the Second Workshop on the Management of Replicated Data, November 1992.

[7] B. R. Badrinath, T. Imielinski and A. Virmani, "Locating strategies for Personal Communication Networks," In IEEE GLOBECOM 92 Workshop on networking of personal communications applications, December 1992.

[8] Michael J. Carey, Michael J. Franklin, Miron Livny, Eugene J. Shekita, "Data Caching tradeoffs in client-server DBMS architectures," Proc. of the 1991 ACM SIGMOD, May 1991, pp. 357–376.

[9] Danny Cohen, Jonathan B. Postel, Raphael Rom, "IP addressing and routing in a local wireless network," Manuscript, July 1991.

[10] Daniel Duchamp, Steven K., Gerald Q. Maguire, "Software technology for wireless mobile computing", IEEE Network Magazine, November 1991, pp. 12–18.

[11] Daniel Duchamp, Neil Reynolds, "Measured performance of a wireless LAN", Columbia University, October 1992.

[12] David Gifford, John Lucassen, and Stephen Berlin, "The application of digital broadcast communication to large scale information systems," IEEE Journal on selected areas in communications, Vol 3, No. 3, May 1985, pp.457–467.

[13] Michael J. Franklin, Michael J. Carey, and Miron Livny, "Global memory management in client-server DBMS architectures," Proc. of the 18th International conference, August 1992, pp. 596–609.

[14] John Ioannidis, Dan Duchamp, Gerald Q Maguire, "IP-based protocols for mobile internetworking," In SIGCOMM 91, September 1991, pp. 235–245.

[15] John Ioannidis, Gerald Q Maguire, "The design and implementation of a mobile internetworking architecture," In USENIX Winter 1993 technical conference January 1993.

[16] James Kistler and M. Satyanarayanan, "Disconnected operation in the CODA file system," ACM Transactions on Computer Systems, Vol 10, No 1, February 1992, pp. 3–25.

[17] Y. Rekhter and C. Perkins, "Optimal routing for mobile hosts using IP's loose source route option," Internet Draft, October 1992.

[18] Samuel Sheng, Ananth Chandrasekaran, and R. W. Broderson, "A portable multimedia terminal for personal communications," IEEE Communications Magazine, December 1992, pp. 64–75.

[19] Carl D. Tait and Dan Duchamp, "Service interface and replica management algorithm for mobile file system clients," Proceedings of the Parallel and Distributed Information Systems Conference, December 1991.

[20] David J. Goodman, "Trends in Cellular and Cordless Communications," IEEE Communications Magazine, June 1991.

[21] David J. Goodman, "Cellular Packet Communications," IEEE Transactions on Communications, VOL. 38. NO 8. August 1990.

[22] David J. Goodman and Binay Sugla, "Signalling system draft," Unpublished manuscript.

[23] Kathleen S. Meier-Hellstern, Eduardo Alonso, and Douglas Oniel, "The Use of SS7 and GSM to support high density personal communications," Third Winlab workshop on third generation wireless information networks, April 1992, pp. 49–57.

[24] L J Ng. et. al. "Distributed architectures and databases for intelligent personal communication networks," Proc of the ICWC, June 1992.

[25] C. N. Lo, R. S. Wolff and R. C. Bernhardt, "An estimate of network database transaction volume to support universal personal communication services," Submitted to the 1st International conference on Universal Personal Communications.

[26] M Satyanarayanan, "Scalable, secure, and highly available distributed file access," IEEE Computer, VOl 23, No. 5, May 1990, pp. 9–21.

[27] Ouri Wolfson and Sushil Jajodia, "Distributed algorithms for dynamic replicated of data," 11th ACM PODS, June 1992, pp. 149–163.

[28] Hiromi Wada et.al., "Mobile computing environment based on internet packet forwarding," 1992 Winter USENIX, January 1993.

[29] Gio Wiederhold, "Mediators in the architecture for future infromation systems," Unpublished manuscript.

Here we list data management consequences of having a restricted energy resource:

- Tradeoffs between memory and channel access: Memory consumes power but so do channel accesses; in particular, transmissions. From the point of view of power consumption, for a given computation, is it better to execute the computation on the mobile client or on the fixed server. How should data be partitioned between the client and the server?

- The role of data compression as a power saving tool. Compressed data uses less memory and communication channel but takes additional CPU cycles (and hence palmtop power) to decompress. It is twice as expensive to decompress than it is to compress. What are the tradeoffs here?

- The role of data broadcast as a "listen only" data access mode in saving power consumption. Listening to the broadcasted data saves power by not having to transmit requests for data. However, the receiver needs to be "active" to listen to the broadcast. If the receiver is smart it can be tuned to become active at the appropriate time to receive the broadcast. How to design protocols to support this kind of data access?

5 Conclusions

Management of data in a massively distributed environment of *mobile* computing offers new challenging research problems. We have identified those challenges and formulated a number of open problems.

Data management offers new challenges both at the global, network, level as well at the local computing platform of a palmtop computer. Global data management on the network level is mainly of interest to the network carriers and providers and have to be resolved before the infrastructure for mobile computing can be built. Here, the scale of the system and mobility of its parts are unheard off and the current network infrastructure is simply not capable to handle mobility both in terms of scale and communication protocols. This is still the main *technological* obstacle before the vision of universal mobile computing can be realized.

Still the traditional question "if we build it, will they come?" remains unanswered. The positive answer will critically depend on the applications which will run on the palmtops. Local data management will be the center of activity for the software vendors offering new software tools running on the palmtops. To support future applications the lower level systems software will have to be built first; it will handle various levels of disconnection, power management and finally provide new functionalities necessary to access broadcasted data.

The case for a wireless connection can be made on the basis of offered flexibility and mobility. What are the "killer" applications that depend on flexibility and mobility? Many suggest "mail enabled applications" such as those targeted by Notes from LOTUS corporation. These applications will be targeted towards collaboration of users who are on the move. Other classes of applications include local information services (local yellow pages) which will provide a much more detailed service than that of the current telephone yellow pages and application support for the mobile work force. And in general, mobile wireless computing could be another added convenience that users want. As in any revolution, the mobile computing has its enthusiasts and opponents. Without taking sides or trying to make predictions we conclude that the concept of mobile computing offers challenges and opens new research problems.

References

[1] Arup Acharya and B. R. Badrinath, "Delivering multicast messages in networks with mobile hosts," Submitted for Publication, October 1992.

[2] Rafael Alonso, Daniel Barbara, and Hector Garcia-Molina, "Data caching issues in an information retrieval system," ACM TODS, Sept. 1990, pp. 359–384.

[3] Baruch Awerbuch and David Peleg, "Concurrent online tracking of mobile users", Proc. ACM SIGCOMM Symposium on Communication, Architectures and Protocols, October 1991.

[4] T. Imielinski and B. R. Badrinath, "Querying in Highly distributed environments," In the Proceedings of the 18th VLDB, August 1992, pp. 41–52.

[5] T. Imielinski and B. R. Badrinath, "Querying Locations in Wireless environments," In Wireless Communications: Future Directions, Kluweir Academic Publishers. October 1992.

[6] B. R. Badrinath and T. Imielinski, "Replication and mobility," In the Proceedings of

are weakly connected to the server through a wireless channel. Broadcast information could either be actual data, invalidations, or even control information such as lock tables or logs. Depending upon the what is broadcasted, appropriate schemes can be developed for maintaining consistency (appropriately defined) of data of a distributed system with mobile clients. Given the rate of updates, the tradeoff is between the periodicity of broadcast and the divergence of the cached copies that can be tolerated. The more the inconsistency tolerated the less often the updates need to be broadcasted.

Given that client disconnections will be frequent and may be of long duration, how can wireless connectivity be used to mitigate the ill-effects of disconnection? What kind of a transaction model is appropriate here? For total disconnection, CODA [16] uses optimistic methods. For weak or elective disconnection, what is the appropriate type of concurrency control scheme? With weak connections, the cost of connection (bandwidth), power restrictions, and resources available on the mobile will play a role in policy decisions such as: when to invalidate, when to write-back, how to batch requests among others. What tradeoffs in terms of resource restrictions should be considered in deciding the above policies?

4 New Access Methods - Research Issues

New Access methods are motivated by wireless transmission medium and by power restrictions on the palmtop. Wireless broadcasting will provide data "in the new form" that is literally "in the air." limitations on battery life will define a new cost measure on the data access and consequently warrant new solutions.

4.1 Wireless Medium

Since the cost of broadcasting over the wireless does not depend on the number of users who are listening, wireless medium provides an excellent platform for broadcasting information to a massive number of users. There are several examples of queries which are asked repetitively by huge number of users in some local area. Local traffic information, stock market data, local sales, events, emergencies, news services are all examples of information that would rather be broadcasted than provided "on a demand" basis. Broadcasting is simply much more economical in terms of transmission costs, which is especially important for wireless medium due to bandwidth restrictions. Broadcasting over wireless can be viewed as another memory medium - public "air" memory with the latency of access due to periodicity of broadcasting.

We see the following functionalities of the data broadcasting service:

- The notion of an index "channel" providing a directory to the information broadcasted. User terminals will listen to the index channel to determine how to tune to the proper information on the "data channel" selectively without listening to the data channel (channels) all the time. This is analogous to TV or Radio program which provides a timetable of the program and allows us to watch, listen or record selectively.

- Analogous to the rating system, (Nielsen's rating of the average number of households tuned to a particular TV program) periodically decisions should be made whether to keep broadcasting a particular piece of information or rather make it available on a "on demand" basis. If the demand for the particular data item drops it should no longer be broadcasted.

- We could view broadcasting as purely passive "read only" form of communication. We could also envision more interactive scenarios. For example after broadcast of data about available tickets users may place orders for them. The next broadcast should then reflect in some way the new state of the system resulting from such updates.

4.2 Energy Efficient Data Management

Due to a rather slow progress expected in the battery technology (in terms of the battery lifetime), available energy on the mobile platform is going to be one of the major resource constraint. Storage, especially non-volatile storage such as disk or CD ROM consumes significant amount of power. Further, transmitting data consumes an order of magnitude more power than just receiving data. Hence, for a given amount of available energy, the tradeoff is between the amount of data that can be accessed locally and the amount of data will that will be requested to be processed remotely and delivered later. Another factor to consider is the processing speed. Here again, longer the latency that can be tolerated in processing, lower the energy consumed. Thus, processing speed, storage cost (in terms of power), and amount of data transmitted and received and the tolerable latency will be factors in considering various aspects of data access and data organization.

to the new location of the user. This is particularly important if we want to maintain *performance transparency*, i.e., guarantee similar access and latency times independent of location. Notice also that the users will move in this environment "together" with patterns of read/write activity. In general, mobile environment requires dynamic replication schemes[27] that will replicate data on the basis of changing "centers of activity." In general, data should move closer to the active readers. If such readers relocate, then the data should "follow them" as opposed to following less active readers whose movement makes little difference in the data usage pattern. Preliminary work in this direction has been reported in [6].

3 Disconnection - Research Issues

Mobile terminals will be frequently switched on and off to save power or due to relocation. Apart from total disconnection, we need to distinguish between various *degrees* of connection: the weak connection (over a low bandwidth radio channel) and the strong connection (over the fixed network). Disconnections will have to be handled differently from failures or crashes that occur in traditional distributed system environments. Compared to failures or crashes, disconnections are elective in nature and can be prepared ahead of time as a part of the disconnection protocol. Similarly, transition from a strongly connected state to a weakly connected state may require a separate "weak disconnection" protocol. Let us now look more closely at the features that need to be supported by various disconnection protocols.

3.1 Cache Consistency

If the mobile user has cached a portion of the shared database then he may request different levels of *cache consistency*. While strongly connected he may afford strict consistency. On the other hand, while weakly connection he may resort to a more flexible approach of *weak consistency*; here the cached item is a quasicopy of the database item[2]. Each type of connection may have a different degree of cache consistency associated with it; the weaker the connection the "weaker" the maintainable level of consistency. In fact, the level of consistency may be part of the *type* system - as part of the protocol, the user while checking out a database item will specify the desired level of consistency of that data item. If the cached data is also going to be updated by the user, then appropriate concurrency control schemes are needed; especially for totally disconnected states. There are a number of possibilities which fill the spectrum between a fully pessimistic scheme which requires obtaining a lock for the entire duration of disconnection and a fully optimistic scheme with no need for locks but possible aborts upon "reintegration"[16]. For example, to avoid a lock being taken away "forever", a variant of the pessimistic scheme in which locks (lease locks) have only a specific duration (i.e., they are valid only for a specific period of time) may be useful. Further, a user may move after requesting a lock. In this case, the server has to search for the user when the lock is actually granted. Such a user, can be declared to be in the *lock waiting* period and be obliged to inform the server about his position more often. This way, when the lock is granted the server spends less effort in finding the user. Finally, locks may be granted only for a predefined scope of changes: for example one could only change his account by up to 10% of its value. Such locks may not be entirely exclusive, other users may operate on such data item with the understanding that its value carries a possible *error*.

3.2 Hand Off and Recovery

What should happen when the user moves moves or switches off in the middle of a transaction's execution? One approach would be for the mobile user, before switching off, to give instructions to the local host on completing the partially executed transaction. For instance, the mobile user may want to buy shares of AT&T stock if it reaches a new high today. He may then leave his "will" in the form of an active rule (trigger) at the local host.

If the user wants to continue the execution of a transaction after he has reached a new destination, then different mechanisms are necessary, If we assume that the log can be safely stored on the mobile that does not constitute a major problem; however it is becoming clear that for recovery purposes it is not a good idea to store the log of transaction on the mobile platform. Hence, we will require that the log be stored on the local server. In this case, as part of the move, the user should write a record <I Have Moved> possibly with the destination (if known) into the log. Such a record in the log will be something less than commit but it will allow the next site to "take it from there." In other words, when the user arrives at the new site he will continue from that point.

Finally, wireless broadcasting seems to be a powerful way of propagating the changes (or cache invalidations) to a massive number of users who

Disconnection is another important issue: mobile terminals will be often disconnected (switched off) mostly as a power saving measure. The main distinction between disconnection and failure is its *elective*[1] nature - disconnections can be treated as *planned* failures - which can be anticipated and prepared. There may be various *degrees* of disconnection ranging from total disconnection to weak disconnection. Weak disconnection occurs when a terminal is connected to the rest of the network via low bandwidth (may be intermittent) wireless channel.

The need for new data access modes stems from two factors: 1) the presence of the wireless medium which can be effectively used for broadcasting large amounts of data to a massive number of users and 2) the power constraints on the palmtop which may lead to a different "energy efficient" data access protocols and algorithms.

Finally, scale is another important factor - here it refers to the massive size of the potential set of users. Issues here cover massive distribution of services and their organization, organization of mediators and information brokers, general questions of knowledge representation and partition of knowledge among different objects in the system.

In terms of data management, we distinguish between *global* and *local* data management. Global data management deals with problems at the network level; these include locating, addressing, replicating, broadcasting etc. Local data management deals with problems at the end user level or the palmtop level; these include energy efficient data access, management of degrees of disconnection and query processing.

These new research problems are consequences of the unique physical characteristics of the computing environment. In particular:

- Small size and weight of the portable terminals

  This contributes to mobility and portability but puts significant restrictions on the interface through the limitations on the screen size and keyboard

- Bandwidth restrictions on the wireless link

  Wireless connection is a decisive factor contributing to mobility. However, bandwidth limitations imposes severe restrictions on the volume of data that can be transferred. Also since the cost of broadcasting over wireless does not depend on the number of users, the medium itself is well suited to provide broadcasting as a main method for dissemination of information.

- Power restrictions on the palmtop platform

  Contributes to frequent disconnections and the need for energy efficient access methods.

[1] Term coined by Dan Duchamp of Columbia University

2 Mobility - Research Issues

2.1 Locating Users

Locating users who are on the move and often to locations which are remote from home is a challenging task. In general, it is unrealistic and unnecessary to track locations of all users all the time. Hence, a database which stores locations of users will often be imprecise in terms of the exact user's location. For instance, a user's location may only be updated when the user crosses a border between two different areas or zones as opposed to updates on crossing a small cell. This, in general, will save on the number of location updates that the moving user will have to perform but will put an additional burden on the search process if the exact location of the user is sought. Since, the location of a user is not completely known, search (paging) over some limited area (zone) would have to be applied. In [4], we demonstrate the effectiveness of a general search process which is a combination of a database look up and paging (additional communication).

2.2 Queries on location dependent data

Examples of such queries include "where is the nearest doctor", "find the shortest route to the hospital" or "what is the number of taxi cabs in the area." If the location information stored in the database is incomplete, then in order to obtain a precise answer, a combination of database search and additional data acquisition during the run time of the query will often be necessary. Therefore a new model of query answering, which includes *data acquisition* during the run time of the query, will have to be developed. Again, [4, 7] show some preliminary results in this direction.

2.3 Replicating Information

Replication is simply a way by which the system ensures *transparency* for mobile users. A user who has relocated and has been using certain files and services at the previous location wants to have his environment *recreated* at the new location. It may involve just a migration of the state variables representing the ongoing process from one machine to another, it may also lead to the establishment of a new replica somewhere closer

Data Management for Mobile Computing

Tomasz Imielinski and **B. R. Badrinath**
Department of Computer Science
Rutgers University
New Brunswick, NJ 08903
e-mail:{imielins,badri}@cs.rutgers.edu

Abstract

Mobile Computing is a new emerging computing paradigm of the future. Data Management in this paradigm poses many challenging problems to the database community. In this paper we identify these new challenges and plan to investigate their technical significance. New research problems include management of location dependent data, wireless data broadcasting, disconnection management and energy efficient data access.

1 Introduction

The rapidly expanding technology of cellular communications, wireless LAN, wireless data networks, and satellite services will give mobile users capability of accessing information anywhere and anytime. In the near future, tens of millions of users will carry a portable (palmtop, laptop) computer (often called personal digital assistant (PDA) or personal communicator) with wireless connection to a worldwide information network. Coming years will most likely be the decade of *mobile* or *nomadic* computing.

This vision poses new challenging problems to the database community. How is the mobility of users going to affect data distribution, query processing and transaction processing? What is the role of wireless digital medium in the distribution of information? How to query data broadcasted over the wireless? What is the influence of limited battery life on data access from a mobile palmtop terminal?

The purpose of this paper is to investigate answers to these questions and identify which of them are technically the most challenging.

Below we briefly present the general architecture of the future Personal Communication Network (PCN) system. We follow then with a summary of the major research problems that need to be explored in this context.

1.1 General Architectures

The Personal Communication Network (PCN) of the future will provide a wide variety of information (voice, data, multi-media) services to users regardless of their location. General architecture of such network is still very much under debate but it is clear that it will include and extend existing infrastructure such as: the cellular (future microcellular) architecture (Analog and Digital cellular phones) capable of providing voice and data services to users with hand held phones, the wireless LAN; a traditional LAN (e.g., ethernet) extended with a wireless interface to service small low powered portable terminals capable of wireless access, specialized service oriented architectures such as those providing data broadcasting over unused portions of radio FM, private wireless data networks, or satellite services (paging) for users with special terminals. Some services or applications may use more than one wireless infrastructure (e.g., e-mail using the cellular to the internet and then to the wireless LAN and vice versa).

1.2 Research Objectives

We have categorized research challenges in data management for mobile wireless computing into roughly four orthogonal categories:

1. Mobility
2. Disconnection
3. New Data Access Modes
4. Scale

All the above categories are almost completely orthogonal: *Mobility* is a behavior which can manifest over both the fixed network as well as the wireless network. Issues such as addressing, data replication and migration will be addressed here.