# Improving Performance of World Wide Web by Adaptive Web Traffic Reduction

Achuthsankar S. Nair, and J. S. Jayasudha

Open Science Index, Industrial and Manufacturing Engineering Vol:2, No:5, 2008 publications.waset.org/526.pdf

***Abstract*****—**The ever increasing use of World Wide Web in the existing network, results in poor performance. Several techniques have been developed for reducing web traffic by compressing the size of the file, saving the web pages at the client side, changing the burst nature of traffic into constant rate etc. No single method was adequate enough to access the document instantly through the Internet. In this paper, adaptive hybrid algorithms are developed for reducing web traffic. Intelligent agents are used for monitoring the web traffic. Depending upon the bandwidth usage, user's preferences, server and browser capabilities, intelligent agents use the best techniques to achieve maximum traffic reduction. Web caching, compression, filtering, optimization of HTML tags, and traffic dispersion are incorporated into this adaptive selection. Using this new hybrid technique, latency is reduced to 20 – 60 % and cache hit ratio is increased 40 – 82 %.

***Keywords*****—**Bandwidth, Congestion, Intelligent Agents, Pre-fetching, Web Caching.

## I. Introduction

THE exponential rate of growth of World Wide Web (WWW) has led to an increase in Internet traffic and degradation in user-perceived latency. With the continuing growth of Internet, web users are suffering from the network congestion and server overloading. Content simplification and data compression techniques are used to reduce the amount of bytes sent over the network. Since these techniques cannot be used for much reduction in bandwidth consumption, alternate schemes such as web caching, web cache sharing, web server pushing, browser initiated server pushing are developed. But these techniques have not yet attained the satisfactory results. Web caching and web cache sharing are unsatisfactory techniques due to dynamic documents, CPU overhead, memory spending for caching, administrative overhead etc.

In web server pushing technique, it is difficult for a web content provider to know the proper place to push documents. In the browser initiated server pushing, the server does not know what is in the client cache. Without considering the client's cache, the web server pushes more images than needed. The user interacts only with the web browsers to fetch the web pages. So the browser can store the history of each user behavior and it can take intelligent decision depending upon the user behavior and traffic condition with the help of intelligent agents. In the proposed adaptive traffic reduction technique, intelligent agents are used at the server and client side to assist them to take intelligent decision on reducing the size of the document for each web request.

The remainder of this paper is organized as follows. The related research works for improving network performance by reducing web traffic are discussed in section 2. Section 3 outlines the existing web traffic reduction techniques. Section 4 describes the uses of web traffic reduction. Section 5 describes the new adaptive web traffic reduction technique and algorithms. In section 6, the new adaptive technique is incorporated into a browser and its performance is compared with some existing browsers. Finally, section 7 concludes the paper.

## II. Related Works

A web-based management system for network monitoring is proposed in [1]. The techniques for improving the network performance by traffic reduction are discussed in [2]. Wisconsin Adaptive Web Assistant (WAWA) is developed in [3] that assist a user in locating specific, current and relevant information on the World Wide Web. The server side overhead can be reduced if the number of requested files is reduced. Multiple files such as HTML text and image files will be packed in an object package file for efficient transmission. At the receiving side, a web browser should unpack the files [4]. Knowledge of HTML is not necessary to create a web page. The HTML markup generated by other applications is not of the same standard as hand-coded markup. The consequences of such bloated HTML are unnecessarily increase storage costs, transmission costs, download times and browser rendering times.

Optimization of HTML generated by WYSIWYG programs is discussed in [5]. For optimizing web content delivery, web server accelerator is used in [6]. The web server accelerator resides in front of a web server and improves server performance by efficiently delivering cached responses while leaving the role of content generator to the web server. Traffic dispersion technique is discussed in [7]. In traffic dispersion, burst traffic is divided into many sub-bursts that are transmitted in parallel through multiple paths which do not share any physical links and are sequenced at the destination. Intelligent browsing for multimedia applications has been developed in [8]. The learning agents attempts to learn what item the user is searching for by watching the user's normal browsing actions. The learning agent builds a profile of the user's search pattern. The inferred goal provides a reliable way of estimating the relevance of a multimedia item to the agent's actual search goal [8]. A heuristic bandwidth allocation method is proposed in [9] for managing the

Manuscript received September 29, 2006.

Achuthsankar S. Nair is working as Hon. Director, Centre for Bioinformatics, University of Kerala, Trivandrum, India.

J. S. Jayasudha is working as Asst. Professor, Department of Computer Science, SCT College of Engineering, Trivandrum, India (phone: 919495376533, 91944311000, e-mail: jayasud@rediffmail.com).

Open Science Index, Industrial and Manufacturing Engineering Vol:2, No:5, 2008 publications.waset.org/526.pdf

bandwidth dynamically by considering the correlation characteristics.

Cache stores cacheable requests and responses for handling new requests. If a new request that is same as a stored request arrives, then cache can supply the stored response rather than accessing the resource from the web server. The web caching and web cache sharing schemes are useful for latency reduction, bandwidth conservation and disconnected operation. If a group of clients are topologically close and under common administrative control, then the administrator could install one or more proxy caches in front of the clients for forming a cluster to lower the client-perceived latency. The methodologies for client cluster identification are discussed in [10]. The various web caching techniques, caching protocols and caching architectures are discussed in [11] - [17]. To extend the effectiveness of HTTP caches, techniques such as cooperative caching, pre-fetching, partial transfers, delta encoding, cache based compaction and HTML macros are developed [14, 15]. A detective browser is proposed in [18], it can immediately determine whether the requested content is dynamic or secured. If so, the browser will bypass the proxy and forward the request directly to the web server.

## III. Web Traffic Reduction Techniques

Web traffic reduction techniques are necessary for accessing the web sites efficiently with existing network facility. It is costlier to use infinite bandwidth in all organizations. Many studies show that the web caching has the maximum limit of cache hit ratio of 50%. But pre-fetching can improve the hit ratio to 60% or even more than 80% [19]. But pre-fetching techniques increases the web traffic for pre-fetching the anticipated sites. Various techniques available for web traffic reduction are given below.

### *A. Content Simplification*

Web designers can use common sense to reduce page complexity or special tools can be used to optimize image coding. But some data such as medical images, broadcast quality videos and executable software cannot be simplified without loss of meaning. Optimization of HTML tags can be done for reducing the size of the web page to be transmitted [5]. But the content simplification and optimization techniques cannot be used for much reduction in web traffic.

### *B. Compression*

Redundant bits within a single transfer can be reduced using compression techniques. Existing general purpose compression algorithms provide significant size reductions. Several compression techniques such as Huffman coding, LZ, LZW, JPEG, MPEG, H.261, H.263 are available for text, image, audio and video compression [20]. Compression ratio depends on the compression algorithm and size of the file to be compressed. These techniques can reduce only temporal or spatial redundancy. But it can't reduce the frequent transmission of the same file through the network.

### *C. Web Caching*

A cache is a facility that stores cacheable requests and responses. Subsequent requests can be satisfied from the cache instead of accessing the objects from web server. The cache can operate on a client or server or on an intermediate system. In proxy caching, proxy server stores cacheable responses to the URL requests, subsequent requests for the same URLs yields cache hit. The URL request from the client is forwarded through proxy servers. The proxy servers accept the URL request from the client and it checks in its own local cache for local hit, if requested object is not available there, it forwards that request to the web server. In web cache sharing, for each URL request, the proxy first checks in its local cache, if there is no local hit, then that proxy checks in other proxies cache for remote hit. If there is no remote hit, then it forwards that request to the web server. Upon receipt of that document, the requested proxy server stores it in cache and returns the document to the client. Due to dynamic documents, CPU overhead, memory spending for caching and administrative overhead, web caching techniques are considered as unsatisfactory techniques. But when the network bottleneck such as congestion is compared, these overhead are negligible.

### *D. Web Server Pushing*

Web server pushes some or all of the documents to some place near the client site. When the client accesses a document, it will go to a site that is nearer to the client [21]. This approach reduces the latency in accessing the documents. In this technique, it is difficult for a web content provider to know the proper place to push documents.

### *E. Browser Initiated Server Pushing (BIP)*

In BIP, upon receiving a HTTP request, the server actively pushes embedded contents if the permission is given by the client [21]. By means of the pushing mechanism, the HTML web page will be down loaded in one RTT if the embedded contents and the HTML web page are on the same server. This approach reduces the download latency for web pages and to improve web server resource utilization. In the browser initiated server pushing, the server does not know what is in the client cache. Without considering the client cache contents, the server pushes much more images than actually needed.

### *F. Load Balancing Techniques*

In this approach, if any server is over loaded, its jobs are shared by other under loaded servers. Load balancing systems monitor the health of the available servers and make decisions to route the traffic for optimizing the performance and availability. This ensures that the users will be connected to the most available server, providing excellent and predictable quality of service to the end user [22]. Many traffic management and load balancing techniques are discussed in [23]. An adaptive load balancing scheme for web servers is discussed in [22]. An adaptive multi agent coordination algorithm is proposed in [24] for performing distributed dynamic load balancing. A load cluster management system based on SNMP and web technology is discussed in [25]. A dynamic load balancing algorithm is proposed in [26] for improving throughput. Even if several load balancing algorithms are available, it is difficult for distributing the work among the existing servers and balancing the work among them.

Open Science Index, Industrial and Manufacturing Engineering Vol:2, No:5, 2008 publications.waset.org/526.pdf

### *G. Intelligent Agents*

Intelligent agent monitors user actions and dynamically takes decision to access web sites by analyzing the traffic on the internet. The intelligent agent monitors the web traffic and it reports the status of bandwidth usage to the browser [27]. The browser parses the HTML page and the agent makes prediction about future references. If the traffic reported by the intelligent agents are less than the threshold, then the browser can pre-fetch the predicted references before it is actually referenced, potentially resulting in much lower latencies. Using intelligent agents delay in accessing the network, bandwidth consumption and network bottlenecks can be reduced.

### *H. Bandwidth Management*

Network bottleneck such as congestion can be avoided by managing the bandwidth properly. If infinite bandwidth is available, there is no need for bandwidth management. It is very costlier for implementing infinite bandwidth in each organization. If congestion occurs there will be packet loss and these packets are to be retransmitted for achieving a reliable data transfer. Bandwidth management is necessary for reducing congestion and web traffic. Intra-domain bandwidth management in differentiated services network is discussed in [28]. Dynamic bandwidth management scheme for a wireless network consisting of heterogeneous computers and devices is proposed in [29]. A two tier resource management model for the internet is proposed in [30]. An algorithm for bounding the bandwidth of a web crawler is proposed in [31].

### *I. Intelligent Router*

Intelligent router can route the traffic efficiently by reducing the network bottlenecks such as congestion. Intelligent router is proposed in [32] to implement distributed applications in a heterogeneous environment. Intelligent routers can dynamically route the web traffic and perform reliable service even if some routes are failed. It can take dynamic decisions for selecting the route depending on the current work load on the network. Layer 3 switches are available for directing the request for dynamic contents to the web servers and other requests to the cache servers.

## IV. Uses of Web Traffic Reduction

Reduction of latency, reduction of bandwidth usage and alleviating network bottlenecks such as congestion are the main benefits of web traffic reduction. Intelligent agents help to access the desired web sites and make the surfing easier. With the help of intelligent agents the web server can negotiate the capabilities of the browser and it can take intelligent decision by considering the browsers capabilities and traffic condition. By reducing the traffic in the existing network, the documents can be accessed in real time. The extra cost that is to be spent for replacing the existing network can also be avoided. The existing network is enough to satisfy the user's request by managing the web traffic efficiently.

## V. Adaptive Web Traffic Reduction Technique

In this new adaptive technique, intelligent agents are maintained at the client and server side for monitoring the web traffic. Adaptive web traffic reduction technique is a hybrid technique and combines the existing techniques such as web caching, web cache sharing, pre-fetching, traffic dispersion, compression etc. Traffic reduction algorithms are developed for combining these techniques. Web traffic reduction is achieved in adaptive manner by monitoring the user's preferences and bandwidth usage. In this technique, intelligent agents monitor the bandwidth usage and select the best techniques at the client side and server side to achieve maximum traffic reduction. Since it is a hybrid technique, efficient bandwidth utilization and more web traffic reduction are achieved. Schematic block diagram of the functions of intelligent agents at web server and browser side is shown in Fig. 1.

### *A. Web Traffic Reduction at Client Side*

The intelligent agents help the browser to make it intelligent. The user interacts only with the web browsers to fetch the web pages. So the browser can store the history of each user behavior and it can take intelligent decision depending upon the user behavior and traffic condition. Agents at the client side monitor the user's preferences and bandwidth usage, then takes decision dynamically for reducing latency and web traffic. The client side algorithm for web traffic reduction is given below.

**Traffic Reduction Algorithm at Browser side**

```
1.  For each URL request
a)  If the request is secured or dynamic, forward the
    request to the web server
b)  Else if the request is static, checks in cache
      (i) If that document is available in browser cache,
          fetch it from browser cache
     (ii) Else check in local proxy cache, if local hit,
          fetch it from the local proxy cache
    (iii) Else check in remote proxy cache, if remote
          hit, fetch it from that cooperating proxy
     (iv) Else forward that request to the web server,
          upon receipt of that document, store it in local
          proxy cache and browser cache.
2.  Intelligent agents monitor the web traffic
a)  If bandwidth usage is less than a threshold value say
    x1, allow all types of traffic
b)  If bandwidth usage is between threshold values say
    x1 and x2, do not allow pre-fetching
c)  If bandwidth usage is greater than a threshold values
    say x2, do not allow pre-fetching and cache
    refreshing. Display only links of the images on all
    documents. Fetch images only if the user clicks a
    particular link for displaying the image.
3.  When uploading large files, intelligent agents prompt
    the user to choose compression techniques.
4.  When searching, the intelligent agent's identify the
    user's preferences from log information and display
    the related sites first.
```

### *B. Web Traffic Reduction at Server Side*

The intelligent agents at the server side negotiate with the web browser about its capabilities. Depending upon the capabilities of the browser, the intelligent agents take dynamic

Open Science Index, Industrial and Manufacturing Engineering Vol:2, No:5, 2008 publications.waset.org/526.pdf

decision about sending the documents in compressed form. The proposed algorithm for web traffic reduction at server side is given below.

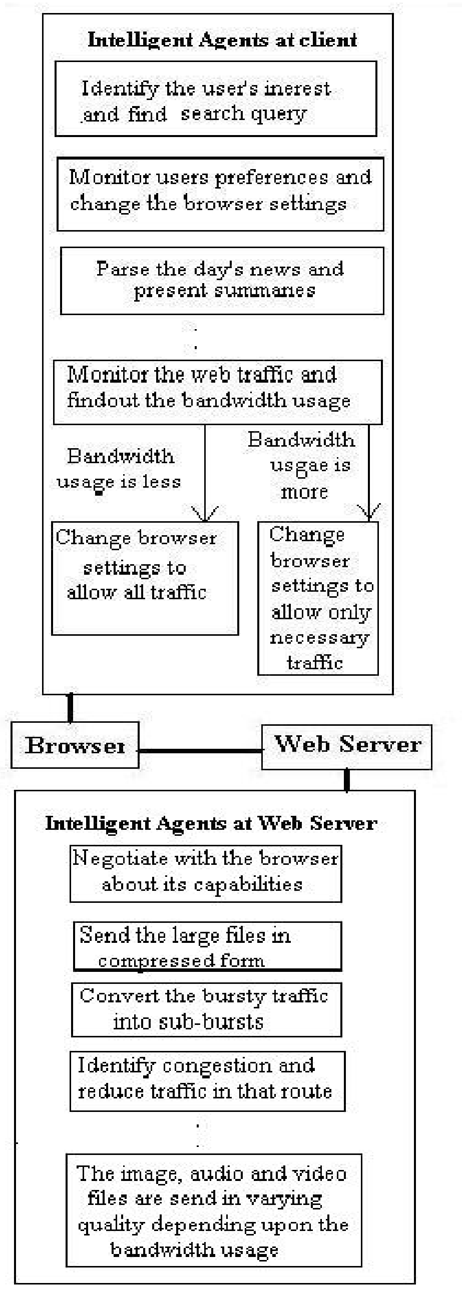
Intelligent Agents at client
Identify the user's inerest and find search query
Monitor users preferences and change the browser settings
Parse the day's news and present summanes
Monitor the web traffic and findout the bandwidth usage
Bandwidth usage is less
Bandwidth usgae is more
Change browser settings to allow all traffic
Change browser settings to allow only necessary traffic
Browser
Web Server
Intelligent Agents at Web Server
Negotiate with the browser about its capabilities
Send the large files in compressed form
Convert the bursty traffic into sub-bursts
Identify congestion and reduce traffic in that route
The image, audio and video files are send in varying quality depending upon the bandwidth usage

Fig. 1 Schematic block diagram of the functions of Intelligent Agents at Web server and Browser side

### Traffic Reduction Algorithm at Web Server side

1. Intelligent agents monitor the bandwidth usage
   (a) If the bandwidth usage is less than a threshold say x1, send good quality image, audio and video files.
   (b) If the bandwidth usage is between threshold values say x1 and x2, send medium quality image, audio and video files.
   (c) If bandwidth usage is greater than a threshold value say x3, send only low quality image, audio and video files
2. Intelligent agents identify the large files and send it in compressed form.
3. Intelligent agents separate the static pages from dynamic pages and make the static pages as public to cache it in browser side.
4. Intelligent agents identify congestion and reduce the traffic in that route.
5. Intelligent agents convert the burst nature of the traffic into sub bursts.
6. Intelligent agents optimize the HTML markup generated by WYSIWYG programs. Only optimized HTML is delivered to the browser

## VI. Simulation and Discussion of Results

We have simulated the adaptive traffic reduction technique by implementing a new browser called dynamic pre-fetching browser (DPB) with pre-fetching, web caching, traffic monitoring capabilities. The intelligent agents help the browser to have these capabilities to reduce latency and to increase response time. We have analyzed this new technique by simulating the environment in a medium sized network containing 90 nodes and trace driven data are collected. We have compared the percentage bandwidth usage of our dynamic pre-fetching browser with Internet Explorer (IE) and Netscape Navigator.

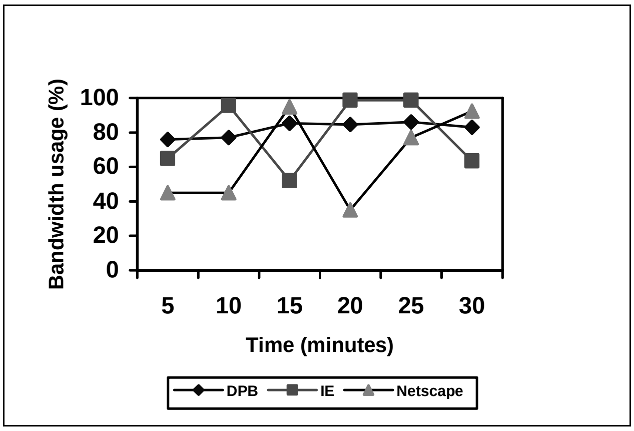

Fig. 2 Percentage use of Existing Bandwidth by Dynamic Pre-fetching browser (DPB), Internet Explorer (IE) and Netscape Navigator

We have observed that the bandwidth usage is almost constant using adaptive traffic reduction algorithms by DPB browser as shown in Fig. 2. Without considering the current network traffic, the other browsers attempts to fetch more objects than the network can afford. But our new technique makes the browser to monitor the current network traffic and user's preferences and helps the browser to utilize the maximum bandwidth available effectively. We have observed that the average bandwidth usage using our new browser is about 70 to 90 % of the maximum bandwidth available of the

Open Science Index, Industrial and Manufacturing Engineering Vol:2, No:5, 2008 publications.waset.org/526.pdf

current network. Intelligent agents help the browser to monitor the bandwidth usage and adjust the pre-fetching of subsequent links to make constant traffic and to avoid the network bottleneck such as congestion. We have also observed that, by incorporating the traffic reduction algorithms in the browser, latency is reduced at a range of 20 – 60 % and cache hit ratio increased 40 – 82 % as shown in Fig. 3 and 4.

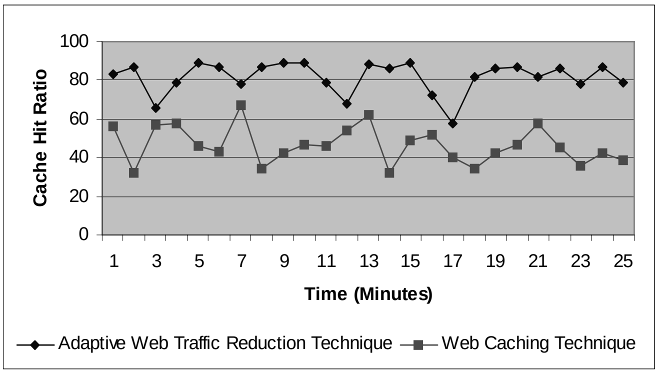

Fig. 3 Analysis of Cache Hit Ratio by Adaptive Reduction Technique and Web Caching Technique

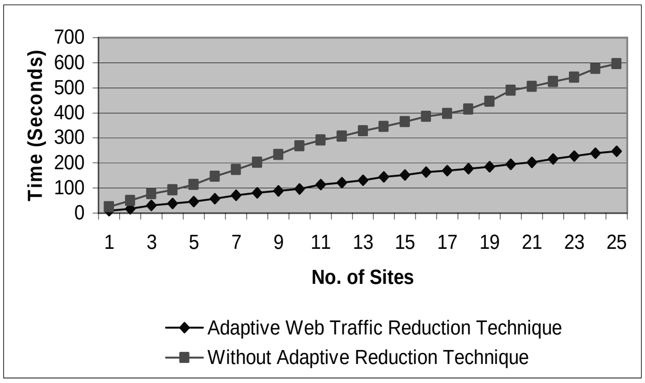

Fig. 4 Analysis of Latency by Adaptive Reduction Technique and Web Caching Technique

## VII. Conclusion

The intelligent agents help the browser and web server to monitor the bandwidth usage and user's preferences. In this paper an adaptive technique is proposed for reducing web traffic and to access the web sites efficiently. The proposed algorithms at client side and server side are efficient to reduce the web traffic in adaptive manner. The simulation results show that this adaptive technique maintains almost constant traffic and provides effective bandwidth usage. Since it is a hybrid technique, latency is reduced to 20 – 60 % and cache hit ratio is increased 40 – 82 %. New technologies are to be developed to reduce web traffic without increasing the cost spent for networking.

## References

[1] Jae-Won Choi and Kwang-Hui Lee, "A Web-based Management System for Network Monitoring", *IEEE workshop on IP operations and management (IPOM 2002)*, 2002, pp. 98-102.

[2] Zhen Sheng Guo, Yan Zhuang, "Improving Network Performance by Traffic Reduction", *International Conference on Information, Communication and Signal Processing*, 1997, pp. 1226-1230.

[3] Jude Shavlik, Susan Calcari, Tina Eliassi-Rad and jack Solock, "An Instructable Adaptive Interface for Discovering and Monitoring Information on the World-Wide Web", *Proceedings of International Conference on Intelligent user interfaces, ACM*, 1999, pp.157-160.

[4] Hiroshi Fujinoki, Kiran K Gollamudi, "Object Packaging – Web Response Time Reduction for slow and busy web servers", *Proc. of the 27th Annual Conference on Local Computer Networks, IEEE Computer society*, 2002.

[5] Jacqueline Spiesser, Less Kitchen, "Optimization of HTML Automatically generated by WYSIWYG Programs", *Proceedings of the ACM SIGPLAN 2004 Haskell Workshop*, 2004, pp.80-91.

[6] Dongjun Shin, Kern Koh, "Optimizing Web Content Delivery using Web Server Accelerator", *25th Australasian Computer Science Conference*, 2002, pp.233-239.

[7] Fumio Ishizaki, "Study on reduction of total bandwidth requirement by traffic dispersion", *In Proceedings of International conference on ATM, IEEE*, 2001, pp.285-289.

[8] Chris Drummond, Dan Ionescu, Robert Holte, "Intelligent Browsing for Multimedia Applications", *Proceedings of Multimedia, IEEE*, 1996, pp. 386-399.

[9] Yen-Wen Chen, "Experimental Study of Internet Traffic Modeling and Bandwidth Allocation", *IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, 2001, pp.587-590.

[10] Balachander Krishnamurthy, Jia Wang, "On Network – Aware Clustering of Web Clients", *Proceedings of ACM SIGCOMM*, 2000, pp. 97-108.

[11] Hossam Hassanein, Zhengang Liang and Patrick Martin, "Performance Comparison of Alternative Web Caching Techniques", *Procceedings of the seventh International Symposium on Computers and Communications, IEEE*, 2002.

[12] C. Mala and J.S. Jayasudha, "Web Cache Sharing Techniques", *Proceedings of 6th International Conference of High Performance Computing, Asia*, 2002.

[13] M. Liu, F. Wang, D. Zeng, L.Yang, "An Overview of world wide Web Caching", *International conference on Systems Man and Cybernetics, IEEE*, 2001, pp.3045-3050.

[14] Greg Barish and Katia, "World Wide Web Caching: Trends and Techniques", *IEEE Communication*, May 2000, pp. 178-185.

[15] Jeffrey C. Mogul, "Squeezing More Bits Out of HTTP Caches", *IEEE Network*, May/June 2000, pp. 6-12.

[16] Hykyung Bahn, "A Shared Cache Solution for the Home Internet Gateway", *IEEE Transactions on Consumer Electronics*, Vol.50, No.1, Feb 2004, pp. 168-172.

[17] Li Xiao, Xiaodong Zhang, Artur Andrzejak, Songqing Chen, "Building a Large and Efficient Hybrid Peer to Peer Internet Caching System", *IEEE Transactions on Knowledge and Data Engineering*, Vol 16, No.6, June 2004, pp.754-769.

[18] Songqing Chen and Xiaodong Zhang, "Detective Browsers : A Software Technique to Improve Web Access Performance and Security", *Proceedings of 7th International workshop on Web content caching and Distribution (WCW' 02)*, 2002.

[19] Shi L, Gu Z, Wei L and Shi Y. "Popularity-based selective markov model". In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, 2004, pp. 504-507.

[20] Fred halsall, Mutimedia Communications, *Pearson Education*, 2002.

[21] Wenting Tang, Matt W. Mutka, "Intelligent Browser Initiated Server Pushing", *IEEE International Conference on Performance, Computing, and Communications Conference*, 2000, pp.17-23.

[22] James Aweya, M. Ouellette, D.Y Montuno, B. Doray, K. Felske, "An Adaptive Load Balancing Scheme for web servers", *International Journal of Network Management*, 2002, pp.3-39.

[23] http://www.zenus.com/solutions/zlb/indexfull.html

[24] David Vengerov, Hamid R Berneji, Alex Vengerov, "An adaptive coordination among fuzzy reinforcement learning agents performing distributed dynamic load balancing", *Proceedings of 11th IEEE International Conference on Fuzzy Systems*, 2002, pp.179-184.

[25] Myung-Sup kim, Mi-Jeong Choi, James W. Hong, "Highly available and efficient load cluster management system using SNMP and web", *Proceedings of Network Operations and Management Symposium, IEEE*, 2002, pp.619-632.

[26] Hemant B. More, Jie Wu, "Throughput Improvement through Dynamic Load Balance", *IEEE southeast con'94*, 1994, pp.339-342.

[27] Jayasudha J.S, Achuthsankar S. Nair, "An Intelligent Browser for Web Traffic and Latency Reduction", *Proceedings of International*

*Conference of Trends in Industrial Measurements and Automation (TIMA)*, 2004.

[28] S. Jha, M. Hassan, P.Nanda, N. Ahmed, "Intra-domain Bandwidth Management in Differentiated Services Network", *IEEE Conference on Local Computer Networks*, 2000, pp.326-327.

[29] Samarth H. shah, K. Chen, K. Nahrstedt, "Dynamic Bandwidth Management for single-hop Ad-hoc Wireless Networks", *Proceedings of first IEEE international Conference on Pervasive Computing and Communications*, 2003.

[30] A. Terzis, L. Wang, J. Ogava, L. Zhang, "A two-tier Resource Management Model For The Internet", *Proceedings of Global Telecommunications Conference, IEEE*, 1999, pp. 1779-1791.

[31] M. Diligenti, M. Maggni, F.M. Pucci, F. Scarselli, "Design of a Crawler with bounded bandwidth", *Proceedings of WWW-2004, ACM*, 2004, pp. 292-293.

[32] C. Daniel Wolfson, Ellen M. Voorhees, M.M. Flatley, "Intelligent Routers", *Proceedings of 9$^{th}$ International Conference on Distributed Computing Systems, IEEE Computer Society*, 1989, pp. 371-376.

Open Science Index, Industrial and Manufacturing Engineering Vol:2, No:5, 2008 publications.waset.org/526.pdf