

# Differences in Cost and Benefit of Prefetching in Circuit-Switched and Packet-Switched Networks

*Michael Angermann*

Institute of Communications and Navigation\*,  
German Aerospace Center (DLR),  
P.O. Box 1116, D-82230 Wessling, Germany  
Michael.Angermann@dlr.de, Fax: + 49 8153 28 1871

## Abstract

**Prefetching of data is frequently discussed candidate technology for mitigation of negative effects caused by low-bandwidth and high-delay connections, especially common in mobile wireless scenarios. An inevitable side-effect of prefetching is increased overall transferred data volume. In this paper this side-effect is investigated and set in relation to the achieved reduction in user waiting time. Simulations on request/response level have been performed to investigate the distinct influence of prefetching in circuit-switched and packet-switched networks on network costs. Simulation results are presented and discussed. Results show that while prefetching can achieve the desired reduction of waiting time for both circuit-switched and packet-switched networks, these gains are coupled with pronounced increase in network cost for packet-switched networks but with slight decrease in network costs for circuit switched networks.**

## 1 Introduction

Prefetching of content is a technique with the potential to reduce the long waiting times that users usually experience when using data services in today's mobile wireless networks. The perception of a slow network is not only caused by the limited transport capacity ("bandwidth") of these networks but also by their substantial latency due to sophisticated and complex channel-coding schemes that require long interleavers and time-consuming digital signal processing. In ex-

treme cases roundtrip times up to seconds are a direct consequence. Applications with high interactivity like web-browsing therefore suffer from waiting times after each click, even for small responses.

While prefetching has been applied for certain applications like web-browsing with classical fixed dial-up connections to the internet it has also been controversially discussed for its unwanted side-effect of increasing the load of the involved HTTP-servers. However, the maximum potential load caused by any of today's mobile wireless clients seems negligible compared to the heavy load from clients connected via digital subscriber line (DSL) or LANs. For the application of prefetching in mobile scenarios its influence on the user's bill is nevertheless of considerable relevance.

In this paper we want to use the results of simulations to investigate and compare the benefits and costs of prefetching for both circuit-switched (CS) and packet-switched (PS) networks. We make this distinction between CS and PS for their typically distinct structure of network costs. While the use of CS networks is commonly charged on the basis of the duration of a connection ("airtime"), operators of PS networks use the ability to charge for the amount of transferred data ("volume").

Thorough analysis of performance modelling of prefetching has been performed by Tuah [1]. Adverse effects of prefetching have been analyzed and countermeasures proposed by Crovella in [2]. In [3] probabilistic fundamentals of prefetching have been analyzed. In contrast to the short-term prefetching investigated in the paper at hand Venkaterami et al have discussed long-term prefetching in [4]. Other valuable resources on the topic are [5],[6] and [7].

We would like to point out the opinion that despite multiple similarities and interactions between caching and prefetching, research in these two fields has different problems to solve – for the scenarios of interest here – for two main reasons: a) prefetching typically retrieves documents a few seconds up to a few minutes before

---

\*This work was done within the *Heywow* Project. More Information can be found at <http://www.heywow.com>

the user requests them. Usually this is close enough to the instant the user is going to request them that a major concern of research in caching, the problem of stale documents, occurs only very rarely. b) mobile devices are limited not only in connectivity but also in memory resources. Large caches are therefore difficult to maintain<sup>1</sup>. Prefetching, as described here, usually keeps only those few documents in memory that have a high probability to be requested by the user.

The main benefit and the metric to rate the performance of prefetching is the reduction of waiting time for the user. As we will see, this benefit does not always come for free, but has repercussions on costs. In order to solve certain decision or optimization problems it seems necessary to include the user’s waiting time into a common cost function together with network costs and energy consumed on the device. In [8] this approach has been proposed. In the paper at hand we are aware of the necessity to formulate a common cost function but evade it by presenting quantitative results of the saved user time and the necessary costs. In our further work we will have to continue at this point and formulate a sensible cost function in order to implement systems capable of autonomously deciding in the user’s interest.

## 2 System Model

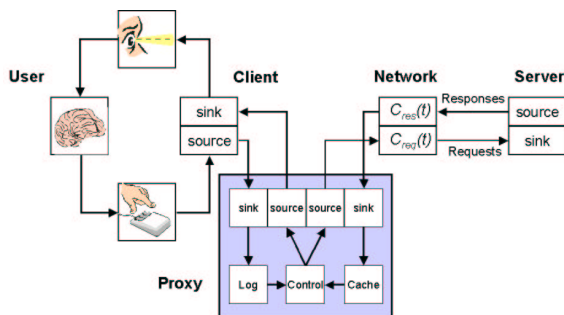


Figure 1: System model overview

With the intention to model a system used for general request/response oriented applications a system comprised of standard components for web-browsing using the HyperText Transfer Protocol (HTTP) was chosen. Depending on the user’s selection of a document the primary request of this document is sent to the server (typically the request for a .html-file). After arrival of the corresponding response the secondary requests (typically requests for embedded elements such as pictures, sounds, animations) are issued. The moment all

<sup>1</sup>This argumentation does by no means intend to discard caching in favor of prefetching, since the reuse of an already transferred document is of course the most network resource efficient technique.

responses that belong to the document have arrived, a timer with the viewing time for this particular document is started. The primary request for the next document is sent the moment the viewing time is finished. This process is illustrated in Fig. 2. We used the empirical distributions derived by Mah [9] to generate the sizes of primary and secondary requests, the number of secondary request per document as well as the viewing times. The same model has already been used in [3].

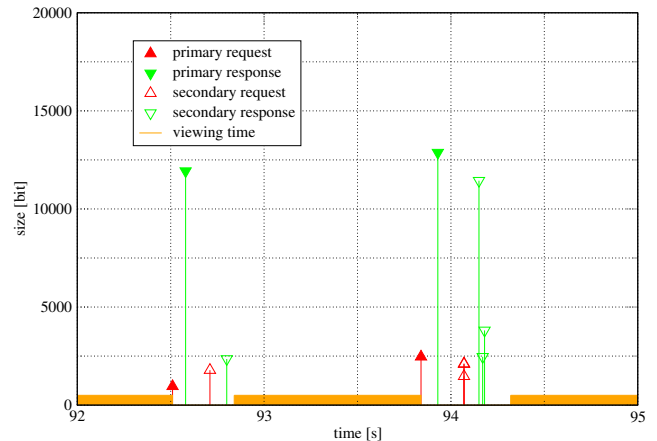


Figure 2: Snapshot of sample HTTP requests and responses and viewing times

For the purpose of simulating the effects of prefetching we had to extend this model to account for the selection process of documents.

The sequence of user decisions that select the documents is assumed to be the driving random process of the system. Each document is assumed to have a number of  $N$  candidate successors. The probability of these candidates to become selected is distributed according to Zipf’s Law. Zipf’s Law states that the probability of the  $i$ -th most likely event is proportional to  $1/i$ . This is also called the *strict* Zipf’s Law. Many interesting experiments show a slight modification of this law. They can be more adequately modelled with a probability proportional to  $1/i^\alpha$  for the  $i$ -th most likely event. The value  $\alpha$  then typically takes a value of less than unity. For  $\alpha = 1.0$  this modified law is equivalent to the strict Zipf’s Law [10].

The proxy (see Fig. 1) is assumed to know of the  $N$  successor candidates and their probability with respect to the currently viewed document. In section 5 we will briefly describe how a real proxy can obtain a close approximation of this “genie”-knowledge.

The moment the viewing time starts, the proxy uses its knowledge and starts to retrieve the objects of the candidate documents. The proxy immediately stops these speculative retrievals whenever any of two possible events occur: a) the viewing time has ended and

a non-speculative request is issued by the client; or b) no more documents with a probability higher than a pre-adjustable threshold  $p_{th}$  are available. Requests and responses are transported by a network with constant transport capacity  $c$  and constant delay  $T_\delta$ . Up-link and downlink channels are independent and have equal capacity and delay. All servers are parametrized with the constant delay  $T_S$ . Their response is created  $T_S$  after the instant the request has fully arrived, independently of the response size.

As mentioned in the introduction we differentiate between circuit-switched and packet-switched networks. This distinction is only made for the access network of the mobile wireless network since its costs are assumed to dominate. For circuit-switched networks we assume the time a user needs to sequentially work through a given plan (a given sequence of documents and viewing times) to linearly determine the total network costs. These costs are independent of the volume of the transferred data. For packet-switched networks the reciprocal situation is assumed: Only transferred volume (regardless of their having been retrieved speculatively or not) linearly determines the total network cost. The length of the period of time the network is in use does not influence network costs.

We will later discuss these costs in light of the waiting time the user has to endure while following a certain plan.

### 3 Simulation and Results

The simulation has been designed for fast and flexible adaptation to prefetching controller strategies, traffic and cost models as well as network models. For the results presented here, the simulation consists of client, proxy, network and server. Traffic is modelled at request/response level using Mah’s empirical distributions. Modifications have been made to limit the maximum viewing time to 300 seconds for the purpose of achieving meaningful results in reasonable simulation time. Influences of these modifications on the results regarding the performance of prefetching are negligible since all candidate documents have usually been transferred within 300 seconds. The simulation operates in a “closed loop”, where the arrival times of responses determine the beginning of the viewing times and therefore the timing of the following requests.

#### 3.1 Single Shot Simulation

The candidate documents are drawn from a distribution following Zipf’s law with parameters  $N = 7$  and  $\alpha = 1.0$ .

Two separate channels for requests and responses (up-link and downlink) are modelled with identical parameters. The transport capacity of each channel is chosen

to be constant at  $1 \cdot 10^5$  bit/s. Each channel introduces 0.1 seconds of delay. The server delay is fixed at 50 ms. Only two values for  $p_{th}$  are used: The no-prefetching case is simulated by setting  $p_{th} = 1.0$ , whereas for unrestrained prefetching  $p_{th}$  is set to 0.0.

The user plan consists of 100 documents that are identically chosen for the prefetching and no-prefetching scenarios.

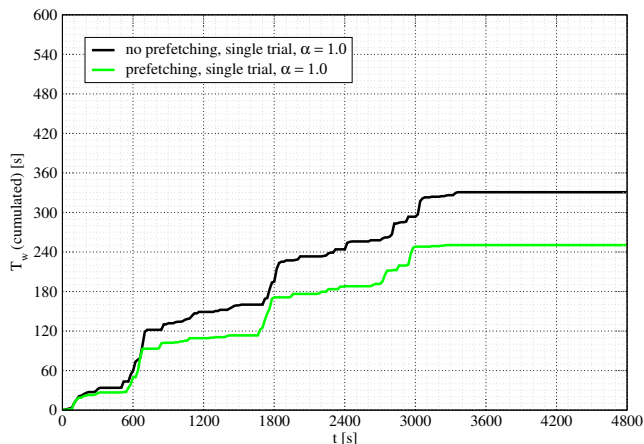


Figure 3: Influence of prefetching on user waiting time, identical for CS and PS, 100 documents, single trial

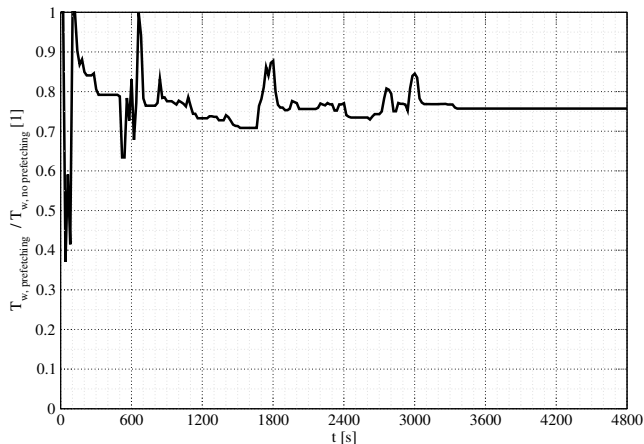


Figure 4: Ratio of user waiting time, 100 documents, single trial

In Fig. 3 the influence of prefetching is depicted. For the chosen parameters it takes the user 3343 seconds to complete his plan consisting of 100 documents if no prefetching is applied. The same plan is executed in 3263 seconds when prefetching is applied. After these durations no more documents are requested, resulting in the flat segment on the right of the curves. We see the cumulated waiting time the user has to endure with and without prefetching. While executing his plan without

prefetching the user has to *wait* for 330.7 seconds. This can be reduced to 250.4 seconds when prefetching is applied. In this trial prefetching reduced the waiting time by 80.3 seconds or 24.3%. The temporal development of this relative gain is depicted in Fig. 4.

It is interesting to note that the sporadically occurring increases of the ratio are caused by the fact that in the prefetching case the user has to wait for a shorter time for some larger document that add up to the cumulated waiting time. These effects are only temporary, as in the case without prefetching the user will also request these documents.

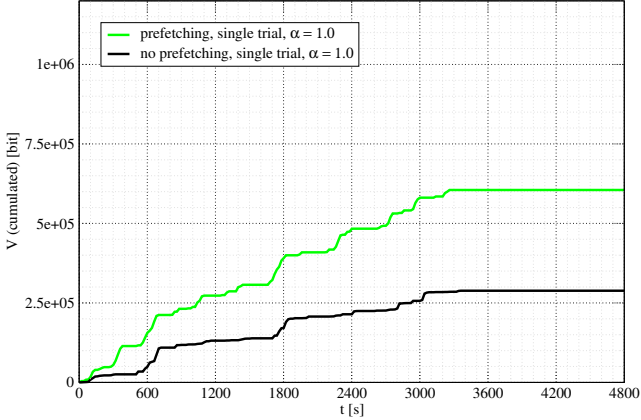


Figure 5: Influence of prefetching on network cost for packet-switched network, 100 documents, single trial

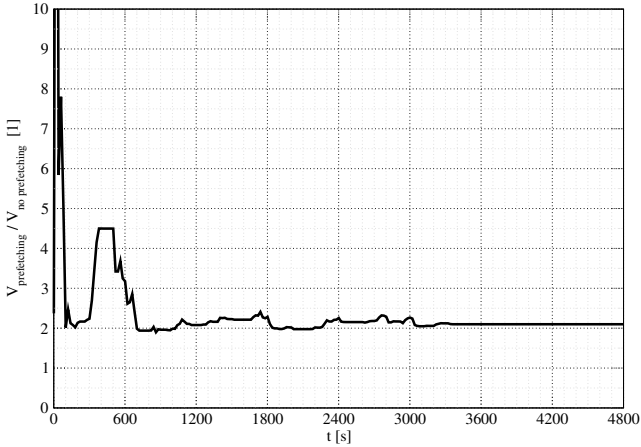


Figure 6: Ratio of network cost, 100 documents, single trial

The achieved reduction of user waiting time is accompanied by an increase in transferred data volume  $V$ . In Fig. 5 the absolute values of the cumulated transferred volume are shown. Here, prefetching results in the upper curve. The cumulated volume for speculative and non-speculative requests and responses nec-

essary for transferring the 100 documents amounts to  $6.05 \cdot 10^7$  bits. Whereas without prefetching only non-speculative requests and responses have to be transported and amount to  $2.88 \cdot 10^7$  bits. The ratio is plotted in Fig. 6. We can clearly see that more than twice the amount of data had to be transported.

For the packet-switched network this directly corresponds to an increase in cost of the same factor. In contrast to this, the costs for a circuit-switched network are reduced. The total time necessary to execute the plan (sequence of documents) has been slightly reduced by 2.4%, which directly results in the same amount of saved network costs.

### 3.2 Monte Carlo Simulations

The results are very sensitive to the outcomes of the individual random trials that generate the documents' requests and the viewing times. This is a well known fact that occurs in simulations of application-layer traffic. For the purpose of obtaining insight into how a system performs *on the average* the method of Monte Carlo simulations can be applied. The following results are obtained by averaging over 10 trials. All trials have been initialized with different and random seeds of the random generators. To keep the high variation of the individual trials in mind their outcomes are also included in the respective diagrams. All system parameters are chosen identical to the single shot simulations. In Fig. 7 the average waiting time of 10 trials with and without prefetching is shown in addition with the outcomes of the individual trials. We can see the strong inter-trial variations. The respective ratios of the the individual trials and the ratio of the means waiting time are shown in Fig. 8.

In the single trial experiment described above we seem to have picked a random sequence of documents and waiting times that resulted in a fairly representative perceived performance of the system. For the average over trials we see a reduction in mean cumulated waiting time from 383.5 seconds without prefetching to 290.7 seconds when prefetching is applied. This reduction of 24.1% comes with an increase of almost exactly 100% in transferred data volume ( $3,43 \cdot 10^7$  bit to  $6,88 \cdot 10^7$  bit), which is equivalent to the increase in network cost for a packet-switched network scenario (see Fig. 9 and Fig. 10). The mean time to complete the plan of 100 documents is reduced from 3284 seconds to 3190 seconds by prefetching, resulting in an equivalent reduction of 2.8% in network cost for the circuit-switched network scenario.

According to [3] the reduction in waiting time and the increase in transferred volume depends on the randomness of the documents. If we assume a distribution of the probabilities according to Zipf's Law, this randomness is adjusted with  $\alpha$ . In the results presented so far we have used the threshold probability  $p_{th}$  only to

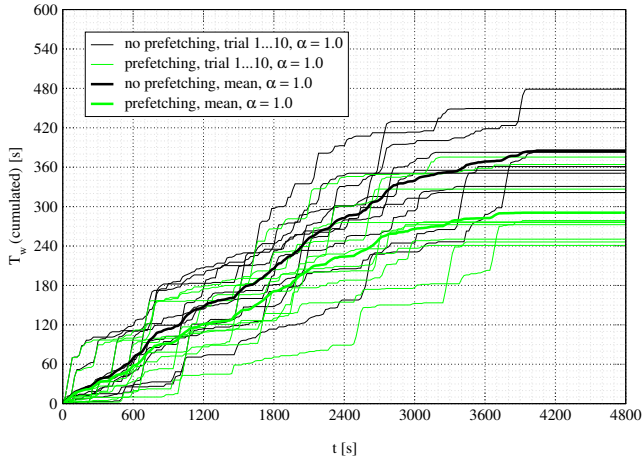


Figure 7: Influence of prefetching on user waiting time, 100 documents, average over 10 trials

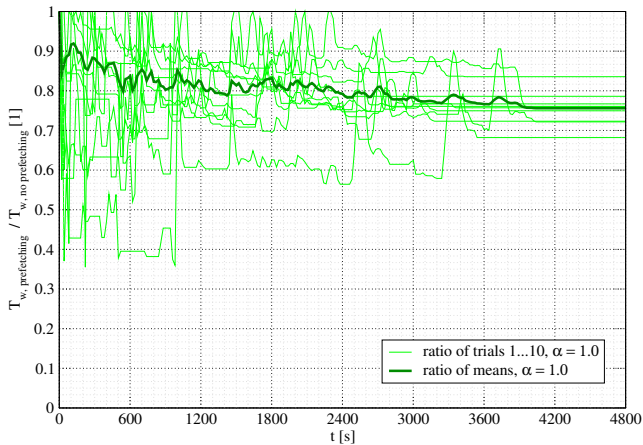


Figure 8: Ratio of user waiting time, 100 documents, average over 10 trials

select between the two extreme cases of *no prefetching* ( $p_{th} = 1.0$ ) and *unrestrained prefetching* ( $p_{th} = 0.0$ ). By adjusting  $p_{th}$  to intermediate values between 0.0 and 1.0 we can smoothly tune the degree of prefetching. To investigate the influence of  $\alpha$  and  $p_{th}$  simulations for various  $\alpha$ 's have been performed with decreasing threshold probabilities. In Fig. 11 the influence of the degree of prefetching on network cost and reduction of waiting time is depicted. The results have been normalized with respect to the no prefetching case of the individual  $\alpha$ .

We see the expected result that the benefits of prefetching grow with increasing  $\alpha$ .

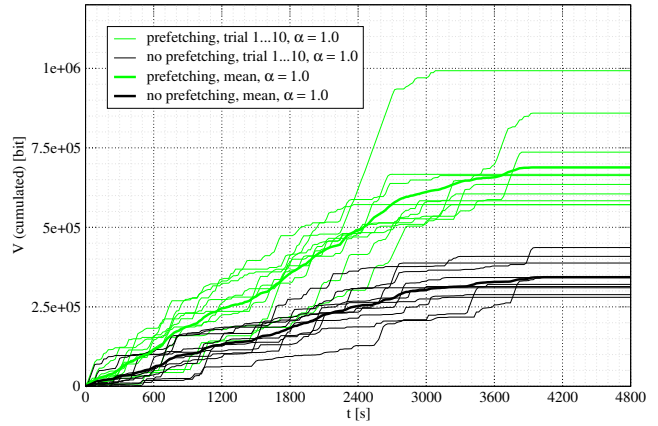


Figure 9: Influence of prefetching on network cost for packet-switched network, 100 documents, average over 10 trials

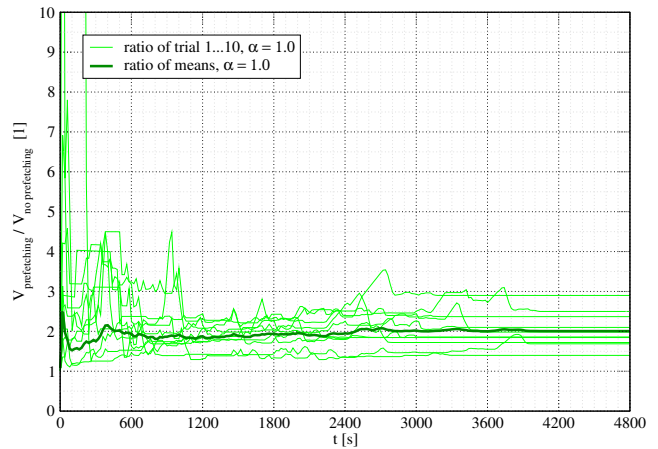


Figure 10: Ratio of network cost, 100 documents, average over 10 trials

## 4 Conclusions

The simulation results permit the conclusion that whenever a network resource is exclusively assigned to a user, as is the case for the duration of a connection in a circuit-switched system, prefetching is a sensible technique that increases the perceived performance of the system and conserves its resources. More differentiation is necessary for the packet-switched case. Whenever a network that is capable of dynamically assigning and sharing its resources among multiple users is operating near its capacity any form of prefetching will have adverse consequences, due to the overproportional consumption of these resources by prefetching. However, prefetching does increase overall system performance also for the packet-switched case whenever network resources are not fully employed. This conclu-

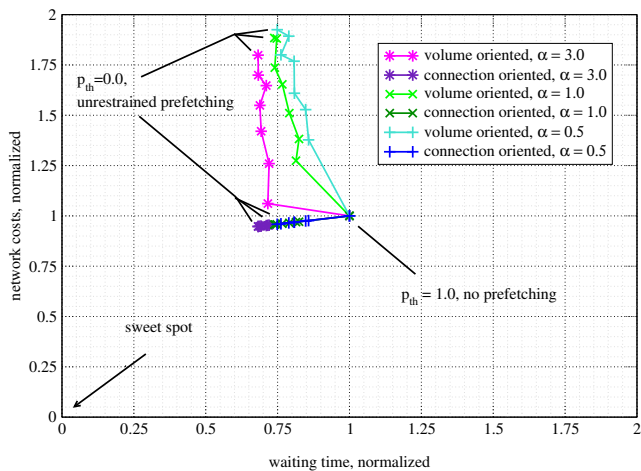


Figure 11: Influence of  $\alpha$  and  $p_{th}$  on network costs and waiting time, 100 documents, average over 10 trials

sion also suggests that dynamical pricing schemes that react on the instantaneous load of the network might be an interesting approach to stimulate prefetching in places/times when the network has free capacities. This would reduce the load in places/times with higher load and increase perceived network performance even without the need for assigning penalties for use of the network during higher load.

We would like to emphasize our view that the application of prefetching offers an additional degree of freedom for decision problems of optimum utilization of system resources. This degree of freedom can be visualized by the mock-up display presented in [8] and reproduced in Fig. 12 for the reader's convenience. The relations between the threshold probability, the reduction of waiting time and the network costs govern the degrees of freedom of the sliders.

## 5 Further Work

In the presented simulation we made the assumption that the proxy has knowledge of the candidate documents and their probabilities. For realistic scenarios this knowledge cannot be obtained by a single proxy, as it is not able to foresee the statistics of documents previously unrequested by its assigned client. Could we combine the knowledge of a substantial number of proxies this deficiency would be remedied. We therefore suggest an architecture as depicted in Fig. 13. The proxies send log messages of the client's requests to a central server component. This central server component is then capable of acquiring sufficient statistics in order to reply with a list of probable candidate documents. Based on this information the proxy starts to speculatively retrieve these documents from the original HTTP-servers. This architecture can be advan-

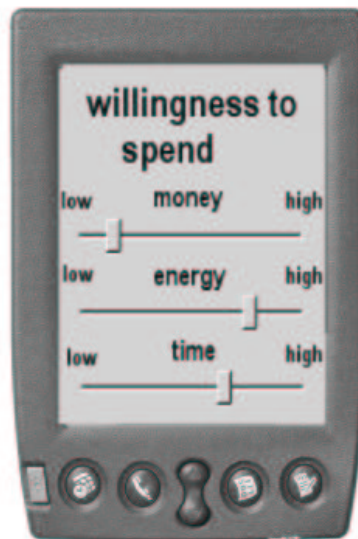


Figure 12: Mock-up display from [8]

tageously combined with the split-proxy concept presented in [11] which additionally allows for handover between distinct networks e.g. Bluetooth, WLAN or GSM on the application level.

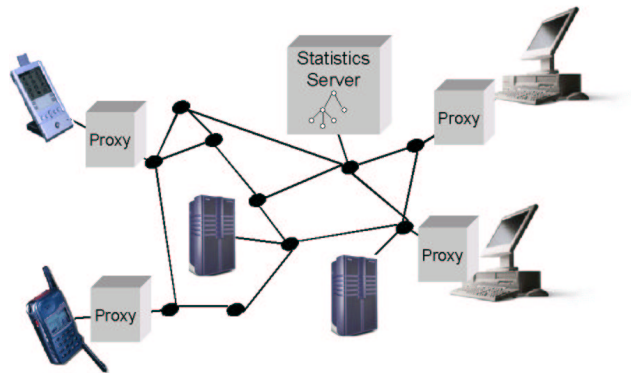


Figure 13: Cooperating prefetching proxies with central statistics server

Initially we plan to use a diagnostic proxy that communicates with a central server for the purpose of obtaining higher order statistics on the sequence of documents that users create when browsing especially in mobile scenarios. It is an important goal of our work to derive flexible models with adjustable parameters from these traces. The influence of the inter-document statistics on the ability to estimate document probabilities will be investigated to further assess the capabilities of prefetching in real world applications.

Field trials that are taking place in spring 2003 within the Heywow project will be used to test prefetching strategies under real world conditions and obtain quan-

titative results on their performance as well as valuable user feedback on perceived drawbacks and benefits.

*IEEE Conference on Mobile and Wireless Communications Networks (MWCN 2002)*, Stockholm, September 2002.

## References

- [1] N. J. Tuah, M. Kumar, and S. Venkatesh, "A performance model of speculative prefetching in distributed information systems," in *Intl. Parallel Processing Symposium, San Juan, Puerto Rico, 1997*.
- [2] M. Crovella and P. Barford, "The network effects of prefetching," in *IEEE Infocom, San Francisco, CA, 1998*.
- [3] M. Angermann, "Analysis of speculative prefetching," *ACM Mobile Computing and Communications Review*, vol. 6, pp. 13–17, April 2002.
- [4] A. Venkataramani, P. Yalagandula, R. Kokku, S. Sharif, and M. Dahlin, "The potential costs and benefits of long term prefetching for content distribution," Tech. Rep. TR-01-13, Department of Computer Science, University of Texas at Austin, June 2001.
- [5] L. Fan, P. Cao, W. Lin, and Q. Jacobson, "Web prefetching between low-bandwidth clients and proxies: Potential and performance," in *Proceedings of the Joint International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS '99)*, Atlanta, GA, May 1999.
- [6] T. M. Kroeger and D. D. Long, "Exploring the bounds of web latency reduction from caching and prefetching," *Proceedings of the USENIX Symposium on Internet Technologies and Systems, Monterey, California*, December 1997.
- [7] Z. Jiang and L. Kleinrock, "An adaptive network prefetch scheme," *IEEE Journal on Selected Areas in Communications*, April 1998.
- [8] M. Angermann and J. Kammann, "Cost metrics for decision problems in wireless ad hoc networking," in *Proceedings of the IEEE CAS Workshop on Wireless Communications and Networking, Pasadena, CA*, September 2002.
- [9] B. A. Mah, "An empirical model of HTTP network traffic," in *INFOCOM (2)*, pp. 592–600, 1997.
- [10] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and zipf-like distributions: Evidence and implications," in *Proceedings of the IEEE Infocom 1999*, pp. 126–134, 1999.
- [11] J. Kammann and T. Blachnitzky, "Split-proxy concept for application layer handover in mobile communication systems," in *Proceedings of the 4th*