Method for Performance Analysis of Web Caching Hierarchical Structures

MARCO ANTONIO S. BARBOSA, CARLOS AUGUSTO P. S. MARTINS
Programa de Pós-Graduação em Engenharia Elétrica
PUC-Minas
Avenida Dom José Gaspar 500, Coração Eucarístico, Prédio 3.
BRASIL
marco@pucminas.br, capsm@pucminas.br http://www.ppgee.pucminas.br

Abstract: - The use of the web-caching technique has been widely spread out with the objective of reducing the impact of some problems caused by the vertiginous growth of the Internet. The main characteristic of this technique is to keep the objects of the Internet next to the user and consequently, reducing the reply latency and the traffic in the communication channels. An important property of the web-caching servers is the cooperation´s capacity that results in increasing the performance in comparison to isolated servers. However, in most cases, the amount of alternatives to establish this cooperation is large and the choice of the best option may be complex and arduous. Our work propose a method for the analysis of performance of hierarchical structures of web-caching servers based on an empirical model, due to difficulty, in some cases, to perform such analysis with analytical models or through measurements in real environments. We also present a validation of the proposed method, applied to a real case study.

Key-Words: - web caching, web caching hierarchical structures, web, caching. performance analysis.

1 Introduction

In the last years, the Internet has evolved of amazing form. Its acceptance was very fast and is diverse the developed applications for that if it transformed access environment into gigantic to the information and a communication [22].

The diversity of services (documents, images, files, electronic commerce, video, voice, etc.) and the specialization of the applications has promoted still more the growth vertiginous of the Internet that had its beginning motivated for strategic questions of the army American north and soon after that counted on the participation of some universities [8] [12] [22].

This very fast growth promoted the sprouting of some problems, mainly due to raised demand generated for the users and the specialization of the applications, that the current structure is not being capable to take care of with efficiency. Between these problems we can show: high latency in the answers for the users, reduction in the availability of the information, incorrect use of resources, high load in the servers, etc. [8]. As proposal for reduction of these problems appeared the idea of the implementation of cache servers of web objects, or still, Web Caching servers.

Our research has as main goal to present a method based on laboratorial experimentation that assists the performance analysis of hierarchic or distributed structures of Web Caching servers for one real or hypothetical situation.

2. Web-Caching

Web Caching is one technique that manages the solicitations made for net navigators and temporarily stores web objects fetch with the goal to prevent future requests to the sources server of these objects. Therefore, Web Caching server is a computational system that executes this manages, normally dedicated and also known as Proxy [7] [8] [36].

Web caching is effective because many resources are requested frequently by some users, or is repeated times for a specific user. This characteristic is known as reference locality [2].

The main characteristics evaluated in web caching are divided in some segments as can be observed in references [14] [15] are: contribution between servers [24] [36]; performance indexes [7] [14] [15] [17] [26], pre-fetch [13] [16] [33], routing [4] [9] [15] [27] [31], replacement [14] [15] [20] [21] and mechanisms of coherence [14] [15]. The determination of each one of these characteristics can represent a problem, being able to result increasing in the latency of reply to the user.

One property of the web caching servers is the capacity of cooperation that can be established between two or more servers. This capacity of cooperation allows the project architectures have to be more elaborate with the objective to increase the efficiency of this technique of cache. This cooperation can be implemented with the hierarchic, distributed, or in mixing concepts.

The establishment of cooperation between web caching servers has some advantages and disadvantages. We can enumerate the following advantages: load balancing (each server stores a set of objects); reduction in the time of reply to the user when the object is located inside of the structure; rise of the hit ratio; reduction of requests to the sources servers; etc. A disadvantage in this cooperation is that in determined requests inside of the structure the time of reply to the user can be bigger than in the cases of a direct fetch from source, due the delay accumulated in the requests between the servers. [5] [11] [14] [15] [30].

In the distributed architecture of web Caching is not defined servers in intermediate levels [14] [15] [23] [32]. All servers meet in the same level and collaborate between itself. This cooperation is established with use of the protocols: Internet Cache Protocol (ICP) [32] [34] [35]; Summary [15]; Hash [25].

In the hybrid project, the servers of web caching cooperate between itself in the same level using caching distributed and between levels using hierarchic web caching. Some servers who use ICP are a typical example.

The determination of the structure most efficient for an environment can be complex and we do not identify a method that has guided the analysis of different structures with experimentation in laboratory as the proposal of this article.

3 Method for Performance Analysis of Web-Caching Hierarchical Structures

We can identify three different ways to carry through the performance analysis of hierarchic web caching servers: through an analytical model, measurement in real environment and by experimentation in laboratory (simulation) [29].

In this work, we looking for a method that can be used in different situations that involve the performance analysis of hierarchic structures of web-caching servers carried through using the context of boarding of experimentation in laboratory. This method is more specific than the proposal of performance analysis of structures Client/Server showed in MENASCÉ & ALMEIDA [18], but it follows the same paradigm Client/Server. We can characterize our method being more specific for dealing with common and important elements the web-caching servers that the structures Client/Server are not found in all, beyond we will define stages directed to the object of our studies.

The considered method is based in stages/actions (blocks), models (circles) and relationships (fine arrows) (Fig. 1). The diagram presented in Fig. 1 that indicates the possibility of implementation in software the routine, facilitating the performance analysis in varies cases.

As well as described on MENASCÉ & ALMEIDA 1998 [18], the two methods are based in the use of three models: workload (M1), performance (M2) and cost (M3). The methods show differences in the stages/actions that precede the formulation of the models, whereas considers in method the stages are more specific considering the environment in study.

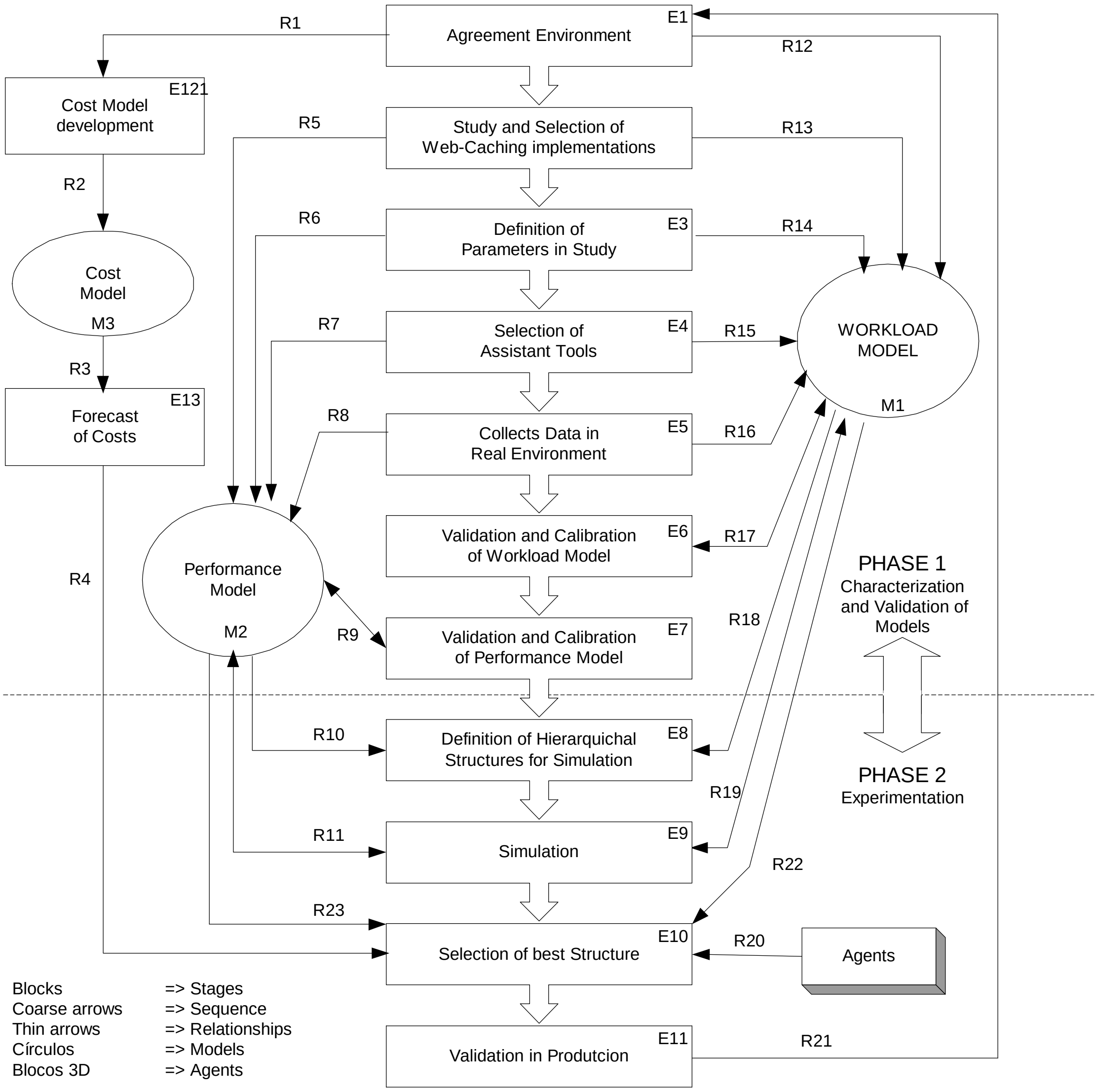
The method is divided in two phases, in the first phase at the characterizations and validations of the models are made second phase at the experimentations in laboratory.

3.2 Characterization and Validation of the Models

In this phase the workload, performance and cost models are defined, following some stages that will be described shortly at follow [5].

In the stage of agreement of the environment (E1) the responsible in it carry out to analysis of performance should: know the environment under study;

Fig. 1 - Method for Performance Analysis of Web Caching Hierarchic Structures



Agreement Environment
E1
R1
R12
Cost Model development
E121
Study and Selection of Web-Caching implementations
R5
R13
R2
Definition of Parameters in Study
E3
R6
R14
Cost Model
M3
Selection of Assistant Tools
E4
R7
R15
WORKLOAD MODEL
M1
R3
Forecast of Costs
E13
Collects Data in Real Environment
E5
R8
R16
Validation and Calibration of Workload Model
E6
R17
R4
Performance Model
M2
PHASE 1
Characterization and Validation of Models
Validation and Calibration of Performance Model
E7
R9
R18
Definition of Hierarquichal Structures for Simulation
E8
R10
PHASE 2
Experimentation
R19
Simulation
E9
R11
R22
R23
Selection of best Structure
E10
R20
Agents
Validation in Produtcion
E11
R21
Blocks => Stages
Coarse arrows => Sequence
Thin arrows => Relationships
Círculos => Models
Blocos 3D => Agents

understand the waited objectives and results; understand all the parameters that influence the environment; study and analyze the main elements that compose the hierarchic systems of web-caching; and consider sazonalidade questions.

In the stage of study and choice implementations of web caching (E2) they should be verified: available implementations; consider characteristics of performance of each implementation; cooperation mechanisms; implementations used for other institutions that allow connection; and availability of resources.

The stage of definition of parameters under study (E3) has purpose of define which will be the parameters under analysis. The goal is to manipulate such parameters and to follow the behavior of the hierarchic structure completely. The parameters normally evaluated are: hit ratio in numbers of requests and bytes, space to storage the consulted objects, amount of memory RAM, relation of cooperation, consumption of communication link, and rules of replacement, etc. The selection of the parameters must be in accordance with the objectives and waited results of performance for the environment in study.

During the selection of assistant tools (E4) it should be chosen two kinds of tools: a for extraction of information of the log files and another one for perform simulations in laboratory. These tools will be used in some stages of the method.

The stage of data's collects in real environment has as objectives verify the behavior of access to Internet from the institution, permitting an adequate formulation of the models that will be utilized during the simulation. This behavior can be evaluated through the analysis of log files from the own institution in case, in others institutions, using available information in others analysis, or generating synthetic data’s in laboratory. During this stage is necessary: esteem the workload that the system can come to suffer; verify the trend of growth or reduction of the load from historical data; analyze performance aspects; and analyze the future strategies, plans and its results in the load of the system.

The stage of validation and calibration of the workload (E6) is one of the most important stage of the proposed method. In this stage, the definite workload should be calibrated until be validated with the objective to approach it in the situation found in the real environment, eliminating possible variations caused by simplifications normally applied to the model.

As well as the stage (E6), the validation and calibration of the performance model (E7) should calibrated the defined performance model, verifying metrics and eliminating possible variations.

The workload is a representation of the real load of the system studied. This characterization should identify the basic components, choose the parameters of characterization, it collect data and classify the parameters [18]. The main considerations in the choice of the workload are: the service exercised by the workload; the level of detail; to representativeness; timeliness; levels of load; impact of others components or resources; and repeatability [1] [2] [3] [4].

The performance model (M2) is used to establish metric of performance of a Client/Server system in function of three categories of parameters: workload, basic software and hardware parameters [18]. The result of this model includes forecasts of the times of reply of the system, throughput, and use of the resources of the system, among others.

The cost model (M3) should enter the investments in the hardware, software, resources of telecommunications, support, among others, allowing to establish commitments of cost x benefit of the evaluated hierarchic structures [10] [18].

3.3 Experimentation

The experimentation phase is the phase that applies the definite models previously in simulated hierarchic structures in laboratory, allowing support for planned performance analysis.

Some stages compose this phase and the first one that it must be executed is the stage of definition of the hierarchics structures for simulation (E8) whose the goal is to propose different hierarchies structures, with viable implantations, that will be simulated and submitted to the performance analysis.

The stage of simulation (E9) has for objective to simulate all the hierarchies structures proposals (E8), applying the workload proposed (M1.R19.E9) and evaluating the metric ones of performance defined in the performance model (M2.R11.E9).

The stage of selection of the best structure (E10) must proceed with the analysis of the results gotten in the previous stage and select the best structure using criteria based on: metric of the performance model (M2.R23.E10); forecasts of costs (E13.R4.E10); influence of the Agents component. (Agentes.R20.E10); and parameters of workload (M1.R22.E10).

Finally it suggests itself to execution from the stage of validation in production (E11) that consists put in operation to structure selected in real environment and analysis the behavior and performance of the environment, about real workload from the institution, and it observe if the results approach of the values simulated.

3.4 Agents

In our method, we call of agents the different groups of people who are directly associates to the problem of performance analysis of a web caching hierarchic structure. We can divide these groups in: users, technical administrators and financial administrators.

Our concern is to present the existence of conflicts between these different groups and those differences must be considered in the stage of selection of the best structure (E10).

4 Validation of the Method

With the objective to validate the method proposed we executed a study of case with the academic net of the PUC-Minas. This study of case considered four campi: in the main campus has two servers (PUC and DCC) and the link Internet; e three remote units (IEC, Betim and BH2), each one with a server of web-caching and a data communication channel with the main campus.

The implementation web-caching chosen was the SQUID [36], but we also evaluate the Microsoft Proxy server [19]. We define as parameters under study the hit ratio in requests, the hit ratio in bytes and the function of the servers inside of the hierarchic structure. Or still, we apply variations only in the function of the servers inside of the hierarchic structure and follow the behavior of the hit ratio in number of requests and bytes.

To assist the experimentations we select the Calamaris [6] for analysis the log files and the Web-Polygraph [28] for simulation the experimentations in laboratory.

We collect logs of the structure in operation during 3 months. In a first moment we collect logs with servers arranged as presented in structure 1 in Fig. 2, after a period of two months we change the arranged of the servers for structure 2 of Fig. 2.

We shape our workload with three types of objects: HTML, images and download files. Each one with three characteristics: size, recurrence and cachable (it indicates the object percentage that can be stored). The objects respected the following distribution respectively: 8.5Kb, 23%, 90%; 4.5Kb, 75%, 80%; 225Kb, 2%, 20%. The characterization of the workload has as objective to approach the results of the environment simulated to the real environment. The real environment presented in average of hit ratio of 53,88% in requests and hit ratio of 32,45% in bytes.

We execute some simplifications in the workload in order to reduce the experimentation time and considering that our intention was validate the method proposed and not finds a structure with the best efficiency for the PUC-Minas.

The validation identified the simplifications realized, mainly due the lack of resources to simulate each one of servers and by limitations detected in the tool chosen for simulation.

The performance model was formulated in function of the amount of bytes direct fetch from sources, in this case the process that simulated the existence of the web servers. The performance model considers the sum total of bytes direct fetch from sources, and determines that the best structure is that one with lesser value of sum. This model it was validated considering that the Structure 2 used for collection of data is more efficient than structure 1 also evaluated.

We define 5 structures hierarchic for analysis, these are presented in Fig. 2.

Each structure was simulated 3 times to reduce errors caused for limitations of the simulation tool and the results obtained in function of the performance model are presented in Table 1.

Following the performance model the structure 5 presented the best performance, we waited that structure 3 had presented better resulted, this did not occur due the saturation in the computational resources used in the experimentation.

Assuming that structure 2 had been selected, we can considers the stage of validation in production as being the stage of collection of data, we would observe great distortion in the results what it would justify a return to the beginning of the method for new characterization of models and experimentation's.

The method was validated, all the stages was followed in accordance with the orientations of the proposed method and the results has reached the objective.

Fig. 2. Structures simulated

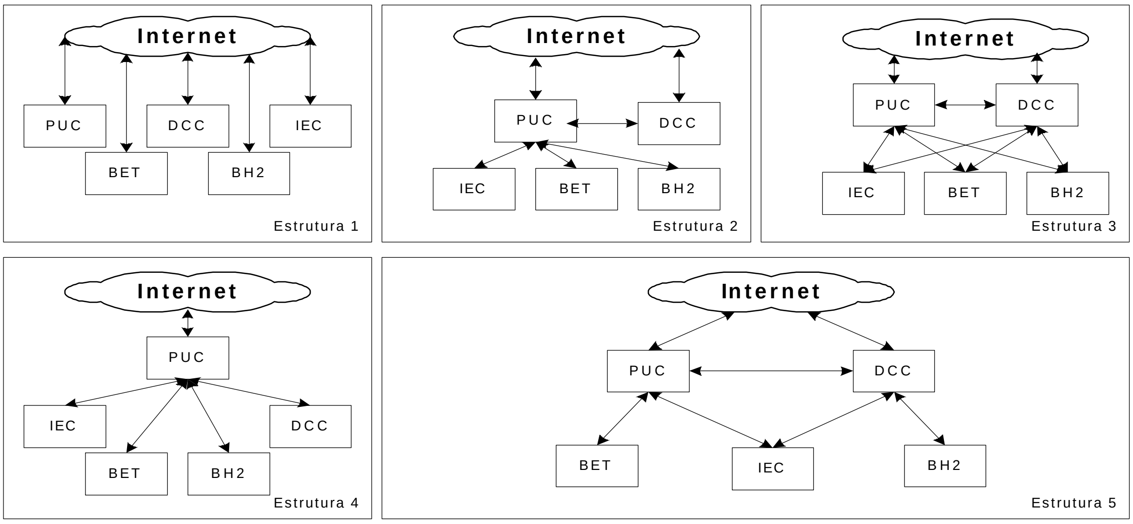


Table 1 – Medium values of sum of bytes direct fetch from sources in simulations.

Server	Structure 1	Strucuture 2	Strucuture 3	Strucuture 4	Strucuture 5
PUC	514832	1542527	970191	1889283	990742
DCC	508273	488544	1038755	107301	989569
IEC	503648	123789	90373	102780	91600
Betim	503129	118622	86305	101067	96774
BH2	508362	123914	91949	100523	95750
SUM	2538244	2397396	2277573	2300954	2264435

5. Conclusions

When we elaborate this article, we concern in presented a more specific method to deal with the problem of performance analysis of web caching hierarchic structures. This method inherited many stages of methods of Client/Server structure performance analysis showed in MENASCÉ [18]. The stage of Collects of Data (E5) and Validation in Production (E11) can extend for weeks until the behavior of the structure stay less committed by variations not waited during the period of measurement.

It is common to find restrictions in some cases for liberation of physical resources, what it hinders the adequate execution of all stages suggested in the proposed method. Considering this, the experimentation phase can be harmed and the laboratory environment cannot reproduce with representativeness form the real structure. In this case, we suggest the use of software's that simulates the physical equipment existence that does not exist.

In the impossibility of execution of all the stages described in the method, we believe that some could be excluded, since that the responsible by the analysis know the implications that this decision can cause. A simplified method must count at least with the following stages: Agreement of the Environment (E1); Definition of Parameters in Study (E3); Selection of assistant tools (E4); Definition of Hierarchic Structures for Simulation (E8); Simulation (E9); Selection of the Best Structure (E10). Beyond the necessity of definition of the performance and workload models. The stage of Collects of Data (E5) in real environment was removed, despite of its importance, because many assistant tools (E4) already indicate standards workload as suggestion for simulation. These standards workload also allow exclude the stages of validation of the models (E6 and E7) by the fact to already present calibrated models.

The presented method was applied to a study of case in the academic net of the PUC-Minas, it was possible validate each one of the stages and consequently validate the method completely. The complete description of the method and the results of the case can be verified in BARBOSA [5].

We suggest as future works the verification of the cover of the proposed method and its automation in software, allowing execute faster the process of performance analysis.

References:

[1] AGGARWAL, C., WOLF, J.L., YU, P.S., *Caching on the World Wide We***b**, IEEE Transactions on Knowledge and Data Engineering, Janeiro-Fevereiro 1999, Volume: 11, 14p.

[2] ALMEIDA, Virgilio A.F., BESTRAVOS, A., CROVELLA, M.Oliveira, *Characterizing reference locality in the www*, In Proceedings Of IEEE Conf. On Parallel and Distributed Systems (PDIS'96), Miami Beach, Flórida, Dezembro 1996.

[3]ARLITT, M. F., WILLIAMSON, C. L., *Web Server workload characterization: the search for invariants*, In Proc. Of ACM SIGMETRICS, Philadelphia, PA, Abril 1996.

[4]ASAKA, T., MIWA, H., TANAKA, Y., *Distributed Web caching using hash-based query caching method*, Control Applications, 1999. Proceedings of the 1999 IEEE International Conference on, 1999, volume 2, 26p.

[5] BARBOSA, Marco A. S., *Proposta de Análise de Desempenho de estruturas hierárquicas de servidores Web-caching*, Tese de Mestrado, Programa de Pós-Graduação em Engenharia Elétrica, PUC-Minas, Belo-Horizonte, Maio 2002.

[6] BEERMANN, Cord, *Calamaris*, http://cord.de/tools/squid/calamaris.

[7] DAVISON, Brian D., *Online Web Caching Resources*. http://www.web-caching.com.

[8] DAVISON, Brian D., *A Web caching primer*, IEEE Internet Computing, Julho-Agosto, 2001, volume 5, 8p.

[9] DYKES, S.G., JEFFERY, C.L., DAS, S., *Taxonomy and design analysis for distributed Web caching*, Systems

Sciences 1999, HICSS-32 International Conference on Proceedings of the 32nd Annual Hawaii, 1999, 11p.

[10] FONSECA, Erik Luiz S., *Hierarquias de Servidores Proxy Cache WWW: Instrumentação e Análise de Desempenho*, Universidade Federal de Minas Gerais, Março, 1999.

[11] HOWARD, J. et al., *Scale and performance in a Distributed File System*, ACM Trans. Computer Systems, Fevereiro 1998, volume 6, número 1, 31p.

[12] KENYON, C., *The evolution of Web-caching markets*, Revista Computer, Novembro 2001, volume 34, 3p.

[13] KRISHANA, P., VITTER, Jeffrey, *Optimal Prediction for Prefetching in the Worst Case*, A shortened version appeared in Proceedings of Fifth Annual ACM-SIAM Symposium on Discrete Algorithms, Janeiro, 1994, 10p.

[14] LAI, Guanpi, LIU, Mingkuan, WANG, Fei-Yue, ZENG, D., *Web caching: architectures and performance evaluation survey*, 2001 IEEE International Conference on Systems, Man, and Cybernetics, 2001, volume 5, 6p.

[15] LIU, Mingkuan; WANG, Fei-Yue; ZENG, D., YANG, Lizhi, *An overview of world wide web caching*, 2001 IEEE International Conference on Systems, Man, and Cybernetics, 2001, volume 5, 6p.

[16] MARKATOS, E. P., CHORNAKI, C. E., *Top-10 Approach to Prefetching the Web*, in Proceedings of the Eighth Annual Conference of the Internet Socienty (INET'98), Geneva, Switzerland, Julho, 1998.

[17] MEIRA, Wagner Jr., FONSECA, Erik L. S., MURTA, Cristina Duarte Murta, ALMEIDA, Vírgilio A F, *Analyzing Performance of Cache Server Hierarchies*, Proceedings of the XVIII International Conference of the Chilean Society of Computer Science, IEEE Computer Society, Los Alamitos, CA, 1998, 8p.

[18] MENASCÉ, Daniel A., ALMEIDA, Virgílio A.F., DOWDY,Larry W., *Capacity Planning and Performance Modeling*, Primeira Edição, Nova Jersey: Editora Prentice Hall, 1994.

[19] MICROSOFT, Microsoft Proxy Server. http://www.microsoft.com/isaserver/evaluation/previousversions/default.asp

[20] MURTA, Cristina Duarte. *Modelo de Particionamento de Espaço para Cache da World Wide Web*, Tese Doutorado, Universidade Federal de Minas Gerais, 1999.

[21] MURTA, Cristina Duarte, ALMEIDA, Virgílio A.F., *Using performance maps to understand the behavior of Web caching policies*, Proceedings The Second IEEE Workshop on Internet Applications, 2001, 7p.

[22] PBS, *Life on the Internet*, http://www.pbs.org/internet/timeline/

[23] POVEY, D., HARRISON, J., *A distributed Internet Cache*, in Proc. 20th Australian Computer Science Conf., Sydney, Australia, Fevereiro 1997.

[24] RODRIGUEZ, Pablo, SPANNER, C., BIERSACK, E.W., *Analysis of Web caching architectures: hierarchical and distributed caching*, IEEE/ACM Transactions on Networking, Agosto, 2001, volume 9, 15p.

[25] ROSS, K., *Hash Routing for Collections of shared Web Cachês*, IEEE Network Magazine, Novembro, 1997, 37p.

[26] ROUSSKOV, AIex, SOLOVIEV, Valey, *On Performance of Caching Proxies*, Conferência ACM SIGMETRICS'98, Junho 1998.

[27] ROUSSKOV, Alex, SOLOVIEV, Valey, *A Performance Study of the Squid Proxy on HTTP/1.0*, World Wide Web Journal, WWW Characterization and Performance Evaluation, 1999.

[28] ROUSSKOV Alex, WESSELS, Duane, WEB POLYGRAPH, Proxy Performance benchmark, http://polygraph.ircache.net/

[29] SHARP, Edward, *Caching Performance Realities*, www.web-caching.com, Janeiro, 2001.

[30] SCHAWARTATZ, M.F. et al., CHANKHUNTHOD, A. et al., DANZIG, P.B. et al., NEERDAELS, C. et al., " *hierarchical internet object cache*, in Proceedings 1996 USENIX Technical Conference, San Diego, CA, Janeiro 1996.

[31] SELVAKUMAR, S., PRABHAKAR, P., *Implementation and comparison of distributed caching schemes*, Proceedings. IEEE International Conference on Networks 2000 (ICON 2000), 2000, 9p.

[32] TEWARI, Renu, DAHLIN, Michael, VIN, Harrick M., KAY, Jonathan S., *Design Considerations for Distributed Caching on the Internet*, IBM T.J. Watson Research Center Department of Computer Sciences Cephalapod Proliferationists, 1999.

[33] VOELKER, G. M., ANDERSON, E. J., *Implementing Cooperative Prefetching and Caching in a Glogally-Managed Memory System*, in Proceedings of the 1998 ACM SIGMETRICS Conference on Performance Measurement, Modeling and Evaluation, Junho, 1998.

[34] WESSELS, D., CLAFFY K., *Application of Internet Cache Protocol (ICP)*, version 2. Networking Group, RFC 2187, Setembro 1997.

[35] WESSELS, D., CLAFFY K., *Internet Cache Protocol (ICP)*, version 2. Networking Group, RFC 2186, Setembro 1997.

[36] WESSELS, Duane, ROUSSKOV, Alex, CHISHOLM, Glenn, *SQUID*, http://www.squid-cache.org/