimate of Network Database Transaction

Volume to Support Universal Personal Communication Services," *Proc. of 1st Int'l Conf. on Universal Personal Commun.*, Dallas, TX, Sept. 1992, pp.236-241.

[MA92] K.S. Meier-Hellstern and E. Alonso, "The Use of SS7 and GSM to Support High Density Personal Communications," *Proc. of IEEE ICC'92*, Chicago, IL, June 1992, pp.1689-1702.

[MJ94] S. Mohan and R. Jain, "Two User Location Strategies for Personal Communications Services," *IEEE Personal Communications Magazine*, Vol.1, No.1, 1994, pp.42-50.

[R93] M. Rahnema, "Overview of the GSM System and Protocol Architecture," *IEEE Communication Magazine*, April 1993, pp.92-100.

APPENDIX A

Justification for the Approximate Performance Model

We provide here justification for the approximation approach by use of independent M/G/1 queues. First of all, as shown in in Figure 3, no message or response return to the same node (database or link) in the proposed performance model. For example, when a query cannot be processed successfully by the RDB when the associated customer is not visiting the foreign region, it will be forwarded to the HDB for further processing. Then, the response for the failed F-query will be routed directly back to the switches overseas, not to the RDB. Similarly, responses to forwarded H-queries are routed directly to the U.S. switches involved in the call setup. As the message flow in our model does not involve any feedback or over-taking, the average query and update response times can be obtained by summing delays on the individual nodes. In fact, with identical exponential service times for query and updates at each node, all message flows in our model are Poisson. This has been our basic motivation and justification for using M/G/1 queues as approximations in this paper.

When query and update service times are not identical, our modeling results are not exact. However, as only part of departing messages (e.g., failed F-queries or forwarded H-queries) are forwarded to another node, it can be shown that such "thinning process" tends to make the traffic flow close to Poisson. To see that, let $\{N_t\}$ and $\{X_t\}$ be the total number of departing messages from a node and that for those input to the next node during time interval $(0,t]$, respectively. As messages of different types arrive at a node randomly and they are processed on a FCFS basis, each departing message has a certain probability to be forwarded to the next node. Assume that a fraction p of all departing messages are forwarded. Then, we have

$$\mathrm{E}[X_t]=p\mathrm{E}[N_t] \tag{A-1}$$

and

$$\mathrm{var}[X_t]=p^2\mathrm{var}[N_t] + p(1-p)\mathrm{E}[N_t] \tag{A-2}$$

By definition, the index of dispersion, (which measures the burstiness of a message flow, [give a reference here] for X_t is $I[X_t]=\mathrm{var}[X_t]/\mathrm{E}[X_t]$. Thus, we obtain from (A-1) and (A-2) that

$$I[X_t] = pI[N_t] + 1 - p. \tag{A-3}$$

Furthermore,

$$R \equiv \frac{I[X_t]}{I[N_t]} = p + \frac{1-p}{I[N_t]} \tag{A-4}$$

Thus, $R<1$ (or $R>1$) if $I[N_t]>1$ (or $I[N_t]<1$). Hence, the "thinning process" makes the forwarded traffic flow random in the sense that $I[X_t]$ is closer to 1 (i.e., Poisson process) than $I[N_t]$ is. As a consequence, the fact that part of departing messages from a node are forwarded to the next node in our model also helps justify the M/G/1 approximations.

The approximation approach can be further justified for the parameter settings considered in our numerical study. As the service times are assumed to be exponentially distributed with different means in Section 5, the quality of our approximations depends on the ratio of their rates. When one type of traffic, either queries or updates which arrive from Poisson sources, dominates (e.g., the ratio is greater than 10), according to Eq.(9b) in [K79], the departure process is very close to Poisson, which corresponds to $I[N_t] \approx 1$ in (A-4). It is indeed the case in Figure 4 and 5 of Section 5 where query messages dominate as $R_q = 10$ and $R_g \geq 20$. Thus, our approximation approach is appropriate. In addition, as the average query and update processing time are 5 and 10 msec, respectively, again by Eq.(9b) in [K79], the squared coefficient of message interdeparture time from the RDB or HDB is about 1.2 for the parameters for Figure 6 to 8, which is close to 1 for Poisson departure already. Furthermore, as pointed out above, the thinning process of forwarding part of messages to the next node will make the traffic flow further close to Poisson.

APPENDIX B

Performance Analysis of Partial Replication

Given all the definitions of Section 4.1, Eq. (1) still holds, and we now turn to the derivation of average response times for the partial replication scheme. Given that the query arrival rate is λ_q, a fraction $\alpha_h(1-\phi) + \alpha_f\phi + (1-\alpha_f)\frac{\phi}{N_q}$ of the queries are handled by the RDB. In addition, among the updates at the rate λ_u, a fraction ϕ of them is generated in the foreign region, assuming that updates are generated at the same rate in home and foreign regions. Combining the query and update load, the server occupancy of the RDB is thus

$$\rho_r = \left\lfloor \alpha_h(1-\phi) + \alpha_f\phi + (1-\alpha_f)\frac{\phi}{N_q} \right\rfloor \lambda_q X_q + \phi\lambda_u X_u. \qquad \text{(2-PR)}$$

By Assumptions 1 and 5, the average response times (i.e., waiting time plus processing time) for queries and updates on the RDB are given by

$$T_{rq} = W_r + X_q \quad \text{and} \quad T_{ru} = W_r + X_u \qquad (3)$$

respectively, where

$$W_r = \frac{\left\lfloor \alpha_h(1-\phi) + \alpha_f\phi + (1-\alpha_f)\frac{\phi}{N_q} \right\rfloor \lambda_q \overline{X_q^2} + \phi\lambda_u \overline{X_u^2}}{2(1-\rho_r)}. \qquad \text{(4-PR)}$$

In the inward direction (from the RDB to the HDB), the link carries failed F-queries and responses to forwarded H-queries. Furthermore, after registration by customers arriving in a foreign region, new customer data including the identification of the RDB is sent to update the customer record on the HDB at

the rate of $\phi\lambda_q/R_g$. (Note that, in contrast to the full replication, updates generated in the foreign region subsequent to the registration are *not* processed by the HDB according to Assumption 8.) Thus, the traffic load on the signaling link in the inward direction is

$$\beta_{in} = \left\lceil \alpha_h(1-\phi) + (1-\alpha_f)\frac{\phi}{N_q} \right\rceil \lambda_q Y_q + \phi\frac{\lambda_q}{R_g} Y_u \qquad \text{(5-PR)}$$

where $Y_q = 8L_q/C$ and $Y_u = 8L_u/C$.

If $\beta_{in} < 1$, we obtain the average response times (i.e., waiting plus transmission time) for query and update messages on the signaling link in the inward direction as

$$R_{in}^q = W_{in} + Y_q \quad \text{and} \quad R_{in}^u = W_{in} + Y_u \qquad (6)$$

respectively, where

$$W_{in} = \frac{\left\lceil \alpha_h(1-\phi) + (1-\alpha_f)\frac{\phi}{N_q} \right\rceil \lambda_q \overline{Y_q^2} + \phi\frac{\lambda_q}{R_g}\overline{Y_u^2}}{2(1-\beta_{in})}. \qquad \text{(7-PR)}$$

To consider the delay on the HDB, we note that the factor $\alpha_h(1-\phi)\lambda_q$ in (5-PR) is the arrival rate of the failed F-queries at the HDB. Combining this with the queries generated in the U.S., the total query arrival rate at the HDB is $(1-\alpha_f\phi)\lambda_q$. In addition, since all updates generated by customers located in the U.S. and the updates corresponding to registrations from customers visiting the foreign region are processed by the HDB, the update total arrival rate is simply $(1-\phi)\lambda_u + \phi\frac{\lambda_q}{R_g}$. As a result, the server occupancy of the HDB is

$$\rho_h = (1-\alpha_f\phi)\lambda_q X_q + \left\lceil (1-\phi)\lambda_u + \phi\frac{\lambda_q}{R_g} \right\rceil X_u. \qquad \text{(8-PR)}$$

By Assumptions 1 and 5, the average waiting time for queries and updates on the HDB is given by

$$W_h = \frac{(1-\alpha_f\phi)\lambda_q\overline{X_q^2} + \left\lceil (1-\phi)\lambda_u + \phi\frac{\lambda_q}{R_g} \right\rceil \overline{X_u^2}}{2(1-\rho_h)}. \qquad \text{(9-PR)}$$

The traffic load on the link in the outward direction includes forwarded H-queries, F-query responses and downloaded records as indicated in Figure 3, if record download is used as the method for setting up replicated records on the RDB. By the definition of the read-registration ratio R_g, the rate of record downloading to the RDB is $\phi\lambda_q/R_g$. In addition, the rate of forwarded H-queries is $(1-\alpha_f)\frac{\phi}{N_q}\lambda_q$, and as for F-query responses, their rate is equal to that of the failed queries from the RDB, $\alpha_h(1-\phi)\lambda_q$. Combining all these messages, the link occupancy in the outward direction is

$$\beta_{out} = \left\lceil \alpha_h(1-\phi) + (1-\alpha_f)\frac{\phi}{N_q} \right\rceil \lambda_q Y_q + \phi\frac{\lambda_q}{R_g} Y_d \ , \qquad \text{(10-PR)}$$

If $\beta_{out} < 1$, by assumptions 3 and 5, the average response times for query (response) messages and downloaded records on the signaling link in the outward direction are

$$R^q_{out} = W_{out} + Y_q \quad \text{and} \quad R^d_{out} = W_{out} + Y_d \qquad (11)$$

respectively, where

$$W_{out} = \frac{\left\lceil \alpha_h(1-\phi) + (1-\alpha_f)\frac{\phi}{N_q} \right\rceil \lambda_q \overline{Y_q^2} + \ + \phi\frac{\lambda_q}{R_g}\overline{Y_d^2}}{2(1-\beta_{out})} \ . \qquad \text{(12-PR)}$$

Based on the delays at the RDB and HDB and the signaling link in inward and outward directions in (3), (6), (9-PR) and (11-PR), we finally obtain the average query response time for the partially-replicated database design as

$$T_q = \left\lceil \alpha_f\phi + \alpha_h(1-\phi) + (1-\alpha_f)\frac{\phi}{N_q} \right\rceil T_{rq} + \left\lceil \alpha_h(1-\phi) + (1-\alpha_f)\frac{\phi}{N_q} \right\rceil \left[R^q_{in} + 2\tau + R^q_{out} \right] \qquad \text{(13-PR)}$$

$$+ (1-\alpha_f\phi)(W_h + X_q).$$

Note that the coefficient for T_{rq} in (13-PR) represents the fraction of queries processed by the RDB, thus incurring the delay T_{rq}. The next term accounts for the failed F-queries and the forwarded H-queries, where both types of queries experience the signaling link delay in both directions and the propagation delay. Finally, since the proportion of queries that can be successfully handled by the RDB is $\alpha_f\phi$, all other queries have to be processed by the HDB one way or the other. This delay is captured by the last term.

The update response time is defined as the time interval from the arrival of an update (at the RDB or the HDB) until the updated data becomes available for call processing, and the average update response time is given by

$$T_u = \phi T_{ru} + (1-\phi)(W_h + X_u). \qquad \text{(14-PR)}$$

List of Figure Captions

List of Table Captions

Table 1. Fractions of Four Call Types

Conference Presentation Information

This paper was presented in part at IEEE ICUPC'96, September 29 to October 2, 1996 in Cambridge, MA and 15th International Teletraffic Congress (ITC), June 23 to 27, 1997 in Washington, D.C.

Biography of Kin K. Leung

Kin K. Leung (S'78-M'86-SM'93) received his B.S. degree with first class honors in electronics from the Chinese University of Hong Kong, Hong Kong, in 1980, and his M.S. and Ph.D. degrees in computer science from University of California, Los Angeles, in 1982 and 1985, respectively.

He attended UCLA under an exchange program between the Chinese University of Hong Kong and University of California. In 1986, he joined AT&T Bell Laboratories in New Jersey. He received the Distinguished Member of Technical Staff Award in 1994 for his research work of performance analysis methodologies and their applications to enhance AT&T product and services. Currently, he is a Technology Consultant at Broadband Wireless Systems Research Department of AT&T Laboratories. His research interests include wireless and computer communication networks, stochastic modeling, and distributed processing and databases.

Biography of Yonatan Levy

Yonatan Levy ....

Global Mobility Management by Replicated Databases in Personal Communication Networks

Kin K. Leung and Yonatan Levy
AT&T Labs
Holmdel, NJ 07733

ABSTRACT

This paper explores the use of replicated databases for management of customer data (e.g., mobility data, call routing logic) in global, intelligent and wireless networks. We propose and analyze two, full and partial, data replication schemes - which are compatible with industry protocol standards - and compare them with the traditional, centralized database scheme. By identifying a set of key teletraffic and mobility parameters, we develop a modeling framework based on queueing models, and apply it to assess the relative performance and merits of these schemes.

Our results reveal that the full replication scheme outperforms the centralized one over a range of parameters. Furthermore, if customers update some of their data frequently (such as location data for highly mobile customers in wireless networks) and each call launches multiple queries into the databases, the partial replication scheme offers further performance improvement.

List of Authors

Kin K. Leung
AT&T Laboratories, Room R-138
791 Holmdel-Keyport Road
Holmdel, NJ 07733

Email: kkleung@att.com
Phone: (908)888-7041
FAX: (908)888-7190

Yonatan Levy *
AT&T Laboratories, Room 1L-224
101 Crawfords Corner Road
Holmdel, NJ 07733

Email: ylevy@att.com
Phone: (908)949-6568
FAX: (908)949-7210

* Please forward future correspondence and galley proof to this author.