Most-Urgent-First Protocol for Real-Time Wireless Channels in a Digital Cellular/PCN Environment

Kang G. Shin Zhihui Huang

Real-Time Computing Laboratory
Department of Electrical Engineering and Computer Science
The University of Michigan
Ann Arbor, MI 48109-2122
(313) 763-0391 (voice); (313) 764-4617 (fax)
Email: {kgshin,zhuang}@eecs.umich.edu

ABSTRACT

Providing mobile clients with real-time access to a shared information base in a digital cellular or Personal Communication Network (PCN) environment is vital to the success of many future Intelligent Transportation Systems (ITS) applications. The well-known limiting factors of such an environment are the relatively low channel bandwidth, a small number of available channels, and a potentially large number of users/clients to serve.

We propose a most-urgent-first (MUF) protocol to solve the problem of accessing the reverse traffic channels where mobile clients compete for use of the channels to convey their requests to the server. We also present a method for evaluating the urgency of each client's request for traffic information. The MUF protocol gives more urgent requests higher priority, responds faster to more important inquiries, accommodates more clients in each channel, and allows information sharing among a potentially large number of clients. Moreover, all services can be provided in real-time. Finally, we have developed an UltraSAN model [1] to evaluate the performance of the MUF protocol. Some of the issues related to the operation of the MUF protocol are also discussed.

Index Terms — Digital cellular/personal communication networks, Intelligent Transportation Systems (ITS), wireless real-time communication, Advanced Traveler Information System (ATIS).

1 INTRODUCTION

The advent of digital cellular and personal communication networks (PCNs) has brought new computerized information services into reality. Researchers envision "seamless coverage" of telecommunication services in metropolitan areas, shopping malls, and even throughout a country or a continent or the whole world. However, before realizing such coverage, there are many issues that must be resolved. One of these issues is the "timely" availability of information, since in some

---

The work reported in this paper was supported in part by Loral Incorporated and the US Department of Transportation under Grant No. DTFH61-93-X-00017. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the funding agencies.

critical applications the information that becomes available after a certain pre-specified time limit or a *deadline* is totally useless or may even be harmful. A real-time information service must satisfy the following conditions: (1)The service provider, or server, must collect information and update its knowledge/information base in real-time; (2) The system must deliver clients' requests for information to the server in real-time; (3) The server must process the requests in real-time; and (4) The system must deliver the requested information to the clients in real-time.

In a digital cellular/PCN environment, one can usually assume that the servers are capable of real-time information collection and request processing, both of which depend on real-time communication in wired networks (either point-to-point networks [2] or fast multiple-access local area networks [3]) and the ability to dynamically schedule and execute tasks [4, 5]. The remaining two conditions in the above list, however, have not been addressed adequately. While some researchers studied the wireless part of communication [6–10], only a few of them considered it in the context of real-time communication [6]. Unfortunately, none of them paid much attention to the real-time aspect of a digital cellular/PCN environment.

Digital cellular/PCN communication differs from traditional, wired multiple-access networks (e.g., Ethernet) in that channel bandwidth is small (normally 4800 bps or 9600 bps), the number of channels is limited and shared among many users, and the base station can monitor the access medium (channels) and feed back the status to mobile clients/stations. A base stations also have the capability to direct the mobile stations' actions. In addition to the traditional function of carrying voice or private data traffic, digital cellular/PCNs can carry information that is sharable among a certain class of clients or mobile stations designated by the system. This type of applications has the following distinct features:

- The information requested by a client can be shared among a selected group of clients;
- A client can be passive (i.e., the client can elect to only listen to the medium) or active (i.e., the client can elect to listen and transmit its requests dynamically);
- Requests from a single client are sporadic;
- The requests from different clients can have different degrees of *urgency* or importance;
- The requested information may have different degrees of urgency.

A typical example application is the Advanced Traveler Information System (ATIS) of Intelligent Transportation Systems (ITS) which is concerned with traffic (incidents, constructions, road closures), weather, parking, public transportation schedules, lodging, and other general yellow page functionalities. Most ATIS applications are soft real-time in the sense that a missed deadline may cause inconvenience to clients but will not cause a (catastrophic) system failure.

In this paper, we propose a wireless communication protocol, called the *most-urgent-first* (MUF) protocol, for ITS applications. In Section 2, we formally state the problem and the application environment. Section 3 describes our proposed solution to this problem. In Section 4, we present and summarize the simulation results of the MUF protocol with a stochastic activity network (SAN) model. Section 6 discusses some additional issues that were considered but not treated in depth. The paper concludes with Section 7.

2 PROBLEM STATEMENT

We consider a digital cellular environment in which clients use Time-Division Multiple Access (TDMA) to access the system, such as GSM (Global System for Mobile Communications) [11] and Telecommunications Industry Association's (TIA) IS-54 [12], instead of Code-Division Multiple Access (CDMA) techniques such as IS-95 [13].

2.1 TIA Interim Standard: IS-54

IS-54 is an interim standard for digital cellular operations in the United States. IS-54 uses a scheme to reuse the current 30 KHz analog FM channel (i.e., one single FM voice channel occupies 30 KHz in AMPS) for transmitting three time-divided digital voice signals. Without changing frequency assignments, IS-54 achieves a three-fold increase in the capacity of the cellular system, and provides an evolution path from the current first-generation analog cellular systems (AMPS, Advanced Mobile Phone Services) to digital cellular systems.

A frame in IS-54 is structured as follows. A 40 ms frame consists of 6 time slots, where each of which lasts 6.666667 ms and is 324 bits long. A slot can accommodate as many as 260 bits of user data. A full digital voice/data channel occupies 2 slots in each six-slot frame, which are exactly 20 ms apart in time. From the IS-54 documentation we derive some of the operational parameters for our simulation model (to be discussed later). It also provides the operational model described below.

2.2 Operation of Digital Cellular Systems

A wireless communications system consists of two types of agents: servers and clients. Servers wait for clients' requests and, after processing them, provide the clients with appropriate responses. The servers are usually abstracted as directory numbers that the clients can call. The servers can communicate among themselves to complement each other's capabilities, or to share information or workload as processing nodes do in a typical distributed computing environment. They can be thought of as the base stations of a digital cellular system. The clients are typically mobile stations (e.g., cellular phones and personal communicators) which gain the service by authenticating themselves to the server and deliver RPC-type (Remote Procedure Call) inquiries. A client typically interacts with only one server at any given time. It could switch to a new server if it happens to leave its current server's range (cell), or if it desires to request information the current server doesn't have.

In the system, the clients and the server(s) establish communication through the following three types of channels (see Figure 1):

Access channel: Used to request "access for service"; access channels are shared by clients. Clients may request voice services (e.g., normal cellular phone conversations), connection-oriented data services (e.g., telnet), or client-server data services (e.g., requests for information). We are interested in the last class of services, especially when the data (e.g., traffic congestion status) can potentially be shared among clients. After the request is granted via the *paging channel*, the client will acknowledge through the access channel and switch to the *traffic channel* specified by the server.

Paging channel: The server uses the paging channel to deliver system-related information to the clients, acknowledge the reception of service-access requests, and direct the clients to traffic channels. Normally, there are many access channels associated with a single paging channel. For example, in IS-95, there can be as many as 32 access channels for each paging channel.

Traffic channel: A traffic channel consists of two simplex channels, one in each direction (forward—from the server to the client, and reverse—from the client to the server). Only active clients can transmit on the reverse channel (some clients may elect to passively listen to the forward channel for information). Some active clients may not send a particular inquiry if they have already obtained the requested information by listening to the forward channel. They send inquiries according to the Most-Urgent-First (MUF) protocol to be described later. The traffic channel is thus shared by multiple clients and could potentially provide better service than dedicated channels or simple broadcast channels.

2.3 The Objectives

The objectives of this work are to:

1. provide the requested information to the clients in real-time;
2. provide faster access to more urgent (or perceived to be more urgent) requests;
3. deliver more important information faster, as long as it does not violate the promised quality of service to other requests;
4. accommodate as many clients in one channel as possible, because traffic channels are scarce resources in cellular systems; and
5. make information available to as many clients as possible.

Note that neither a broadcast scheme nor a dedicated channel scheme (where one client occupies one channel no matter how much (or little) data it has to transmit/receive) can satisfy all of these objectives. A broadcast scheme will not work because of the unpredictability caused by the ever changing number of related messages and the maximum delay associated with any given message to a single client. An unlucky client might have to wait for all the messages to be re-broadcast before it receives the desired message, if it “tunes in” immediately after the message has been announced. A dedicated channel scheme will not work for a different reason. Since many clients may request the same information and the information can be shared, this scheme unnecessarily burdens the servers with duplicated requests and stresses the resource (channel) by sending the same information many times. Moreover, under this scheme the server will not be able to serve as many clients as it could if it does not have to serve duplicated requests.

2.4 The Approach Taken

When a client issues an “access for service” request, it includes in its message the intended mode of operation; an *active* mode allows the client to request information via the traffic channel while a *passive* mode will not. We also assume that the server may have several traffic channels open to use. Once a client is directed by the server to a shared traffic channel (we will call this “shared data channel”), it will be able to continuously monitor the messages on the forward channel. Some messages on the forward channel that are addressed to specific clients (such as power control orders) will be discarded by the other clients. Otherwise, all clients assigned to the same traffic channel will be able to receive shared information regardless of their operational modes. By allowing this type of passive access, we meet objectives 4 and 5 in Section 2.3.

In addition, we assume that all messages transmitted on the shared data channels follow a certain formatting protocol, which may be application-dependent. For example, requests for traffic information will typically include a field that specifies the location of interest. However, the actual formatting protocol will not be discussed any further in this paper. Since the server has absolute control over the forward channel, objective 3 can be addressed as a message scheduling problem, the subject discussed in Section 5.

The remaining objectives are related to the control of access to the reverse channel and are addressed in Section 3. Note that the reverse channel is slotted and a client will request information only at the beginning of a slot and the request will last only for the duration of that slot. Our basic idea is to have the clients compute urgency indicators for their requests. The clients then compare their indicators with the *system urgency window* (SUW) set by the server. If a client’s urgency indicator falls within the SUW, the client transmits his/her request at the next slot on the reverse channel. The SUW is controlled by the server and will change dynamically to reflect the status of the reverse channel and control the access to it.

3 THE MUF PROTOCOL

Access to the reverse channel is controlled by an *urgency* indicator. This indicator reflects both the time constraint and the importance of the request that cannot be modeled by its deadline or laxity alone. For example, a request for the location information where no alternative routes around the location are available could be more important than a similar request where many alternative routes are available, since failure to provide the information may result in a longer delay to the client in the former case.

3.1 Computation of Urgency

After a client formulates a request, it will dynamically compute an urgency indicator. The computation of the urgency indicator will be different for different types of information requested.

For a demonstrative purpose, we define an example urgency U for traffic incident reports as:

$$U = T \times (1 - H) \times R/P$$

where

T is the expected time for the client to reach the target area that it is inquiring about. It can be calculated by a combination of client's speed and the distance between the client and the target area (on freeways, for example), or the sum of link times for the links that lead to the target area (for local roads);

H is a real number in [0..1) indicating the target area's incident history. A higher value of H is assigned to more incident-prone areas. One can obtain this number from the statistics a public service bureau has had collected over a certain period;

R is a real number in (0..1] specifying the availability of alternative routes around the location where the client's inquiry is being made. A higher R value indicates a higher number of diversion routes. It can also be obtained from a public service bureau.

P is the time between two consecutive slots of the same channel. This value is 20 ms for systems that conform to the IS-54 specification.

Note that H can be defaulted to 0 when the information is not available, and R can be defaulted to 1 when the information is not available, all to make an optimistic request. Other default values, such as a neutral value of 0.5, can also be used.

3.2 The MUF Protocol

The MUF protocol controls access to each shared reverse traffic channel. Clients transmit requests only at the beginning of, and in the time slots assigned to the channel. Since the server monitors the reverse channel continuously, it can determine in real-time the "best" value for the system urgency window (SUW) size. If a slot on the reverse channel is corrupted, the server shrinks SUW and transmits the new SUW size with the next message (padded in a message using the Slow Associate Control Channel (SACCH) capability available in a slot). When the server sees an idle reverse channel slot, it expands the SUW size and then transmits it with the next message (see Figure 3). The forward channel carries responses from the server and other system messages. Each response has associated with it a validity timer which indicates when the information may become invalid.

If a client listening to the forward data channel sees the information it is interested in *and* the validity timer for the information has not expired (or is not expected to expire before the client reaches the target area), it will refrain from sending a new request. The client also monitors the forward channel for the server-controlled SUW and records it (see Figure 2).

If a client decides to send a request, it computes the urgency of the request. If its value falls within the SUW, the client will try to send the request at the beginning of the next time slot. Otherwise, it will wait and repeat the process for the next slot. If a request is sent without any collision, a response to the request will be received by the server via the forward channel after a system-determined, fixed amount of time (*response latency*). This latency is guaranteed by the network and includes any delay associated with the server's communication with other servers to obtain the current/correct information. If a response is not received after the response latency and the client does not receive other messages from the server, the server is declared to have failed.

If two or more clients transmit in the same time slot, a *collision* occurs. The server sends the status of the reverse channel (in this case, collision) through the forward channel so that the clients are notified and can take appropriate actions (see Figure 4).

4 ANALYSIS OF THE MUF PROTOCOL

To evaluate the performance of the MUF protocol, we need: (1) the distribution of the urgency number of clients' requests, (2) the function(s) according to which the urgency changes, (3) the rate at which the clients generate their requests, and (4) the probability distribution of the requests over the set of all possible destinations (locations, stock of a particular company, etc.). This will affect the number of requests that are cancelled as a result of receiving the information in the response to another competing request, in addition to system parameters such as the server range (typically a cell), size, shape, the number of clients on the channel, the rate at which the clients travel, and external (unpredictable) elements such as multipath, highway overpass, building shadow, terrain that may cause transient failures.

We have developed an UltraSAN model [1] to simulate the operation of the MUF protocol based on the following assumptions:

1. The channel timing is the same as that in IS-54.
2. There are only active clients, since no passive client will compete for the slots of reverse channels.
3. The urgency of all requests is assumed to be uniformly-distributed integers in the range of zero and a maximum specified in the model, measured in units of slot interval (20 ms).
4. The urgency of each request is monotonically decreasing with time. This may not reflect the real-world situations where a request's urgency may change randomly due to the movement of clients and the structure of the road networks instead of changing linearly.
5. The transmission delay is assumed to be the same for all clients. This can be thought of as the maximum delay that can be tolerated by the client farthest away from the server.
6. A client will not generate a new request until the previous request is transmitted or timed out.

There are three parameters in the model for each channel: the number of clients assigned to the channel, the rate at which the clients generate requests, and the urgency range of the requests. The model is simulated under three sets of parameters: (1) the client request rate is 0.005 or 0.050 (representing the rate of 5 requests per second and 1 request per 20 ms, respectively); (2) the number of clients in the channel ranges from 1 to 20.

Performance data has been collected for the average number of pending requests in the channel, the probability of a request timing out, and the average channel utilization, all against the number of active clients using the channel for each simulation scenario.

As can be seen from Figure 5, a higher request generation rate results in a higher number of pending requests in the system. For the case where the rate is 0.05 (or one request per 20 ms), this number is roughly equal to the number of clients. This is because the maximum request transmission rate is only one per 20 ms. In the case where the rate is 0.005, the number of pending requests is smaller. The size of maximum SUW does not have much impact as expected.

Figure 6 shows the dependency of the MUF protocol on the allowed range of urgency numbers. The rate of request generation does not show any notable effect because the assumption of one outstanding request per client is enforced.

Figure 7 shows that utilization depends more on the allowed range of urgency indicators than on the request generation rate. Smaller urgency range provides better utilization, with the same number of clients. This is because there are probabilistically fewer requests in a small window with a larger range. (The utilization is defined as the portion of time the channel carries uncorrupted clients' requests.)

Note that the offered load L on the channel can be expressed as: $L = R \times C$ where R is the number of outstanding requests and C is the channel capacity. The results thus obtained have shown good performance of the proposed scheme in controlling and coordinating access to reverse traffic channels, especially when the range of urgency indicators is relatively large.

5 FORWARD TRAFFIC CONTROL

A forward (shared) channel carries responses from the server and other server-generated messages such as SUW size and client power control commands. The server schedules the responses based on urgency. For example, if the result is affirmative (an incident has occurred), it is deemed more urgent. This scheduling must be deadline-driven since the server has to respond within the response latency specified by the system.

In fact, the server schedules the responses based on urgency if all deadlines are observed. Otherwise, more urgent responses may be scheduled at the expense of missing the deadlines of less urgent responses as explained below. The server maintains two separate queues: the messages that are important (e.g., incident reports) are entered in queue 1 (Q1), the messages that are of less significance are entered in queue 2 (Q2). Each queue is sorted in ascending order of message laxities (the latest time for a message to be transmitted without missing its deadline). A message is chosen for transmission as follows:

- If Q1 is not empty, the message at the head of Q1 is transmitted.
- If Q1 is empty and Q2 is not empty, the message at the head of Q2 is transmitted.
- If both Q1 and Q2 are empty, a null message is transmitted.

6 RELATED ISSUES

Discussed below are some of the issues that will affect the operation of the proposed scheme.

6.1 Access to the Access Channel

In Section 3, we described a protocol for the clients in a traffic channel who compete for the server. But we did not mention how a client is admitted to a shared traffic channel. It turns out that the access channel for the type of applications we are interested in can also be controled in the same manner as the traffic channels.

Since the access channel is shared among clients requesting different services such as voice, continuous data or client-server data, each with a different timing constraint or no constraint at all, no single heuristic access control seems to satisfy the requirements of all clients. For example, laxity-based access control will mandate all clients to keep laxities for their request-for-access messages, while laxity is not an integral part of normal dialing process and it may not be the best heuristic for clients requesting traffic information.

Typically access control to an access channel is slotted ALOHA randomized among all clients, with the help of a persistent backoff which varies with the type of messages — access requests or responses to paging. A few measures can be taken to give access priority to clients of certain classes (e.g., those who pay a larger premium). For example, the system may instruct all or some of the clients of a certain class to use certain slots of the channel to transmit their requests or responses to paging messages. The system may have them use an exclusive access channel and compete among themselves. In either case, access control to the access channel is reduced to the control of access to a traffic channel and can be dealt with, as described earlier in the MUF protocol or other suitable access control schemes.

In any case, the client that sends a request will wait a fixed amount of time for a reply from the server through the paging channel. If the timer expires, the client knows that a collision has occurred. It then uses a persistent scheme or a collision resolution scheme adopted by the system to continue the access attempt. If a reply is received before the timer expires, the client will follow the instruction from the server and switch to the designated traffic channel.

6.2 Periodic Guaranteed Accesses to the Access Channel

Some clients may wish to make periodic access to the server with delay bound guarantees. This can be accomplished by assigning clients access channel slots with the required timing guarantees when each client first registers with the server. This problem is in effect reduced to that of assigning tasks with timing constrains to a processor. Existing solutions in the literature can be used. Problems may occur when the client moves out of the cell of the current server. This can happen when the client is idle (idle handoff) or when the client is active (normal handoff). In either case, the server of the cell the client is moving into must again arrange the same type of access slots for the client, which it may or may not be able to do. This particular problem is solvable because of the “centralized” control that can be exercised by the server. However, further work is needed for the details of such a solution.

6.3 Handoff

Since the application introduces new information about the client/server relation (whether the client is passively listening or actively transmitting), the handoff procedure between cells should be modified accordingly.

6.3.1 Idle Handoff

Idle handoff occurs if the client is idle when the handoff takes place. If the client has established a periodic, real-time access pattern with the current server, the current server will have to notify the potential new server so that the new server will make the necessary resource reservation for the client. If no server can satisfy the requirements of the client, the client will be notified by the new server.

6.3.2 Busy Handoff

A busy handoff happens if the client is active when the handoff takes place.

- If the client has established a service pattern with the old server, the transition is the same as in the case of idle handoff, except that the client is immediately assigned to a traffic channel in which it can transmit its requests.
- If the client did not request a pattern to be established, the new server will assign the client to an existing shared channel. If there is no such channel, the server will find a new channel and assign the client to it. If this attempt fails, the server will notify the client and drop the client.
- If the client operates in passive mode, the new server will assign the client to an existing shared channel. If there is no such channel, the server will notify the client as such and the client will be dropped.
- If the client is a normal user (not a client to our application), the transition is the same as described in, for example, IS-54.

7 CONCLUSIONS

We have examined the problem of providing real-time access to information in a digital cellular/PCN environment. The general limiting factors in this environment are the relatively small bandwidth of the channels, the small number of available channels and the potentially large number of users. These factors distinguish our problem from the others commonly seen in general multiple-access networks.

We have proposed the most-urgent-first (MUF) protocol to solve part of the problem based on a new concept of *urgency*. MUF is designed to control access to reverse traffic channels where the clients compete to have the system deliver their requests to the server. A method for the evaluation of urgency for ATIS traffic related information was also given as an example of combining time and the (perceived) importance of requests. An UltraSAN model [1] was developed and simulations performed. The results obtained from the experiments are very encouraging. We have also discussed some of the issues related to the operation of MUF protocol in the presence of other types of applications.

The MUF protocol and the concept of using urgency to control access to multiple-access medium can provide additional leverages not available when we consider time only. The MUF concept can be applied to vehicle-to-vehicle communications in ITS for safety-critical applications. The MUF protocol may also be applicable to Code-Division Multiple Access networks.

There still remain several issues that warrant further study, including:

- A more detailed evaluation of the performance of the MUF protocol is needed. For example, with the increase of the number of clients in one channel, the probability of two or more clients having the same request increases (for a fixed number of possible requests). When one of the clients successfully has transmitted the request and received a response from the server, the other clients with the same request will cancel their own requests thus reducing the contention for the reverse channel.
- Comparison with other protocols in terms of dropped message ratio, and system utilization.

References

[1] W. H. Sanders and J. F. Meyer. A unified approach for specifying measures of performance, dependability, and performability. In *Proc. International Working Conference on Dependable Computing for Critical Applications*, 1989.

[2] Dilip D. Kandlur, Kang G. Shin, and Domenico Ferrari. Real-time communication in multi-hop networks. *IEEE Transactions on Parallel and Distributed Systems*, 5(10):1044–1056, 1994.

[3] James F. Kurose, Mischa Schwartz, and Yechiam Yemini. Multiple-access protocols and time-constrained communication. *Computing Surveys*, 16(1):43–70, 1984.

[4] Kang G. Shin and Chao-Ju Hou. Analytic models of adaptive load sharing schemes in distributed real-time systems. *IEEE Transactions on Parallel and Distributed Systems*, 4(7):740–761, 1993.

[5] Krithi Ramamritham, John A. Stankovic, and Wei Zhao. Distributed scheduling of tasks with deadlines and resource requirements. *IEEE Transactions on Computers*, 38(8):1110–1123, 1989.

[6] Wei Zhao, John A. Stankovic, and Krithi Ramamritham. A window protocol for transmission of time-constrained messages. *IEEE Transactions on Computers*, 39(9):1186–1203, 1990.

[7] Donald C. Cox. Wireless network access for personal communications. *IEEE Communications Magazine*, 30(12):96–115, 1992.

[8] D. J. Goodman, R. A. Valenzuela, K. T. Gayliard, and B. Bamamurthi. Packet researvation multiple access for local wireless communications. *IEEE Transactions on Communications*, 37(8):885–890, 1989.

[9] Allen Salmasi and Klein S. Gilhousen. On the system design aspects of code division multiple access (CDMA) applied to digital cellular and personal communications networks. In *Proc. IEEE Vehicular Technology*, pages 57–62, 1991.

[10] W. C. Wong and David J. Goodman. A packet reservation multiple access protocol for integrated speech and data transmission. *IEE Proceedings-I*, 139(6):607–612, December 1992.

[11] Michel Mouly and Marie-Bernadette Pautet. *An Introduction to GSM*. Michel Mouly and Marie-Bernadette Pautet, France, 1993.

[12] Telecommunications Industry Association. *EIA/TIA Interim Standard: Cellular System Dual-Mode Mobile Station-Base Station Compatibility Standard*, April 1992.

[13] QUALCOMM Incorporated. *Proposed EIS/TIA Interim Standard: Wideband Spread Spectrum Digital Cellular System Dual-Mode Mobile Station-Base Station Compatibility Standard*, April 1992.

LIST OF FIGURES

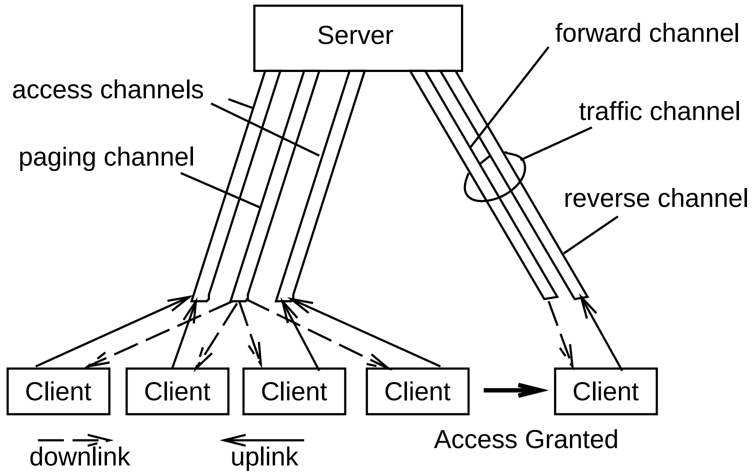

Figure 1: Overview of Digital Cellular Operations

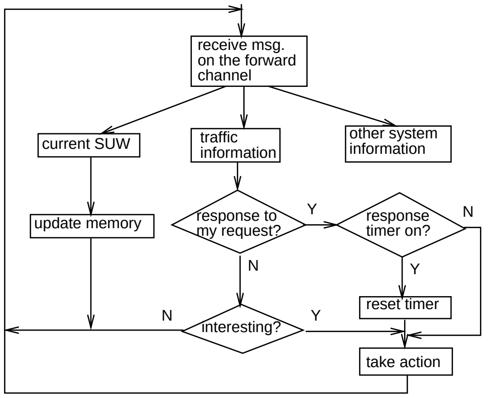

Figure 2: Client Actions on the Forward Channel

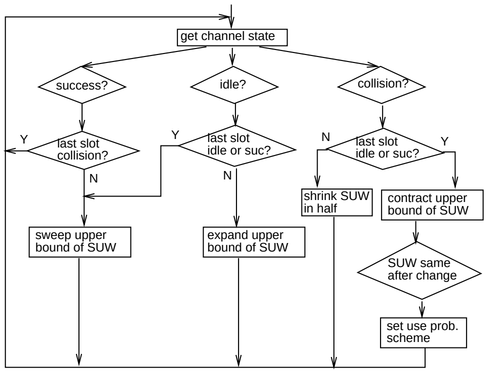

Figure 3: Server Actions on System Urgency Window Size

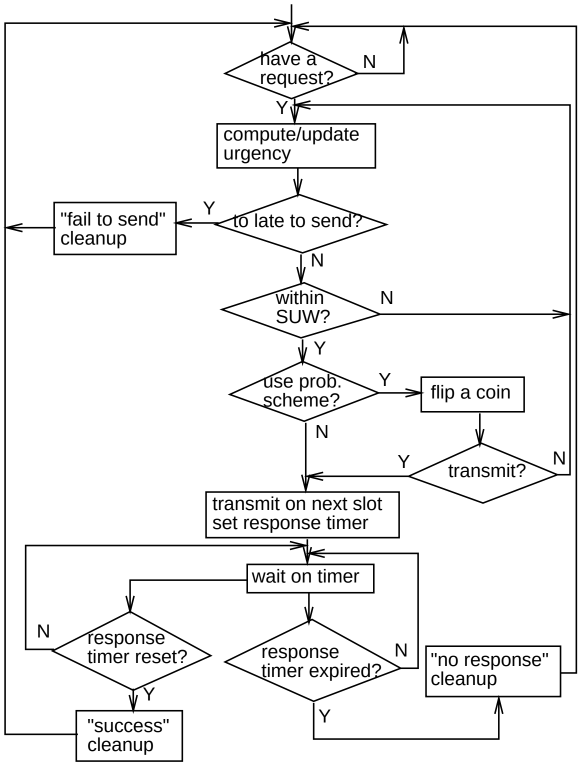

Figure 4: Client Actions on the Reverse Channel

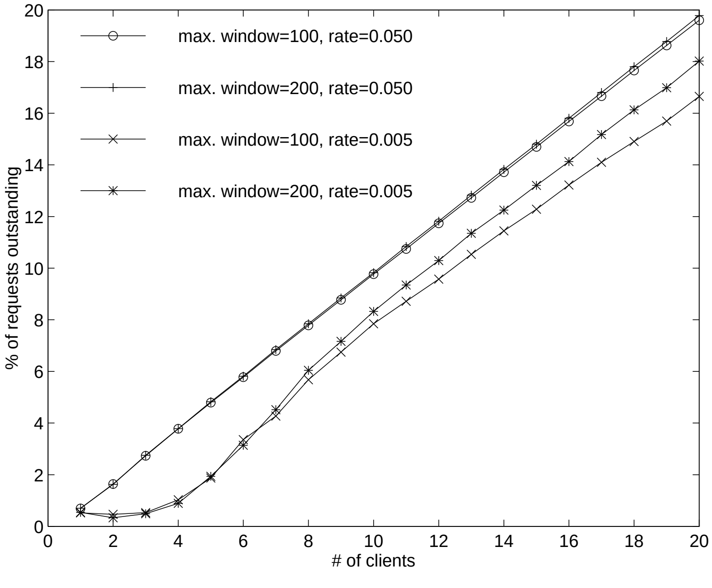

Figure 5: Average Number of Pending Requests

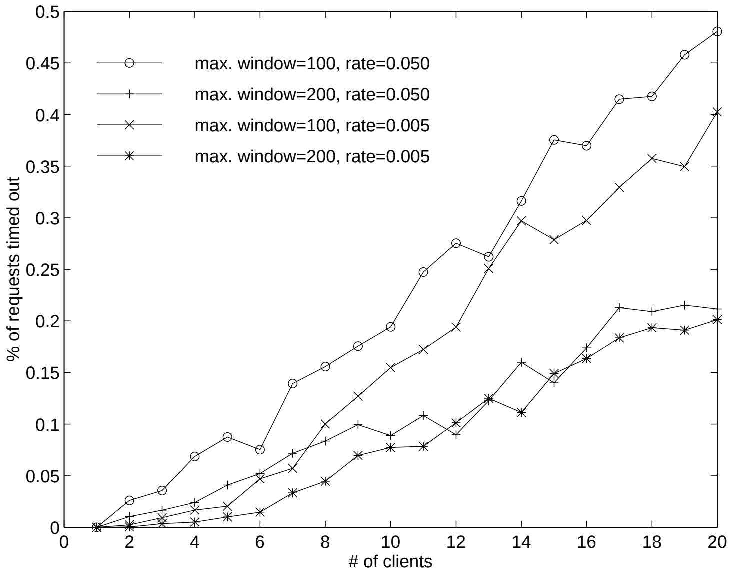

Figure 6: Percentage of Timed Out Requests

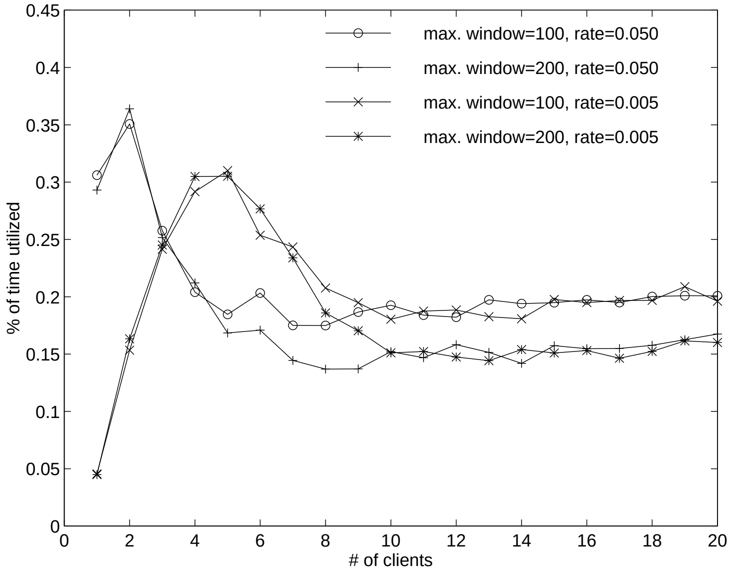

Figure 7: Channel Utilization