Exploring the Benefits of Caching and Prefetching in the Mobile Web

Johann Márquez, Josep Domènech, José A. Gil and Ana Pont
Polytechnic University of Valencia
Valencia, Spain
{jmarquez,jdomenech}@ai2.upv.es;{jagil,apont}@disca.upv.es

***Abstract*—The extensive presence of wireless technologies to access Internet jointly with the massive use of the mobile phones have turned the mobile web into a close reality. Additional to the unquestionable interest of mobile web to provide e-services and information anywhere, it opens new possibilities of crucial importance to bridge the digital divide between the developed world and the developing countries. But the underlying technologies used introduce high latencies that can do very unpleasant the web navigations. For this reason it is important to devote effort to develop new solutions to improve web performance considering the mobility of users. In this paper we present an initial approach to study the benefits that techniques like caching and prefetching can achieve for the mobile web users.**

***Index Terms*—Mobile web, performance evaluation, developing countries, caching, prefetching.**

I. Introduction

In the last decade, wireless communications have shown a enormous growth and many people around the world have embraced these technologies at a remarkable rate. As a consequence, many traditional services of the wired networks like the Web have been exported to the wireless powered by the freedom of mobility.

With the enormous proliferation of mobile devices, such as mobile phones, PDAs, smartphones, the access to the World Wide Web (WWW) via public networks (GSM, GPRS, UMTS,...) has grown exponentially. Accessing the Web via mobile devices is known as the "Mobile Web".

This work has been partially supported by:

- Programme Alβan, the European Union Programme of High Level Scholarships for Latin America, scholarship No.E04D031142BO.
- Spanish Ministry of Education and Science and the European Investment Fund for Regional Development (FEDER) under grant TSI 2005-07876-C03-01.
- Polytechnic University of Valencia under grant PAI no.6164

Today, the majority of wireless service providers in the United States, Europe, and Japan offer wireless Internet and mobility services, and also many web sites offer adapted content for small devices with display, bandwidth, memory and processing power restrictions [1]. This holds the promise of making ubiquitous mobile access to IP-based applications and services a reality.

We consider that the advantages offered by the wireless connectivity services go beyond the mobility itself. It clearly seems that wireless technologies are one of the most promising ways to deliver content and services to those disadvantaged sectors that are not able to easily access to the Internet and its services, especially in the developing countries or rural areas. The mobile web is a potential solution to bridge the digital divide with the deployment of mobile networks all around the world and to deliver important services such as eHealth, eGovernment, eLearning, among others [2].

Despite the "always-on" paradigm that wireless technologies offer, there are some drawbacks to be considered, such as low bandwidth available to the end user who is connected via an outdoor wireless networks, long and variable latencies in document access, temporary disconnections, etc. [3]

In order to improve the web performance over wired and wireless networks, web architecture techniques such as caching and prefetching can be used. The main goal of the caching technique is to reduce the latency as well as the traffic consumption by storing the most popular objects accessed closer to the clients. The prefetching technique is focused on web latency reduction by predicting the next future web objects to be accessed by the user and prefetching them in idle times transparently to the user. So that, if finally the user request them, the objects will be already at the client cache [4]. Both techniques and their benefits have been widely studied on traditional network environments.

In this paper we perform a preliminary study to ex-

plore the benefits of caching and prefetching when they are applied over wireless technologies such as Wireless Fidelity (WIFI), Universal Mobile Telecommunications System (UMTS), and the General Packet Radio Service (GPRS), where the high latencies are an important drawback.

The remainder of this paper is organized as follows: Section II describes related work. Some basic concepts about web caching and web prefetching in general are addressed in Section III. A preliminary study of the latency in diverse wireless technologies is shown in Section IV. The caching and prefetching experiments performed, the environment and the simulation tools used are presented in Section V. Finally Section VI summarizes the main conclusions.

II. Related Work

Web caching is one of the most effective techniques to alleviate server bottleneck and reduce network traffic, thus decreasing the latency perceived by web users. This technique has been widely explored and used. Wang *et al.* in [5] surveys the caching studies taking into account many issues such as caching architectures, replacement policies, cache routing, dynamic caching, fault tolerance, security, etc. Their study compiles different representative research work showing that caching can improve the web performance achieving latency reduction between 23 % to 60 %.

There are many papers in the open literature that study the benefits of prefetching techniques applied to the World Wide Web [6], [7], [8]. Others have made interesting proposals to improve its performance by adapting it to the current web scenario [9] or have even proposed how to use it in real world without modifying the HTTP standards [10].

There are studies suggesting that caching and prefetching, working in a collaborative manner, improve the web performance reaching higher boundaries [7], [11], [12], [13].

But, the vast majority of that work has been done considering wired environments, and there are only few attempts to study the effect of prefetching over wireless networks. An early attempt to apply caching and prefetching techniques has been presented by Fleming *et al.* in [14]. They propose a web architecture that uses an intermediary multithreaded prefetching proxy in a wired and wireless scheme with a narrow bandwidth available. Their proposal reduces the document download time by up to 30%.

Jin *et al.* in [15] study caching and prefetching in an integrated system for wireless local area networks (WLAN) in a University campus environment, taking into account a prediction algorithm based on sequence mining and performing a context-aware prefetching as well as a profit-driven caching replacement policy.

Liang *et al.* in [16] presents a study about multi-user prefetching applied to a two-tier heterogeneous wireless network, introducing the effect of the roaming into the UMTS/Wifi two-tier network. On the other hand, Jiang and Kleinrock [17] consider different networks performing prefetching based on parameters including network capacity and network cost.

III. Caching and Prefetching Overview

Web caching is a technique that takes advantage from the web object's temporal locality to reduce the perceived latency and bandwidth consumption. The most accessed web objects are stored (cached) close to the client-side to avoid requesting them again to the original web servers.

The prefetching technique takes advantage of the spatial locality shown by the web objects to reduce user's perceived latency. It is based on two main components: The ***Prediction engine*** and the ***Prefetching engine***.

The prediction engine is in charge of making predictions about the future user accesses. It usually processes the user request patterns to perform the predictions. The prediction engine can be set in different elements of the Web architecture. When it is set at the client it makes use of the user accesses pattern to perform predictions [6], [18]. When the engine is set at the proxy it takes advantage of the multi-user and multi-server information gathered at this element to do the predictions [19], [20]. Finally, if the engine is located at the server it makes predictions based on multi-user accesses to the same domain [9], [21], [22].

The prefetching engine prefetches the predictions made by the prediction engine and is usually set at the client side, considering its available resources, to prefetch. Nevertheless, the prefetching engine can be also set at the proxy, working transparently to the clients [23]. Furthermore, the prefetching engine can be set at the server side *pushing* the web objects to proxies or clients when there is a collaborative scheme among them [24].

Latency can be understand as the waiting time since the user requests a Web page or object until it can be completely displayed. Kroeger *et al.* in [7] divide the total latency into two latencies: *internal* and *external* taking into account the use of an intermediary proxy. Figure 1 illustrates this concept.

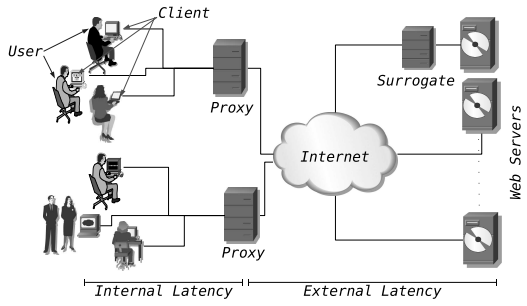

Fig. 1. Generic Web Architecture

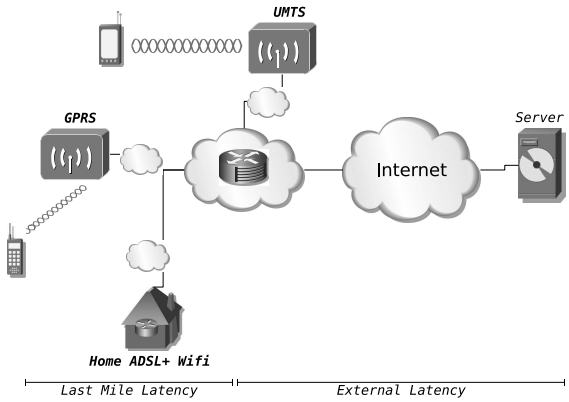

Fig. 2. Mobile Web Access Scheme

Domenech *et al.* proposed a metric taxonomy to evaluate the prefetching performance considering the predictions, the resource usage and the latency [25]. This study demonstrates the importance of the use of the *latency per page* and *latency per object* metrics to fair evaluate this technique since depending on the metric used, the results may not only vary but also reach opposing conclusions. The main conclusions of this work can be extended to the caching performance evaluation.

Despite the prefetching's potential [13], [26] its application has not been as much spread out as caching, due to the higher bandwidth consumption and even HTTP modification that first proposals required [27]. Nevertheless, current studies demonstrate that this technique can be implemented in real scenarios adapting commercial products and without modifying the HTTP standard protocol[28].

IV. Wireless Latencies

The wireless networks present an intrinsic latency which is considerably higher than for wired networks. These high latencies are important factors to be considered when measuring the performance of the network.

In this work we consider the scenario where a web user is connected to Internet through a wireless technology, like WIFI, UMTS, and GPRS. The current Internet providers supply the wireless access as far as a main host at the base station system (BBS), that acts like a router, then the communication follows the same path independently of the original connectivity used. Figure 2 shows this scheme. Here we can observe that there are two types of latencies: the last mile latency and the external latency. We consider the last mile latency as the latency generated in the network section between the client host and the last hop within the service provider network (i.e. core router), while the external latency is considered as the latency between the last hop and the destination server host. Consequently, the overall network latency is the total amount of both latencies as Equation 1 shows.

$$(Latency_{Total} = Latency_{Lastmile} + Latency_{External}) \quad (1)$$

In order to identify and measure the latency generated by the wireless connectivity of the overall network latency we have performed several experiments. These experiments traced and measured the Round Trip Time (RTT) of 64 KB ICMP Packets from a source client host to diverse destination server hosts varying the underling technology connectivity (WIFI, UMTS, GPRS). To make a fair comparison we have used the same telecommunication service provider for all the experiments under the same conditions.

Table I presents the latencies measured from a host, geographically located at Valencia (Spain), to server hosts located at each place shown in the first column. The group of columns represents the underlying technology used in the last mile where each first column shows the last mile latency, the second column presents the total latency (Eq. 1) and the third column represents the percentage of the last mile latency with respect to the total latency.

As we can see in Table I, depending on the networking technology used, the last mile latency could represent from 5% up to 84% of the total latency. In the Ethernet case, the last mile latency represents generally only a small portion of the overall latency since this wired technology offers higher bandwidth and data transmission speed in comparison to the wireless. Unlike UMTS and GPRS show higher last mile latencies due to the wireless intrinsic issues such as interferences and noise.

Once we have established the importance of the network technology used over the latency perceived, we will

TABLE I
LAST MILE LATENCIES

Technology	*Ethernet*			*WIFI*			*UMTS*			*GPRS*		
Target latencies	**Last mile [ms]**	**end-to-end [ms]**	**%**	**Last mile [ms]**	**end-to-end [ms]**	**%**	**Last mile [ms]**	**end-to-end [ms]**	**%**	**Last mile [ms]**	**end-to-end [ms]**	**%**
USA - Alaska	40	250	***16***	43	257	***17***	391	571	***68***	844	1489	***57***
USA - East	69	270	***24***	74	288	***26***	358	738	***49***	962	1444	***67***
USA - West	37	247	***15***	44	254	***17***	395	469	***84***	823	1013	***81***
Cuba	35	705	***5***	44	715	***6***	280	1063	***26***	734	1939	***38***
Bolivia	40	369	***10***	43	375	***11***	410	704	***58***	771	1089	***71***
Spain	37	83	***45***	42	90	***47***	361	445	***81***	887	1114	***80***
South Africa	32	401	***8***	44	417	***10***	389	614	***63***	918	1623	***57***
Russia	38	127	***30***	42	135	***31***	297	388	***77***	881	1061	***83***
Japan	95	341	***28***	106	352	***30***	393	588	***67***	1252	1491	***84***
China	30	589	***5***	43	603	***7***	394	833	***47***	795	1434	***55***
Australia	56	507	***11***	76	613	***12***	310	714	***43***	810	1239	***65***

study how web caching and prefetching can benefit web users by decreasing the final user latency perceived.

V. EXPERIMENTS

This Section presents the experiments performed as well as the environment conditions used.

A. Web Architecture

Our set of experiments evaluates the prefetching and caching performance within diverse prefetching schemes:

Client-Server Scheme: As most research works use a client-server scheme in order to study the prefetching, we set the framework to simulate different clients accessing a web server. The prediction engine is set at the web server and the prefetching engine runs at the client side. We set the framework as shown in Figure 3. The clients are connected to the wireless public network varying the underlying wireless technology offered by the base station subsystem (BSS) and this latter to wired Internet to reach the server.

Proxy Cache Scheme: To study the effects of the caching and the prefetching techniques working in a collaborative manner, based on the scheme shown by Figure 3, we have added to the BSS an intermediary proxy acting as a server cache able to perform predictions since the prediction engine is set at this component. The clients remain as wireless clients (with the prefetching engine set on them) accessing to multiple servers.

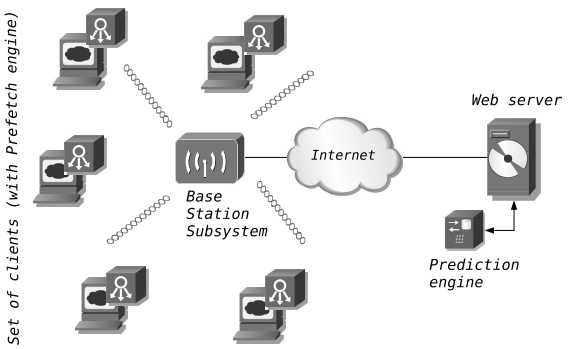

Fig. 3. Simulation Architecture

B. Simulation Framework

To perform our experiments we use the framework presented and described in [29]. This discrete-event based simulator is a flexible tool to study, reproduce, check and compare the performance of prefetching and caching techniques at any element of the Web architecture. It is a trace-driven simulator able to simulate the real user behavior, offering full result statistics and performance indexes with a low cost in terms of resource consumption.

C. Workload

For our experiments we have used three traces collected from a main Squid Proxy server at the Polytechnic University of Valencia (Spain). Two traces represent the real user accesses to two popular Spanish news

TABLE II
WORKLOAD TRACES

Scheme	***Client-Server***		***Proxy-Cache***
Trace	**Elpais**	**Marca**	
Year	March 9-12, 2007		
No. of Accesses	505868	423559	7324698
No. of Pages	20253	29942	1326033
Avg. objects per page	24	14	4.52
No. of Sessions	2586	1999	
No. of Users	892	1180	7987
Bytes transfered (GB)	1.48	2.06	107.3
Avg. objects size (KB)	3.08	5.10	15.87
Avg. page size (KB)	77.08	75.93	71.73
Avg. HTML size (KB)	30.55	14.52	8.82
Avg. image size (KB)	1.93	4.38	17.12
No. of Servers	1	1	28978

servers (www.elpais.com, www.marca.com). Their main characteristics are show in Table II in columns one and two. These traces were obtained by filtering the web server accesses from the trace shown in the third column. Both traces are used to drive the experiments in the Client-Server scheme. The third column shows the characteristics of the trace used to drive the experiments in the Proxy Cache scheme. This trace present the total multi-user accesses to multi-servers.

D. Performance indexes

To evaluate the results we use the following indexes:

- $\nabla Latency_{Page}$:The latency per page ratio is the ratio of the latency per page that prefetching achieves to the latency with no prefetching.
- $\nabla Latency_{Object}$: Same as $\nabla Latency_{Page}$ but measured per objects.
- $\Delta Traff_{bytes}$: The amount of total traffic over useful traffic that is the traffic generated by user´s requests.
- $Precision$: The ratio of prefetch hits to the total number of objects prefetched.
- $Recall$: The ratio of prefetch hits over all the user request. This metric is the prediction index that better explains the latency per page ratio.

E. Prediction algorithms

In order to study the prefetching for the current web structure, we have used the *Double Dependency Graph (DDG)* prediction algorithm, which presents a better cost-benefit relationship than others [9].

DDG keeps track of the dependences among the objects accessed by the user on a graph and takes into account the current Web structure by distinguishing two classes of dependences: dependences to an object of the same web page and dependences to an object of another page. The graph has a node for every object that has ever been accessed and an arc from node A to another B only if there has been an access to B within *w* accesses after A. DDG generates the predictions applying a *cutoff threshold* to the weight of its arcs that leaves from the node of the last user's access.

We have set the secondary cutoff threshold = 0.3 and vary the primary cutoff threshold from 0.2. to 0.8 to study conservative as well as aggressive prefetching.

F. HTTP Cacheability

We have set the proxy to store only the *"cacheable"* objects bypassing any response that comes from a request with dynamic characteristics (i.e. asp, php queries,...). Considering the amount of transferred bytes of the Proxy-Cache Scheme trace, the cache size was set to infinite.

G. Performance Evaluation

In order to make a suitable evaluation of the caching and prefetching gain it is important to tackle these techniques from the user's point of view and make use of the cost-benefit analysis.

Client-Server Scheme: Figures 4 and 5 present the results of the experiments using the *trace Marca* and the *trace Elpais* respectively. They show the benefits achieved by the prefetching for each underlying technology. The *x* axis shows the *traffic ratio* measured in bytes while the *y* axis shows the *latency per page ratio* when applying prefetching.

Each curve represent a wireless technology and each point on the curves is obtained from one experiment considering a specific threshold parameter.

Both figures show that prefetching reduces up to 20% the user's latency perceived with the cost of increasing the traffic up to 26%. Comparing the curves (underlying technologies), the technologies with higher last mile latency get higher benefits when applying the prefetching.

To analyze the prefetching performance in more detail, we have taken one experiment from each trace and observed the prefetching indexes and the relationship among technologies. The chosen experiment uses a DDG primary threshold = 0.2 for an aggressive prefetching.

Table III presents the results of the experiments done using both traces. Analyzing the results for the trace

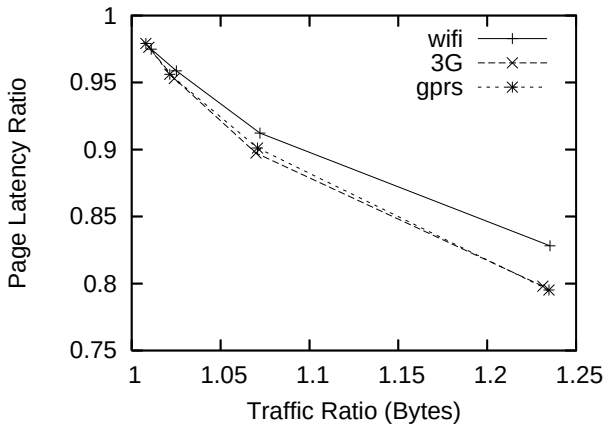

Fig. 4. Trace Marca: Prefetching Cost-Benefit relationship

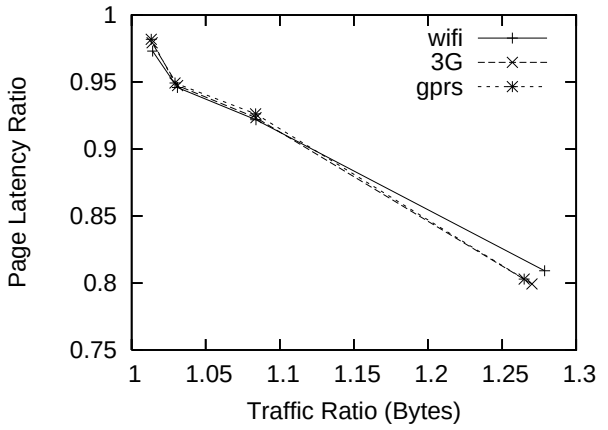

Fig. 5. Trace Elpais: Prefetching Cost-Benefit relationship

Marca we can observe that GPRS and UMTS results show a higher page latency reduction even with a lower precision and recall, reaching up to 20.20 % of user perceived latency reduction because these technologies present higher last mile latencies. Consequently, the latency reduction ratio among these technologies are up to 7.78% in the case of GPRS in comparison to WIFI and up to 2.17% when comparing GPRS against UMTS as table IV shows. A similar analysis can be done for the trace *Elpais*.

Proxy-Cache Scheme: To study caching and prefetching working together, we use the proxy-cache scheme described in section V-A. The prediction engine is set at the proxy using the DDG prediction algorithm fed with multi-user and cross-server patterns. The primary threshold is 0.2 and it gives hints only for *"cacheable"* objects.

Table V presents the results of the experiments with both techniques. Caching presents the highest latency

TABLE III
EXPERIMENTS RESULT

Metric	**WIFI**	**UMTS**	**GPRS**
Marca			
$\nabla Latency_{Page}$ [%]	17.18	20.20	20.20
$\nabla Latency_{Object}$ [%]	14.61	14.40	14.31
$\Delta Traff_{bytes}$ [%]	23.53	23.13	23.47
$Precision$ [%]	39.22	38.9	38.53
$Recall$ [%]	21.3	20.3	19.67
Elpais			
$\nabla Latency_{Page}$ [%]	19.08	20.07	19.71
$\nabla Latency_{Object}$ [%]	18.43	18.00	17.42
$\Delta Traff_{bytes}$ [%]	27.87	27.01	26.49
$Precision$ [%]	45.55	45.38	45.39
$Recall$ [%]	12.12	18.6	17.97

TABLE IV
PAGE LATENCY REDUCTION RATIO AMONG WIRED & WIRELESS TECHNOLOGIES

$\nabla Latency_{Page}$	**Ethernet**	**WIFI**	**UMTS**	**GPRS**
Marca				
Ethernet	1			
WIFI	1.91	1		
UMTS	6.84	3.59	1	
GPRS	14.83	7.78	2.17	1
Elpais				
Ethernet	1			
WIFI	1.82	1		
UMTS	6.22	3.42	1	
GPRS	13.28	7.29	2.13	1

reduction for any technology used. This reduction is due to the massive storage of the caching whereas prefetching among the clients and the proxy only adds up to 2% of extra latency reduction.

We conclude from the set of experiments and its results, that caching and prefetching techniques offer an interesting latency reduction to the users. The applicability and performance of each technique not only lies on scheme and architecture issues but also on the underlying networking technologies issues.

We clearly observed that prefetching performs better in a client-server scheme since a successful prefetched document reduces the total latency whereas the proxy cache scheme reduces only the internal latency of those "cacheable" documents. Nevertheless, since wireless

TABLE V
PROXY-CACHE SCHEME RESULT

Caching	**WIFI**	**UMTS**	**GPRS**
$\nabla Latency_{Page}$ [%]	26.61	32.73	31.93
$\nabla Latency_{Object}$ [%]	31.01	38.14	37.21
Caching & Prefetching			
$\nabla Latency_{Page}$ [%]	27.90	34.32	33.48
$\nabla Latency_{Object}$ [%]	32.73	40.29	39.28

technologies presents higher latencies in comparison to the wired, predicting at the proxy contributes to improve the Web performance.

VI. CONCLUSIONS

The mobile Web offers the possibility to bring services such as eHealth, eLearning, etc, to developing countries thus reducing the digital divide.

In the emerging wireless technology we found an important research area to apply caching and prefetching techniques since mobile web presents high intrinsic latencies in comparison to wired network. We have identified a high percentage of the latency that represents the wireless connectivity in the whole latency.

Through a wide range set of experiments applying caching and prefetching techniques over different wired and wireless connectivity technologies we have demonstrated that both web techniques improve the performance of the mobile Web. We conclude that depending on the scheme and strategy applied caching and prefetching could reach different boundaries but they have a high potential for reducing the user's perceived latency in the Mobile Web.

REFERENCES

[1] A. Adya, P. Bahl, and L. Qiu, "Characterizing alert and browse services for mobile clients," in *Proc. of the 2002 USENIX Annual Technical Conf.*, 2002, pp. 343–356.

[2] S. Boyera, "The mobile web to bridge the digital divide," in *IST-Africa 2007*, Maputo, Mozambique, May 2007.

[3] T. S. Loon and V. Bharghavan, "Alleviating the latency reduction and bandwidth problems in www browsing," in *Proc. of the 1st USENIX Symposium on Internet Technologies and Systems*, Monterey, USA, 1997.

[4] M. Rabinovich and O. Spatscheck, *Web Caching and Replication*. Addison Wesley, 2002.

[5] J. Wang, "A survey of web caching schemes for the internet," *ACM Computer Communication Review*, vol. 29, no. 5, pp. 36–46, 1999.

[6] A. Bestavros, "Using speculation to reduce server load and service time on the www," in *Proc. of the 4th ACM International Conf. on Information and Knowledge Management*, Baltimore, USA, 1995.

[7] T. M. Kroeger, D. D. Long, and J. C. Mogul, "Exploring the bounds of web latency reduction from caching and prefetching," in *Proc. of the 1st USENIX Symposium on Internet Technologies and Systems*, Monterey, USA, 1997.

[8] T. Palpanas and A. Mendelzon, "Web prefetching using partial match prediction," in *Proc. of the 4th International Web Caching Workshop*, San Diego, USA, 1999.

[9] J. Domenech, J. A. Gil, J. Sahuquillo, and A. Pont, "DDG: An efficient prefetching algorithm for current web generation," in *Proc. of the 1st IEEE Workshop on Hot Topics in Web Systems and Technologies*, 2006.

[10] B. de la Ossa, J. Sahuquillo, J. A. Gil, and A. Pont, "Web prefetch performance evaluation in a real environment," in *IFIP/ACM Latin America Networking Conf.*, 2007.

[11] J. Xu, J. Liu, and B. L. X. Jia, "Caching and prefetching for web content distribution," *Computing in Science & Engineering*, vol. 06, no. 4, pp. 54–59, 2004. [Online]. Available: citeseer.ist.psu.edu/754047.html

[12] Y. Jiang, M.-Y. Wu, and W. Shu, "Web prefetching: Costs, benefits and performance," in *Proc. of the 7th International Workshop on Web Content Caching and Content Distribution*, Boulder, USA, 2002.

[13] J. Domenech, J. Sahuquillo, J. A. Gil, and A. Pont, "The impact of the web prefetching architecture on the limits of reducing user's perceived latency," in *Proc. of the 2006 IEEE/WIC/ACM International Conf. on Web Intelligence*, 2006.

[14] T. B. Fleming, S. F. Midkiff, and N. J. Davis, "Improving the performance of the world wide web over wireless networks," in *Proc. of the Global Telecommunications Conf.*, Phoenix, USA, 1997.

[15] B. Jin, S. Tian, C. Lin, X. Ren, and Y. Huang, "An integrated prefetching and caching scheme for mobile web caching system," in *Proc. Eighth ACIS International Conf. on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing SNPD 2007*, vol. 2, July 30 2007–Aug. 1 2007, pp. 522–527.

[16] B. Liang and S. Drew, "Multiuser prefetching with queuing prioritization in heterogeneous wireless systems," in *QShine '06: Proc. of the 3rd international conference on Quality of service in heterogeneous wired/wireless networks*. New York, NY, USA: ACM, 2006, p. 34.

[17] Z. Jiang and L. Kleinrock, "Web prefetching in a mobile environment," *Personal Communications*, vol. 5, no. 5, 1998.

[18] D. Duchamp, "Prefetching hyperlinks," in *Proc. of the 2nd USENIX Symposium on Internet Technologies and Systems*, Boulder, USA, 1999.

[19] L. Fan, P. Cao, W. Lin, and Q. Jacobson, "Web prefetching between low-bandwidth clients and proxies: Potential and performance." in *Proc. of the ACM SIGMETRICS Conf. on Measurement and Modeling Of Computer Systems*, Atlanta, USA, 1999, pp. 178–187.

[20] C. Bouras, A. Konidaris, and D. Kostoulas, "Predictive prefetching on the web and its potential impact in the wide area." *World Wide Web*, vol. 7, no. 2, pp. 143–179, 2004.

[21] S. Schechter, M. Krishnan, and M. D. Smith, "Using path profiles to predict http requests," in *Proc. of the 7th International World Wide Web Conf.*, Brisbane, Australia, 1998.

[22] V. N. Padmanabhan and J. C. Mogul, "Using predictive prefetching to improve World Wide Web latency," in *Proc. of the ACM SIGCOMM '96 Conf.*, Stanford University, USA, 1996.

[23] W.-G. Teng, C.-Y. Chang, and M.-S. Chen, "Integrating web caching and web prefetching in client-side proxies," *IEEE Transactions on Parallel and Distributed Systems*, vol. 16, no. 5, pp. 444–455, 2005.

[24] A. Bestavros and C. Cunha, “Server-initiated document dissemination for the WWW,” *IEEE Data Engineering Bulletin*, 1996.

[25] J. Domenech, J. A. Gil, J. Sahuquillo, and A. Pont, “Web prefetching performance metrics: A survey,” *Performance Evaluation*, vol. 63, no. 9-10, 2006.

[26] C. Bouras, A. Konidaris, and D. Kostoulas, “Efficient reduction of web latency through predictive prefetching on a wan.” in *Proc. of the 4th International Conf. on Advances in Web-Age Information Management*, Chengdu, China, 2003.

[27] M. Crovella and P. Barford, “The network effects of prefetching,” in *Proc. of the IEEE INFOCOM’98 Conf.*, San Francisco, USA, 1998.

[28] B. de la Ossa, J. A. Gil, J. Sahuquillo, and A. Pont, “DELFOS: the oracle to predict next web user’s accesses,” in *IEEE 21st International Conf. on Advanced Information Networking and Applications (AINA’07)*. IEEE, 2007, pp. 679–686.

[29] J. Marquez, J. Domenech, J. Gil, and A. Pont, “A web caching and prefetching simulator,” in *Proc. of SoftCOM 2008 International Conf. on Software, Telecommunications and Computer Networks*, Split, Croatia, 2008.