

Revisiting Web Server Workload Invariants in the Context of Scientific Web Sites

Anne M. Faber, Minaxi Gupta, and Camilo H. Viecco

Computer Science Department, Indiana University, Bloomington, IN, U.S.A.

{anmcleve, minaxi, cviecco}@cs.indiana.edu

Abstract

The Web has evolved much from when Arlitt and Williamson proposed the ten Web workload invariants more than a decade ago. Many diverse communities now depend on the Web in their day-to-day lives. A current knowledge of the invariants for the Web is useful for performance enhancement and for synthetic Web workload generation. Invariants can also serve as a useful tool for detecting anomaly and misuse, a new dimension of Web usage arising from the change in trust assumptions in the Internet in the recent years. Focusing on scientific Web servers, we revisit the Web server workload invariants and find that only three out of the ten invariants hold as-is. We investigate appropriate revisions to the invariants that do not hold and also propose three new invariants for scientific Web servers.

1 Introduction

Over the past several decades, the Web has grown from a research experiment utilized by a handful of scientists to an integral part of our economy and everyday lives. Today, it is diverse both in the kinds of data it serves, and in terms of the communities that use it to access data.

The continued growth in Web traffic has been a perpetual motivation to improve it. One of the ways to improve Web performance is to understand and characterize Web server workloads. This exercise aids in designing better caching strategies in the Internet, or at the server which can improve performance for the clients. Also, Web workload characterization studies can assist in generating synthetic workloads for experimental purposes.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SC2006 November 2006, Tampa, Florida, USA
0-7695-2700-0/06 \$20.00 ©2006 IEEE

Many studies have been done in the area of Web workload characterization with the above motivations in mind. These studies (for example [Padmanabhan and Qui 2000; Bent et al. 2004; Arlitt and Williamson 1996; Crovella and Bestavros 1997; of Dynamic Web Content 2003]) have gone a long way in understanding the characteristics of Web servers from many different application domains, such as educational Web servers, news Web servers, and Web servers used for e-commerce. In fact, the seminal work by Arlitt and Williamson [Arlitt and Williamson 1996] helped establish ten invariants for the Web from an extensive study of 6 different Web servers.

The work presented in this paper is motivated by several observations. First, the Web has evolved much since Arlitt and Williamson coined the ten invariants more than a decade ago. We were interested in understanding if those invariants still hold for today's Web and if new invariants now exist for the Web. Second, none of the Web workload characterization studies that we are aware of considered scientific Web sites, especially the kind that host a back end database in order to support user queries. We were interested in finding out if the invariants hold for scientific Web sites and in understanding if the scientific Web sites differed in major ways from other types of Web sites. The third motivation for our work relates to the change in trust assumptions in today's Internet. Denial-of-service (DoS) attacks, spam, and malicious crawlers are just some of the examples of threats that the Internet faces today. Many of these threats directly affect the Web. An understanding of the *normal* workload at a Web server and knowledge about the profile of its clients could help in developing tools to prevent damage to the Web server from unwanted misuse, like the kind inflicted by malicious Web crawlers.

Our goal then was three-fold: 1) find out if the invariants coined by Arlitt and Williamson hold true for today's Web and in particular for scientific Web sites 2) examine how the invariants have evolved for scientific and other Web sites, and 3) investigate new invariants for both scientific and non-scientific Web sites that could assist in anomaly and misuse detection.

To achieve our goal, we collected four sets of workloads, three from scientific Web servers, and one from a

Invariant	Results from [Arlitt and Williamson 1996]	Our results (all data sets)	Our results (scientific data sets)
1. Success Rate	Success rate for lookups at server: about 88%	Success rate for lookups at server: about 75%	Same as for all data sets
2. File Types	HTML and image files account for 90-100% of requests	HTML and image files account for 60-80% of requests	HTML and image files account for < 70% of requests and dynamic content accounts for 20-30% of requests
3. Mean Transfer	Mean transfer size \leq 21 Kilobytes	Mean transfer size \leq 38 Kilobytes	Same as for all data sets
4. Distinct requests	Less than 2.1% of the requests are for distinct files	Up to 20% of the requests for distinct files	Same as for all data sets
5. One Time Referencing	About one third of the files and bytes accessed in the log are accessed only once	About one third of the files and bytes in the log are accessed only once for the non-scientific data set	At least two-thirds of the files and bytes accessed in the log are accessed only once
6. Size Distribution	File size distribution is Pareto with $0.4 < \alpha < 0.63$	File size distribution is log-normal	Same as for all data sets
7. Concentration of References	10% of the files accessed account for 90% of the server requests and 90% of the bytes transferred	10% of files accessed account for > 80% of the server requests and > 80% of the bytes transferred	Same as for all data sets
8. Inter-reference Times	File inter-reference times are exponentially distributed and independent	File inter-reference times are exponentially distributed and independent	Same as for all data sets
9. Remote requests	Remote clients account for 75-99% of the accesses to the server and 75-99% of the bytes transferred	Remote clients account for 81-99% of the accesses to the server and 73-99% of the bytes transferred	Remote clients account for at least 95% of the requests and the bytes transferred
10. Wide Area Usage	Web servers accessed by thousands of domains, with 10% of the domains accounting for \geq 75% of the usage	Web servers accessed by thousands of domains, with 10% of the domains accounting for \geq 69% of the usage	Web servers accessed by thousands of domains, with 10% of the domains accounting for \approx 70% of the usage

Table 1: Comparison of invariants.

Invariant	Our results (all data sets)	Our results (scientific data sets)
1. Weekday versus Weekend Usage	>74% of usage occurs on weekdays	>80% of usage occurs on weekdays
2. Diurnal Usage	Web servers usage is heavier during the day.	Same as for all data sets
3. Client Concentration	25% of clients are responsible for >75% of requests.	Same as for all data sets

Table 2: New invariants.

departmental server. An overview of the workloads we examined is presented in table 3. The key findings of our work include: 1) only three of the ten invariants defined by Arlitt and Williamson hold for both the scientific and non-scientific Web sites considered, 2) two additional invariants hold for the non-scientific Web site but not for the scientific Web site, indicating that perhaps a different category of invariants exists for the scientific Web sites, 3) the defining characteristics of the scientific Web sites are: the presence of a large percentage of dynamic content, a large fraction of files that are accessed only once, and lesser concentration of clients in terms of domains. We propose revisions to the invariants that do not hold for scientific sites and also propose three new invariants that could be used for anomaly and misuse detection at scientific Web sites. An overview of our findings about the original invariants is given in table 1. Table 2 presents the three new invariants.

The rest of this paper is organized as follows. In section 2 we discuss related work. Section 3 describes the Web servers we studied and how our data was collected. Section 4 presents an analysis of our data in terms of the ten workload invariants outlined in [Arlitt and Williamson 1996] and section 5 derives our three new workload invariants. Finally, section 6 offers concluding remarks.

2 Related Work

Many workload characterization studies have been done but as far as we know, none have focused specifically on a Web server hosting a scientific application. Other studies in this area have focused on commercial, academic, and government Web servers.

Our work is motivated by the seminal work done by Arlitt and Williamson [Arlitt and Williamson 1996]. Their research on workload characterization focused on finding invariants among different Web servers. That is, they looked at several different Web servers and tried to find characteristics shared by all. They looked at access and error logs from six different Web servers: three from academic environments, two from scientific research institutions, and one from a commercial Internet provider. Some invariants discovered were a similar success rate among the Web servers (about 88%), a similar mean transfer size, and the dominance of HTML and image files over other file types being transferred. These and the other invariants are summarized in table 1. Our study differs from Arlitt and Williamson’s study in the type of server being studied. While they studied two Web servers from scientific research institutions, these Web servers were geared more toward then general public. For example, the NASA Web server contains information about NASA’s current projects, latest findings, and other topics the general public may find interest-

ing. One of our goals then, was to see how the scientific Web servers compared to those studied by Arlitt and Williamson. That is, do the invariants that were discovered hold for a Web server hosting a scientific application? The comparison is complicated by the fact that their study was done nearly a decade ago. This brings up the question of whether the invariants discovered back then are still invariants today and whether it even makes sense to look for them in a modern Web server workload. We address this issue by examining logs from a computer science department Web server in addition to studying the scientific Web servers.

Another study done by Padmanabhan and Qiu examined the workload of the Web server for MSNBC, a popular news site [Padmanabhan and Qui 2000]. They looked at both the content logs and the access logs for the MSNBC Web server to examine the file creation/modification dynamics and also the access patterns. Some of their key findings were: (a) modifications to a file usually change very little, (b) a small subset of the files is modified repeatedly, (c) file popularity follows a Zipf-like distribution, and (d) there is significant temporal stability in file popularity. Our investigation, again, differs from this study in the type of Web server being examined. While it is natural that the content of a news Web site will change frequently and the popularity of particular content will change over time and possibly eventually be removed, this is not necessarily true for a Web server hosting a scientific application.

Bent et al studied the workload of a large server farm [Bent et al. 2004]. They examined the workload of 3,000 sites, all hosted by the same ISP. The purpose of their study was to examine the characteristics of the Web sites from the server’s point of view and to determine if the Web sites could benefit from subscribing to a content delivery network. In particular, they examined the cookie usage and cacheability of content for the Web sites studied. The results showed that use of content delivery networks could improve the performance of most of the Web sites studied. Our study is done on a much smaller scale and focuses on a completely different kind of Web server.

Self-similarity in Web traffic has been studied in [Crovella and Bestavros 1997]. Work in [of Dynamic Web Content 2003] models the characteristics of dynamic Web content. Other studies have been done [Acharya et al. 1998] [Almeida et al. 2001] [Cheshire et al. 2001] [Harel et al. 1999] [Padhye and Kurose 1998] [Cherkasova and Gupta 1998] that examined the workloads of multimedia Web servers for both universities and corporations. Like our study, these examine the workloads of a specific type of server to guide the provisioning of servers and resources. The main goal of [Almeida et al. 2001] was to identify the important parameters for generating synthetic workloads.

Item	Flybase 1	Flybase 2	ReciprocalNet	CS
Access Log Duration (days)	73	63	60	32
Access Log Start Date	March 12, 2004	Oct 22, 2004	March 27,2005	February 28, 2005
Access Log Size (MB)	963.8	910.8	285.04	491.19
Total Requests	4,528,542	4,414,043	1,269,997	4,150,935
Average Requests/Day	62,035	70,064	21,502	136,190.87
Total Bytes Transferred (MB)	89,444.8	93,349.6	12,664.9	107,429.1
Average Bytes/Day (MB)	1,225.3	1,481.7	210.7	3,539.0

Table 3: Access log characteristics (raw data).

This is also one of our goals.

A representative piece of work that utilizes the results of workload characterization studies to enhance performance can be found in [Doyle et al. 2001]. Work in [Barford and Crovella 1998] studies generation of representative synthetic Web workloads.

3 Data

The four sets of access logs used in this research were obtained from the following three Web servers: 1) Flybase [Fly], a Web server serving the database of Drosophila Genome, 2) ReciprocalNet [cry], a Web server hosting the molecular structures database and 3) the Indiana University Computer Science Departmental Web server. While Flybase and ReciprocalNet primarily serve researchers and students in their respective scientific communities, both inside and outside Indiana University, the Computer Science Departmental Web server contains general information about the department for general public and department personnel.

For all the Web servers studied, the Apache [apa] access logs contained information about each request encountered by the server. This information included: the client IP address, the time the request was serviced, the URL of the requested page, a status code, and how many bytes were transferred. We created a Perl script to load each set access logs into a mysql [mys] database. The flexibility of the database allowed us to test for different usage patterns. Table 3 summarizes the overall characteristics of our raw data. It contains two snapshots of the Flybase Web server usage (referred to as Flybase 1 and Flybase 2 respectively), one snapshot of the ReciprocalNet Web server usage (referred to as ReciprocalNet), and one snapshot of the Computer Science Department Web server usage (referred to as CS).

3.1 Flybase Data Set

Flybase [Fly] (*Fly Database*) is the front end Web site to a database of genetic and molecular data for Drosophila. This database is used extensively for genetic research due to the parallels that can be drawn

between the Drosophila genome and the human genome. Flybase is hosted by the Genome Informatics Lab at Indiana University through dedicated Web servers and has mirror sites at University of Cambridge (UK) and the National Institute of Genetics (Japan).

Client requests to the Flybase Web server fall in two major categories: static and dynamic. Static content is stored on the Web server and is returned immediately. Dynamic content is typically the result of a user query on some aspect of the Drosophila genome. It requires the server to query its back end database and build a Web page dynamically based on the result of the user query. This page is then returned to the client.

We obtained two 5-month apart snapshots of the Flybase access logs. As shown in table 3, the first snapshot, Flybase 1 was for 73 days, from March 12th through May 23rd in 2004 and the second snapshot, Flybase 2 was for 63 days from October 23rd through December 24th 2004. Flybase 1 witnessed an average of 62,035 requests per day and transferred 1.2GBytes on an average day. The main motivation to obtain another snapshot of the Flybase Web server was to understand the evolution of the site. This is evident from the changes observed in the Flybase 2 log. As shown in table 3, the average number of requests per day in Flybase 2 was 70,064 and correspondingly, an average of 1.5GBytes were transferred in a day. This is a 13% increase in the average number of requests per day and a 20% increase in the corresponding bytes served over a 5-month period.

3.2 ReciprocalNet Data Set

ReciprocalNet [cry] is a distributed database containing information about molecular structures. Researchers from thirteen crystallography laboratories around the world contribute molecular structures to this database. Chemistry researchers, teachers, students, and general public can query this database using a distributed search engine supported by all participating sites.

Just like Flybase, ReciprocalNet also contains both static and dynamic information. It uses a visual approach to relay information about the molecular structures and allows its users to search for a specific

Item	Flybase 1	Flybase 2	ReciprocalNet	CS
Access Log Duration (days)	73	63	60	32
Total Requests	3,709,111	3,564,258	866,586	2,948,822
Distinct Requests	552,293	555,368	66,415	169,118
Average Requests/Day	50,810	56,575	16,015	96,826
Total Bytes Transferred (MB)	88,804.2	93,719.5	11,627.0	105,703.9
Mean Transfer Size (bytes)	25,105	27,572	14,069	37,587
CoV of Transfer Size	18.79	19.39	22.38	17.56

Table 4: Access log characteristics (reduced data containing successful requests).

molecule or categories of molecules, such as biochemical molecules, or minerals and gems. Once a molecular category is selected, a page is dynamically generated by the Web server that gives a 3D image of its structure, chemical formula, history and uses, and also provides other links that contain more information. The 3D image of the molecule is displayed by a Java applet that allows the user to click on the image and move it around to view it from different angles.

As table 3 shows, we obtained the ReciprocalNet access logs for a period of 60 days beginning on March 27, 2005. The Web server encounters an average of 21,502 requests per day and transfers 210.7MBytes per day.

3.3 Computer Science Department Data Set

The Computer Science Department Web server contains information on the students, faculty, and staff in the department and is accessed both by personnel within the department and outside. The primary reason to obtain a snapshot of this Web server usage was to be able to compare the usage of scientific Web servers similar to Flybase and ReciprocalNet with that of an informational site like CS.

As shown in table 3, we obtained CS access logs for a period of 32 days beginning February 28, 2005. The average requests per day for this site were 136,191 and an average of 3.5GBytes of data was transferred each day.

4 Revisiting the Original Invariants

This section presents the analysis of each of our four data sets in terms of the ten workload invariants identified in [Arlitt and Williamson 1996]. The overview of our conclusions is presented in table 3.

4.1 First Invariant: Success Rate

The status codes, defined in [Fielding et al. 1999], are three digit codes that indicate the server’s status after

Response Code	Flybase 1	Flybase 2	ReciprocalNet	CS
Successful	81.9%	80.7%	68.2%	71.0%
Not Modified	5.9%	5.9%	16.5%	11.7%
Found	4.2%	5.1%	1.2%	10.0%
Unsuccessful	8.0%	8.3%	14.0%	7.2%

Table 5: Server responses.

its attempt to interpret and satisfy an HTTP request from a client. An accepted method is to classify the status codes into four categories: *successful*, *found*, *not modified* and *unsuccessful*. A Web server returns a status code *successful* if the request is for a valid document, the client has permission to access the document, and the Web server is able to return that document to the client. The status code of *not modified* is returned if the client already has a copy of the requested document cached and that copy is current. If the requested document has a new location, either temporary or permanent, the Web server returns a status code of *found* along with the updated URL. Finally, a status code of *unsuccessful* is returned if an error is encountered at either the client or the server. Some possible error conditions include: a request for a nonexistent document, a request for a page the client does not have permission to access, and unavailability of the Web server.

The first invariant Arlitt and Williamson [Arlitt and Williamson 1996] observed in their study was that 78 – 93% of the status codes (an average of about 88%) indicated a successful transfer of a requested document. They also found that *not modified* status codes ranged from 4 – 13% (an average of about 8%) of the total requests, about 2% of the status codes indicated that the requested documents were *found*, and another 2% requests were *unsuccessful*.

We found this invariant to be much different for all our data sets. As shown in table 5, the *successful* status codes range from 68.2 – 81.9% for our datasets. This yields an average success rate of 75.5%, as against the 88% observed by Arlitt and Williamson [Arlitt and Williamson 1996]. The biggest contributor to this deviation arises from the difference in *unsuccessful* requests, which ranged from 7.2 – 14%, an average of 9.5%. Most

of the unsuccessful requests had status codes that indicated that client tried to request a forbidden document or requested a document that did not exist. While the CS data set had a much higher percentage of the status codes that indicated that the document was *found*, little else differed in this data set compared to those from scientific sites that could be attributable as a *trend*.

Item	HTML	Images	Video	Dynamic	Formatted	Other
Flybase 1						
% Requests	38.0	23.5	0.03	28.5	0.1	9.0
% Bytes transferred	26.2	4.2	1.9	34.2	0.2	33.2
Flybase 2						
% Requests	43.4	23.1	0.04	29.8	0.9	2.7
% Bytes transferred	33.8	3.8	2.7	42.1	0.3	17.2
ReciprocalNet						
% Requests	11.0	47.1	1.9	20.2	8.0	11.7
% Bytes transferred	15.1	19.2	2.2	11.6	7.3	44.6
CS						
% Requests	33.7	44.3	0.06	6.0	4.9	10.5
% Bytes transferred	21.9	47.5	4.8	2.1	10.7	10.9

Table 6: Document types and sizes.

We investigate the reasons behind this difference in the status codes in the subsequent analysis. For the rest of our analysis, we only consider those requests that successfully returned the desired content to the client. The summary characteristics of this reduced data set are given in table 4.

4.2 Second Invariant: File Types

Arlitt and Williamson [Arlitt and Williamson 1996] sorted the documents into the seven categories: *HTML*, *images*, *video*, *audio*, *formatted*, *dynamic*, and *other* and found that 89 – 99% of the client requests were for HTML and image documents combined. This was the second invariant in their study.

The Web servers we studied provide data to users in a variety of formats. For this reason, we observed a large variety of file types in our data. Placing postscript and PDF files in the ‘formatted’ category and code files like CGI scripts and Java files in the ‘other’ category, we obtained a much different distribution of file types for our data sets. Our results are summarized in table 6. The Flybase 1, Flybase 2, and ReciprocalNet access logs did not contain any references to audio files, so that category is omitted for these three data sets.

In contrast to Arlitt and Williamson’s observation, we found the percentage of HTML and image requests to be significantly lower, ranging from 58-78%. Also, images accounted for a disproportionately lower percentage of bytes transferred for scientific Web sites. Since their study was done in the early days of Web, Arlitt and Williamson predicted that this file type invariant would change over time. However, they expected the change to be due to the growth in the use of video and audio files. Our data sets indicate a slight increase in

the presence of video files for all four datasets and an increase in the percentage of audio files for the CS data set while the audio files are absent in all three scientific data sets. The primary contributor to the deviation in the file type invariant is the increased use of dynamic content, especially for the three scientific data sets.

Arlitt and Williamson’s second invariant could perhaps be revised to accommodate the CS data set because HTML and image comprise 78% of requests for this data set. However, extending their invariant to the scientific data sets would be a stretch. This is due to the extensive search functionality over diverse criteria that is common to many scientific Web sites, which leads to the strong presence of dynamic content (20 – 30% for our scientific data sets). We hypothesize that a new second invariant exists for scientific sites: *Dynamic content comprises 20–30% of total requests and HTML and images contribute less than 70% of requests for scientific Web sites*. Further research into the access patterns of other prominent categories of Web servers like those used to deliver news or by e-commerce sites is required to understand how changes in file types have evolved the second invariant for the entire Web.

Before we end the discussion on file types, we note that the growth in the *other* content type category has also contributed significantly to the deviation from the previously observed invariant on file types. We observed that the designers of scientific Web sites tended to define their own file extensions for some content and many URLs did not contain a file extension at all. For such cases, we looked at the actual URL in the application in order to classify the content. Much of this involved manual inspection and the end result was that we were able to classify 95–99% of the requests. The rest were placed in the *other* category. This is noteworthy because for scientific data sets the *other* category consumed a disproportionate amount of bandwidth (3 – 12% of the requests caused the servers to transfer 17 – 45% of bytes).

4.3 Third Invariant: Mean Transfer Size

The next invariant we looked at relates to the mean number of bytes transferred per request. The information about bytes transferred is recorded in the access log after the Web server has transferred the file to the client. Arlitt and Williamson observed [Arlitt and Williamson 1996] the mean transfer size to be 6 – 21 KBytes. They used this observation to define the mean transfer size < 21 KBytes as the third invariant.

Our data sets differed for this invariant as well. The mean transfer sizes for our data sets ranged from 14 – 38 KBytes, as shown in table 4. In particular, the CS data set had the highest mean transfer size of 38 KBytes. Since the deviation in this invariant is contributed most by the type of Web sites studied by Arlitt and Williamson (and not the scientific Web sites that host a

Item	Flybase 1	Flybase 2	ReciprocalNet	CS
Distinct Requests/Total Requests	15%	16%	8%	6%
Distinct Bytes/Total Bytes	13%	14%	7%	7%
Distinct Files Accessed Only Once	65%	69%	82%	33%
Distinct Bytes Accessed Only Once	65%	68%	81%	33%

Table 7: Statistics on distinct documents.

back end database), we conclude that this invariant has changed for today’s Web.

Table 4 also gives the coefficient of variation (CoV) of the mean transfer size. Arlitt and Williamson found that it varied from 3 – 11%, which indicates a high degree of variability in the transfer size. We found even more variation in the transfer sizes, with the CoV ranging from 17.6 – 22.4%. This is substantiated by the fact that 15% of the successful requests for Flybase 1 (24% for ReciprocalNet and 20% for CS) resulted in only one KByte or less being transferred. The increased mean transfer size and greater variation in transfer sizes can both be explained by changes in the type of documents being transferred, which was discussed in section 4.2.

4.4 Fourth Invariant: Distinct Requests

Caching could play an important role in enhancing the performance of a Web server. If a document is requested by multiple clients, it can be cached at the server or in the network to reduce client latency. To understand the possibility of caching for Web content, Arlitt and Williamson investigated the the client requests for *distinct* files and also the corresponding bytes transferred. They found for their traces that 0.3 – 2.1% of requests and 0.4 – 5.1% of the bytes were for distinct documents. They summarized this finding in the form of the fourth invariant which said that less than 3% of the requests are for distinct files.

We observed that for some of the files in our logs there were many different transfer sizes recorded. Without modification logs for the Web content and the back-end database, it is hard to tell whether this is the result of content modification or a partial transfer of content. For simplicity, we decided to assume that each distinct URL represented a distinct file. However, we accounted for various transfer sizes for each file by using the average number of bytes per file while calculating the distinct bytes accessed.

Our results, summarized in table 7, show a significant deviation from this invariant as well. The deviation is most pronounced for scientific Web site traces. In particular, for our four data sets, we observed that 6 – 16% of the requests and 7 – 14% of the bytes are for distinct files. This points to the limitation of the effectiveness of caching in today’s Web. We believe this difference can

be attributed to the presence of a high percentage of dynamic content, which leads to many documents being requested just once. The latter phenomenon, commonly referred to as *one time referencing*, is discussed in section 4.5.

4.5 Fifth Invariant: One Time Referencing

The fifth invariant is related to the fourth invariant discussed in section 4.4. It examines the percentages of files and bytes that are requested just once in the trace, a phenomenon referred to as *one time referencing*.

Arlitt and Williamson [Arlitt and Williamson 1996] observed that 23 – 42% of the distinct documents and 14 – 42% of the distinct bytes were observed only once in their data sets. This led them to conclude that approximately one third of the files and bytes are accessed just once, their fifth invariant. Just like the fourth invariant, this invariant also has implications for the effectiveness of content caching.

Our results for this invariant were very interesting. As shown in table 7, we found that this invariant held for CS data set, but not for the other three Web server scientific traces. In particular, 65 – 82% of files and 65 – 81% of the bytes for scientific traces are requested only once. The corresponding numbers for the CS data set were 33% each. This difference in behavior between the CS and scientific Web server traces can be attributed to the the presence of a higher percentage of dynamic content for the latter. Since the CS trace is closer in nature to the Web traces studied by Arlitt and Williamson, this finding leads us to conjecture that perhaps the fifth invariant holds in today’s Internet as well. However, it does not hold for the scientific Web servers. It appears that a more appropriate fifth invariant for the Web servers catering to scientific community would be: *at least two third of the files and bytes are requested just once*.

4.6 Sixth Invariant: Size Distribution

The next invariant Arlitt and Williamson [Arlitt and Williamson 1996] found in their data sets is related to the sizes of the files referenced at each site. They found that the file size distribution matches well with the Pareto distribution with $0.40 < \alpha < 0.63$. They also observed that most transferred files are in the range of

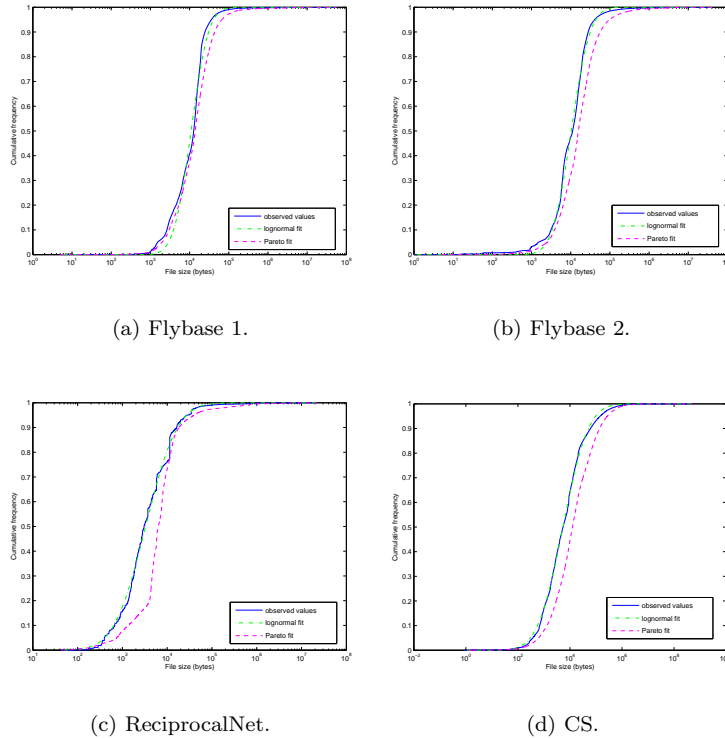


Figure 1: Distribution of file sizes.

0.1 to 100KBytes. We found that the latter observation held well for all four of our data sets as well, in that that majority of the transferred files had sizes in the range of 1 to 100KBytes. However, their sixth invariant did not hold for our data sets.

Figures 1(a), 1(b), 1(c), and 1(d) show the cumulative file size distributions for Flybase 1, Flybase 2, ReciprocalNet, and CS respectively. We show both a Pareto fit to our data and also a lognormal fit. We found that except for Flybase 1, where both Pareto and lognormal distributions fit almost identically, in general the lognormal distribution fitted better for all data sets. The α values for the Pareto fits shown in figures 1(a), 1(b), 1(c), and 1(d) range from 0.63 – 1.17. The lowest end of these α values is in accordance with Arlitt and Williamson’s findings and belongs to the CS data set.

4.7 Seventh Invariant: Concentration of References

Not all Web documents are equally popular. Arlitt and Williamson [Arlitt and Williamson 1996] captured the non-uniform referencing behavior by defining *concentration of references*. They sorted the list of distinct files into decreasing order based on how many times they were accessed. A plot of cumulative frequency of re-

quests versus the fraction of the total file referenced revealed another invariant. The seventh invariant, which relates to the fourth and fifth invariants discussed in sections 4.4 and 4.5 in terms of its impact on potential caching strategies says that 10% of files accessed account for 90% of requests and 90% of bytes transferred on an average. Arlitt and Williamson based it on the observation that 10% of the files accessed accounted for 80 – 95% of requests.

Our traces also showed significant non-uniform referencing behavior. For example, the root page for Flybase was accessed three times as many times as any other document since most users go through that page to get to other pages. In figures 2(a) and 2(b), we plot the cumulative frequency of requests against the fraction of the files referenced and percentage of bytes transferred respectively. We find that 10% of the requests account for 80 – 95% of the requests (an average of 83%) and 77 – 93% of the bytes transferred (an average of 83%) for our traces. We conclude that this invariant holds true for today’s Web with some revision.

4.8 Eight Invariant: Inter-Reference Times

The eighth invariant proposed by Arlitt and Williamson [Arlitt and Williamson 1996] was that file inter-reference times are exponentially distributed

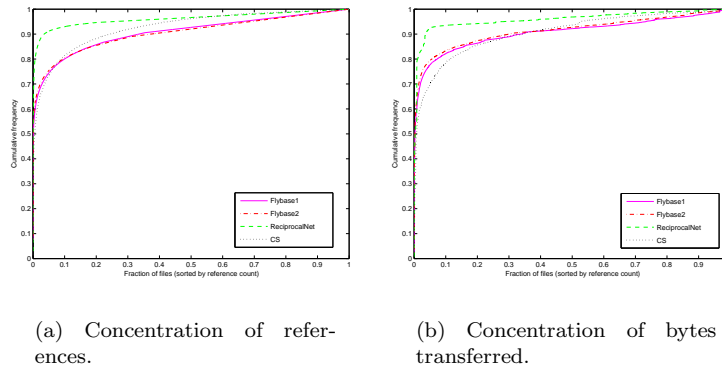


Figure 2: Concentration of references and bytes transferred (all data sets).

and independent. The inter-reference time is the time that elapses between subsequent requests for a file. For example, if file A is accessed at time i and then again at time j , then the inter-reference time is $j - i$. The inter-reference times for our data sets are shown in figures 3(a)- 3(d). On all servers, most documents were referenced at shorter intervals (minutes and hours) rather than days.

We found that our inter-reference times were exponentially distributed for three of our four data sets. Figures 3(a)- 3(d) show the cumulative distribution of the observed times along with the best exponential fit. As can be seen in figure 3(c), the ReciprocalNet data did not fit the exponential distribution well. Since the other sets are exponentially distributed, we conclude that the eighth workload invariant still holds in the Internet today.

4.9 Ninth Invariant: Remote Requests

The last two invariants are related to the number of remote requests versus local requests and the wide area usage. It was shown in [Arlitt and Williamson 1996] that the majority of requests to the Web servers come from remote clients, rather than local clients. In particular, Arlitt and Williamson observed that remote clients accounted for 54 – 99% of requests and 63 – 99% of the bytes transferred. Though one of their six traces was an exception, they concluded the following ninth invariant: remote clients account for $\geq 70\%$ of the requests to a Web server and $\geq 60\%$ of the bytes transferred.

We defined a local client to be a client with an IP address belonging to Indiana University and classified all others as remote clients. We observed that only 0.7% of the Flybase 1 clients (460 out of the total 63,535) were local. As shown in table 8, this resulted in 98% of the Flybase 1 requests coming from remote clients. Correspondingly, 97% of the bytes transferred were trans-

Local Hosts				
Item	Flybase1	Flybase2	ReciprocalNet	CS
% All Requests	2%	1%	2%	19%
% All Bytes	3%	1%	5%	27%
Remote Hosts				
Item	Flybase1	Flybase2	ReciprocalNet	CS
% All Requests	98%	99%	98%	81%
% All Bytes	97%	99%	95%	73%

Table 8: Geographic distribution of requests.

ferred to remote clients. These numbers increased to 99% for Flybase 2. Also, the corresponding numbers for ReciprocalNet were 98% and 95% respectively. The CS trace had the most local requests and bytes transferred and the corresponding percentages were 81% and 73% respectively.

Overall, the ninth invariant holds for all four of our traces. In fact, the scientific Web traces had much higher percentage of remote requests and bytes transferred compared to the CS traces. This leads us to conclude that a tighter ninth invariant for scientific Web servers may be stated as: *remote clients account for at least 95% each of the requests and bytes transferred for scientific Web sites.*

4.10 Tenth Invariant: Wide Area Usage

The final workload invariant we examined involves the wide-area usage of Web servers. Arlitt and Williamson [Arlitt and Williamson 1996] found that their Web servers were accessed by thousands of domains, with 10% of the domains accounting for 75–85% of usage. This led them to coin the tenth invariant which states that 10% of the domains account for $> 75\%$ of usage.

In order to see whether our data sets conform to the tenth invariant, we classified the clients into domains

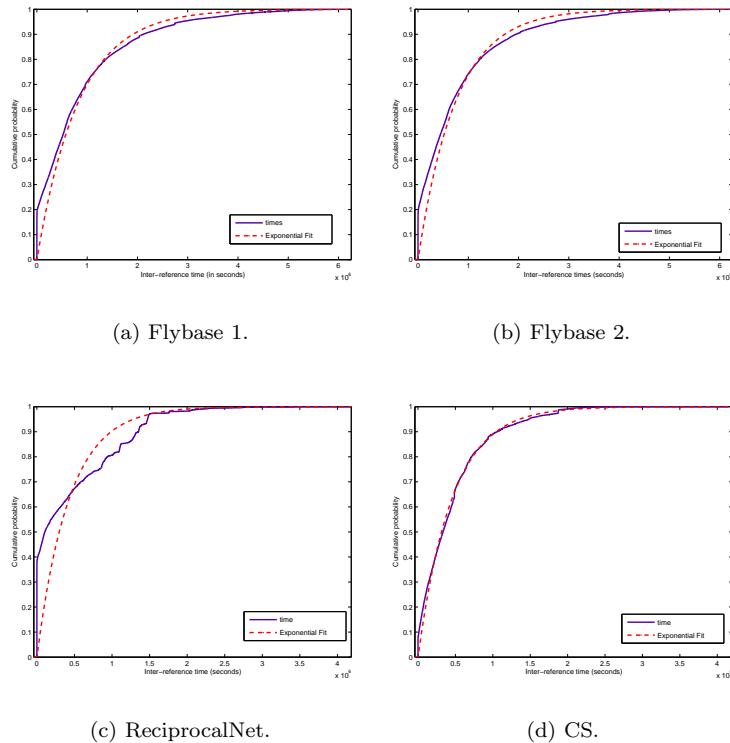


Figure 3: Inter-reference times.

based on their IP addresses¹ and determined which domains sent the most requests to the Web server. Our results, shown in table 9 differed somewhat from this invariant. While 10% of the domains accounted for 92% of the usage for both Flybase 1 and CS, the corresponding numbers for Flybase 2 and ReciprocalNet were only 69% and 70% respectively. The difference in the wide area usage for Flybase 1 and Flybase 2 data sets can perhaps be attributed to the change in client profile due to the increase in the popularity of the Flybase Web site over the two snapshots we analyze in this paper². This is corroborated by the observation that the usage was distributed more evenly among domains for Flybase 2. The top ten percent of domains still account for the majority of usage, but not nearly as much as was seen in the Flybase 1 logs.

We hypothesize that though Arlitt and Williamson’s tenth invariant holds for CS-like Web sites, it needs to be revised for scientific sites. More research is needed to determine if 10% of the domains account for $\approx 70\%$ of the usage for scientific sites a new appropriate tenth invariant.

¹The domain name was not always present in our Apache logs. Sometimes, only the IP address was present but we were able to do successful reverse DNS lookups for 95% of the IP addresses.

²as shown in table 3, Flybase 2 saw an increase of 10% in the average number of requests per day.

Data Set	Wide area usage
Flybase1	92%
Flybase2	69%
ReciprocalNet	70%
CS	92%

Table 9: Usage by top 10% of domains (all data sets).

Another interesting aspect about the scientific Web sites relates to the domains of their clients. We expected that most clients would access Flybase and ReciprocalNet from “.edu” top level domain (TLD), but that does not seem to be the case. Most of the top domains for Flybase were ISPs for home and small business users, indicating that many users access Flybase from home. Also, many clients are based in Europe and Japan. This is surprising, considering that Flybase has mirror sites in the United Kingdom, France, and Japan which may them a better response time. Its possible that they do not know that the mirrors exist, or that using the Indiana University Flybase does not actually cost much in terms of latency. Similarly, only 30% of the top domains accessing ReciprocalNet were from “.edu” TLD, of which Indiana University had the most usage. As was the case with Flybase, the top domains consisted

Data Set	Weekday Usage	Weekend Usage
Flybase1	84%	16%
Flybase2	86%	14%
ReciprocalNet	80%	20%
CS	74%	26%

Table 10: Weekday versus weekend usage (all data sets).

mainly of ISPs for home and small business users.

5 New Invariants

We have shown that only a few of the workload invariants previously discovered still hold as-is. The majority must be modified to accommodate different kinds of workloads and changes in Internet traffic in general, especially for scientific workloads. This section describes the three new invariants we derived with particular emphasis on the scientific Web sites. The first describes the weekday versus weekend usage of Web servers, the second explores the presence of diurnal pattern in client accesses, and the third involves the concentration of usage among clients.

5.1 Weekday versus Weekend Usage

We examined each of our four data sets to determine if weekday and weekend usages of our Web sites differed. In general, we found that Tuesday was the busiest day three scientific data sets in terms of the number of requests. Surprisingly, it was the least busy day for the CS data set. Wednesday was the busiest day for CS and Saturday was the least busy day for all three scientific data sets. Overall, we found the number of requests to be higher on the weekdays than weekends, especially for the three scientific data sets.

Table 10 shows the average weekday and weekend usages for all four data sets and figures 4(a)- 4(d) show the requests per day for each of the four data sets.. Notably, the CS data set shows the least difference between weekday and weekend usage. However, for scientific data sets, a much higher percentage of the requests occur on weekdays. Based on these results, we propose a new invariant for the scientific Web servers: *> 80% of scientific Web site usage occurs on weekdays.*

5.2 Diurnal Usage

The second invariant we studied relates to the day time versus night time usage of Web servers. One would expect that the Web sites would get used more heavily during the day than at night. To determine this conclusively, we looked at requests from the United States, Mexico, and Canada and defined daytime to be

Data Set	Usage by Top 25% of Clients
Flybase	92%
ReciprocalNet	75%
CS	85%

Table 11: Client Concentration (All data sets).

7 : 00AM – 7 : 00PM Central Time. This time interval was chosen because it encompasses the start of the work day on the east coast (8 : 00AM) and the end of the work day on the west coast (5 : 00PM). By limiting ourselves to only clients from the United States, Canada, and Mexico we minimized the effect of different time zones on our results.

Figures 5(a), 5(b), and 5(c) show the number of requests during the day and night for each day of the logs for Flybase 1, ReciprocalNet, and CS data sets respectively. We notice that while the CS data set shows several cross-over points, the scientific data sets clearly show the diurnal usage pattern. Hence, we conclude a new invariant for the scientific Web servers: *usage is diurnal for scientific Web sites.*

5.3 Client Concentration

Our third new invariant delves deeper into the wide area usage invariant proposed by Arlitt and Williamson [Arlitt and Williamson 1996] (the tenth invariant described in section 4.10) by examining the concentration of references by particular clients rather than entire domains. The question we were trying to answer was: are client requests evenly distributed among clients or do a small percentage of the clients issue the majority of the requests?

We discovered that the latter is true. Our results, summarized in table 11, showed that 75 – 92% of the total requests come from only 25% of the clients. For this invariant we looked at Flybase 1 and Flybase 2 together so they are shown in table 11 as simply “Flybase”. We hypothesize that the client concentration invariant for the Web is: *25% of the clients account for at least 75% of the requests.*

Looking at the client concentration also revealed that a large number of requests are issued by crawlers. Out of curiosity, we looked at the number of requests issued by Google [goo] crawlers in particular and found that they account for 2 – 4% of the total requests! Under heavy demand, the client concentration invariant could be a useful invariant to rate-limit the requests served to crawlers.

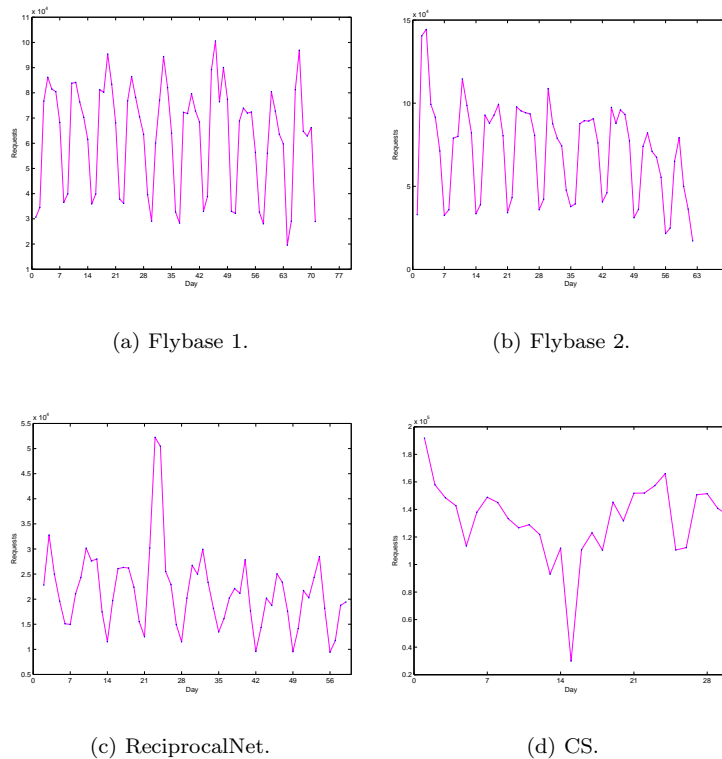


Figure 4: Requests per day (all data sets)

6 Concluding Remarks

This study examined the ten Web invariants coined by Arlitt and Williamson more than a decade ago in the context of scientific Web servers and suggested revisions where applicable. With the goal of anomaly and misuse detection in mind, it derived three new invariants for the scientific Web sites.

Many future directions arise out of this work. In general, the workloads examined here were not sufficient to answer questions like: *what invariants exist today for the entire Web, if any?, do different invariants exist for different categories of Web servers (scientific, e-commerce, news)?* More workloads from a variety of Web servers are required to answer these questions.

Also, further research is required to understand how invariants can be utilized to achieve the goal of anomaly and misuse detection. In particular, it remains to be investigated how tools can be developed to use the knowledge about invariants such as client concentration, request arrival rate, file types requested etc. We plan to pursue these avenues in our future work.

Acknowledgments

We would like to acknowledge the help of the following colleagues in acquiring various data sets: Don Gilbert for Flybase logs, John Bollinger and Randy Bramley for ReciprocalNet logs, and Rob Henderson for CS logs.

References

- ACHARYA, S., SMITH, B., AND PARNES, P. 1998. Characterizing user access to videos on the world wide web. In *SPIE/ACM MMCN*.
- ALMEIDA, J., KRUEGER, J., EAGER, D., AND VERNON, M. 2001. Analysis of educational media server workloads. In *ACM NOSSDAV*.
- Apache Web page. <http://www.apache.org/>.
- ARLITT, M. F., AND WILLIAMSON, C. L. 1996. Web server workload characterization: The search for invariants. In *ACM SIGMETRICS*.
- BARFORD, P., AND CROVELLA, M. 1998. Generating representative web workloads for network and server performance evaluation. In *ACM SIGMETRICS*.

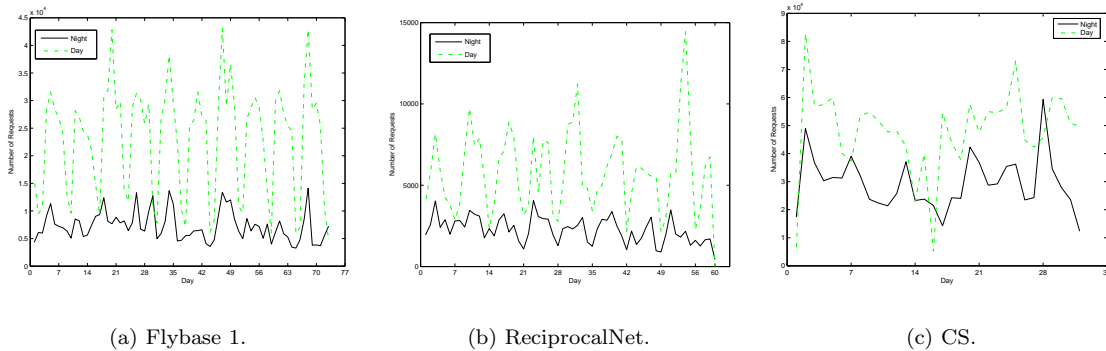


Figure 5: Diurnal usage.

BENT, L., RABINOVICH, M., VOELKER, G. M., AND XIAO, Z. 2004. Characterization of a large website population with implications for content delivery. In *ACM WWW Conference*.

CHERKASOVA, L., AND GUPTA, M. 1998. Characterizing locality, evolution, and life span of accesses in enterprise media server workloads. In *ACM NOSS-DAV*.

CHESHIRE, M., WOLMAN, A., VOELKER, G. M., AND LEVY, H. M. 2001. Measurement and analysis of a streaming media workload. In *USENIX Symposium on Internet Technologies and Systems (USITS)*.

CROVELLA, M. E., AND BESTAVROS, A. 1997. Self similarity in world wide web traffic: Evidence and possible causes. *IEEE Transactions Of Networking* 5, 6, 835–846.

ReciprocalNet Web page. <http://reciprocalnet.org/>.

DOYLE, R. P., CHASE, J. S., GADDE, S., AND VAHDAT, A. 2001. The trickle-down effect: Web caching and server request distribution. In *International Workshop on Web Caching and Content Distribution*.

FIELDING, R., GETTYS, J., MOGUL, J., FRYSTYK, H., MASINTER, L., LEACH, P., AND BERNERS-LEE, T., 1999. Hypertext transfer protocol – http/1.1. IETF RFC 2616, June.

Flybase Web page. <http://www.flybase.net/>.

Google Web page. <http://www.google.com/>.

HAREL, N., VELLANKI, V., CHERVENAK, A., AND ABOWD, G. 1999. Workload of a media-enhanced classroom server. In *IEEE Workshop on Workload Characterization*.

Mysql Web page. <http://dev.mysql.com/>.

OF DYNAMIC WEB CONTENT, M. O. C. 2003. W. shi and e. collins and v. karamcheti. *Journal of Parallel and Distributed Computing* 63, 10, 963–980.

PADHYE, J., AND KUROSE, J. 1998. An empirical study of client interactions with continuous-media courseware server. In *ACM NOSS-DAV*.

PADMANABHAN, V., AND QUI, L. 2000. The content and access dynamics of a busy web site: Findings and implications. In *ACM SIGCOMM*.