An SPN-Based Integrated Model for Web Prefetching and Caching

Lei Shi¹,² (石磊), Ying-Jie Han² (韩英杰), Xiao-Guang Ding¹ (丁晓光), Lin Wei² (卫琳) and Zhi-Min Gu¹ (古志民)
¹Department of Computer Science and Engineering, Beijing Institute of Technology, Beijing 100081, P.R. China
²School of Information Engineering, Zhengzhou University, Zhengzhou 450002, P.R. China
E-mail: {shlei, hanyj}@zzu.edu.cn; xgding@163.com; weilin@shengda.edu.cn; zmgu@x263.net

Abstract The World Wide Web has become the primary means for information dissemination. Due to the limited resources of the network bandwidth, users always suffer from long time waiting. Web prefetching and web caching are the primary approaches to reducing the user perceived access latency and improving the quality of services. In this paper, a Stochastic Petri Nets (SPN) based integrated web prefetching and caching model (IWPCM) is presented and the performance evaluation of IWPCM is made. The performance metrics, access latency, throughput, HR (hit ratio) and BHR (byte hit ratio) are analyzed and discussed. Simulations show that compared with caching only model (CM), IWPCM can further improve the throughput, HR and BHR efficiently and reduce the access latency. The performance evaluation based on the SPN model can provide a basis for implementation of web prefetching and caching and the combination of web prefetching and caching holds the promise of improving the QoS of web systems.

Keywords stochastic Petri nets, web prefetching, web caching, performance evaluation

1 Introduction

The WWW continues its remarkable and exponential rate growth. The rapid growth in the number of users, servers and network traffic has seen a corresponding increase in network loads and user response times. Currently, the most effective solutions to improving the retrieval rate of large distributed documents are web caching and prefetching.

The heart of web caching is the replacement policy. If a newly arrived object is about to be placed in the cache and there is no room left, some cached objects would be evicted in order to make room for the newcomer. Typical replacement algorithms are LFU, LRU, GD-Size, GDSF, etc.

Web prefetching is also called active caching, it can predict future requests by analyzing access history of users and fetch these objects into a local cache before they are requested. Studies have shown that web prefetching can reduce the perceived access latency, smooth the traffic load of the network and improve the quality of services, thus improve the performance of the WWW[¹-⁴].

Most of the studies reported that the hit ratio of caching is from 30% to 50%, while prefetching can improve it up to 60% or even more. Web caching and prefetching should ideally be integrated so that the limited resources such as memory and bandwidth can be managed more efficiently. Researches and applications of the integrated prefetching and caching can be found in many areas such as file system, mobile computing, Internet, and P2P, especially the semantic P2P[⁵,⁶].

Intuitively, prefetching can have a better performance than those without prefetching mechanism. Previous studies in web prefetching focus on building access models and evaluating the performance of such models in predicting future accesses. Although these models are important, they lack the consideration of theoretical analysis based on integrated model for prefetching and caching.

Petri nets theory is exceptionally well suited for designing, modeling, analyzing and evaluating the performance of computer systems that have concurrent, synchronous, distributed, parallel, and/or stochastic processing features[⁷,⁸]. Stochastic Petri nets is a powerful graphical and mathematical tool of system modeling. The theory of SPN makes use of simulation technique to calculate the performance metrics in terms of system performance model.

Our work tries to make performance evaluation of web prefetching and caching quantitatively. For the first time as far as we know, we make use of SPN to model the integrated web prefetching and caching system, and based on which, the metrics, access latency, throughput, HR and BHR are compared and analyzed.

2 Related Work

Although web caching has been already widely used in the WWW, the benefits from web caching are becoming limited due to the rapid changes of network resources, while web prefetching can optimize the World Wide Web in many aspects. The combination of web prefetching and caching holds the promise of improving the QoS of web systems.

In 1994, Griffioen et al.[²] paid attention to the modeling of web prefetching and caching on file system. The research assumed that prefetching and caching share the same cache space, and showed that integrated web
prefetching and caching can improve the performance of cache system.

In 1995, Cao et al.\cite{3} presented a model of integrated web prefetching and caching on file system, and based on which, she made performance study and simulative validation. Simulations illustrated that the integrated model could reduce the elapsed times of the applications by up to 50%.

In 2001 and 2003, Yang et al.\cite{4,9,10} presented an integrated architecture for web object caching and prefetching. The web object prediction model was built by mining the frequent paths from past web log data, and prefetching algorithm named Pre-GDSF was implemented. Experimental results show that integrated web prefetching and caching system can have a better performance than those without prefetching mechanism.

Inspired by Yang’s work\cite{9}, Fig. 1 depicts the architecture of an integrated web prefetching and caching model (IWPCM) that consists of six parts.

- **Cache Manager**: provide the application interface to the cache itself. It is the center of the cache control, which is to coordinate the cache components and maintain web objects in memory.

- **Metadata**: describe the data of the cached objects.

- **Replacement Manager**: choose a set of objects to be evicted from the cache, whenever cache space is needed, e.g., to store a recently arrived object. Typical implementations for replacement managers include LRU, LFU, GDSize, GDSF, etc.

- **Memory**: include Normal Cache Memory and Prefetch Memory. The former is for normal caching usage, while the latter serves as prefetch buffer.

- **Prediction Model**: based on the previous work in the web prefetching field, prediction algorithms can be classified into the following categories\cite{1}. (1) Pattern-based algorithm usually predicts web pages by means of Markov models or hidden Markov models. Prediction by Partial Match (PPM) belongs to this category. (2) Frequency-based algorithm, such as popularity-based algorithm, keeps a list of popular documents and maintains it by the requests of clients. (3) Semantic-based algorithm predicts forthcoming web pages from semantic meaning. Compared with the former two models, this approach is relatively difficult to realize, but it may be a promising one in the future due to the mechanism.

- **Prefetching Manager**: used to preload web documents into the prefetch buffer. Apparently, prediction model provides the information of prefetching items; meanwhile, Prefetching Manager includes prefetching control mechanism in order to balance the reduced latency and the potential increase in network and resource loads.

The flowchart of the IWPCM can be described as follows.

1. The request issued by a client for object O is sent to the Cache Manager.
2. Cache Manager checks if O exists in Metadata, if exists, called hit, go to (3), if not, called miss, go to (4).
3. Take object O from Normal Cache Memory and go to (10).
4. Cache Manager sends the request to the web server.
5. Object O is retrieved from the web server.
6. Replacement Manager calls Replacement Algorithm.
7. Replacement Algorithm collects the information of object O in cache and makes the replacement decision.
8. Replacement Algorithm makes the decision of whether receiving object O or not.
9. If object O is rejected, go to (10), if not, cached in the Normal Cache Memory.
10. Update the Metadata.
11. Cache Manager returns requested object O to the Client.
12. Prefetching Manager gets the prediction information from the Prediction Model.
13. Prefetching Manager sends prefetching requests to the web server and controls the prefetching process.
14. Predicted objects are retrieved.
15. Prefetching Manager caches the prefetched objects in Prefetching Cache.

3 SPN Model for IWPCM

Firstly, we focus on web prefetching and caching on a client/web server architecture, and then choose two schemes to discuss. Caching only model (CM), which means without prefetching, taking actions only as a result of requests, and IWPCM, which refers to the combination of prefetching and caching, making use of local information to determine the objects to preload.

An SPN model for IWPCM is designed after considering the following points.

1. Model the system as a multi-user system consisting of a single web server. Multiple users share the network link, and issue requests for objects from the web server.