Wayne Stone. WEBSITE LOG ANALYSIS: Case Study of USA Cycling's Website. A Master's Paper for the M.S. in L.S. Degree. November, 2002. 78 pages. Advisor: Gregory B. Newby.

Numerous articles on the reasons for web log analysis exist. Much of the web-log analysis literature deals with how to collect data, technical aspects, and how to select the appropriate software for collecting the data; it will be the aim of this paper to create a user profile for USA Cycling's website by using WebTrends software to analyze web-log files. After the user profile has been developed, it will be shown that the web-log analysis of USA Cycling's website can be used to make daily and long term decisions about the its functionality. In addition, this paper will cover the basic issues of web-log analysis as well as exploring the practical application for USA Cycling. To accomplish these tasks, USA Cycling's web-logs were analyzed from August 1999 to April 2002 using WebTrends log analyzing software and key questions were developed based on observations and sent to USA Cycling for clarification.

Headings:

Internet – USA Cycling.

Internet -- Statistics.

Use studies -- Internet.

Web sites -- Case studies.

World Wide Web -- Statistics.

WEBSITE LOG ANALYSIS:
CASE STUDY OF USA CYCLING'S WEBSITE

By
Wayne Lawrence Stone

A Master's Paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Library Science.

Chapel Hill, North Carolina

November, 2002

Approved by:

_____
Advisor

ACKNOWLEDGMENTS

The following people were instrumental in the completion of this project.

Jennifer Winn	**Greg Newby** SILS Professor	**Nikki Warren**	**"Siggy"**
Kelly Walker USA Cycling	**Danny Smith** USA Cycling Summer Intern	**Frank and Shirley Stone**	

TABLE OF CONTENTS

ACKNOWLEDGEMENTS iii

LIST OF TABLE vii

LIST OF GRAPHS viii

INTRODUCTION 9
- Background: USA Cycling and Its Purpose 11
- USA Cycling's Information Services 12
- USA Cycling Website 12
- Web Logs at USA Cycling 13

LITERATURE REVIEW 13
- Questions for the paper and research 14
- Measuring and Characterizing web traffic 15
- Web Background 16
  - Origins 16
  - The Next Step: ARPANET, TCP/IP, 1980s and Beyond 16
  - State of the Web: Web Trends and Traffic 17
  - Semantic Components of the Web-URI, HTML, and HTTP 18
  - Terms and Concepts 19
- Web Traffic and Server Log Analysis 20
  - How- website log analysis? 21
  - Problems with Web Log Analysis 21
- Web Traffic Measurement 22
  - Motivation for measurement-the whom 22
  - Content creators 23
  - Web-hosting companies 23
  - Network operators 24
  - Web/networking researchers 24
- Measurement Techniques 25
  - Server logging 25
  - Proxy logging 26
  - Client logging 26
  - Packet monitoring 27
  - Active measurement 28

Proxy/Server logs 29
Common Log Format (CLF) 29
Extended Common Log Format (ECLF) 29
Preprocessing Measurement Data 30
Drawing Inferences from Measurement Data 31
Limitations of HTTP header information 31
Ambiguous client/server identity 31
Inferring user actions 32
Detecting resource modifications 32
METHODOLOGY 33
WebTrends Report 33
USA Cycling Contacted 37
WebTrends Data Compiled 37
Data Analyzed 38

WEBTRENDS 38
Pros 38
Cons 39

FINDINGS 39
Average User's Time 40
Authenticated Users 41
IP Identification of Top 20 Non-authenticated Users 42
Most Requested Web Pages 43
User's Time Spent 44
Least Requested Web Pages 45
Top Exit Pages 46
Top Entry Pages 47
Number of Pdfs Download 48
Average Number of User Sessions per Hour 50
Average User Session by Day 51
Hits Failed 52
Origins of User's Sessions 53
Cached Hits 54
Bandwidth (kBytes Transferred) 55
Top Referring Websites 56
Top Search Engines Used 57
Web Browser Usage 58
Most Popular Types of Platforms Used 59
CONCLUSION 61
Profile 61

Recommendations 62

BIBLIOGRAPHY 63

APPENDICES 65

APPENDIX A: Screen Shots of USA Cycling's Website 65
USA Cycling's Homepage 65
Mountain Biking 66
Road Cycling 67
Cyclocross 68
Track 69

APPENDIX B: Measurement Case Studies 70
SILS server log study-1995-1996 70
Saskatchewan server log study 70
British Columbia 71

APPENDIX C: Authenticated Users' Rank 72

APPENDIX D: Un-Authenticated Users' Rank 73

APPENDIX E: USA Cycling and Select User's Server Info 74

APPENDIX F: Correspondence with USA Cycling 76

APPENDIX G: Questionnaire 77

APPENDIX H: Search Engine Statistics for August 2002 78

LIST OF TABLES

Table 1: Common Log Format 29

Table 2: Extended Common Log Format 30

Table 3: Authenticated Users Overall Usage 41

Table 4: USAC's website visitors identified by IP address and relative Activity 42

Table 5: Least Requested Web Pages Ranked First and Second 46

Table 6: Number of Pdfs Download from USA Cycling Website 49

Table 7: Key Metrics of the Saskatchewan Study of Server Logs 70

Table 8: Authenticated User's Rank from August 1999 to April 2002 72

Table 9: Un-Authenticated Users of the USA Cycling Website
Identified by IP Address from August 1999 to April 2002 73

Table 10: HTTP header from 161.58.123.16 75

Table 11: IP Address and NSLOOKUP Information of USA Cycling's Users 75

LIST OF GRAPHS

Graph 1: Average User's Time Spent on USA Cycling.Org Website in Minutes 40

Graph 2: Most Requested Web Pages on the USA Cycling.Org Website from August 1999 to April 2002 43

Graph 3: User's Time Spent on a Most Popular Web Page in Seconds 45

Graph 4: Top Exit Pages Identified based on User's Session 47

Graph 5: Top Entry Pages in Terms of User Sessions for the USA Cycling.Org Website 48

Graph 6: Average Number of User Sessions per Hour for the USA Cycling.Org Website from August 1999 through April 2002 51

Graph 7: Average User Session by Day for the USA Cycling Website from August 1999 to April 2002 52

Graph 8: Hits Failed on USA Cycling's Website 52

Graph 9: Origins of User's Sessions for the USA Cycling Website 54

Graph 10: Cached Hits as Percent 55

Graph 11: Bandwidth (kBytes Transferred) from August 1999 through April 2002 for the USA Cycling.Org Website 56

Graph 12: Top Referring Websites to the USA Cycling.Org Website based on User Sessions from August 1999 to April 2002 57

Graph 13: Top Search Engines Used to Access USA Cycling's Website from August 1999 to April 2002 58

Graph 14: Web Browser Usage for USA Cycling.Org as a Percent from August 1999 to April 2002 59

Graph 15: Types of Platforms Used to Access USA Cycling.Org from August 1999 through April 2002 in Percents 60

INTRODUCTION

Sports information is an integral part of the American way of life. There are numerous websites that are dedicated to sports information and not only mainstream sports. In an effort to promote the sport of cycling both recreationally and competitively, USA Cycling (USAC) the governing body of competitive cycling, maintains a website that provides related information concerning competitive cycling events and news plus links to recreational websites.

In the recent past, flyers and magazines mailed to USAC members were the main method of disseminating cycling competition and news information. But as the Internet developed, so has the need to make cycling information available over the web. Cycling information has "come of age" as the Internet has developed over the past ten years. By distributing competitive cycling information over the web, it can be accessed by members and non-USAC members. Essentially, the information is more quickly distributed and reaches a larger target audience than was possible before using flyers and magazines. Suppose a non-USAC member searches the web for road cycling, they will likely see USA Cycling's website as one of their hits and possibly click on it.

The Internet has become a significant source for information about cycling. There are how-to guides, product information and reviews, online stores, and discussion forums about anything dealing with cycling. As with most sports, there is a competition factor that exists within cycling. In the United States, the governing body of competitive cycling is USA Cycling. They are headquartered in Colorado Springs, CO and are a

nonprofit organization whose goal is to promote local, regional, national and international competitive cycling. USA Cycling has a website that provides information about competitive cycling on all levels. A discussion of cycling is not the intent of this paper; it is the intent of this paper to look at USA Cycling's web statistics and compare web usage statistics and generate a general user profile and draw conclusions about the state of their website and how information is disseminated.

USA Cycling's website is information gateway for competitive cycling. Every USA Cycling event is permitted and information about those races can be found on USA Cycling's website. There are races that are not permitted and do not appear on USA Cycling's website. Any statistics or information pertaining to those races will not be covered in this paper. Before the advent of USA Cycling's website, races were promoted by using flyers and magazines. This type of advertising is not as far reaching as the web because it potentially did not reach non-USAC members. Now, when race promoters are granted a permit, their race will be entered into USA Cycling's event database, which is searchable by event type or date. Thus race promoters can reach a larger audience than before and hopefully attract more racers to an event.

Cycling is not a "big-time" sport. It typically only gets airtime on the television during the Olympics, The Tour de France, the X-Games and copycat X-Games on various cable channels. When cycling does receive airtime in the United States, it is typically events that include freestyle maneuvers and big-air jumping. Cycling is much more than that. There are five main racing disciplines that USA Cycling governs and provides information about on their website: road cycling, mountain biking, cyclocross, track and BMX. Within each of these disciplines, there are different events. For

example, mountain bike events might be made up of the following events: cross country, downhill, mountain-cross and/or dual slalom. A mountain biking race does not have to include all these events, but it might. In addition to race information, USA Cycling's website serves as an outlet for racing news and results throughout the US and the world. There are links to race information around the world. During big cycling events like the Tour de France, there are daily updates and results and links to other websites for race news and information. USA Cycling's website is the main source for US racing news and it focuses and spotlights US racers in international events. There are several cycling websites that provide daily updates and reports about all types of cycling, but they typically are centered around the European cycling community.

Background: USA Cycling (USAC) and Its Purpose

USA Cycling's website (www.usacycling.org/about) states "[it] was organized in 1920 as the Amateur Bicycle League of America and was incorporated in New York in 1921. In 1975, the name was changed to the United States Cycling Federation. In 1995, a new organization, USA Cycling, was incorporated in Colorado, and on July 1, 1995, the two corporations merged, with USA Cycling being the umbrella corporation. Since the creation of the modern bicycle, the United States has been a dominant force in cycling competition. Before World War II, cycling was second only to baseball as a national sporting pastime. Following a period of decline in the 1950s and '60s, cycling regained its popularity and today is the fastest-growing amateur participation and spectator sport. Studies show that more than 99 million Americans are active in cycling. Research further indicates that these people spend more than billion annually to participate in the sport of cycling, and that these expenditures will likely double over the next several years."

USA Cycling's Information Services

"USA Cycling is the official cycling organization recognized by the United States Olympic Committee (USOC) and is responsible for identifying, training and selecting cyclists to represent the United States in international competitions. USA Cycling, doing business as the USCF, NORBA and USPRO, controls nearly two-dozen major events each year and issues permits for up to 3,000 more." On USA Cycling's website there is information about the five major racing discipline they manage. The five disciplines include road, mountain bike, cyclocross, bmx, and track racing. "The major activities of USA Cycling ensure the ongoing development and safe participation in the sport of cycling."

USA Cycling Website

USA Cycling's website (see Appendix A for screen shots) is the gateway to cycling information in the United States. The homepage (www.usacycling.org) has news and information about upcoming or recent cycling events. Typically, there are three to five stories with links to more information. In addition to news, the homepage offers five main links on the left side of the web page in a vertical orientation. Each discipline has click-able symbol that links to the main web pages for road, mountain biking, BMX, track and cyclocross. From each discipline's web page, a user can connect to another discipline by clicking on the appropriate symbol. The right hand side of the homepage has links to information about "Latest Updates" which includes information about membership, rulebooks, and member benefits; "USAC Programs" which include links for coaching, mechanics, and colligate racers; and "Miscellaneous Links" that allow users

to access a variety of links that contain information about the Under 23 team, career information and athlete bios.

Web Logs at USA Cycling

No raw web-logs were analyzed for this paper. WebTrends data from August 1999 to April 2002 was reviewed and selected data was compiled into spreadsheet format and analyzed.

LITERATURE REVIEW

Yu and Apps (2000) define validity as the extent to which the measurement measures what it intends to measure. They insist for log files to be valid, the following conditions must exist:

There has to be a defined range of data for the study-some type of time frame, the selection and implementation of the appropriate logging program must be used, a suitable analysis package must be employed, there must be an integration of the data into a formal analysis package, the variables in the analysis package should be defined, assessments of the validity of the measurements should be taken into account, and the use descriptive and inferential stats must be used to describe the data.

In addition, from the perspective of user studies, a log file is essentially an instrument for data collection just as a questionnaire or interview and the value of the data is directly related to the design of the experiment.

Questions for the paper and research

What are the main reasons a web user visits USA Cycling? What layers of information are they looking for and how easy is it for them to find this information? The study sports information and website analysis is much like the study of why people visit websites in general. The difference being that (most) USA Cycling website users are seeking information pertaining to competitive cycling. Therefore, USA Cycling must maintain a functional website that supplies the sought after information. There are two main questions with subset questions that seem to arise when websites and sports intersect.

1. What do USA Cycling users want to know?
    a. How do they go about gathering the information?
    b. What information is important to them? Events? Forms? News about racers?
2. How does the USA Cycling's website function as an information provider?
    a. Is the website usable?
    b. What do the server logs indicate about visitors to the website?
    c. Do cycling enthusists seek information in other formats? Magazines? Newsletters? Television?

It is the intent of this literature review to introduce the reader to the basics of web log analysis. It will be the goal of the author to determine the following questions about the USA Cycling's website by analyzing historical WebTrends data for the website. By compiling the historical data, it is the author's intent to generate an USA Cycling user profile. The following information will be compiled to create a user profile and answer the subsequent questions:

- Average user's time spent on USA Cycling's website,
- Most and least requested web pages on the USA Cycling,
- User's time spent on the most popular pages,
- Top exit pages identified based on user's session,
- Top entry pages in terms of user sessions for the USA Cycling's website,

- Average number of user sessions per hour for the USA Cycling,
- Hits failed on USA Cycling's website,
- Origins of user's sessions for the USA Cycling website,
- Cached hits as percent,
- Bandwidth (kBytes transferred),
- Top referring websites to the USA Cycling.Org website,
- Top search engines used to access USA Cycling's website,
- Web browser usage for USA Cycling, and
- Types of platforms used to access USA Cycling's website.

Measuring and Characterizing web traffic

The rise of the electronic environment over the past 30 years has given rise to new user study techniques that include web log files analysis. Yu and Apps (2000) indicate, "log file data represent a major thrust of evidence in an area where hard data has been in short supply."

Web proxies and servers create logs as a routine part of part of performing HTTP transactions. Measurements can also be collected passively by monitoring links in the network or actively generating requests to targeted servers. Since early days of the web, researchers and protocol designers have analyzed measurement data to characterize web traffic and evaluate techniques for improving web performance. Web performance depends on how user access patterns interact with the underlying protocols and software components. Measurement and analysis of Web traffic have also played a crucial role in the development of benchmarks for comparing different proxy and server implementations. The first chapter of *Web Protocols and Practice: HTTP/1.1, Networking Protocols, Caching, and Traffic Measurement* identifies 3 main steps in monitoring web traffic: monitoring web transfer, generating the measurement records, and preprocessing the data in preparation for analysis. Then, it identifies four areas of major study when considering web traffic analysis-client, proxy and serving logging plus

packet monitoring and active measurement. Before we examine the meat of web traffic measurements, a brief overview of the web's origins is in order.

Web Background

Krishnamurthy and Rexford (2001) indicate that Tim Berners-Lee first proposed the web in 1989, his vision was that the web would be a universe of information accessible via networked computers. The web has become an intuitive graphical interface that allows users to look through a compilation of web pages by clicking on links free from format or location worries. The web allows users to search for information, send/receive electronic mail (email) and conduct business transactions. Essentially, the web is a networked application that links users via computers around the world.

Origins

Krishnamurthy and Rexford (2001) state that Vannevar Bush's proposal in 1945 for *Memex* marks the beginning point for the web. Bush suggested that a mechanical retrieval device, *memex*, could store information (books, records, and communications) and this information could be retrieved quickly and efficiently. The *memex* essentially enlarged one's memory. Bush was worried about the speed at which information was being produced (through publications) would outpace the speed at which human could access the information.

The Next Step: ARPANET, TCP/IP, 1980s and Beyond

Bush's article set the stage for large scale indexing of text and multimedia resources. In 1965, Ted Nelson coined the term "Hypertext." He described hypertext as nonsequential writing that presents information as a "collection of linked nodes." In the

mid 1960s, ARPANET was conceived as a way for researchers to share information with each other via supercomputer connectivity. The United States Department of Defense was interested and by the late 1960s there were efforts to standardize the information network communication protocols. During the 1970s, the scientific community used ARPANET to exchange information, connect to remote machines, email, and copying files between machines. By the end of the 1970s, many universities and research organizations around the world could communicate through ARPANET. TCP/IP protocols were finalized in 1980. Berners-Lee was influenced by hypertext and wanted to link information on the CERN, the European Laboratory for Particle Physics near Geneva. In 1989, Berner-Lee's proposal was called "Enquiry Within" and was written a decade earlier. Several other systems that searched and accessed information over the Internet already existed such as FTP, Gopher, Archie, WAIS (Wide Area Information Servers). FTP allows users to retrieve and store files on servers and is password protected. In the 1970s and 1980s, it was the main means of distributing software and large documents over the Internet. By 1990, FTP was responsible for over half the Internet traffic.

State of the Web: Web Trends and Traffic

Krishnamurthy and Rexford (2001) state that in 1991 the first browser and server appeared and by the start of 1993 there were fifty servers. In the December of 1993, Marc Andreesen and Eric Bina wrote the Mosaic browser and it was introduced in the spring of 1993. By the end of 1993, there were 500 servers and the web accounted for one percent (1%) of the traffic on the Internet. The explosive growth of the web was due to the graphical interface of the Mosaic browser. In the late 1990s, the web was

responsible for 75% of the traffic on the Internet. Initially, the web was to provide public access to information, but this quickly changed. Many companies and entrepreneurs used the web to directly market customers and some companies use the web as an internal information network (salary/benefits/policies) for employees to access.

Web usage in the United States of America has increased 811% from 18 million in 1995 to 164 million in 2002 according to Nielsen Net Ratings. Many factors have contributed to the increase, including, but not limited to more personal computers in the home and work place and a shift from paper to electronic communications and commerce. In some instances, the web has replaced traditional methods of information gathering. According to a survey conducted by ESPN.com (an all sports media network), men ages 18-34 spend more time surfing the net (12.2 hours) than watching television (12.1 hours) and the number one reason is to seek sports information.

The rise of the web and specific information seeking behaviors triggered researchers to begin investigating web/information-seeking behaviors of people. Currently, there are two main methods in which to study web usage, usability test and web server logs. When used together, they can provide insight to the "hows" and "whys" of web usage. Independently, they function well, but do not provide the whole picture. Meyer (2000) has suggested that server logs provide much information in the way of data, but little about analysis.

Semantic Components of the Web-URI, HTML, and HTTP

There are three main semantic components of the web: Uniform Resource Identifiers (URI), Hypertext Markup Language (HTML), and Hypertext Transfer Protocol (HTTP). Berners-Lee indicates that a Uniformed Resource Identifier (URI)

identifies a “web resource”. A URI is a universal naming mechanism for web resources. The URI points to a “black box” were the request methods are recognized and a response is produced. A URI is a formatted string like http://www.unc.edu/~stonw/raceteam.htm. An URI typically consists of the following three parts: HTT, the protocol for communicating with the server; www.unc.edu, the name of the server and ~stonw/raceteam.htm, the resource at the server. Hypertext Markup Language (HTML) provides a standard representation for documents in ASCII format and HTML was derived from the generalized Standard Generalized Mark-up Language (SGML). HTML applications allow authors to format text, reference images, and embed hypertext links with a document. Hypertext Transfer Protocol (HTTP) is a “standard, well-defined” communication method for web components denote Krishnamurthy and Rexford (2001). They contend that HTTP is the most common way information is transferred on the web and defines the format and meaning of the messages that are exchanged between web components. HTTP defines the syntax of the code and how each line should be understood. HTTP is a *request-response* protocol; the client sends a request message and the server replies with a response message.

Terms and Concepts

There are several terms that are standard within the world of the web. This section will be used to designate a general definition for each of the following terms that appear in *Web Protocols and Practice: HTTP/1.1, Networking Protocols, Caching, and Traffic Measurement:*

- Content: the exchange of HTTP messages provides web users with access to resources,
- Software: web transfers include the exchanges between clients, intermediaries, and servers,

- Underlying Network: the Internet provides the backbone for communications between web components,
- Standardization: helps ensure that web components can communicate with each other,
- Web traffic and performance: software and network efficiency impacts users perception of web transfers; analysis yields information that help improve efficiency and
- Web applications: web caching and multimedia are two factors that affect web performance and user experience (2001).

Web Traffic and Server Log Analysis

Krishnamurthy and Rexford (2001) maintain that there are three main steps in web traffic measurement-monitoring web traffic from a location, generating web traffic measurement in some format, and processing the records for analysis. Zhang indicates the following issues that need to be answered before undertaking a web server analysis: What established the need for the analysis? What are the objectives and information requirements? What are the evaluative data sources and design sampling procedures?

And finally, what is the analysis and how to apply the results? Yeadon (2001), a web coordinator in Great Britain, states that web tracking services and software enable the collection of information about a website and create virtual "footprints" of visitors to the website. It is through the virtual "footprint" that web log analysis tries to understand characteristics about users. Hochheiser and Shneiderman (2001) suggest that understanding user's visit patterns is "essential for effective design" of websites that include on-line communities, government services, digital libraries, and electronic commerce. Yu and Apps (2000) essentially agree with Zhang and state that to understand user's visit patterns from server log file studies, the study must go through five stages; *first, planning data collection; second, collecting data; third, processing log files; fourth, determining the validity of measurement; and fifth, deriving descriptive and*

inferential statistics. When these have been completed, a more in-depth picture of users can be established and design or redesign of a particular website can take place.

How- website log analysis?

Yeadon (2001) points out that website statistics provide information about usage over time, popularity of certain pages, guide design and uncover navigation problems. These are the most pertinent to web design teams and webmasters wanting to maximize a visitors experience on a website. Yeadon (2001) suggests that following information can be collected from web pages and server logs: *Basic*-page based counters that display the number of hits on a page; these provide the least information; *Intermediate*-third party services that gather and report on site usage; the collection agencies logo has to be on page and may distract from visitors experience; and *Advanced*-computer software packages that collect and analyze server produced log files-Web Trends or Analog.

Problems with Web Log Analysis

Yu and Apps (2000) insist the following problems are inherent with log files and can cause confusion when trying to interpret user behavior:

1. Web caching,
2. Application of ambiguous usage measurements,
3. Log files can get large and unwieldy,
4. Lack the flexibility and adjustability of human eyes during observations-not knowing where the user looks,
5. Duplicates information from users repeated log –ins,
6. Tells what user does, but does not tell why-other contextual information must be collected, and
7. Several other variables including-frequency of use, breadth of use, time of use, use of functions and features can limit the yield of web server log analysis.

Nicholas (2000) and Meyer (2000) agree with Yu and Apps and propose that the idea that server logs are data and not an analysis of data. They claim that web server logs

yield a lot data, but not offer any type of full depth analysis of the visitor. They recommend combining server logs with usability testing. Web caching can skew results significantly, Nicholas (2000) and Meyer (2000) suggest that that 32% to 55% of web pages are cached by the browser and are not recorded in the server web logs. Zawitz (1998) notes, "server logs and their measures were designed originally to measure and managed server traffic and not to analyze the use/effectiveness of websites." The key word is *analyze* and its interpretation is up for debate. Another noted problem is the time of the day when web server logs are analyzed. Certain times of the day would yield much higher or lower web usage. In addition, time of the year can have significant impacts of log server analysis. A university library's web page would probably receive more use during the academic year versus use patterns in the summer months. To adjust for this, most studies have used a one to two year time period to analyze web server logs. Randomizing the time interval sample over a one to two year time interval seems appropriate and looks to give suitable results. In addition, they add that "hits" do not necessarily reflect user's interest. In fact, to reach a desired page, one might have to navigate through many "pre-pages" before arriving at the desired web page.

Web Traffic Measurement

Motivation for measurement-the whom

Nicholas (2000) and company state web server log analysis helps demonstrate the huge investments (of time and money) are worthwhile, help develop and redevelop site, assist marketing departments in their planning, and satisfy sponsors/investors and attract new ones. They also state that transaction logs describe what searches and what time searches were entered. Yu and Apps (2000) interject that log files record user behavior at

the same time as the user interacts with the system. Yeadon (2001) adds the following reasons why website statistics can be useful:

1. Give you indications of usage over time,
2. Help with server hardware upgrade choices (if needed),
3. Can demonstrate the need to keep jobs in shrinking budget times,
4. Indicate what pages are hot and those that are used less,
    a. Hint for hot/quick links,
    b. What links to put on home page,
5. By knowing browser type, the webmaster can make site the most accessible and appealing to the audience,
6. Interpret visitors navigation methods (positive and negative)-what seems to be a "logical" route to a particular piece of information-might be an arduous slog for the visitor, and
7. Can indicate what times would be good for server maintenance when traffic is low.

Content creators

Krishnamurthy and Rexford (2001) contend that web content creators can use web traffic measurements to know how long visitors visit a website and how many pages visitors download. In addition, they suggest that if visitors leave after visiting one or two pages the website might need reorganization or more interesting material. High latency, "the time between the initiation of an action and the first indication of a response" or low throughput might cause the web designer to redesign the website for telephone modem visitors.

Web-hosting companies

In *Web Protocols and Practice: HTTP/1.1, Networking Protocols, Caching, and Traffic Measurement,* the authors assert that web-hosting companies can compile web traffic statistics-bytes transferred- for websites hosted for billing purposes or for deciding how to allocate server resources for each site. For example, a site that receives a large amount of hits during the day (a commercial site) could share server space with a site that receives a lot of hits at night (an entertainment site). Web traffic stats can be useful for

comparing web server software and benchmarking server performance. Measuring traffic could inform web-hosting companies whether or not surrogate servers could handle request.

Network operators

Krishnamurthy and Rexford (2001) declare that companies with local area networks might benefit by installing a caching proxy and measuring what percentage of requests can be handle by the shared cache. An Internet Service Provider (ISP) could do much the same by monitoring web traffic and estimating the amount of bandwidth that could be saved by using cache proxy in a local network. "Measurement of web traffic can help the network provider identify the most popular websites among its users and to track the latency in transferring content to these sites." Popular web sites that have poor connections might necessitate the allocation of more resources for that site. Since the web is responsible for most traffic on the Internet, web traffic measurements are useful in the testing of network equipment like routers and assessing the load on Domain Name System (DNS) servers.

Web/networking researchers

The research community for evaluating the performance of web protocols has used web traffic measurements and software components state Krishnamurthy and Rexford (2001). In addition, web traffic characterization has been an active research area since the early days of the web. Web traffic has been a valuable research tool in the development of HTTP. Analysis of measurement has helped in the decision to make persistent connections the default behavior of HTTP/1.1 servers. Cache replacement,

cache validation, and prefetching are web-caching techniques that were developed by measuring and studying web traffic patterns.

Measurement Techniques

Krishnamurthy and Rexford (2001) declare web browsers, proxies, and servers can generate logs as part of handling requests. Also, hints of web traffic can be monitored passively through link monitoring and router information.

Server logging

Krishnamurthy and Rexford (2001) affirm that web servers typically generate logs as part of client processing and each log relates to HTTP requests handles by the server. Typical information yielded includes information about requesting client, the time requested, the request and the response message. Server logs have given site administrators a vehicle to examine access patterns of clients to a certain set of resources. Some problems with server logging include lack of detailed information, meaning recording the header of each request would impose a significant overhead. Most logs record the request method, Request-URI, and response code. In addition, time is not an exact measurement, but rather when the request was received and when the server started or finished the requests.

Each entry in a server log includes information about the client responsible for the request like clients' IP address or hostname. However, associating request is a difficult proposition because of proxies, shared client machines, and dynamic IP address assignments. Proxies can generate requests on behalf of multiple users making it difficult to determine single user requests. Organizations typically have shared computing platforms with separate accounts for users-the client IP address is not a unique identifier

in this situation. IP address can change overtime for a certain machine. When users connect to the Internet via modem, ISPs assign IP address to clients based on what is available in the pool of IP addresses.

Proxy logging

Krishnamurthy and Rexford (2001) express that web proxies create logs as normal operations, cover a wide range of requests for web sites and can be more detailed than server logs if the proxy is located near the requesting client. The first proxy in the chain from the user to the origin server can distinguish between requests of different users. Distinguishing between the users can be valuable for studying access patterns. Proxy logs include requests that are satisfied by the proxy's cache and origin server would never see or record a particular request. Also, proxies can help determine the relative popularity of a site and help direct web caching policies.

There are some disadvantages of proxy logs-proxies do not see the requests satisfied by web browser caches or other proxies closer to the client. The proxy does not record requests to any particular server and this makes it difficult to determine request rates for popular sites and resources. Web proxy might also be quite "homogenous" based on the set of clients in terms of geographical location and bandwidth. In addition, commercial institutions typically do not make proxy logs public knowledge.

Client logging

Krishnamurthy and Rexford (2001) suggest client logging has the potential to provide detailed user browsing patterns. The following could be recorded by client logging: a "timestamp" could be recorded for various request/response exchanges, the browser can record user request that never turn into HTTP-including request satisfied by

the browser's cache and keyboard/mouse operations, and the browser can determine when a request has been aborted by pressing the "STOP" button-this would never be recorded by the origin server.

In contrast to server and proxy logging, there is no standard for browser log formats. Popular browsers do not generate logs by default, but need to be modified and distributed to users. The source code for popular browsers is not typically available and to understand user patterns, a large study would need to be conducted using the modified browser. Another alternative that Krishnamurthy and Rexford (2001) put forth is to run a proxy server on the client's machine and configuring the browser to make request to the proxy. A typical proxy would know what requests were satisfied by the browser's cache, therefore the browser's caching capacity would have to be disabled. By forcing the browser to generate HTTP request might negatively affect performance, which might affect the user's attitude towards browsing.

Packet monitoring

In Krishnamurthy and Rexford's book, they indicate that logs collected at the application level have no or little information about network activity. They suggest that packet monitoring can produce "detailed traces of web activity at the HTTP, TCP and IP levels." Packet monitoring does not affect the performance of the web, therefore the users do not experience any "slowdowns" and it can provide an exact "timestamp" on the request/response timeline. Packet monitoring can help analyze aborted HTTP transfers that are difficult to understand by using web logs.

Packet monitoring does not portray request that were satisfied by proxy servers or HTTP messages that have been encrypted in Secure Socket Layer (SSL). There are

hardware considerations also. The packet monitoring system must be able to capture the data, process it, and store it why link speed increases. Processor and memory limitations plus disk speed make it challenging to monitor high-bandwidth links. Packet monitoring is much more costly than client, proxy or server logging.

Active measurement

Krishnamurthy and Rexford (2001) assert that using client, proxy, server logs, and packet monitoring to study user performance has two main problems. First, all the HTTP measurement methods are taken at a single location which makes it difficult to determine the user's experience and to breakdown components of a delay. Second, these measurement techniques monitor transfers "in the wild" and there is no control over when these request occur. An alternative method of collecting measurement data is by employing an active method. This is when a user sends a request and that information about the response is recorded such as a timestamp and HTTP headers. When one conducts an active measurement, there are three key issues to contend with. First, where should the modified user agent be located? Client/server performance varies considerably depending on relative location. A Russian web user visiting a website hosted in North American using a telephone modem would not have the same experience as a Canadian using a cable modem connection visiting the same website. Second, what type of request to generate? Web sites are hosted on a multitude of hardware platforms, software, network connectivity, and popularity of the site. Third, what measurements to collect? The information collected in an experiment has a direct bearing on what performance issues can be studied.

Proxy/Server logs

Krishnamurthy and Rexford (2001) indicate that most proxies and servers generate logs as normal operating procedures. Each log entry represents a single request/response pair and includes other fields that correspond to the requesting client, the time of the request, and the HTTP request/response message. There are no standards for log format and interpretation varies also.

Common Log Format (CLF)

The most common log format is the NCSA Common Log Files (CLF) says Nicholas (2000). TABLE 1, below, shows seven fields for CLF and gives their basic meaning.

Table 1: Common Log Format

FIELD	MEANING
Remote Host	Hostname or IP address of requesting client
Remote Identity	Account associated with connection on client machine
Authenticated User	Name provided by user for identification
Time	Date/time associated with request
Request	Requested Method, Requested-URI, and protocol version
Response Code	Three digit HTTP response code
Content Length	Number of bytes associated with the response

* Table extracted from *Web Protocols and Practice: HTTP/1.1, Networking Protocols, Caching, and Traffic Measurement(2001).*

Extended Common Log Format (ECLF)

Web Protocols and Practice: HTTP/1.1, Networking Protocols, Caching, and Traffic Measurement's denotes that Extended Common Log Format (ECLF) represents additional fields that might be captured by server logs. Useful fields include the following: User agent, Referrer, Request processing time, proxy request header size, and proxy response header size. These additional fields could help fill in some gaps that the CLF missed. TABLE 2, below, shows other fields that may be included in log format.

Table 2: Extended Common Log Format

FIELD	MEANING
User Agent	Information on user agent software
Referrer	URI from which Request-URI was obtained
Request Processing time	Time spent processing the request
Request header size	Number of bytes in the request header
Request body size	Number of bytes in the request body
Remote response code	Response code from the server
Remote content length	Size of the response from the server
Remote response header size	Size of the response header sent by the server
Proxy request header size	Size of the request header sent to the server
Proxy response header size	Size of the response header sent to the client

* Table extracted from *Web Protocols and Practice: HTTP/1.1, Networking Protocols, Caching, and Traffic Measurement(2001).*

Preprocessing Measurement Data

Web Protocols and Practice: HTTP/1.1, Networking Protocols, Caching, and Traffic Measurement identifies that large volumes of data can accumulate from web logs and must be organized in some fashion before being analyzed. There are three steps that occur in the preprocessing stage.

1. Parsing measurement data to find any erroneous data. CLF server logs do not require sophisticated parsing compared to packet parsing. The parsing code can identify fields that have invalid date in the fields and generate a log that is in more manageable form. Nicholas (2000) suggests parsing is a sequence of operations that replace, add, change or delete characters in a file. They contend that the parsing function is like the "find and replace" function within word processing software.
2. Filtering measurement data to remove any unnecessary fields. Filters can delete information that is not useful in analyzing the data. Filtering might be set-up to remove records that are based on invalid fields such as timestamps that do not fit into the window of time of interest.
3. Transforming measurement data into a format that is open to analysis. The request-line field in CLF could be separated into three fields (the request method, Request-URI, and protocol version) for easier interpretation. IP address could be converted to a hostname by a DNS query.

Drawing Inferences from Measurement Data

This section discusses the techniques for drawing inferences about HTTP headers; client/server identity, user actions, and resource modifications despite the limitations of server log data.

Limitations of HTTP header information

Krishnamurthy and Rexford (2001) indicate that server logs do not capture all the header information. They point out that most logs depict the request line and response code, but not the header fields. The Request-URI can be used to deduce the *Content-Type*; files that end in .htm or .html are likely to be HTML files and files that end in .jpeg or .gif are likely to be images. A *cgi-bin* Request-URI usually corresponds to a script and *Post* requests characteristically refers to HTML forms. Other response codes such as *206 Partial Content* and *304 Not Modified* refer to a client requesting a subset of a resource and the request included valid information respectively.

Ambiguous client/server identity

An IP address logged by a website is not a unique identifier purport Krishnamurthy and Rexford (2001). A single user may submit requests on behalf of multiple users or a single client may browse the web from multiple clients IP s. In addition, multiple valid request refer to the same resource and identifying that unique resource can be difficult. Consider that www.flow.com/ and www.flow.com/index.html typically refer to the same file.

Inferring user actions

Determining when and how a particular user action event occurs is crucial in web traffic measurements conclude Krishnamurthy and Rexford (2001). Identifying user clicks is important to understanding user behavior and is classifying clicks is a difficult task. The "time field" from a server log can be a practical way to determine "sequence of requests by a user." The time between requests can indicate how long a user is visiting a page within the website. Estimating the requests from user clicks is important for studying the user's experience. As users become more accustomed to a particular web site, the layout, there may be less time between clicks or if a visitor is clicking on the links randomly-can not differentiate between the two.

Detecting resource modifications

Krishnamurthy and Rexford (2001) state that web resource modifications need to be noted also. When a web resource is created, modified or deleted, it most be noted to completed understand a user's experience of the web site. Therefore, statistics on how often a website changes may be just as important as the web logs. Typically, web traffic measures do not include modifications to websites. HTTP headers, response size, and timestamps may help infer modified web resources. One way to study modified websites/resources is to compare the *Last-Modified* headers of successive responses for the same source. By comparing the difference in time, inferences can be made about modifications. Suppose a *Last-Modified* header had a timestamp of 2 PM and another *Last-Modified* header of the same resource had a timestamp of 4PM. One could infer that the resources had been modified within the last two hours. Another header, *Content-Length,* can also provide conservative approximations of modifications to the resource.

METHODOLOGY

USA Cycling uses WebTrends to analyze its raw web logs. WebTrends was founded in 1993 and their website can be found at www.webtrends.com. WebTrends can handle multiple log formats and costs $299. Schultz (1997) reports that WebTrends is unique because it can support different types of Netscape Web and proxy servers along with NCSA, IBM, and Novell web servers. In addition, Schultz (1997) says log-file processing is where WebTrends stands out above the rest of the web-log analyzers. WebTrends users can define report characteristics before and after processing. For example, you can define WebTrends to identify IP address to domain names and store the information in a database instead of having to reread the log files. WebTrends has limits, as do other log-file analyzers. WebTrends readily admits that only way to measure unique users to a website is to require visitors to log in with a username and password (Bauer, 2000). Therefore, "unique session" data generated by WebTrends and other log-file analyzers is an estimation (Warren 2002).

WebTrends Report

USA Cycling uses WebTrends to analyze their web-logs. Each month's web-log data was imported into WebTrends and outputted into tables and graphs in HTML format so that monthly trends could be easily recognized. WebTrends can be customized to accommodate the needs of the user and USA Cycling has selected to examine the following categories in the WebTrends Reports:

> **General Statistics:** The User Profile by Regions graph identifies the general location of the visitors to your Web site. The General Statistics table includes statistics on the total activity for this server during the designated time frame.
>
> **Most Requested Pages:** This section identifies the most popular Web Site pages and how often they were accessed. The average time a user spends viewing a page

is also indicated in the table.

Least Requested Pages: This section identifies the least popular pages on your Web site, and how often they were accessed.

Top Entry Pages: This section identifies the first page viewed when a user visits this site. This is most likely your home page but, in some cases, it may also be specific URLs that users enter to access a particular page directly. The percentages refer to the total number of user sessions that started with a valid Document Type. If the session started on a document with a different type (such as a graphic or sound file), the file is not counted as an Entry Page, and the session is not counted in the total.

Top Exit Pages: This section identifies the most common pages users were on when they left your site. The percentages refer to the total number of user sessions that started with a valid Document Type. If the session started on a document with a different type (such as a graphic or sound file), the file is not counted as an Entry Page, and the session is not counted in the total.

Single Access Pages: This section identifies the pages on your Web site that visitors access and exit without viewing any other page. The percentages refer to the total number of user sessions that started with a valid Document Type. If the session started on a document with a different type (such as a graphic or sound file), the file is not counted as an Entry Page, and the session is not counted in the total

Most Downloaded Files: This section identifies the most popular file downloads of your Web site. If an error occurred during the transfer, that transfer is not counted.

Most Submitted Forms and Scripts: This section identifies the most popular forms or scripts executed by your server. WebTrends counts any line with a Post command or a Get command with a "?" as a form or script, and shows only successful hits.

Most Active Organizations: This section identifies the companies or organizations that accessed your Web site the most often.

Top Authenticated Users: This section identifies the true name and relative activity level of the users logging onto a server that requires user name and password.

Top Users: This section identifies the IP address and relative activity level of the most active visitors to your web site.

Most Active Countries: This section identifies the top locations of the users of your site by country. The country of the user is determined by the suffix of their domain name. Use this information carefully because this information is based on where the domain name of the visitor is registered, and may not always be an accurate identifier of the actual geographic location of this visitor (for example, while a vast majority of .com domain names are from the United States, there is a small minority of domain names that exist outside of the United States.)

Summary of Activity by Day: This section outlines general server activity, comparing the level of activity on weekdays and weekends. The Average Number of Users and Hits on Weekdays are the averages for each individual week day. The Average Number of Users and Hits for Weekends groups Saturday and Sunday together. Values in the table do not include erred hits.

Activity Level by Day of Week: This section shows the activity for each day of the week for the report period (i.e. if there are two Mondays in the report period, the value presented is the sum of all hits for both Mondays.) The Total Weekdays line indicates the number of hits occurring Monday through Friday of the report period. The Total Weekends line indicates the number of hits occurring Saturday and Sunday of the report period. Values in the table do not include erred hits.

Activity Level by Hour of the Day: This section shows the most and the least active hour of the day for the report period. The second table breaks down activity for the given report period to show the average activity for each individual hour of the day (if there are several days in the report period, the value presented is the sum of all hits during that period of time for all days).

Technical Statistics and Analysis: This table shows the total number of hits for the site, how many were successful, how many failed, and calculates the percentage of hits that failed. It may help you in determining the reliability of your site.

Forms Submitted By Users: This section shows the number of successful form submissions compared to the number that failed. WebTrends considers anything with Post command as a form.

Client Errors: This section identifies the type of errors which were returned by the Client accessing your server.

Page Not Found (404) Errors: This section identifies "Page Not Found" (404) errors which occurred on your server.

Server Errors: This section identifies by type the errors which occurred on your server.

Most Downloaded File Types and Sizes: This section identifies the download file types and the total kilobytes downloaded for each file type. Cached requests and erred hits are excluded from the totals.

Organization Breakdown: This section provides a breakdown by types of organizations (.com, .net, .edu, .org, .mil, and .gov.) This information can only be displayed if reverse DNS lookups have been performed.

North American States and Provinces: This section breaks down Web site activity to show which of the North American States and Provinces were the most active on your site. This information is based on where the domain name of the visitor is registered, and may not always be an accurate representation of the actual geographic location of this visitor. This information can only be displayed if reverse DNS lookups have been performed.

Most Active Cities: This section further breaks down your Web site's activity to show which cities were the most active on your site. This information is based on where the domain name of the visitor is registered, and may not always be an accurate representation of the actual geographic location of this visitor. This information can only be displayed if reverse DNS lookups have been performed.

Bandwidth: This section helps you understand the bandwidth requirements of your site by indicating the volume of activity as Kbytes Transferred.

Most Accessed Directories: This section analyzes accesses to the directories of your site. This information can be useful in determining the types of data most often requested.

Top Referring Sites: This section identifies the domain names or numeric IP addresses with links to your site. This information will only be displayed if your server is logging this information.

Top Referring URLs: This section provides the full URLs of the sites with links to your site. This information will only be displayed if your server is logging the referrer information.

Top Search Engines: The first table identifies which search engines referred visitors to your site the most often. Note that each search may contain several keywords. The second table identifies the main keywords for each search engine.

Top Search Keywords: The first table identifies keywords which led the most visitors to your site (regardless of the search engine). The second table identifies, for each keyword, which search engines led visitors to your site.

Most Used Browsers: This section identifies the most popular WWW Browsers used by visitors to your site. This information will only be displayed if your server is logging the browser/platform information.

Netscape Browsers: This section gives you a breakdown of the various versions of Netscape browsers that visitors to your site are using.

Microsoft Explorer Browsers: This section gives you a breakdown of the various versions of Microsoft Explorer browsers that visitors to your site are using.

Visiting Spiders: This section identifies all robots, spiders, crawlers and search services (i.e. Alta Vista, Lycos, and Excite) visiting your site.

Most Used Platforms: This section identifies the operating systems most used by the visitors to your Web site.

Hit: An action on the Web site, such as when a user views a page or downloads a file.

USA Cycling Contacted

In May 2002, USA Cycling was contacted via email as a potential candidate for this user profile case study. Danny Smith, USAC Intern, responded and sent the WebTrends data to be analyzed in June 2002. USAC sent thirty-three months of data covering the time period from August 1999 to April 2002. The data was sent in HTML format. The data included all the categories listed in the *WebTrends Reports* section.

WebTrends Data Compiled

The HTML data that USAC sent needed to be converted to spreadsheet format to be analyzed. The data was meticulously entered into Excel spreadsheets from July 2002 to mid-August 2002 to develop trends over a two and half year period. Once all the data was in spreadsheet format, it was used to create tables and graphs for the thirty-three month period.

Data Analyzed

From late August 2002 to late September 2002, the data was analyzed and reworked in order to create a profile of an USA Cycling website user. It was the overall intent of this study is to develop a user profile of a typical user by analyzing trends in USA Cycling's web user data. *Appendix B* is an example of the first two pages of a monthly WebTrends's output. The data was put into spreadsheets to examine it over the thirty-three month period in order to see larger trends than what the HTML WebTrends report showed. The cycling season is much like any other sport with an off-season and a competitive season. To gain a better understanding of the typical USA Cycling user, more data would need to be examined. The two and half years of data provided an overall view of the typical USAC website user. Typically, most cycling events take place between the months of March and October, except cyclocross. Cyclocross's season runs from mid-October through February. As with other sports, it was hypothesized that higher website usage would take place during the competitive season. Therefore, it was crucial to examine the logs for a longer duration than a month to view the natural ebbing of the bicycling season.

WEBTRENDS

Pros

WebTrends can generate General Statistics, Resources Accessed, Visitors and Demographics, Activity Statistics, Technical Statistics, Referrers and Keywords, and Browsers and Platform information about users indicates Bauer (2000). In addition, she states that WebTrends can generate tables and graphs for entry pages, exit pages, paths through the site, downloaded files, and forms. The reports can be filtered to exclude or

include particular data about users. WebTrends uses an algorithm in order to distinguish the number of visitors to a website.

Cons

One of the more difficult aspect of all web-log analyzers is gathering the complete picture of the website user. There are many factors that the web-log analyzer does not measure-like user intent and how the user uses the information. These are measured in more controlled type studies that this paper does not intend to venture into. Problem with web-log analyzers, like WebTrends, were covered in the *Problems with Web Log Analysis* section in the *Literature Review*.

FINDINGS

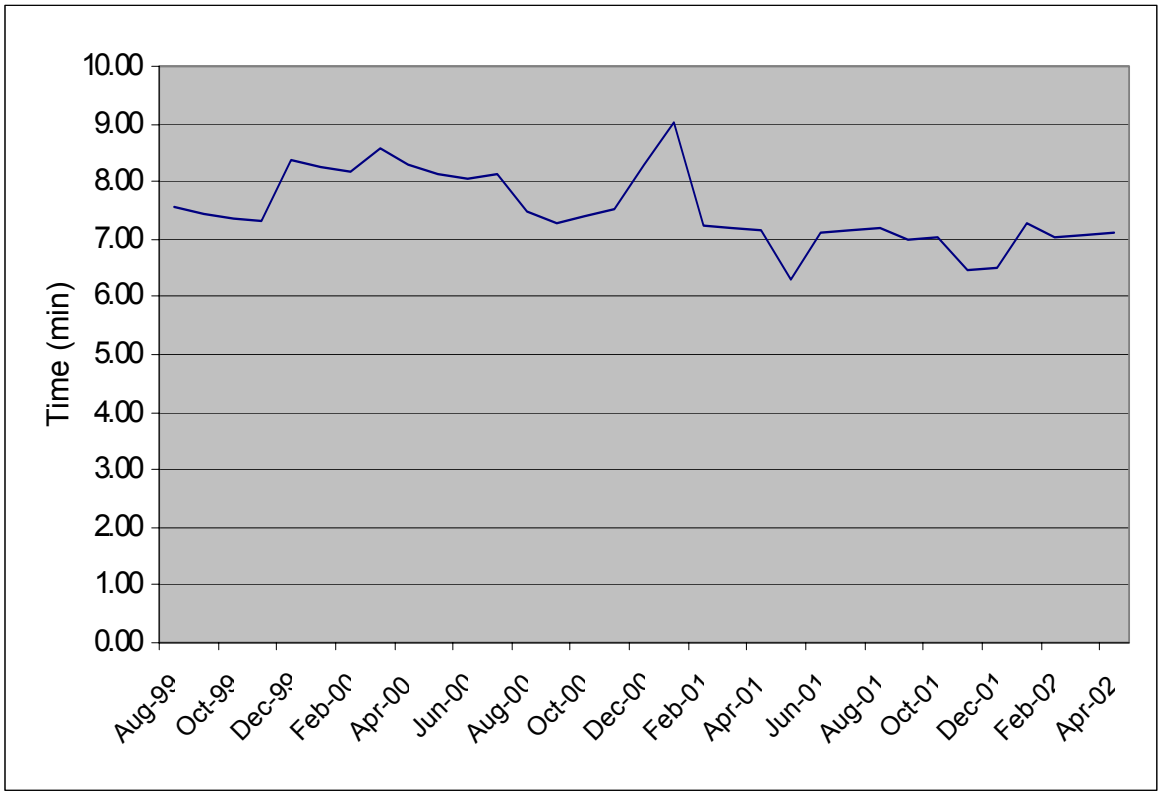
In this section, selected information will be presented and used to compile the USA Cycling User Profile. The following criteria were selected to craft the typical user profile for USA Cycling website users from August 1999 to April 2002:

Average Time Spent on USAC website, Authenticated Users, IP Identification of Top 20 non-authenticated Users, Most Requested Web Pages, User's Time Spent on the Most Popular Web Pages, Least Popular Web Pages, Top Exit Pages, Top Entry Pages, Number of Pdfs Download, Average Number of User Sessions per Hour, Average User Session by Day, Hits Failed, Origins of User's Sessions, Cached Hits as Percent, Bandwidth (kBytes Transferred), Top Referring Websites, Top Search Engines, Web Browser Usage, and Types of Platforms Used.

Average User's Time

Graph 1 shows the average amount of time that a user spent on USA Cycling's website in terms of minutes. Examination of the graph indicates several spikes in the time spent on the website. Three main spikes occurred in March 2000, February 2001, and February 2002.

Graph 1: Average User's Time Spent on USA Cycling Website in Minutes



Essentially, these spikes (more use) represent the beginning of the competitive race season in road and mountain bike disciplines. At the beginning of the season, race dates and event dates are released. These spikes represent an increased use in response to the new information posted to the USA Cycling website. Thus, the start of the race season represents a high use time for the website.

Authenticated Users

Table 3 shows the authenticated users visiting the website by identifying IP address. In addition, Table 3 displays the rank of each authenticated user and their relative activity over the thirty-three month period in terms of how much they access the USA Cycling website. For each ranking, there is an associated percent of use for that login ID.

Table 3: Authenticated Users Overall Usage Ranked and Percent Usage from August 1999 to April 2002

	Rank and Percent (%)		
Authenticates Users	**1**	**2**	**3**
USACYCL	**39**	**24**	**21**
PROMO	**39**	/	/
MHANLEY	**12**	**21**	**12**
TVINSON	**3**	**6**	**12**
JPARSONS	**3**	/	**3**
GHEAGERT	**3**	**18**	**12**
ACOOK	/	**12**	**3**
LSEIDMAN	/	**3**	**6**
TDELP	/	**3**	**6**
RCS	/	**6**	/
DEAN	/	/	**3**
JMILLER	/	/	**6**
TEMP	/	**3**	/
MWISE	/	/	**3**
No User (NA)	/	**3**	**3**

Each user's name was identified by WebTrends and ranked first, second or third for that month's use from August 1999 to April 2002. Then, a tally of how many times that login ID appeared in the first, second, and third position was compiled. Next, the percent according to each ranking was calculated by taking the total number of times in the first, second, or third position and divided by the thirty-three, the total numbers of months. For example, USACYCL appeared thirteen times in the first position and when divided by

thirty-three months, corresponds to thirty-nine percent usage for the thirty-three month time period. Therefore, USACYCL was the number one authenticated user thirty-nine percent of the time from August 1999 to April 2002. The percents are shown in bold for easier detection in the table.

IP Identification of Top 20 Non-authenticated Users

In addition, WebTrends identifies top non-authenticated users and their relative activity.

Table 4: USAC's website visitors identified by IP address and relative Activity

Rank	User's IP Address	Number of Appearance in Top 3
1	**209.107.36.74**	16
2	**209.248.75.38**	6
3	np-serial109.co.verio.net	6
4	24.64.152.223.on.wave.home.com	3
5	adsl-63-198-178-114.dsl.snfc21.pacbell.net	3
6	cx1002002-b.phnx3.az.home.com	3
7	12-237-34-208.client.attbi.com	2
8	**57.68.12.102**	2
9	arthur4.sda.t-online.de	2
10	cache1.lgca.org	2
11	cache-1.sbo.ma.webcache.rcn.net	2
12	dt061n62.maine.rr.com	2
13	dt0d1n2a.maine.rr.com	2
14	modem125115.westman.wave.ca	2
15	**12.21.187.194**	1
16	2.42.50.511	1
17	12-237-224-130.client.attbi.com	1
18	12-237-34-23.client.attbi.com	1
19	**149.149.200.200**	1
20	**192.249.47.9**	1

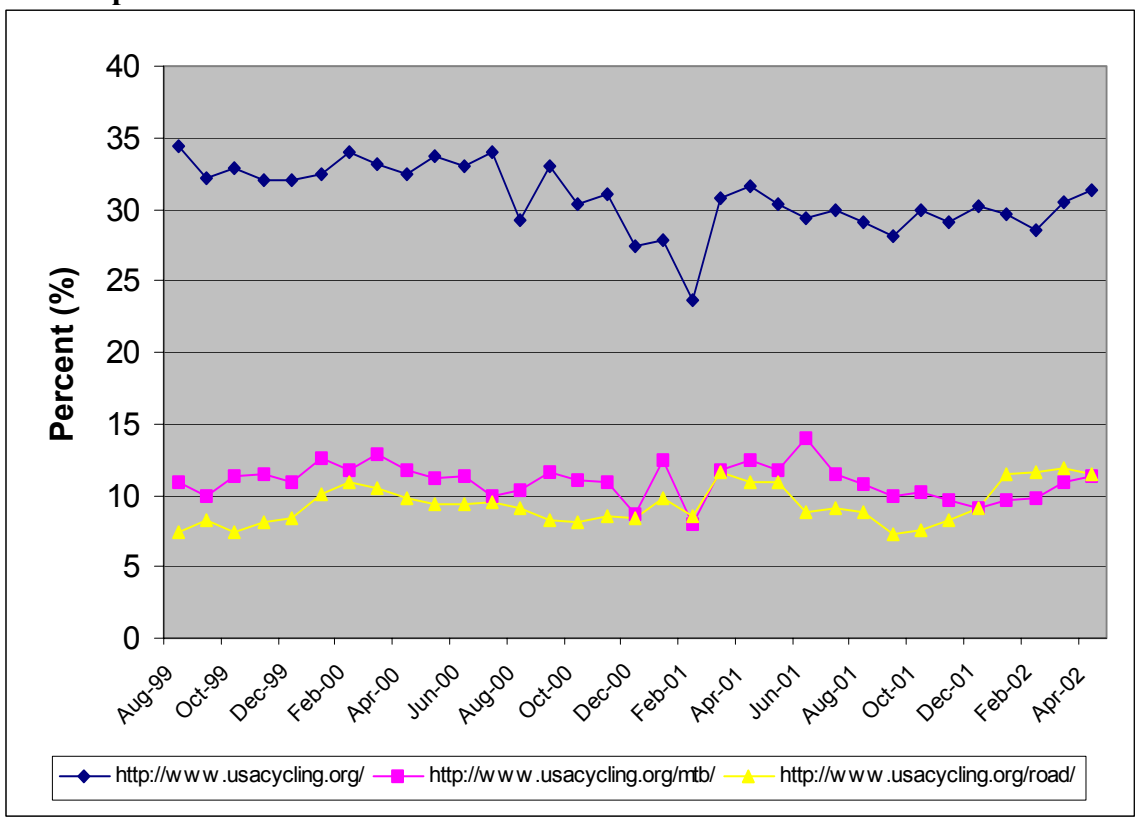
Table 4 shows the IP addresses of the top non-authenticated users for the thirty-three month period. A *NSLOOKUP* was done on the IP address that are in bold to help identify the non-authenticated user. The results are in Table 15 in **APPENDIX E**.

The above results represent the top twenty in terms of top usage for a particular month. For example, 209.107.36.74 appeared 16 times during the 33-month period. The user's IP address had to be ranked in the first, second or third position to be considered as a primary user of the USAC website for this study.

Most Requested Web Pages

Graph 2 displays the most requested web pages on the USA Cycling website in terms of percents. The homepage, www.usacycling.org, ranked consistently as the number one requested web page for the web site with www.usacycling.org/mtb and www.usacycling.org/road following in second and third position respectively.

Graph 2: Most Requested Web Pages on the USA Cycling's Website from August 1999 to April 2002 in Percent



The homepage typically ranks first because it serves as the junction for the rest of the website. From the USA Cycling's homepage, a user can link to the following main pages for different biking and racing disciplines: road, mountain, cyclocross, track and BMX.

Links for the other types of cycling have links from the USA Cycling homepage. Both the mountain biking and road biking are more popular that the other three forms of racing thus one would expect their usage to be higher based on the number of events.

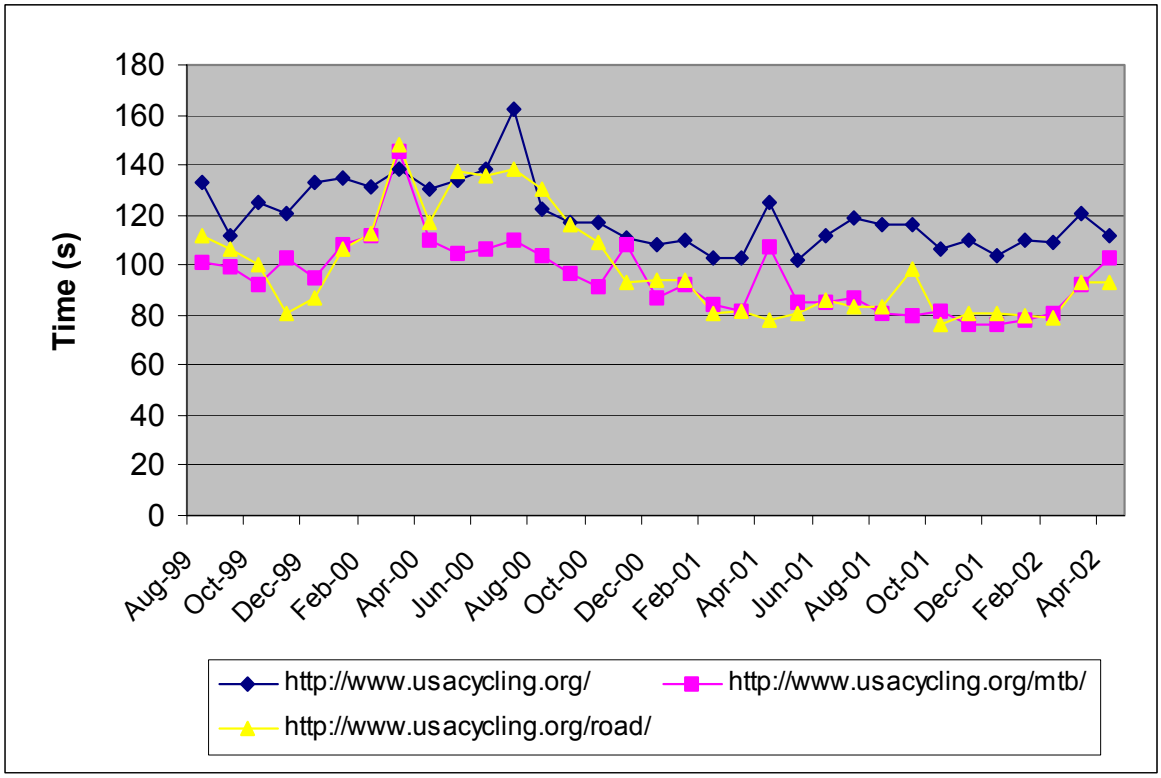
These exceptions are not shown on the graph, but deserve attention and are explained here after. The third most requested web page in December 2000 was www.usacycling.org/cx, which is the main web page for cyclocross. The Cyclocross season dictates its popularity and its popularity can be considered a seasonal phenomenon. The cyclocross race season runs from late October to mid-February. December represents the midseason and in mid-December, the US Cyclocross National Championships are held. To access and search the cyclocross event and race database, users typically navigated through the main cyclocross web page to thus the increased traffic to the cyclocross web page. In January 2001, the membership page was the third most accessed (people getting information about upcoming race season). For February 2001 and June 2001, the www.usacycling.org/mtb site was the second and third most accessed.

User's Time Spent

Graph 3 communicates the amount of time a user spent on the top three requested web page of the USA Cycling website. Several interesting items to note are the spikes in the www.usacycling.org time usage for the month of august for each year. August is a peak time for cycling and racing throughout the US and World. Since USA Cycling

serves as a gateway to news and events about competitive US cycling and August is a peak event month, the higher usage time is appropriate. In addition, there are usage spikes for April on the www.usacycling.org/mtb web page. Typically, by April, several major events and races have occurred and more are added to the database. Therefore, users interested in mountain biking news, events, and races would access the web page more at the start of the season.

Graph 3: User's Time Spent on the Most Popular Web Pages in Seconds



Least Requested Web Pages

Table 5 below represents the pages that were least accessed on USA Cycling's website. They were only hit once and seem to be "rough drafts" of pages that were eventually posted on the website. In order to free up space on the server-these pages should be removed.

Table 5: Least Requested Web Pages Ranked First and Second

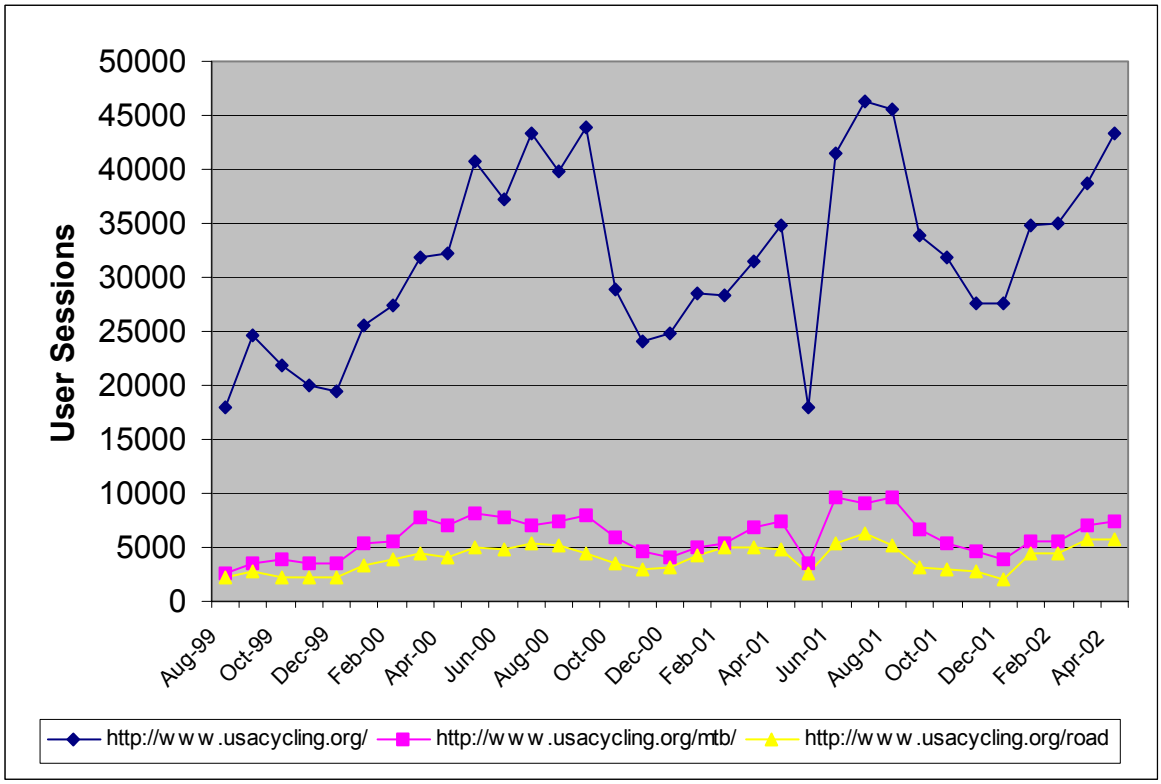
	1		2	
	Web Page Name and Number of Hits		**Web Page Name and Number of Hits**	
Aug-99				
Sep-99	cgi-local	1	track/upload/r_m_rr_tan70.htm	1
Oct-99	cx/news/data/2/data	1	brn/roundup/articles/1/articles	1
Nov-99	track/upload/r_w_crit_1012.htm	1	track/upload/r_w_spr_1718.htm	1
Dec-99	brn/track/articles/1/	1	brn/track/articles/1	1
Jan-00	rankings.html	1	brn/road/articles/2.html	1
Feb-00	cx/news/data/2/data	1	regional/ncca_member.html	1
Mar-00	road/upload/horner.html	1	membership/support/1999_coaching_license.htm	1
Apr-00	road/links/	1	road/links	1
May-00	results/files/9918R483.html	1	results/files/101099R233.html	1
Jun-00	cx/news/data/2/num	1	results/files/2000475R836.html	1
Jul-00	results/files/2000238R850.html	1	results/files/2000295R873.html	1
Aug-00	results/files/20001171R1128.html	1	results/files/2000862R1127.html	1
Sep-00	track/upload/r_w_pts_40.htm	1	results/files/2000705R851.html	1
Oct-00	results/files/2000952R1062.html	1	results/files/2000475R841.html	1
Nov-00	results/files/2000731R1355.html	1	track/upload/r_w_rr_cp34.htm	1
Dec-00	results/files/20001055R1077.html	1	results/files/20001392R918.html	1
Jan-01	road/events/niwot_entry.html	1	results/files/2000306R1314.html	1
Feb-01	track/upload/r_w_rr_4549.htm	1	track/upload/r_m_3k_4549.htm	1
Mar-01	track/upload/r_m_crit_6064.htm	1	mtb/results/results/10/	1
Apr-01	results/files/20001494R1295.html	1	mtb/results/results/12/	1
May-01	results/files/2000261R1015.html	1	results/files/2000261R1014.html	1
Jun-01	mtb/results/results/8	1	results/files/2001349R306.html	1
Jul-01	mtb/results/98results/1/results.43.html	1	brn/roundup/articles/2/	1
Aug-01	results/files/2001503R267.html	1	results/files/2001503R266.html	1
Sep-01	track/upload/r_w_1scr_1012.htm	1	track/upload/r_m_200_6064.htm	1
Oct-01	results/files/2001766R279.html	1	results/files/99510R443.html	1
Nov-01	brn/roundup/articles/11/articles	1	cx/news/data/1/num	1
Dec-01	track/upload/rwc_m_sprint.htm	1	track/upload/r_w_crit_5054.htm	1
Jan-02	track/upload/rwc_tpursuit.htm	1	track/upload/r_m_crit_3034.htm	1
Feb-02	mtb/mtb/	1	track/upload/rwc_m_madison.htm	1
Mar-02	u/ftp/pub/msql/java/tutorial.txt	1	results/files/200131R872.html	1

Top Exit Pages

Graph 4 shows the top exit pages for the USA Cycling website. The number one exit page is www.usacycling.org. This seems both intuitive and counter-intuitive. The www.usacycling.org web page offers information about a variety of topics. A screen shot of www.usacycling.org and other main web pages are provided in **APPENDIX-A**.

These three web pages are also the top used pages in terms of minutes accessed. Each page has news and information about upcoming events. Since the three pages are the top used and top exited-that seems to suggest that users are not going past these three main pages. Does this mean that users are finding the information they need or does it mean they are not finding the information and leaving the site? If users do not find what they are looking for, then typically they will return to search engine's list to explore other websites about cycling.

Graph 4: Top Exit Pages Identified based on User's Session

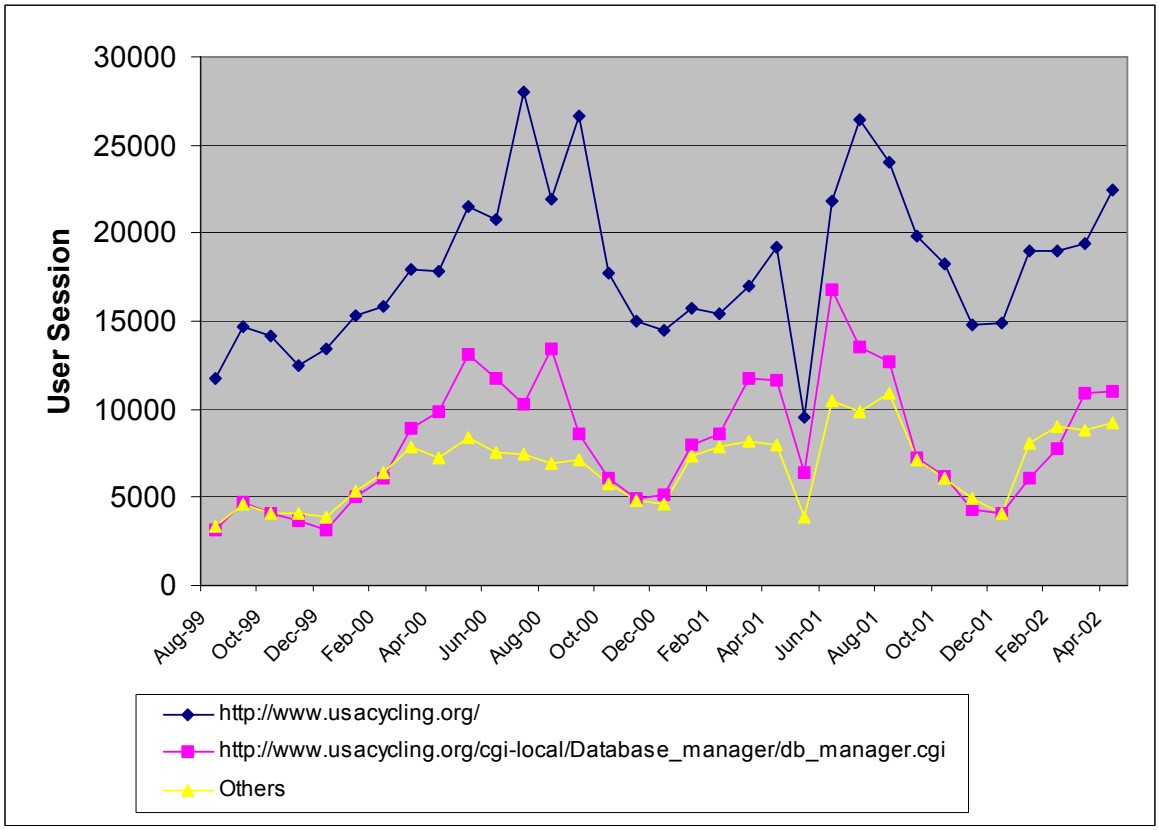


Top Entry Pages

This graph identifies the first page viewed when a user visits this site and typically represents the website's home page. USA Cycling's most viewed page is their homepage. It serves as a reference point to access the other five main pages plus the link

to the race database. The race database web page is the second entry web page on USA Cycling's website.

Graph 5: Top Entry Pages in Terms of User Sessions for the USA Cycling.Org Website



On the www.usacycling.org/cgi-local/Database_manager/db_manager.cgi web page, a user can search for all races regardless of discipline or a user can search for a specific race in a specific state for a specific month by disciplines. This database is not fixed; it grows throughout the year. Though, most races are permitted by February or March of that year.

Number of Pdfs Download

Table 6 shows the first and second ranked downloaded Pdfs from the USA Cycling Website. Several Pdfs are downloaded more than others like the license

application, membership application, and the athlete release form. To race in USA Cycling permitted events, participants must possess the license of that discipline before racing. Most races offer one-day license, but they are triplicate print form and are not downloaded from the computer.

Table 6: Number of Pdfs Download from USA Cycling Website

		% of Total		% of Total
		Downloads		Downloads
	1		2	
Aug-99	road/events/CyclingInsert.pdf	41.84	membership/rules/deadlines.PDF	5.67
Sep-99	upload/olynom.pdf	29.28	road/events/CyclingInsert.pdf	6.2
Oct-99	upload/olynom.pdf	11.24	membership/forms/clubofyear.pdf	6.82
Nov-99	upload/olynom.pdf	12.78	membership/forms/lajorsapp.pdf	9.42
Dec-99	membership/forms/2000licapp.pdf	21.26	upload/olynom.pdf	7.06
Jan-00	membership/forms/2000licapp.pdf	33.4	upload/olynom.pdf	4.28
Feb-00	membership/forms/2000licapp.pdf	37.37	upload/olynom.pdf	4.68
Mar-00	membership/forms/2000licapp.pdf	34.34	membership/forms/standard_form.pdf	8.79
Apr-00	membership/forms/2000licapp.pdf	21.74	membership/forms/Ins_2000_23.pdf	7.3
May-00	membership/forms/2000licapp.pdf	22.02	membership/forms/standard_form.pdf	8.05
Jun-00	membership/forms/2000licapp.pdf	19.21	membership/forms/Ins_2000_23.pdf	6.7
Jul-00	membership/forms/2000licapp.pdf	17.74	membership/forms/Ins_2000_23.pdf	15.42
Aug-00	membership/forms/2000licapp.pdf	20.03	membership/forms/Ins_2000_23.pdf	7.6
Sep-00	membership/forms/2000licapp.pdf	21.21	upload/olynom.pdf	5.24
Oct-00	membership/forms/2001_lic_app.pdf	17.41	membership/rules/2000genrules.pdf	4.94
Nov-00	membership/forms/2001_lic_app.pdf	18.16	ncca2/upload/NCCA_scholarship.pdf	7.53
Dec-00	membership/forms/2001_lic_app.pdf	17.75	membership/forms/2001_lic_app.pdf	5.22
Jan-01	membership/forms/2001_member_app.pdf	30.81	membership/forms/2001_intl_app.pdf	4.57
Feb-01	membership/forms/2001_member_app.pdf	30.64	membership/forms/2001_athlete_release.pdf	8.85
Mar-01	membership/forms/2001_member_app.pdf	14.71	membership/forms/2001_athlete_release.pdf	11.29
Apr-01	membership/forms/2001_member_app.pdf	14.76	membership/forms/2001_athlete_release.pdf	11.99
May-01	membership/forms/2001_member_app.pdf	17.07	membership/forms/2001_athlete_release.pdf	14.14
Jun-01	membership/forms/2001_member_app.pdf	15.27	2001_uscf_champ/masters_road.pdf	12.51
Jul-01	membership/forms/2001_membership_app.pdf	15.53	2001_uscf_champ/jr_esp_road.pdf	9.54
Aug-01	membership/forms/2001_membership_app.pdf	18.46	membership/forms/2001_athlete_release.pdf	12.8
Sep-01	membership/forms/2001_membership_app.pdf	17.02	membership/forms/2001_athlete_release.pdf	8.62
Oct-01	membership/forms/2001_membership_app.pdf	16.57	membership/forms/2001_athlete_release.pdf	9.58
Nov-01	membership/forms/2002_membership_app.pdf	27.06	membership/forms/2002_international_app.pdf	7.87
Dec-01	membership/forms/2002_membership_app.pdf	30.34	membership/forms/2002_international_app.pdf	8.1
Jan-02	membership/forms/2002_membership_app.pdf	32.54	membership/forms/2002_international_app.pdf	8.32
Feb-02	membership/forms/2002_membership_app.pdf	19.59	rulebooks/uscf_rulebook_section1.pdf	13.47
Mar-02	membership/forms/2002_membership_app.pdf	16.92	rulebooks/uscf_rulebook_section1.pdf	14.27
Apr-02	membership/forms/2002_membership_app.pdf	17.47	rulebooks/uscf_rulebook_section3-4.pdf	9.81

Between September 1999 and December 1999, www.usacycling.org/upload/olynom.pdf was the most popular downloaded file. This file is the policy and procedure for determining US cyclist to participate in the 2000 Olympics. It was approved by the USA Cycling Board on 10 May 1999 and then put on

the web site for viewing. The cyclic nature of USA Cycling comes through again. The timing with a particular file or web page depends on the season or the time of the year.

Most racers apply for a yearly license because it is cheaper than buying a one-day license for every event. Yearly licenses can be applied for at races using the same triplicate form or the PDF form can be downloaded from USA Cycling's website and mailed in. The form has check boxes on it that allow the racer to indicate what type of license they are applying for. A license can be applied for at any time of the year. The cost is $40 for each discipline.

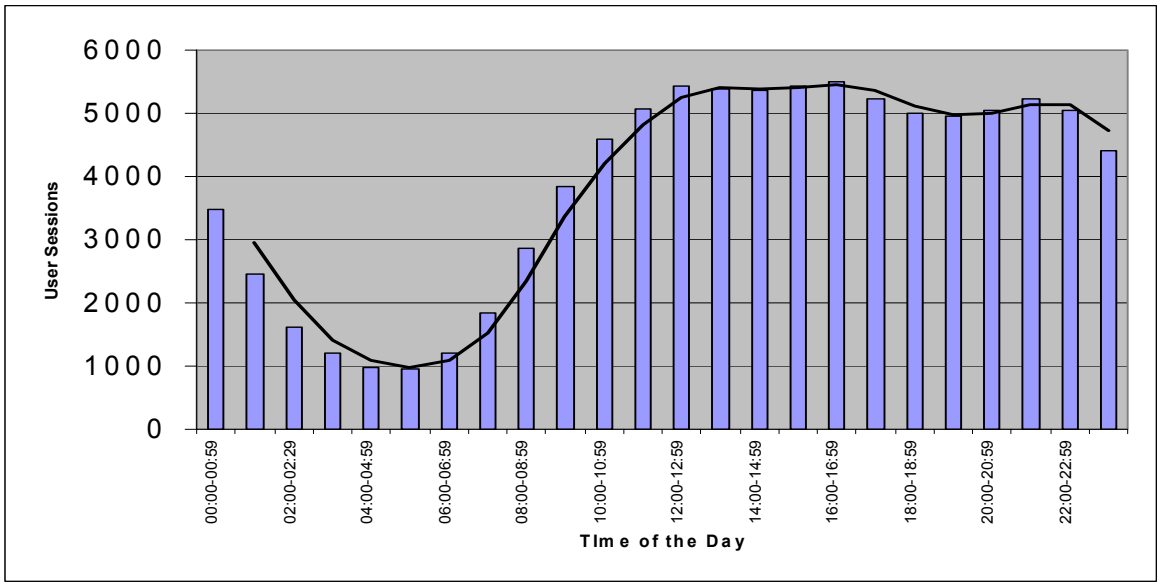
The "membership application" is the 2002 version of the 2001 "license application." The "athlete release" PDF is a form that has to be filled out by each rider before a USA Cycling permitted event. It is a liability form that releases USA Cycling and the promoters for any injuries incurred by racers. Typically, race participants will fill out the form and mail it in to the race promoter to be pre-registered. Pre-registering alleviates waiting in line and late-fees.

Average Number of User Sessions per Hour

This graph identifies average server activity as logged by hour of the day from August 1999 to April 2002. It indicates that USA Cycling website receives it highest use between the hours of 12:00 and 18:00. In addition, usage does not really taper off until the hours of 22:00 and 22:59. After midnight, usage drops until 06:00 from there it begins to climb. One key aspect from this graph is the least amount of usage. From this graph, the lowest amount of usage occurs from 04:00 to 06:00. AM. The time zone that this graph refers to is not known-USAC was unable to provide answers to a questionnaire

that was sent to them in October 2002. The questionnaire can be found in **APPENDIX G.**

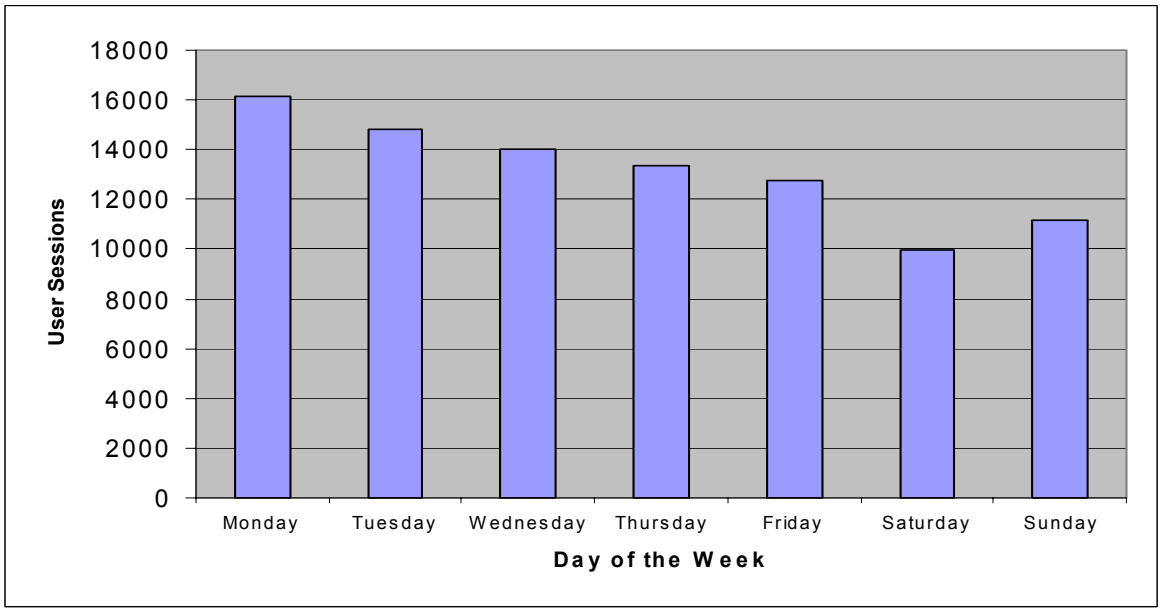
Graph 6: Average Number of User Sessions per Hour for the USA Cycling.Org Website from August 1999 through April 2002



Average User Session by Day

WebSideStory, Inc. purported that Monday is the most popular day of the week that users surf the web and Saturday and Sunday are the least popular (Anfuso 2002). Graph 7 exhibits that USA Cycling's website was most heavily accessed on Mondays from August 1999 to April 2002. Monday popularity is followed by Tuesday, Wednesday, Thursday, Friday, Sunday, and Saturday. Typically, cycling races occur over the weekend-Saturdays and Sundays contributing to the lower usage over the weekend. Accordingly, updates or links to events and races are typically updated from the past weekend of racing are usually made to the website by Monday which plays a role in Monday's usage patterns.

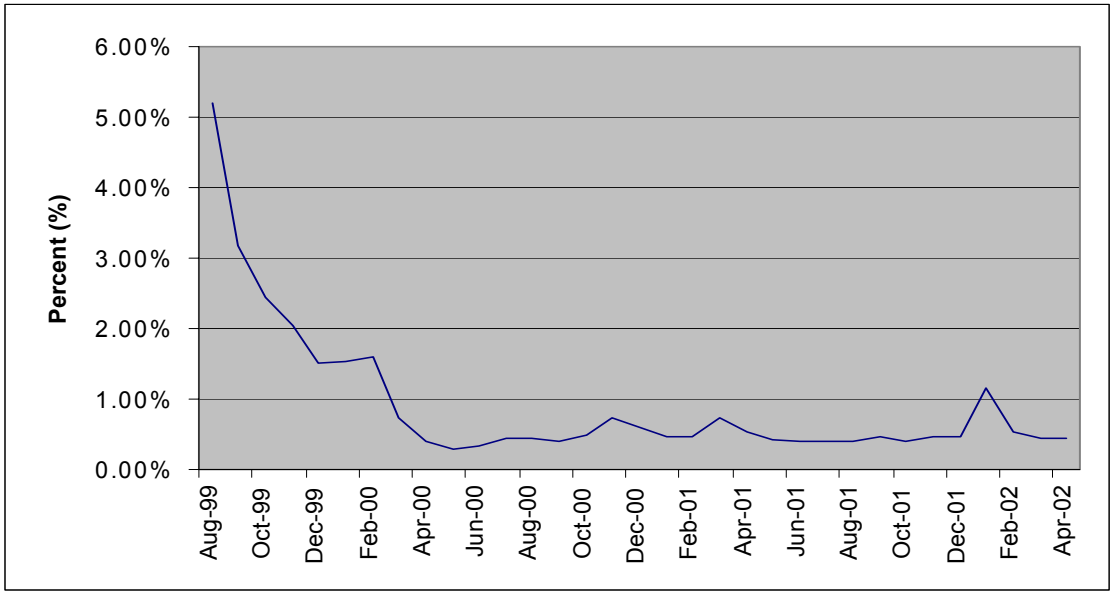
Graph 7: Average User Session by Day for the USA Cycling Website from August 1999 to April 2002



Hits Failed

Graph 7 identifies the number of hits that failed on USA Cycling's website from August 1999 to April 2002. It is interesting to note the sharp decline of hits fails from August 1999 until May 2000.

Graph 8: Hits Failed on USA Cycling's Website

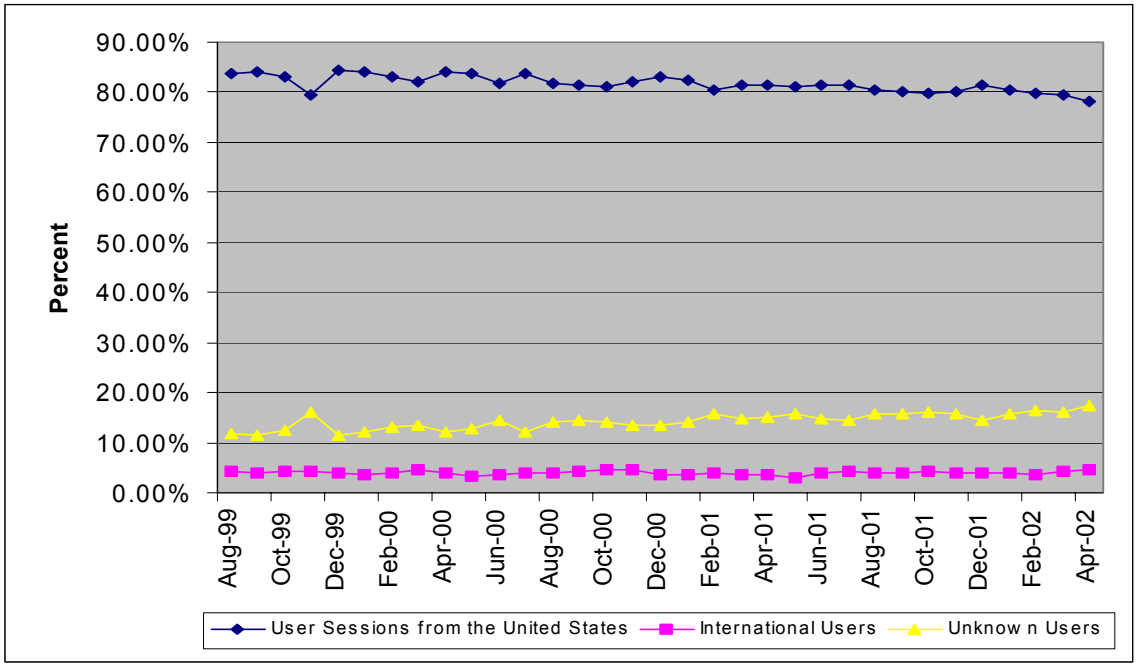


This sharp decline of hits failed raises several questions-was there a change in the server, was a new web site launched, was this the beginning of WebTrends usage, or was there some type of change with other service providers? In general, beginning in April 2000, there were relatively few problems with hits failed. For example, in April 2000, there was a failure rate of 0.44%. In April 2000 there were 1,928,980 successful hits. Therefore, 0.44% of 1,928, 980 equates to 8487 failed hits. This is relatively low failure rate for the amount of hits. Failure rates remained relatively level until a small spike eclipses the 1.0%mark in January 2002. Does this spike represent a new server, new format of web pages or some type of change in the network? What caused the January 2002 spike is unclear and at this point will not be resolved. USA Cycling was unable to answer this questions when posed in October 2002.

Origins of User's Sessions

The origins of user's sessions are shown on graph 8. The graph shows the origin of USA Cycling's users in three categories-United States, International Users, and Unknown Users. From August 1999 to April 2002, the greatest numbers of users are from the United States and range from 80% to 85%. One reason to expect such a high US user rate is because USA Cycling governs competitive cycling in the USA. International competitors are allowed to race in US events, but typically do not frequent the regional and local races; thus most participates on the regional or local level live in the US.

Graph 9: Origins of User's Sessions for the USA Cycling Website

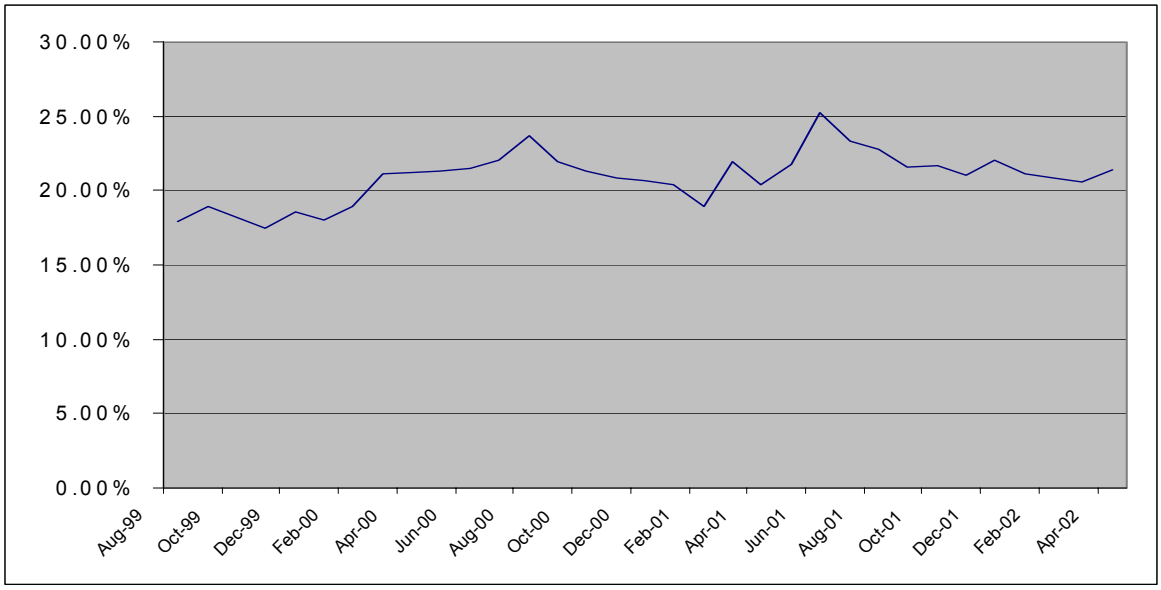


Unknown users comprise the second highest amount of USA Cycling users. Their use ranges from 10% to 17% during this same time frame. The third group is International Users, and they constitute about 5% of use for the same time frame. At national level racing events, there are a significant number of international riders participating in the racing events.

Cached Hits

The amount of cached browser hits is displayed in Graph 9. These represent the amount (percentage) of web pages that were cached by a USA Cycling user's browser. The amount of browser caching ranged from 18% to 25%. Caching can either be thought of as beneficial by speeding up performance resulting in faster web page loading times or troublesome because the cached pages occupy hard drive disk space on the user's computer, thus slowing it.

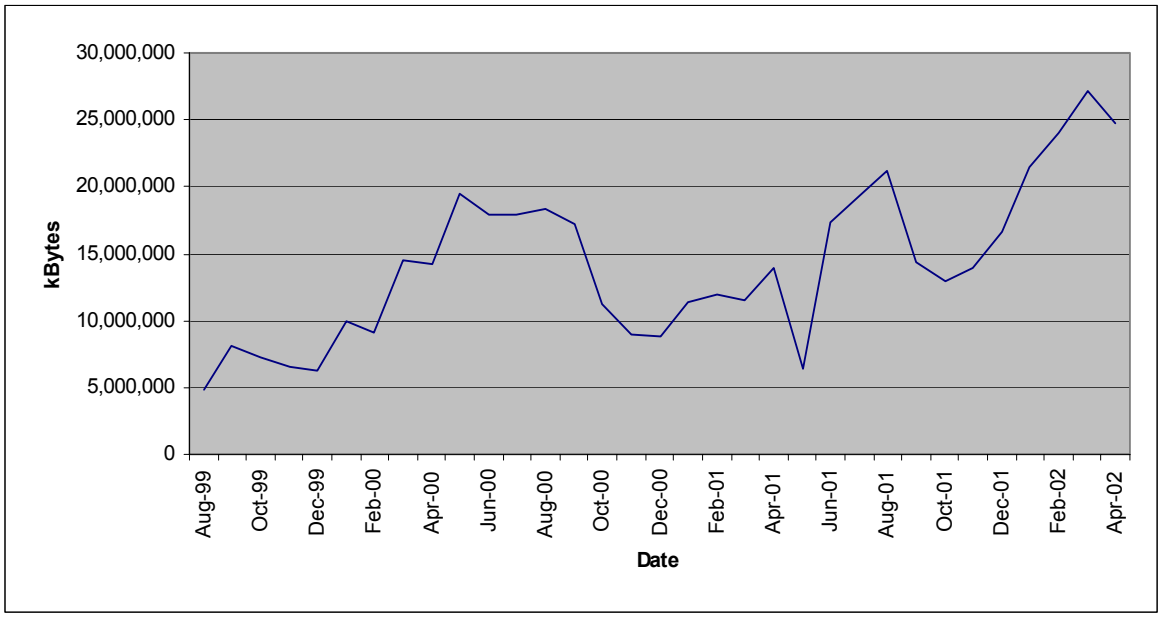
Graph 10: Cached Hits as Percent on the USA Cycling Website from August 1999 to April 2002



Bandwidth (kBytes Transferred)

Graph 10 indicates bandwidth transferred for USA Cycling's website. In general there has been an increase of bandwidth from August 1999 to April 2002. The need for bandwidth seems to fluctuate with the on and off-season for biking. Because most events and races occur from March to October there should be an increase in bandwidth transferred. In general, 2000 and 2001, from February to October, in general, represent the highest bandwidth outputs. Because the race season sloughs off by October, the resulting drop in bandwidth reflects this. Then from January 2001 to August 2001, there is a general increase with a downward anomaly occurring in May 2001. In addition, the bandwidth drops off again in October 2001 and begins rising again by January 2002.

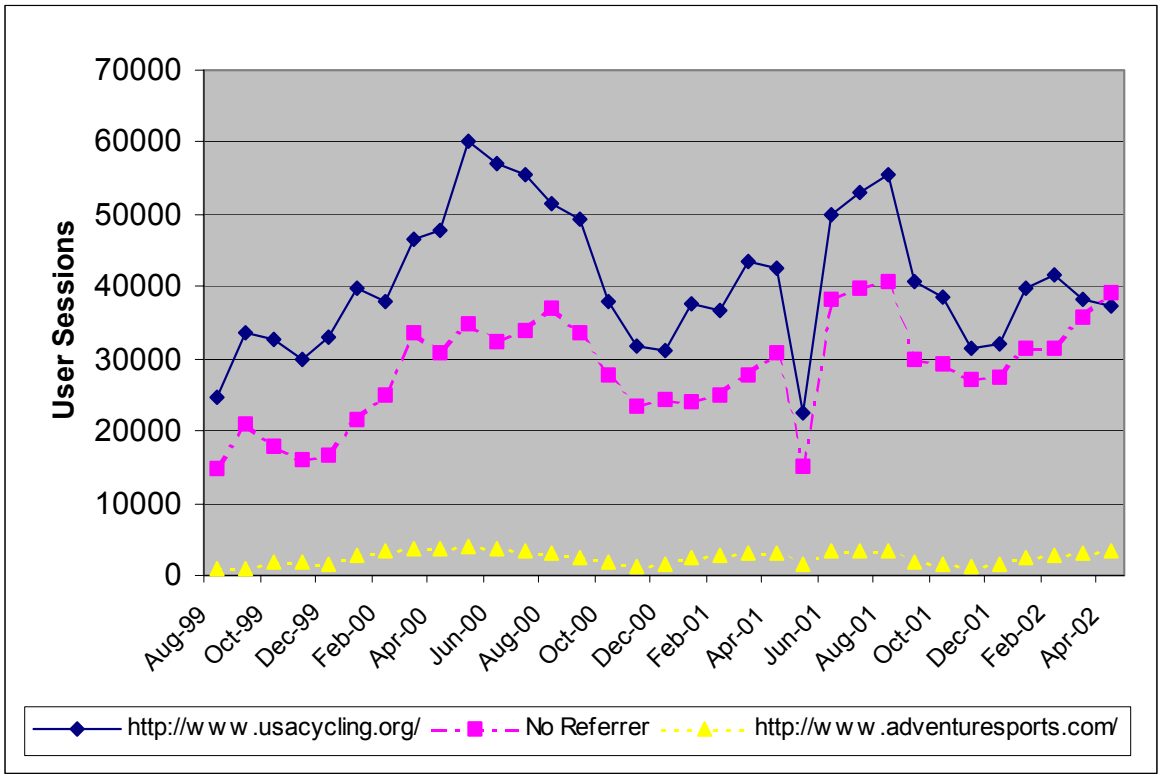
Graph 11: Bandwidth (kBytes Transferred) from August 1999 through April 2002 for USA Cycling's Website



Top Referring Websites

The top referring websites are shown in Graph 11. The top referring website was www.usacycling.org with no referrer and www.adventuresports.com following in second and third respectively. Since www.usacycling.org represents the most accessed page and users spend the most time on it, it follows that it should be the top referring web page. The www.usacycling.org web page functions as the junction to all other web pages within the USA Cycling website. Also, the usage patterns for *Top Referring Websites* are concurrent with the trends from the *Top Entry Page Results, Most Time Spent Results, and the Most Requested Web Pages*. Since competitive cycling's race season is from March to October, these usage trends fit appropriately.

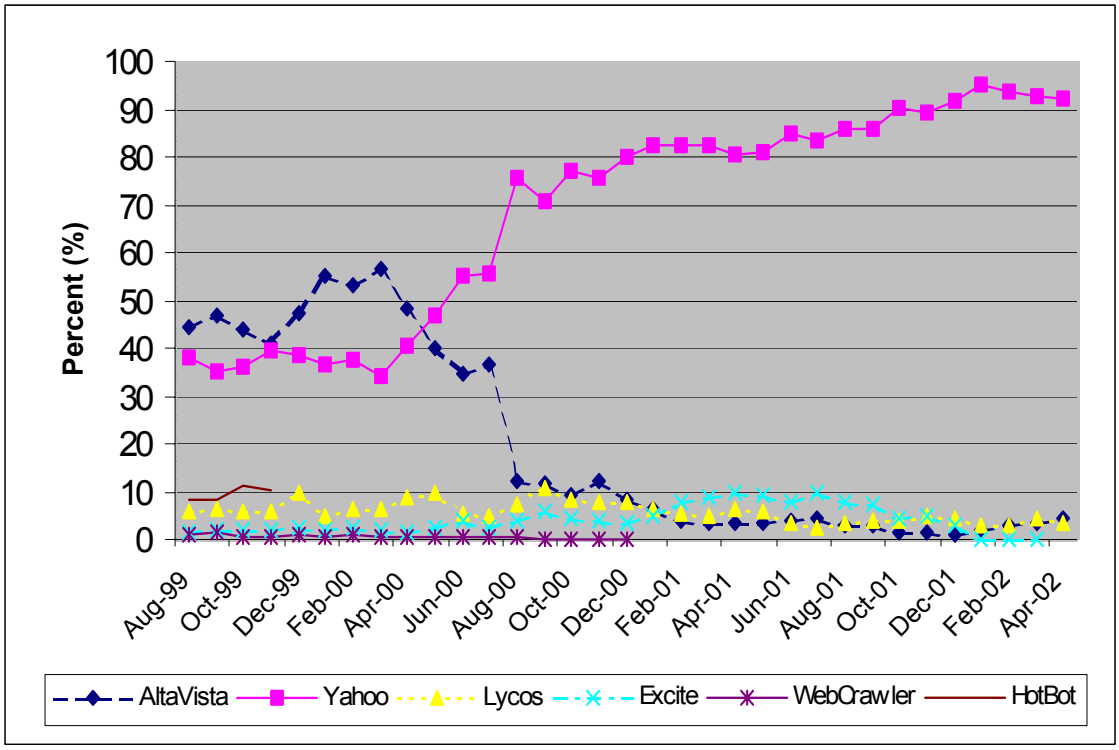
Graph 12: Top Referring Websites to the USA Cycling.Org Website based on User Sessions from August 1999 to April 2002



Top Search Engines

Graph 12, below, show the top search engines used to locate the USA Cycling website. In August 1999, the first month of WebTrends data, *AltaVista* was the most popular search engine used to locate the USA Cycling's website. In less than one year, *Yahoo* caught and surpasses *Alta Vista* as the number one search engine used to locate USA Cycling's website. *Yahoo's* initial usage percentage was approximately 40% in August 1999 and by December 2000, *Yahoo* accounted for 80% of USA Cycling's website hits. By November 2001, *Yahoo* exceeded 90% usage and maintained this rate through April 2002. This was complete domination. According to a report by Danny Sullivan of SearchEngineWatch.com (September 2002), *Yahoo* still maintains a slight edge over *MSN* and *Google*; see **APPENDIX I.**

Graph 13: Top Search Engines Used to Access USA Cycling's Website from August 1999 to April 2002



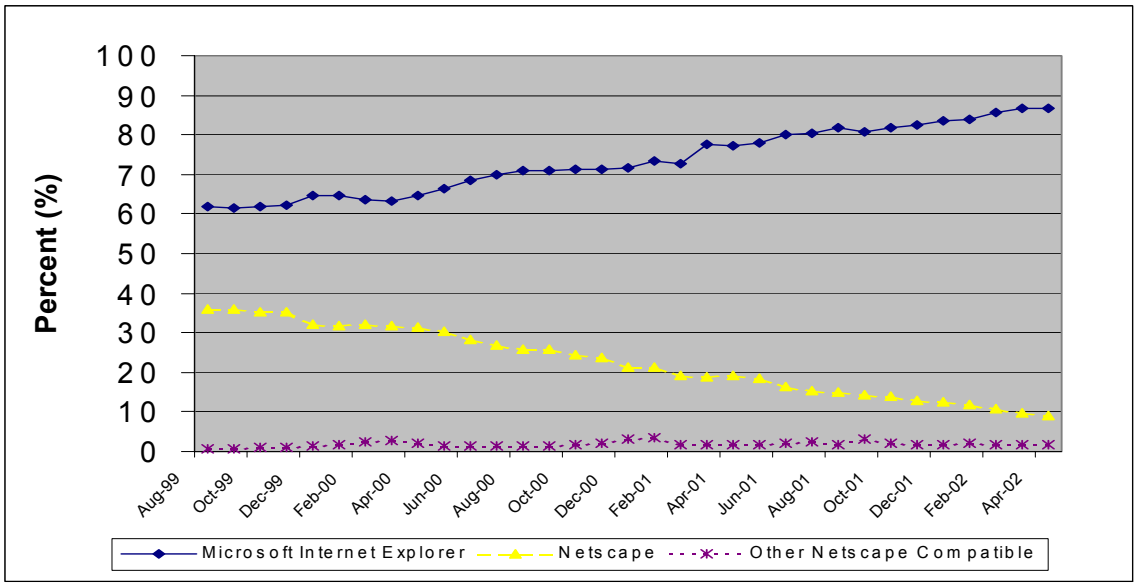
Based on the Survey results in APPENDIX I, it is interesting to note that *MSN* and *Google* do not turn up in the top 3 search engines used.

Web Browser Usage

Graph 13 displays the results of the three most used browsers to view USA Cycling's website from August 1999 to April 2002. In August 2002, WebSideStory, Inc. reported that Microsoft's Internet Explorer (IE) had control of over 90% of the browser market. In terms of USA Cycling, this trend follows the WebSideStory's report. In August 1999, Netscape and Internet Explorer controlled 35% and 62% respectively. Netscape maintained this position for approximately 3 months and then began declining to less than 10% of use by April 2002 among USA Cycling website users. Conversely, Internet Explorer improved their position and by April 2002 USA Cycling users

employed Microsoft Internet Explorer from 1.5 times to 9 times more resulting in almost a 90% usage rate. Microsoft IE increased from 61% to 88% usage over the thirty-three month period as Netscape plummeted from 36% to 9%.

Graph 14: Web Browser Usage for USA Cycling.Org as a Percent from August 1999 to April 2002



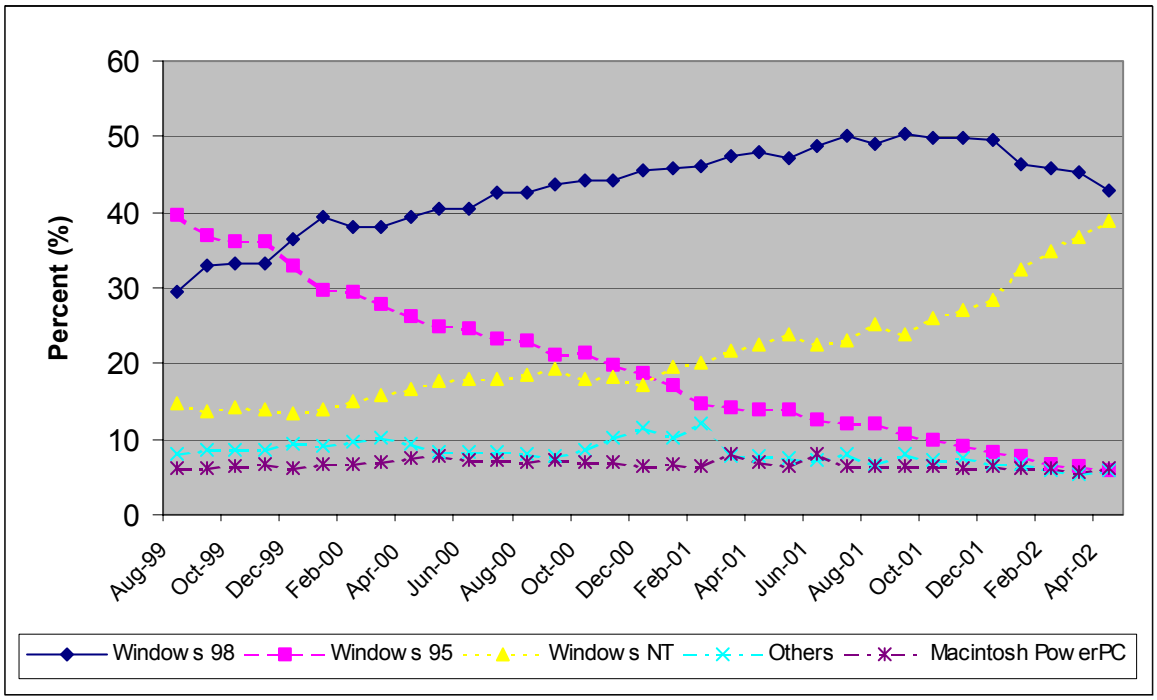
The third most popular browser was "Other Netscape Compatible;" their usage was minimum at 1% and maximum at 2.5%.

Types of Platforms Used

The graph below shows the most popular operating systems/platforms used to access USA Cycling's website from August 1999 to April 2002. For the time period, it demonstrates the changing trends with the release of new platform products as their popularity grows. The graph shows that Microsoft Operating Systems have a firm control for USA Cycling users. The Others and Macintosh Power PC operating systems account for only about 20% of total usage at any one time during the thirty-three month time frame. The graph also shows the change in usage for Microsoft operating systems. There is an increase in the number of Windows 98 users as Windows 95 declines for

most of the time period, then December 2001 Windows 98 starts declining and Windows NT operating systems begin to increase steadily. One can imagine by June 2002, Windows NT had become the number one used platform for USA Cycling website users by extrapolating the trend line.

Graph 15: Types of Platforms Used to Access USA Cycling.Org from August 1999 through April 2002 in Percents



The Windows Operating Systems constitute the bulk of use on the USA Cycling website. From August 1999 to October 2001, combining percentages for Windows Operating Systems accounted for approximately 80% of use. By knowing the platform most used, USA Cycling can optimize performance for those users.

CONCLUSION

It was the intent of this paper to create a user profile for a typical USA Cycling Website visitor. During the analysis, several other thoughts emerged and the results are the recommendations that follow the profile.

Profile

The purpose of this case study was to construct a user profile for the average USA Cycling website visitor. Using the graphs and charts above yield an interesting profile picture of the average USA Cycling user. The average user visits www.usacycling.org first for 1.5 to 2 minutes on Monday between 12:00 and 12:59 and spends 5 to 7 more minutes on the website. The most popular downloaded PDF pertain to membership applications. In addition, the average user uses Microsoft Internet Explore and has upgrade operating software as Microsoft has released it; changing from Windows 95 to Windows 98, and headed towards using Windows NT by the end of the study. The typically user searches *Yahoo* to locate USA Cycling, purposefully and/or unintentionally, and most significant referrer to USA Cycling's website is www.usacycling.org. The average user access the website from March to October, during the race season as shown by the higher bandwidth during this time period. User's Internet browsers typically cache 17% to 25% of pages hit on the USA Cycling and 80% of users are from US with less than 5% being international visits. The USA Cycling website had a low number of hits failed; hits failed dropped from 5% to less than 1% during the thirty-three month study period.

Recommendations

By constructing a user profile, USA Cycling can make choices that involve the direction of their website. The recommendations, below, are based on the user profile and are intended to maximize the user's experience on www.usacycling.org:

- Update event and race results faster-real time in order to relieve some of the Monday web traffic,
- Schedule any routine server maintenance between 4am to 6am-the lowest usage of the day,
- Increase bandwidth options during the heavily accessed times of the day, week and year,
- Consider putting a link on all their web pages that links a user back to the homepage,
- Ensure that the USA Cycling website is pulled from search engine queries that deal with bicycling,
- Incorporate a special promotion with *Yahoo, MSN* and/or *Google* to promote USA Cycling's website,
- Consider adding a site map link to the homepage, and
- Make sure that redundancy is achieved during high use times in case of a problem with one server.

BIBLIOGRAPHY

Anfuso, Dawn. (2002). *Web Gets Heaviest Traffic on Mondays*. Retrieved November 06, 2002, from http://www.imediaconnection.com/content/news/053002c.asp

Bauer, K. (2000, January/February). Who Goes There? Measuring Library Web Site Usage. *Online*, 25-31.

Berners-Lee, T., Fielding, R., and Masinter, L. (1998). *Uniform Resource Identifiers (URI): Generic Syntax. RFC 2396.* Retrieved June 2, 2002, from http://www.rfc-editor.org/rfc/rfc2396.txt

Krishnamurthy, Balachander and Rexford, Jennifer. (2001). *Web Protocols and Practice: HTTP/1.1, Networking Protocols, Caching, and Traffic Measurement*. New York, NY: Addison-Wesley.

Ivey, Sarah. (1996). *Analysis of Use of an Academic Web Server*. Master's Paper. University of North Carolina at Chapel Hill, School of Information and Library Science.

Meyer (2000). Web metrics: too much data, too little analysis. Impact and Evaluation of the Internet: Conference held at Cumberland Lodge, Windsor Great Park, July 1999, *(Aslib, London)*.

Hocheiser, Harry and Shneiderman, Ben. (2001). Using Interactive Visualizations of WWW Log Data to Characterize Access Patterns and Inform Site Design. *Journal of the American Society for Information Science and Technology, 52(4), 331-343.*

Nicholas, David, Huntington, Paul, Lievesley, Nat and Wasti, A. (2000). Evaluating consumer website logs: a case study of The Times/The Sunday Times website*. *Journal of Information Science, 26(6), 399-411.*

Nicholas, D., Huntington, P., Williams, P., Lievesley, N., Dobrowolski, T., and Withey, R. (1999). Developing and Testing Methods to Determine the Use of Web Sites: Case Study Newspapers. *Aslib Proceedings, 51(5), 144-154.*

Nielsen Net Ratings(2002). Retrieved May 30, 2002, from http://www.nielsennetratings.com/

Schutlz, Keith. (1997). Two Tools for Monitoring Your Web Site. (WebManage Technologies' NetIntellect 3.0 and e.g. Software's WebTrends 3.5 Web Site Monitoring Software). *InternetWeek, 687, 60-61.*

Sullivan, Danny. (2002). *Nielsen//NetRatings Search Engine Ratings*. Retrieved Novemebr 07. 2002 from http://searchenginewatch.com/reports/netratings.html.

USA Cycling Website. (2002). Retrieved September 14, 2002, from http://www.usacycling.org

Warren, Nikki. (2002). *Website Log Analysis: Approaches for the Library of the NationalInstitute Of Environmental Health Sciences*. A Master's Paper for the Masters of Science in Library Science. July, 2002. 75 pages. Advisor: Gregory B. Newby.

WebSideStory, Inc. (2002, May). *The Incredible Shrinking Browser –Netscape Share Less Than 4% Worldwide, According to WebSideStory's StatMarket Despite Higher Percentages in Certain Countries, Netscape Continues to Lose Market Share to Microsoft Globally.* Retrieved November 7, 2002 from http://www.websidestory.com/cgi-bin/wss.cgi?corporate&news&press_1_193.

Yeadon, Jane. (2001). Web Site Statistics. *Vine 9(1), 55-60.*

Yu, Liangzhi and Apps, Ann. (2000). Studying E-Journal User Behavior Using Log Files: The Experience of the SuperJournal. *Library and Information Science Research, 22(3), 311-338.*

Zawitz, M.W. (1998, May). *Web Statistics - Measuring User Activity: An Analysis of Bureau of Justice Statistics (BJS) Website Usage Statistics. 1-12. (Bureau of Justice Statistics Publication No. NCJ – 171118).* Retrieved May 1, 2002, from http://www.ojp.usdoj.gov/bjs/abstract/wsmua.htm

APPENDICES

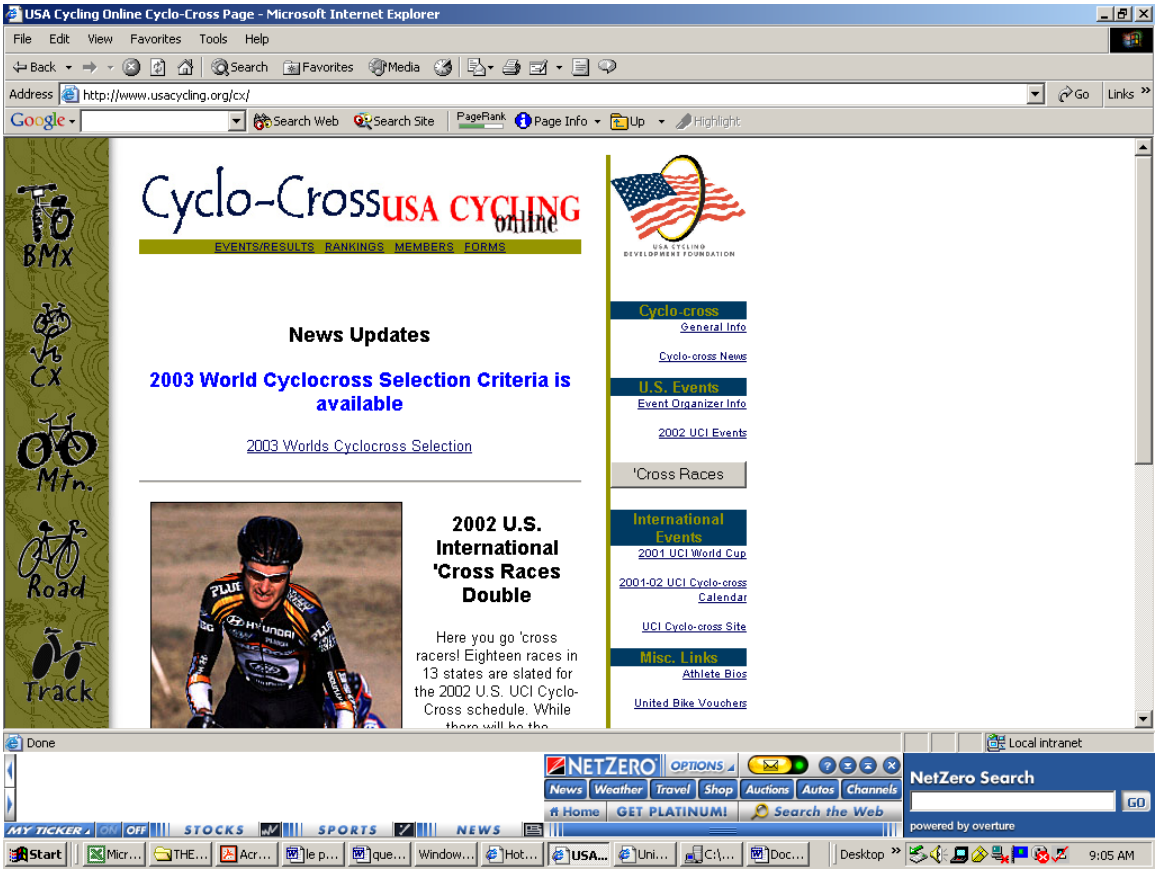
APPENDIX A: USA Cycling Screen Shots

USA Cycling Homepage www.usacyling.org

Mountain Biking Web Page www.usacycling.org/mtb

USA Cycling Road Biking Web Page www.usacycling.org/road/

USA Cycling Cyclocross Web Page www.usacycling.org/cx/



USA Track Cycling www.usacycling.org/track/

APPENDIX B: Measurement Case Studies

SILS server log study-1995-1996

In 1996 at the University of North Carolina at Chapel Hill, Sarah Ivey conducted an analysis of the School of Information and Library Science's academic web server. She investigated a fifteen-month period and used Analog, a freeware log analysis tool, to carryout her study. From her study, she found that there had been a general increase in Internet use as reported by various Internet surveys and that two domains accounted for over 50% of total server traffic. In addition, she found that more than 50% of the pages requested had less than nine hits per month.

Saskatchewan server log study

The goal of the study was to determine the basic properties of web workload on six servers in 1995 express Krishnamurthy and Rexford (2001). Three universities, two scientific organizations, and one ISP represented the six servers. The researchers did not have control over the available fields and TABLE 3, below, shows the information they were able to determine.

TABLE 7: Key Metrics of the Saskatchewan Study of Server Logs

Category	Metric
Basic statistics	Number of distinct Request-URIs Average/median transfer size Frequency of each response code
Access Patterns	Time between successive requests Popularity of each requested resource Time between requests for the same resource
Inferences	Content types of requested resources Average/median resource size Frequency of resource modification Frequency of aborted transfers

*Table is from Krishnamurthy and Rexford (2001) text *Web Protocols and Practice: HTTP/1.1, Networking Protocols, Caching, and Traffic Measurement(2001).*

British Columbia proxy log study

Researchers used seven proxy logs from 1996 to 1997 to test caching under a variety of conditions maintain Krishnamurthy and Rexford (2001). The seven proxy logs represented institutions around the world and included universities, companies, and one national proxy. All seven institutions ran squid proxy servers. Each log included client IP address, the request time, the Request-URI, and the response size. Six of the server logs came directly from the caching proxies and reported on the request action being satisfied or not. The other log entry included *Last Modified* and *Expires* times. Unlike the Saskatchewan study, which focused on statistics, the British Columbia study intended to evaluate the effectiveness of proxy caching. The experiment included different cache sizes, request rates, and cache-coherency policies using simulation. The simulation technique had several advantages. First, tying up seven servers at seven institutions would have been unpractical. Secondly, different cache-coherency policies could be studied that might have not been available on the operational servers. Lastly, the same logs could be used over and over with different proxy configurations.

APPENDIX C: Authenticated Users' Rank

Table 8: Authenticated User's Rank from August 1999 to April 2002

	Autheticated Users		
	1	2	3
Aug-99	usacyc	na	na
Sep-99	usacyc	acook	tvinson
Oct-99	usacyc	acook	jparsons
Nov-99	usacyc	acook	lseidman
Dec-99	usacyc	acook	gheagert
Jan-00	usacyc	gheagert	mhanley
Feb-00	usacyc	lseidman	acook
Mar-00	usacyc	mhanley	lseidman
Apr-00	usacyc	tvinson	tdelp
May-00	mhanley	usacyc	tvinson
Jun-00	usacyc	mhanley	tvinson
Jul-00	mhanley	tvinson	usacyc
Aug-00	mhanley	usacyc	tvinson
Sep-00	tvinson	mhanley	usacyc
Oct-00	jparsons	usacyc	mhanley
Nov-00	gheagert	usacyc	mhanley
Dec-00	mhanley	gheagert	usacyc

	Autheticated Users		
	1	2	3
Jan-01	usacyc	gheagert	tdelp
Feb-01	usacyc	mhanley	Dean
Mar-01	usacyc	tdelp	gheagert
Apr-01	promo	rcs	temp
May-01	promo	rcs	usacyc
Jun-01	promo	temp	usacyc
Jul-01	promo	usacyc	jpeterson
Aug-01	promo	gheagerty	mwise
Sep-01	promo	gheagerty	usacyc
Oct-01	promo	usacyc	gheagerty
Nov-01	promo	usacyc	mhanley
Dec-01	promo	usacyc	gheagerty
Jan-02	promo	mhanley	jmiller
Feb-02	promo	mhanley	jmiller
Mar-02	promo	mhanley	temp
Apr-02	promo	gheagerty	usacyc

APPENDIX D: Un-Authenticated Users' Rank

Table 9: Un-Authenticated Users of the USA Cycling Website Identified by IP Address from August 1999 to April 2002

	Un-Authenticated Users		
	1	**2**	**3**
Aug-99	dt061n62.maine.rr.com	209.107.36.74	209.38.121.114
Sep-99	209.107.36.74	cx1002002-b.phnx3.az.home.com	dt061n62.maine.rr.com
Oct-99	209.107.36.74	cx1002002-b.phnx3.az.home.com	Chub.Stanford.EDU
Nov-99	209.107.36.74	207.51.218.126	dt0d1n2a.maine.rr.com
Dec-99	209.107.36.74	pa-bethelpark4a-234.pit.adelphia.net	modem125115.w estman.w ave.ca
Jan-00	209.107.36.74	cx1002002-b.phnx3.az.home.com	modem125115.w estman.w ave.ca
Feb-00	209.107.36.74	mar31.marriott.com	mar32.marriott.com
Mar-00	209.107.36.74	fw -eth0.racegate.com	208.132.152.33
Apr-00	209.107.36.74	208.132.152.34	resnet5087.resnet.union.edu
May-00	209.107.36.74	dt0d1n2a.maine.rr.com	gnab91h.nab.usace.army.mil
Jun-00	209.107.36.74	24.64.152.223.on.w ave.home.com	cache-1.lnh.md.w ebcache.rcn.net
Jul-00	209.107.36.74	24.64.152.223.on.w ave.home.com	m138-mp1-cvx1c.lee.ntl.com
Aug-00	np-serial109.co.verio.net	24.64.152.223.on.w ave.home.com	209.107.36.74
Sep-00	np-serial109.co.verio.net	AC9F9FCA.ipt.aol.com	fw -us-hou-2.bmc.com
Oct-00	np-serial109.co.verio.net	balt1.fbw .com	adsl-63-193-96-114.dsl.snfc21.pacbell.net
Nov-00	np-serial109.co.verio.net	adsl-63-193-96-114.dsl.snfc21.pacbell.net	default-2.farcpe.cableone.net
Dec-00	np-serial109.co.verio.net	cache1.lgca.org	icmproxy.icmnet.net
Jan-01	np-serial109.co.verio.net	192.249.47.9	cache1.lgca.org
Feb-01	arthur4.sda.t-online.de	arthur4.sda.t-online.de	209.107.36.74
Mar-01	209.107.36.74	AC838E05.ipt.aol.com	crow nmail.crow nintl.com
Apr-01	209.248.75.38	24-168-192-13.ff.cox.rr.com	209.107.36.74
May-01	209.248.75.38	12.42.50.51	f1.airproducts.com
Jun-01	209.248.75.38	ACAA0C93.ipt.aol.com	AC9EE1FD.ipt.aol.com
Jul-01	209.248.75.38	har2-di130.rica.net	12.21.187.194
Aug-01	209.248.75.38	c169503-a.boulder1.co.home.com	cache-1.sbo.ma.w ebcache.rcn.net
Sep-01	209.248.75.38	57.68.12.102	cache-1.sbo.ma.w ebcache.rcn.net
Oct-01	209.248.103.74	57.68.12.102	149.149.200.200
Nov-01	209.248.103.74	siliconpeak2.media3.net	cx1002002-e.phnx3.az.home.com
Dec-01	209.248.103.74	24.125.8.194	adsl-63-198-178-114.dsl.snfc21.pacbell.net
Jan-02	209.248.103.74	216.167.97.169	12-237-34-23.client.attbi.com
Feb-02	209.248.103.74	adsl-63-198-178-114.dsl.snfc21.pacbell.net	12-237-34-208.client.attbi.com
Mar-02	209.248.103.74	adsl-63-198-178-114.dsl.snfc21.pacbell.net	12-237-34-208.client.attbi.com
Apr-02	209.248.103.74	AC953E93.ipt.aol.com	12-237-224-130.client.attbi.com

APPENDIX E: USA Cycling and Select User's Server Info

USA CYCLING NSLOOKUP

Server: www-coastland-49.highertech.net
Address: 66.129.3.49

www.usacycling.org internet address = 161.58.123.16
usacycling.org nameserver = b.ns.verio.net
usacycling.org nameserver = ns1.verio.net
usacycling.org nameserver = t.ns.verio.net
b.ns.verio.net internet address = 129.250.35.32
ns1.verio.net internet address = 204.91.99.140
t.ns.verio.net internet address = 192.67.14.16

here is the **traceroute** result from this host to **161.58.123.16** :

```
 traceroute to 161.58.123.16 (161.58.123.16), 30 hops max, 38
byte packets
  1  shredder-1c.higherbandwidth.net (66.129.3.1)  0.333 ms
0.276 ms  0.253 ms
  2  ru-1.higherbandwidth.net (66.129.0.1)  1.392 ms  0.645 ms
0.729 ms
  3  500.Serial3-11.GW7.ATL1.ALTER.NET (157.130.42.65)  14.231 ms
8.557 ms  5.113 ms
  4  174.at-1-0-0.XL3.ATL1.ALTER.NET (152.63.82.50)  11.200 ms
7.058 ms  8.267 ms
  5  0.so-3-0-0.TL1.ATL1.ALTER.NET (152.63.10.69)  3.746 ms
10.660 ms  6.715 ms
  6  0.so-6-0-0.TL1.CHI2.ALTER.NET (152.63.13.21)  19.506 ms
20.465 ms  25.503 ms
  7  0.so-7-0-0.XL1.CHI2.ALTER.NET (152.63.68.81)  22.191 ms
17.965 ms  25.265 ms
  8  0.so-7-0-0.BR6.CHI2.ALTER.NET (152.63.71.94)  21.996 ms
31.624 ms  35.138 ms
  9  204.255.174.162 (204.255.174.162)  26.144 ms  19.212 ms
32.535 ms
 10  p16-2-0-0.r01.chcgil06.us.bb.verio.net (129.250.5.70)
24.874 ms  21.434 ms  20.150 ms
 11  p16-0-1-1.r20.dllstx01.us.bb.verio.net (129.250.5.85)
39.678 ms  46.012 ms  36.257 ms
 12  p64-0-0-0.r21.dllstx01.us.bb.verio.net (129.250.3.41)
41.174 ms  86.033 ms  43.203 ms
 13  p16-2-0-0.r00.stngva01.us.bb.verio.net (129.250.5.35)
66.885 ms  66.736 ms  75.722 ms
 14  ge-1-2-r0709.stngva01.us.verio.net (192.67.244.245)  53.022
ms  48.219 ms  56.299 ms
 15  ge-26-a0723.stngva01.us.verio.net (192.67.243.117)  65.131
ms  60.357 ms  53.706 ms
 16  www.usacycling.org (161.58.123.16)  68.263 ms  59.576 ms
50.887 ms
```

TABLE 10: HTTP header from **161.58.123.16**

```
HTTP/1.1 500 Internal Server Error
Date: Thu, 07 Nov 2002 19:04:33 GMT
Server: Rapidsite/Apa/1.3.26 (Unix) FrontPage/5.0.2.2510
mod_ssl/2.8.10 OpenSSL/0.9.6e
Connection: close
Content-Type: text/html; charset=iso-8859-1
```

TABLE 11: IP Address and NSLOOKUP Information of USA Cycling's Users

IP Address	OrgNAme	NetRange	NetType	RegDate
209.107.36.74	Verio, Inc.	209.107.0.0 - 209.107.95.255	Direct Allocation	1997-07-23
209.248.75.38	Vanion, Inc.	209.248.64.0 - 209.248.127.255	Direct Allocation	2000-08-25
57.68.12.102	SITA-Societe Internationale de Telecommunications Aeronautiques	57.0.0.0 - 57.255.255.255	Direct Assignment	1993-06-21
12.21.187.194	AT&T WorldNet Services ATT	12.0.0.0 - 12.255.255.255	na	na
149.149.200.200	Tennessee Technological University	149.149.0.0 - 149.149.255.255	Direct Assignment	1991-05-02
192.249.47.9	United Technologies Research Center	192.249.32.0 - 192.249.51.255	na	na

APPENDIX F: Correspondence with USA Cycling

Three attempts to clarify questions about the data were made. There was no reply from the first attempt. The reply from the second attempt follows:

Kelly Walker of USA Cycling wrote “Concerning your questions- right now our IT dept. is knee-deep underway in bringing in a new CMS. They are hammered to put it lightly. When you inquired about some statistical info about USA Cycling and our website I guess I didn't realize that the research would be so detailed and time-intensive. I have been informed by the IT dept. that they simply don't have the extra time to answer these questions. I hope this doesn't come off as being short- it certainly is not meant to- I just have to be honest with you and what we can do.”

The third attempt received a no response.

APPENDIX G: Questionnaire

13 october 2002

When did the USA cycling website come to its present form? Aug 1999?

Did a new server come online in Aug 1999? Did it change from August 1999 to April 2002?

What type of Server is being used now? How many servers are in service?

SEE GRAPH 7: In Aug 1999, over 5% of hits failed, by March 2000 this had declined to less than 1%, what do you attribute this to?

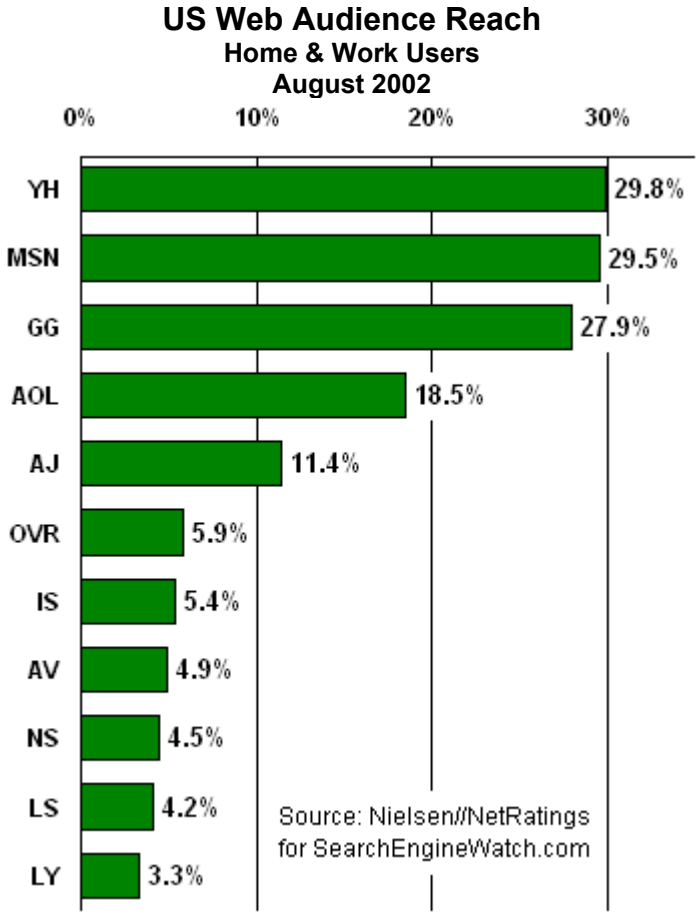
In January 2002, there was a spike in the "hits failed" to over 1%, can you think of any reason this might have occurred?

10-15-2002
Are there quarterly or yearly WebTrends data for USA Cycling's website?

Who are the authenticated users? USA Cycling employees? Race Promoters?

10-16-2002
Are the USACYCL and PROMO authenticated names unique logins or generic logins?

APPENDIX H: Search Engine Statistics for August 2002



KEY: YH=Yahoo, MSN=MSN, GG=Google, AOL=AOL, AJ=Ask Jeeves, OVR=Overture (GoTo), IS=InfoSpace;AV=AltaVista, NS=Netscape, LS=LookSmart, LY=Lycos
For links, see the Major Search Engines and Major Metacrawlers pages.

"The chart [above] shows the most popular search sites in the United States, as based on audience reach for August 2002. Audience reach is the percentage of US home and work internet users estimated to have searched on each site at least once during the month. For August 2002, there were an estimated 122 million total internet users online in the US at work or at home. Only "search specific" traffic is counted toward the figures below. This means that only visits deemed to be search-related were counted in the totals. That helps prevent non-search traffic at portals (such as visits to get email) from polluting the data.