Electronic Commerce Research, 3: 167–192 (2003)
© 2003 Kluwer Academic Publishers. Manufactured in the Netherlands.

# Characterization of E-Commerce Traffic

UDAYKIRAN VALLAMSETTY uday.vallamsetty@amd.com
*Advanced Micro Devices, Sunnyvale, CA, USA*

KRISHNA KANT krishna.kant@intel.com
*Intel Corporation, Hillsboro, OR 97124, USA*

PRASANT MOHAPATRA prasant@cs.ucdavis.edu
*Department of Computer Science, University of California, Davis, CA 95616, USA*

*Abstract*

The World Wide Web has achieved immense popularity in the business world. It is thus essential to characterize the traffic behavior at these sites, a study that will facilitate the design and development of high-performance, reliable e-commerce servers. This paper makes an effort in this direction. Aggregated traffic arriving at a Business-to-Business (B2B) and a Business-to-Consumer (B2C) e-commerce site was collected and analyzed. High degree of self-similarity was found in the traffic (higher than that observed in general Web-environment). Heavy-tailed behavior of transfer times was established at both the sites. Traditionally this behavior has been attributed to the distribution of transfer sizes, which was not the case in B2C space. This implies that the heavy-tailed transfer times are actually caused by the behavior of back-end service time. In B2B space, transfer-sizes were found to be heavy-tailed. A detailed study of the traffic and load at the back-end servers was also conducted and the inferences are included in this paper.

**Keywords:** bussines-to-business (B2B), business-to-clients (B2C), e-commerce servers, front-end and back-end servers, traffic characterization

## 1. Introduction

The explosive popularity of Internet has propelled its usage in several commercial avenues. E-commerce, the usage of Internet for buying and selling products, has found a major presence in today's economy. E-commerce sites provide up-to-date information and services about products to users and other businesses. Services ranging from personalized shopping to automated interaction between corporations are provided by these Web-sites. It has been reported that e-commerce sites generated $132 billion in 2000, more than double of the $58 billion reported in 1999 [18]. Even though the power of the servers hosting e-commerce sites has been increasing, e-commerce sites have been unable to improve their level of service provided to the users. It has been reported that around $420 million has been lost [7] in revenues due to slow processing of the transactions in 1999. Thus it is desirable and necessary to focus on the performance of the servers used in these environments.

There are two main classes of e-commerce sites, Business-to-Business (B2B) and Business-to-Consumer (B2C), providing services to corporations and individual users, respectively. Web sites like Delphi, which provide services to corporations like General Motors come under B2B sites, whereas sites like Amazon.com providing services to general users come under B2C sites. Most of the revenue generated by B2C sites will be during holiday season when the load on the system is maximum. During such periods, aggregate arrival rates that are orders of magnitude higher than normal rates will be observed at servers. This load results in response-times orders of magnitude higher than the normal conditions, causing users to leave the site. Similarly, new products or other events may trigger surges in the B2B environment.

Considering the revenues involved in e-business, availability and performance of the e-commerce servers become the two most important issues. Service agreements with corporations or individual users have to be honored to gain or retain customers. Server overload can seriously compromise the availability and performance of the servers. To avoid these unwarranted situations, admission and overload control schemes have to be implemented at the server. Additionally load balancing is a popular approach for enhancing the performance. For these techniques to be effectively implemented, a good understanding of the workload is required. Due to the complex nature of e-commerce traffic and the access restrictions, there has been limited work reported where e-commerce traffic is modeled for synthetic generation. As a first step towards this goal, aggregate traffic arriving at the server has to be analyzed to understand its behavior. This effort would require a complete characterization of the real workload seen at typical e-commerce sites. Our preliminary work has indicated significant differences in the traffic characteristics of e-commerce and general Web servers.

In this paper, we have analyzed the characteristics of e-commerce traffic and propose techniques for reducing the burstiness in the response-times. Traffic from a B2C and a B2B site is being used for the study. The workload is initially inspected for understanding the diurnal nature of the traffic. Different load periods were identified for both the B2C and B2B environments. These have been found to be complimentary in nature, which may be intuitive. A set of parameters were chosen for each site for each component which would impact the performance of the system to the maximum extent. Statistical tests are then used to prove the self-similar nature of the traffic at different scales. Two different tests are used for validating the results for each of the parameters. It has been observed that the arrival traffic is highly bursty in nature, much more than the burstiness seen in normal Web-traffic [Crovella and Bestavros, 5]. The response-time distribution is found to be heavy-tailed. This has been previously attributed to the heavy-tailed nature of request and response file-sizes. But the behavior of transfer sizes is not heavy-tailed, unlike the general Web-environment. The traffic arriving at the back-end servers is characterized to obtain similar statistics about the impact of burstiness on the system. Also preliminary tests have shown that the back-end utilization is more bursty than the front-end server utilization, the reasons for which are explained later. A correlation is drawn between the behavior of the front-end and the back-end servers under different load conditions. Performance implications from the results of the above experiments will give valuable information for improving e-commerce server performance.

The workload characterization studied in this paper is based on one representative system from each of the environments (B2C and B2B). Considering the difficulty in obtaining this valuable and guarded information from the e-commerce sites, and the fact the sites we have considered are quite busy, the results, although preliminary, could be valuable for future studies on e-commerce workload characterization and server designs.

The rest of the paper is organized as follows. Related work is outlined in Section 2. Section 3 discusses the architecture of e-commerce sites along with a description of the configuration of the sites used for this study. Section 4 discusses the behavior of the workload and the traffic and load characteristics of the front-end and back-end servers. The concluding remarks are sketched in Section 5.

*Note.* For this study data from two popular sites (one B2C and the other B2B) was used. Due to a non-disclosure agreement (NDA), the identities of these sites are not revealed. Throughout this paper the two sites are identified as B2C site and B2B site. Without the NDA, we would not have been able to acquire the data for the study.

## 2. Related work

Although there has been several works reported on the workload characetrization of general Web servers [Bradford and Crovella, 2; Chen, Chen, and Mohapatra, 3; Kant and Won, 11; Menasce and Almeida, 14], only a few studies have been reported on the characterization of e-commerce traffic based on the client behavior. In [Menasce et al., 15] the authors have developed a resource utilization model for a server which represents the behavior of groups of users based on their usage of the site. This model represents the usage of each resource in the Web server. The states of the model can be defined as a specific operation performed by a client at the server. The operations performed by the client can be strictly divided based on the resources used by that operation. This division would generate a model giving information about different resources utilized by the client at any instant for a given session. Resource utilization model for server will have an average usage information over all the clients. This kind of model would help in comparing the performance of two servers given two users at different load conditions. Most of the work done in this area has not looked at the actual traffic from e-commerce servers.

Some studies on e-commerce sites [Diffie, 6; Glushko, Tenenbaum, and Meltzer, 9; Ma, 13] have looked at various features of emerging technologies for providing increased security and accessibility to users. In this environment, the revenue generated has the highest importance and hence most of these studies have concentrated on increasing revenue generation by introducing new applications or technologies which would increase the performance of e-commerce sites based on the revenues generated per transaction.

Not much work has been attempted in e-commerce traffic characterization. The main reason for this shortcoming is the unavailability of representative data. E-commerce sites have highly secure and personalized information in the traces and access logs. Due to the security implications e-commerce sites are reluctant to divulge this information for research purposes. As a result, studies in this field are still in the preliminary stages.

It should also be noted that the existing work reported on e-commerce traffic has been done on the front-end servers only and to the best of our knowledge nothing has been reported on the back-end servers. The back-end servers are the ones which experience the maximum load in an e-commerce environment [4]. We would like to characterize the load on the back-end servers along with a study of the system characteristics collected from system logs in e-commerce sites. This would give an opportunity to study the correlation between the traffic arriving at the front-end and back-end of e-commerce servers. Further such a study would also provide an understanding of the impact of the front-end on the traffic arriving at the back-end.

## 3. E-commerce architecture

A generic organization of e-commerce sites is depicted in Figure 1. E-commerce sites can be broadly classified into two different categories:

- Business to Business (B2B).
- Business to Consumer (B2C).

The main difference between the above two categories of sites lies in the user population accessing these sites. Business-to-Business e-commerce sites serve transactions between different businesses whereas Business-to-Consumer sites serve general users over the Internet.

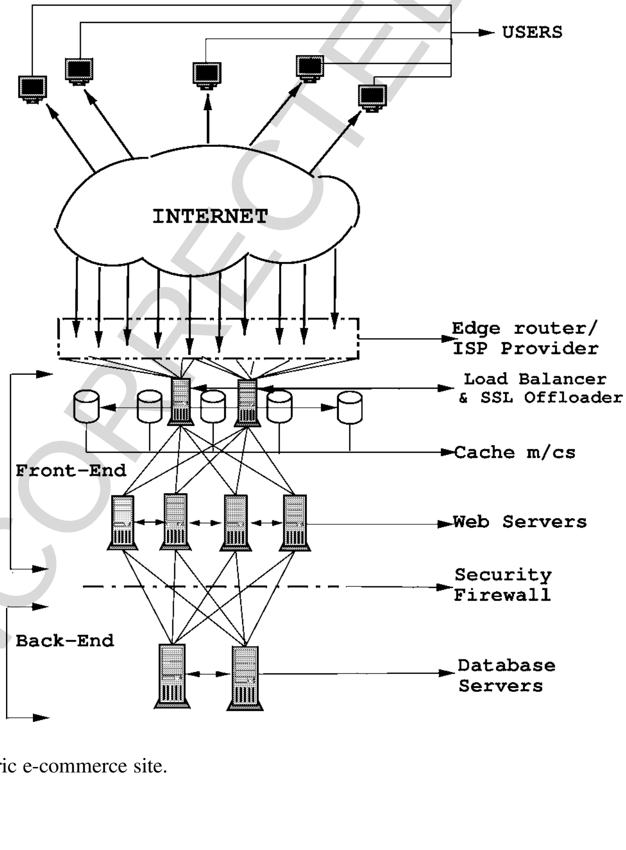

*Figure 1.* A generic e-commerce site.

*Business-to-Business.* One of the main characteristics of this category of sites is the regularity in the arrival traffic. It was observed that heavy traffic comes between 9 am and 5 pm, normal business hours. Regularity does not imply the lack of heavy spikes in the traffic. There will be sustained load on the system either due to seasonal effects or due to the availability of different services at the site. These sites can be categorized by the high amount of buying taking place in them. It has been observed that the percentage of transactions resulting in buying are very high compared to those in B2C environment.

Our preliminary analysis have revealed some very important features of B2B e-commerce sites. In B2B space, the population of users accessing a specific set of servers is known a-priori, along with the kind of transactions that will be taking place. This enables the designers to customize these sites to specific users, for specific transactions. With this information, the response time can be improved keeping the load on the system balanced evenly among all the different servers in the site.

*Business-to-Consumer.* B2C servers are the normal e-commerce sites where any user can get service. The security involved in B2C site is only restricted to any financial transactions involved, whereas in a B2B environment all the transactions are normally done in secure mode. One implication of this is that increased buying in a B2C environment can throttle the system since the designed system does not expect high percentage of buy transactions. Another important characteristic of a B2C site is the very low tolerance to delayed responses. This increases the need to make quality of service more important than providing absolute security for all the transactions, hence security is reserved for transactions involving buying.

### *3.1. System configuration*

In this section we will discuss the configuration of a typical e-commerce site. Since the general configuration of both B2B and B2C servers is similar, a typical configuration is described. The differences observed between the two types of sites will be noted as the discussion progresses. Any e-commerce server environment consists of two main parts, the front-end and the back-end network. The front-end consists of Web and application services accessible by the users over the Internet. Back-end consists of the security firewalls and the database servers.

***3.1.1. Front-end servers*** Typically the front-end servers are comprised of the Web server, application server, server load balancer, and the secure socket layer (SSL) offloader.

Front-end Web servers host the actual site content that clients see on their Web browsers. They serve different types of requests from the clients, comprising of static content, graphics, or dynamic content. In addition, Web servers are the only authorized hosts able to access the back-end database and application services as necessary. The application servers are responsible for the business logic services. The application server will be the most heavily loaded server in the B2C environment. This is due to the heavy traffic of dynamic and secure requests arriving at the server. In a large scale e-commerce site, there will be dedicated application servers, alternatively these servers can be combined with the Web servers or the database servers. The decision is based on how the Web server, business logic, and database services communicate. If the Web servers make many small requests to the business servers then it probably makes sense to move the services closer together. Conversely, if the business servers process lots of data into small results then you can move the business logic closer to data. Additionally, the placement of application servers influences scalability, high availability, and security. In most B2B environments, the application server is separately maintained, both for scalability and security reasons.

Due to the heavy traffic seen by e-commerce servers and also due to the availability requirements, there will be a network of Web servers instead of a single monolithic server at the front-end. This basically improves the scalability and fault-tolerance of the server to any bursts of busy traffic. Load balancers help increase the scalability of an e-commerce site. Load balancing works by distributing user requests among a group of servers that appear as single virtual server to the end user. Its main function is to forward user traffic to the most available or the "best" server that can provide a response to the user. Load balancers use sophisticated mechanisms to detect the best server. These mechanisms include finding the server with the least connections, the least load, or the fastest response times. Ultimately, server load balancing helps maximize the use of servers and improves the response times to end users.

SSL is a user authentication protocol developed by Netscape using RSA data security's encryption technology. Many commerce transaction-oriented Web sites that request credit card or personal information use SSL. The SSL off-loader typically decrypts all http's requests arriving at the server. It should be noted that the link between the front-end and the back-end servers is fully secure. So all the secure transactions are decoded by the security machine at the front-end before being sent to the back-end servers.

***3.1.2. Back-end servers*** The back-end servers mainly comprise of the database servers and the firewall which would protect sensitive data from being accessed by unauthorized clients. These fire walls provide security services through connection control. They are predominantly used when protecting mission-critical or sensitive data is of utmost importance.

The database servers reside in the back-end of the network and house the data for e-commerce transactions as well as sensitive customer information. This is commonly referred to as the data services. The clients do not directly connect to these servers, the front-end Web servers initiate connections to these servers when a client conducts a series of actions such as logging in, checking inventory, or placing an order. Most e-commerce sites scale up their database servers for scalability and implement fail-over clustering for high availability. Partitioned databases, where segments of data are stored on separate database servers, are also used to enhance scalability and high availability in a scale-out fashion.

### *3.2. B2C configuration*

A simplified configuration of the B2C site being used for the study is given in Figure 2. The site comprises of ten Web servers, each one powered by a Intel Quad P-III systems with a 512 MB of RAM. The Web servers run IIS 4.0 HTTP server. This cluster of Web servers is supported by three image servers, each one powered by a Dual P-II system. As can be seen from the figure, the image servers serve both the database servers and the front-end Web servers. For the purpose of our study, the image servers were considered to be in the back-end system. The product catalog server, connected to both the front-end and the back-end, runs an NT 4.0 providing backup and SMTP services to the back-end servers. The LDAP server is connected to the back-end.

There are three different types of database servers at the back-end, the customer database, the membership database and the catalog database. Only the customer database and catalog database are being used for this study. There is very minimal traffic coming to the membership database hence this was not used. Each of the databases, have NT 4.0, running SQL Ent. 7.0 SP1. There are some other component of the site which are not shown in the figure. These are the components which will be used by specific hosts or are used for security or scalability of the site. It has to be observed that the B2C site is very similar to the generic e-commerce site model shown in Figure 1. Since the user population that demand services from a B2C site is not well defined, the site is designed to sustain a variety of load conditions and user behaviors.

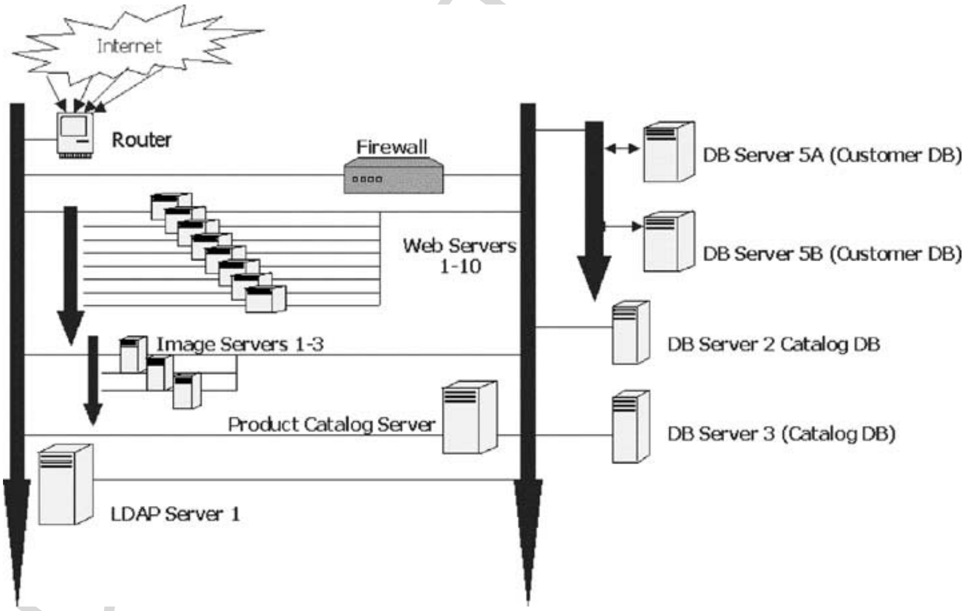

*Figure 2.* Simplified configuration of the B2C site.

### *3.3. B2B configuration*

In the B2B space, the design of e-commerce sites is completely different from their design in B2C space. Here the user population is known a-priori. The transactions being processed by each user arriving at the server is also known with reasonable bounds. B2B sites serve a limited population as opposed to B2C sites which aim at serving the entire Internet. These aspects enable the designers to customize the site to specific user requirements.

Scalability is one of the main issue that has to be taken care of when designing such customized system. So the design is done as a cluster of B2C sites, interconnected to form a large B2B portal. The interconnections between the individual B2C components in the site determine the user population to that site and also the services provided by that site. Figure 3 shows a simplified version of the B2B site being used for the study. Each of the Web servers, can be individually used as a B2C site with its own database and network connection.

An important feature of B2B sites is the accessibility constraints on the users. Here the access to the Web servers is restricted by login/authentications machines which do load balancing along with directing the traffic to the appropriate servers. Another important feature in a typical B2B site is the migration of the application servers from the Web servers closer to the database servers. Because of independent authentication machines,

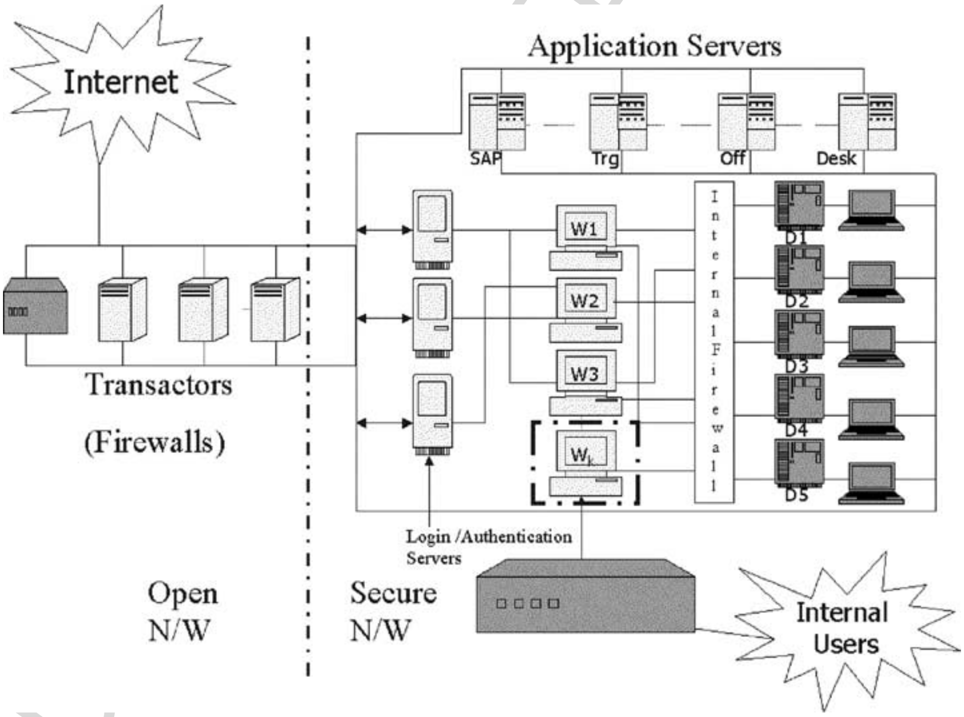

*Figure 3.* Simplified configuration of the B2B site.

traffic to the application servers can be sent directly without loading the Web servers. Even though this would depend on the services provided by the site, application servers are separated due to their usage internally for training etc. and for promotional usage to other corporations.

## 4. Workload characterization

In this study we have analyzed the behavior of e-commerce servers with relation to the behavior of the incoming traffic. Various data was collected at different levels in the system. Web server access logs from the the front-end and the back-end servers were collected at a granularity of 1 sec. This is an application level data giving the load on the httpd. This data will give the characteristics of the traffic arriving at the system, average network bandwidth utilization, and the file transfer rate.

For the system level information, data was collected from the performance logs [16] from all the servers present in the site. This data was collected at a granularity of 5 sec. This data provides information about the I/O bandwidth used, the processor and disk utilization of the system etc. As the data was collected at a constant rate of 5 sec intervals, it was at a higher scale than the logs from the Web servers. However, both the scales are below the non-stationarity time scale used for the analysis.

Data was collected at the server and the performance monitor for an entire day. A weekday is used for data collection since this would represent the average traffic. Additionally, data for a five day period was used to study the average behavior of the traffic over a long period of time.

### *4.1. Characteristics of the workload*

In this work we are looking at Web workload comprising of e-commerce traffic. The main differences between general Web and e-commerce workload are the following.

1. Presence of a high level of Online Transaction Processing (OLTP) activity is observed among the transactions at the server. This is due to the database transactions accruing for every request from the user. Due to security reasons most of the data is present in the database server which is protected by a secure firewall. This prevents the Web server from responding to most of the requests without sending a query to the back-end server.
2. Along with the database activity a large proportion of requests come in secure mode. B2C traffic has lesser secure traffic, B2B sites experience almost complete secure traffic from users. This is due to the heavy security constraints present in industry to industry transactions. Increased amount of secure transactions implies heavy processing at the front-end server. Most of the sites have SSL off-loaders, which do encryption/decryption of requests to reduce the load on the system. This process adds to the response time. Aggregating these transactions with normal transactions increases the variability in the response times observed by the user.

3. The proportion of dynamic requests (that require some amount of processing) is very high, as was expected. In fact, in most e-commerce sites almost all requests are handled as dynamic requests.

### *4.2. Front-end characterization*

A visual inspection reveals the workload at e-commerce sites to be more bursty than normal Web workload. To study this behavior, the following parameters were used, which would have the maximum impact on the behavior of the traffic: arrival process, utilization of the server, response time, request file sizes, and response file sizes.

***4.2.1. Arrival process*** Figures 4 and 5 show the arrival process at the B2C and B2B e-commerce sites. The data shows traffic on a normal weekday with an average arrival

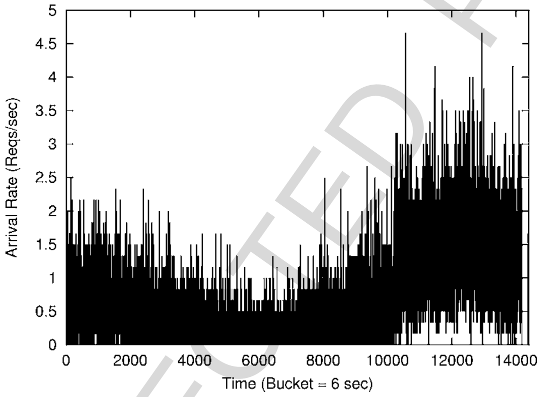

*Figure 4.* Arrival process at B2C site, 6 seconds granularity.

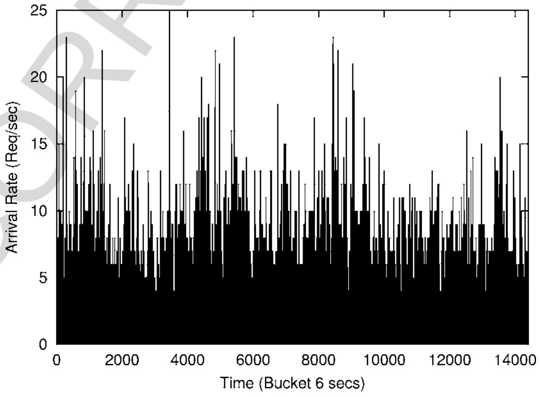

*Figure 5.* Arrival process at B2B site, 6 seconds granularity.

rate of 0.65 requests/sec at the front-end Web server for the B2C site and around 1 req/sec arrival rate at one of the Web servers in the B2B site. A visual inspection reveals the burstiness in the arrival process. The B2C server is a 4P system with an average processor utilization of 6% per processor and disk utilization of 2% during the period starting from 9.00 am till 6.00 pm. The low utilization is typical of e-commerce sites since they are designed for much higher load and sustain a very minimal load during normal working periods. It is the high load periods showing bursts of orders of magnitude more than normal operating parameters which cause concern for better capacity planning and performance analysis of these systems.

Figures 4 and 5 show that the sites have distinct high and low load periods during the course of a day. For the B2C site, busy period starts around 6.00 pm in the evening and ends at around 11.00 pm in the night. Since this is a B2C site serving general consumers, the traffic is heavy during the after-office periods. Distinctive low periods during the morning between 7.30 am to 11.30 am can also be observed. In case of the B2B site, the traffic concentration lies mostly during normal office hours, between 9.00 am and 8.00 pm, which is intuitive. It should be noticed that the graphs show aggregated arrival traffic for the B2B site and the averaged arrival process for the B2C site.

The Arby–Veitch (AV) [Veitch and Abry, 20] estimator test was used for estimating the Hurst-parameter (H-parameter) [Kant, 10] for the arrival time-series. This is known to be a reliable test for workloads with busy periods showing a non-stationary behavior. E-commerce workload is influenced by busy periods caused by different sales promotions and seasonality. Hurst parameter is also calculated using the R/S plot test [Kant, 10]. Reliability of this test under low time-scales for e-commerce traffic is tested by comparing the H-parameters obtained using the two methods.

Figure 6 used the arrival time-series at the B2C site with a granularity of 3 sec. The H-parameter is estimated to be 0.662. This shows that the arrival process at the B2C site is *self-similar* in nature.

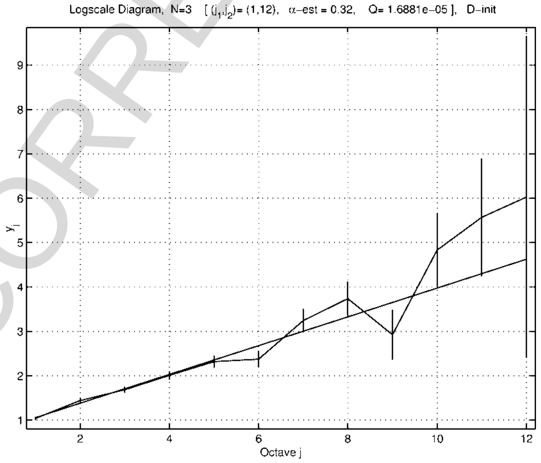

*Figure 6.* AV estimator test for self-similarity for arrival process.

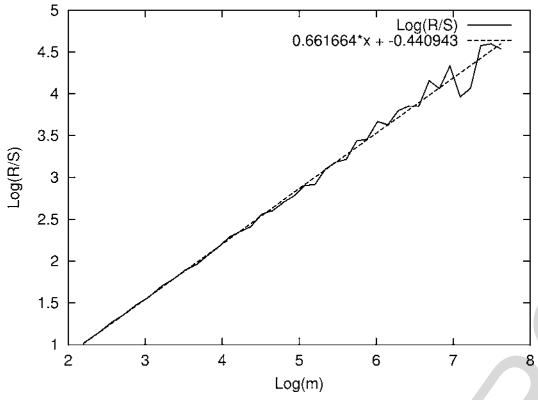

*Figure 7.* R/S plot test for self-similarity for arrival process at B2C site.

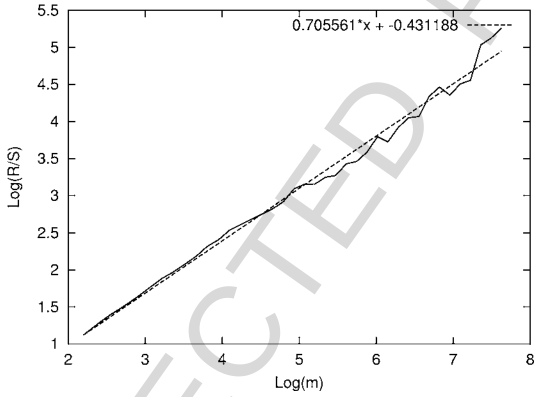

*Figure 8.* R/S plot test for self-similarity for arrival process at B2B site.

In Figures 7 and 8 the log–log plot of the rescaled range ratio is shown for the B2C and B2B arrival traffic, respectively. The Hurst parameter is estimated to be 0.662 using a linear-regression line through the R/S points for the B2C site, which matches the estimation made by the AV-estimator. Similar test was done for the arrival traffic at the B2B site. Using the AV-estimator the H-parameter was estimated at 0.69, whereas the R/S plot gave an estimate of 0.70 for the H-parameter, which is a good approximation. From the above results it can be conjectured that the arrival series does not show any non-stationary behavior at lower time-scales.

***4.2.2. Processor utilization*** Figures 9 and 10 show the utilization of the front-end Web server for the B2C and B2B sites, respectively. As explained earlier, the data is collected between 9.00 am till 5.00 pm at a granularity of 5 secs for the B2C site. For the B2B site the data represents the activity between 10.00 am in the morning till 9.30 am the next day morning. The B2C server sustains a constant load throughout the day, with an average load

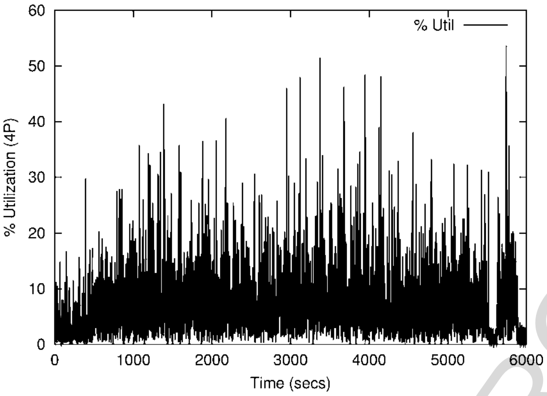

*Figure 9.* Utilization at the front-end Web server (4P), B2C.

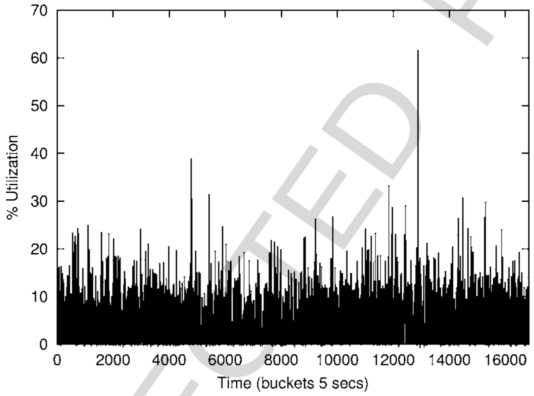

*Figure 10.* Utilization at the front-end Web server (2P), B2B.

of 7% on each of the four processors. High and low load periods can be observed on the B2C server during the course of the day. This behavior is absent in the B2B server. This is due to the a-priori knowledge of the transactions and load from users in the B2B space. B2B sites are customized for specific traffic patterns and a normal traffic would not affect the load on the system to a higher degree. Thus the load on the system appears almost constant even though there is a variation in the arrival rate at the server. The time-series obtained from the utilization was also tested for self-similar behavior. The AV-wavelet based test and the R/S plot test are used for estimating the H-parameter (Figures 11 and 12). The estimated H-parameter is 0.755 using the AV estimator, and 0.77 using the R/S plot test for the B2C site. In the B2B space, the load on the system did not have a high degree of self-similarity. The H-parameter is estimated to be 0.66 using both the AV-estimator and the R/S plot test. Due to a balanced load on the B2B system throughout the duration, the degree of self-similarity is very low. The effect of the arrival process is not seen in the overall load sustained by the B2B server.

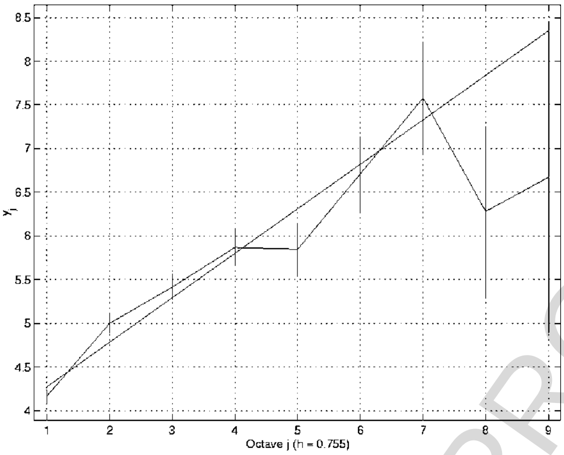

*Figure 11.* AV estimator for the front-end B2C Web server (4P).

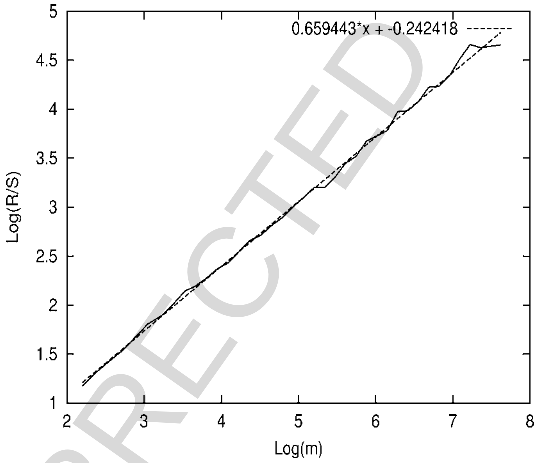

*Figure 12.* R/S estimator for front-end B2B server (2P).

A higher H-parameter implies an increased degree of self-similarity. Utilization is a factor of the response-time and the arrival process. The inherent burstiness in the arrival process is already established in the previous section. Since the H-parameter for the utilization is more than that for the arrival process, it is assumed that the service-time is long-range dependent with a heavy tailed distribution.

***4.2.3. Response time*** In Figure 13, the response time observed by the users over the entire day period is shown for the B2C site. Previous studies [Crovella and Bestavros, 5; Sahinoglu and Tekinay, 19] have concentrated on the study of the heavy-tailed behavior of Web response times. In this work the response-time distribution is converted into a

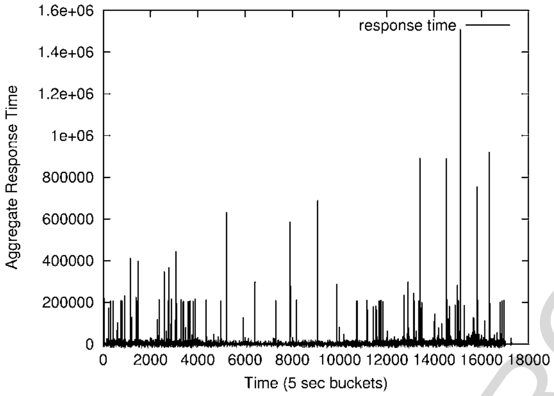

*Figure 13.* Aggregated response time at the front-end Web server (4P).

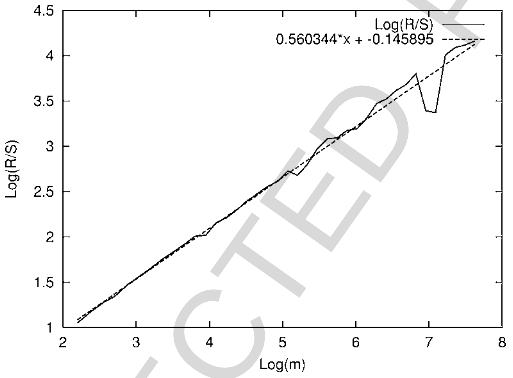

*Figure 14.* R/S test for estimating H-parameter for response time (H = 0.56).

time-series by aggregating the response-times seen for non-overlapping intervals of 5 secs. Even though the times seen are not the actual response times observed by the user, they can be used for time-series analysis. Only a multiplicative factor of 1/5 will be required to get the actual response-times. The time-series obtained is checked for self-similarity and any non-stationary behavior. The AV test and R/S plot test are used for estimating the H-parameter (Figures 14 and 15). As explained earlier, a good estimation of H-parameter is obtained using R/S test only when the time-series is stationary. So both the tests are used for estimating the H-parameter.

Response time is one of the very important performance metrics in the design and analysis of any server system. High burstiness in the arrival traffic implies saturating server queues, leading to high response times. Studies have shown that the 90th percentile response-times can be used for predicting the mean response-time [Krishnamurthy and Rolia, 12]. This measure cannot be used in presence of high burstiness in the response-time distribution. Figure 13 shows response times orders of magnitude higher during the high load periods in the evening. Comparing this graph with the arrival process shown

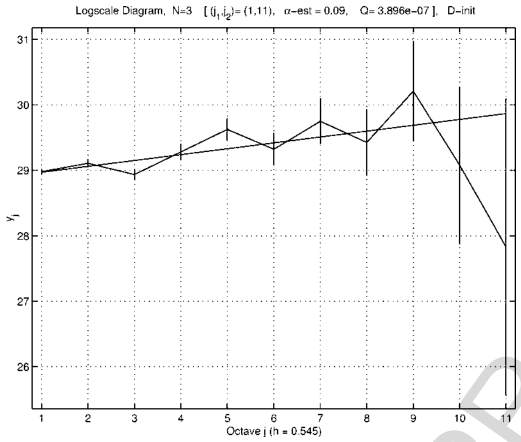

*Figure 15.* AV estimator for the front-end Web server response time (B2C).

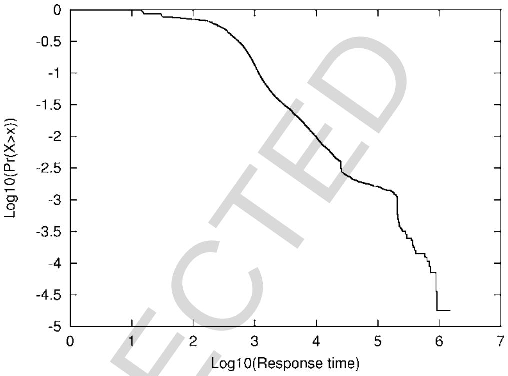

*Figure 16.* LLCD of response-time distribution at the front-end.

in Figure 4, unmistakable correlation can be found between the different load periods. Even though the utilization of the system does not get effected, buffer queue lengths increase thereby increasing the user perceived response times. Increased burstiness impacts the overall response time of the system to a higher extent than the arrival process. This burstiness in the response time is a factor of the back-end data retrieval time and the server processing time.

The response time distribution is examined for the presence of heavy-tailed behavior. In [Crovella and Bestavros, 5] the authors established the heavy-tailed behavior of Web response times. Thus heavy-tailed distribution of Web transmission times implies that the users can observe response times orders of magnitude higher than normal response times during a busy period. The authors in [Crovella and Bestavros, 5] used Log–Log cumulative distribution plots (LLCD) to estimate the tail weight of Web transmission times. Similar method was used to model the response time behavior for the B2C site. In Figure 16 the

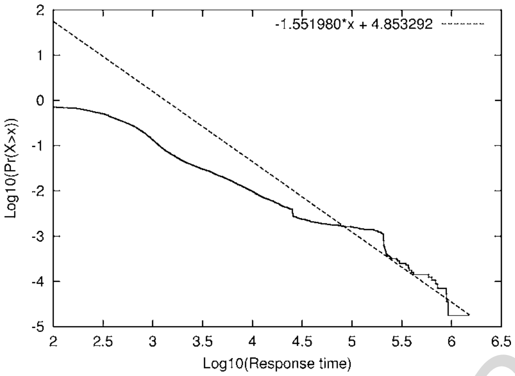

*Figure 17.* Estimated tail weight for the response-time distribution.

LLCD plot of the response-time distribution is shown. This figure shows that for values greater than 2, the distribution is nearly linear indicating a hyperbolic tail. A least-squares fit was made for data points more than 2 giving a slope of $-1.55^1$ as shown in Figure 17. This indicates that $\alpha = 1.55$. This shows that the transmission times are in fact heavy-tailed and can be modeled using a Pareto distribution with $\alpha = 1.55$. Similar result is found with the distribution of response-times in the B2B space. The distribution is found to be heavy-tailed with an $\alpha = 1.58$, for file transfers greater than 1000 secs.

In [Crovella and Bestavros, 5] the authors showed that the distribution of Web transfer times over different sets of data is heavy-tailed with $\alpha = 1.21$. The tail weight appears to be reduced in e-commerce environment. The reduction in tail weight could either be a characteristic of the dataset being used or the inherent behavior of e-commerce traffic.

***4.2.4. Request/response file sizes*** The request and response file sizes in Web environment [Arlitt and Williamson, 1; Crovella and Bestavros, 5] have been studied previously. It was observed that these distributions show a heavy-tailed behavior with a tail weight of approximately $\alpha = 1.06$ for file-sizes greater than 1000 bytes [Crovella and Bestavros, 5]. This was considered one of the main reasons for the heavy-tailed behavior of the Web response times. In e-commerce environment, it has already been shown that transfer times have a heavy tailed behavior with $\alpha = 1.55$. In this section the behavior of transfer size distribution is studied. Figures 18 and 19 show the request and response size distribution over the observation period at the B2C server.

It can be observed that the distribution of transfer sizes is fairly constant in the B2C environment. A visual inspection rules out the possibility of heavy burstiness in the aggregated time-series obtained from the transfer sizes. The distribution of request sizes is further investigated for heavy-tailed behavior using LLCD plots. Figure 20 shows the log-scale plot of the cumulative probability function over the different request sizes observed. The plot appears linear after $x > 2.5$. A linear-regression fit to the points for requests more than 320 Bytes gives a line with slope $\alpha = -4.12$ ($R^2 = 0.947$). The linear fit can be seen in Figure 21. This gives an estimate of $\alpha = 4.12$ thereby indicating that the request size distribution is not heavy-tailed in nature. This result refutes the previous results about

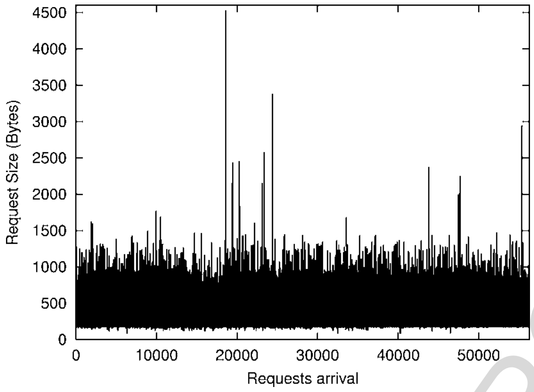

*Figure 18.* Request size distribution over time.

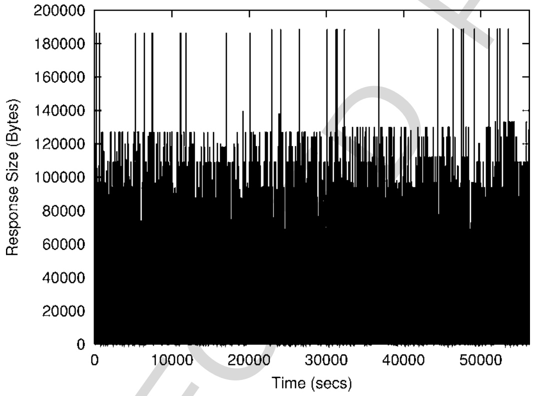

*Figure 19.* Response size distribution over time.

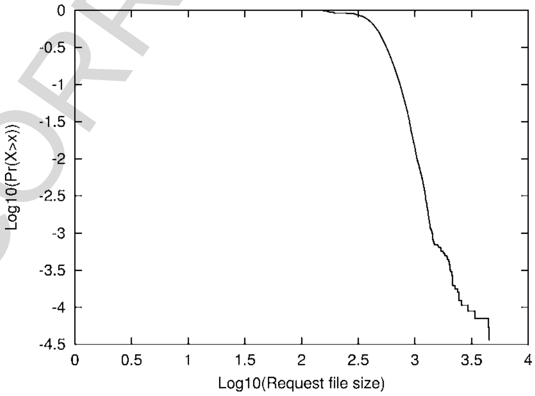

*Figure 20.* LLCD of request size distribution.

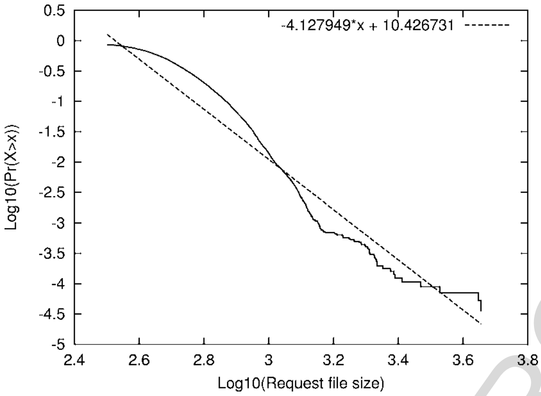

*Figure 21.* Estimated tail weight of request-size distribution.

Web traffic. In [Crovella and Bestavros, 5] the authors found that the requests also follow a heavy tail distribution with $\alpha = 1.16$. Using similar tests, we also infer that the response file sizes do not follow a heavy tailed distribution.

### *4.3. Performance implications*

Previous studies on Web traffic and LAN traffic have attributed the self-similar behavior of network traffic to the aggregation of long-range dependent ON/OFF processes. In e-commerce space, the response-times are found to be heavy-tailed in nature even though the request and response file sizes are almost a constant. The heavy-tailed behavior of response-times in Web environment was believed to be caused by the heavy-tailed behavior of the file transfer sizes in the Web environment. Studies in UNIX file-systems [Floyd, 8; Ousterhout et al., 17] have also indicated the same behavior even though the actual measurements were not made. In e-commerce environment, the transfer sizes do not follow a heavy-tailed distribution as shown earlier in this section. Heavy-tailed behavior of Web transfer sizes are fundamentally caused by the inclusion of image and video files in the overall traffic. Since these files are minimized in e-commerce environment (for reducing the overhead in response times), the behavior of the transfer sizes becomes somewhat intuitive. The lack of large image and video files removes the heavy-tailed nature of e-commerce traffic.

It is observed that the response time is still shows a heavy-tailed behavior in both B2C and B2B space. As explained earlier this implies that the user perceived response-time can increase by orders of magnitude under load conditions. Due to the critical nature of e-commerce applications and also the business model (increasing criticality with the increase in load), it is imperative that the response-times are kept under normal bounds even in high load conditions. In e-commerce environment response-time is dependent on the processing time and the transfer time. Since the file-sizes do not follow a heavy-tailed distribution, it can be safely assumed that the transfer time does not contribute to the variation in the response-time. This shows that the characteristic of the processing time is affecting the response-time to a higher extent than the response size. Also the effect of file-sizes appears to be negligible on the end-end response-times observed. This result contradicts the behavior of response-times for normal Web traffic where the response-size of files can be assumed as a good approximation of the response-time. The difference is that, in Web environment the transfer times consumes most portion of the response-time which is not the case in e-commerce environment due to the different composition of requests. Due to the presence of OLTP type of transactions, it can be assumed that the processing time consumes the major portion of the response time. This is decreasing the dependence of the file sizes on the response time.

With the absence of a major impact in response-time due to the file sizes, the variation in the response-time is attributed to the variation in the response-time seen at the back-end servers. This aspect will be further investigated as part of the back-end analyses.

### *4.4. Back-end characterization*

The most important and sensitive information in e-commerce servers is kept in the back-end servers. It is the back-end servers that perform the business logic for the e-commerce site and are hence the most crucial components of any e-commerce server. In this section the characterization of the behavior of the back-end servers is discussed. The parameters used for doing the characterization depend mostly on the configuration of the site and the purpose of the individual components [Menasce and Almeida, 14] in the back-end. As described earlier, the composition of back-end servers is closely dictated by the business model of the site. So different parameters might be interesting for different sites. In this study the following parameters are used for studying the characteristics of the two sites.

- Processor utilization.
- Disk accesses.

In the B2C site there are four different servers at the back-end. These are: main database server, customer database server, image server, and LDAP server.

The image server and LDAP server are not heavily loaded during the observation period. There is a single burst of traffic to and from these servers when the data is updated daily. This burst is also seen in other back-end databases and will be discussed in detail later in this chapter. The only servers that experience a sustained load throughout the day are the customer database and the main database. These two servers are used for studying the characteristics of the back-end system.

***4.4.1. Processor utilization*** In Figures 22 and 23 the processor utilization of the two back-end servers in the B2C site is shown. It can be observed that the back-end server experiences a sustained load of 10% on average over the entire period. There is a visible peak of almost 100% utilization of the catalog server. This will be discussed later in the section. For the main D/B server, the utilization remains at around 30% for most of the observation period. This shows that the load on back-end servers is higher than on the front-end servers, when compared with Figure 9. Previous studies have speculated that the load on the back-end servers is more regulated due to the presence of the front-end

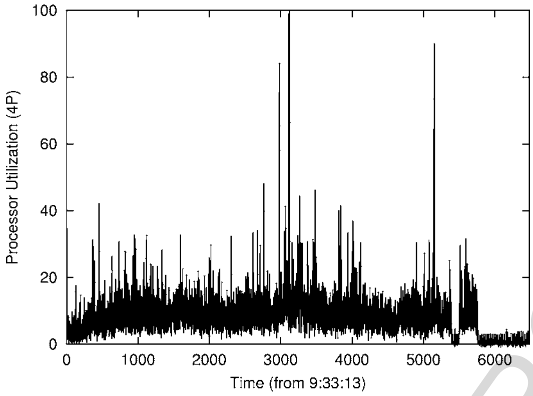

*Figure 22.* Processor utilization of catalog server (5 secs).

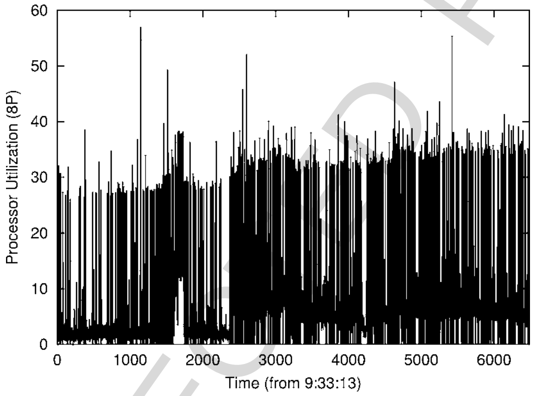

*Figure 23.* Processor utilization of the main D/B server (5 secs).

server. One of the reasons for this speculation is the service time of the front-end server. This either causes a delay or reduces peak of any burst reaching the back-end servers. This behavior of the back-end servers is investigated by looking at the time-series obtained from the utilization of the servers. H-parameter values of 0.87 and 0.77 were obtained for the utilization of the main database server and the catalog server, respectively (see Figures 24, 25). The burstiness observed at the back-end servers is more than the front-end servers (H = 0.77). Similar results have been observed in the B2B space also. The utilization of the database server of the B2B site is shown in Figure 26. It can be observed that the load on the system reaches 100% around the 4000th bucket. This is due to the update activity which takes place periodically in most e-commerce sites. The actual time when this takes place is around 1.00 pm in the night. Similar activity can be seen in the other back-end servers, but nothing can be observed at the front-end servers, as the bulk of the data which needs any maintenance is present in the back-end servers only. Figure 27 shows the Hurst parameter estimation for the utilization time-series of the database server. The

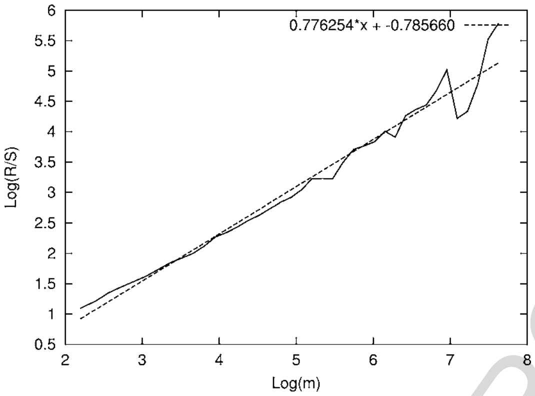

*Figure 24.* Estimation of H-parameter for catalog server (H = 0.77).

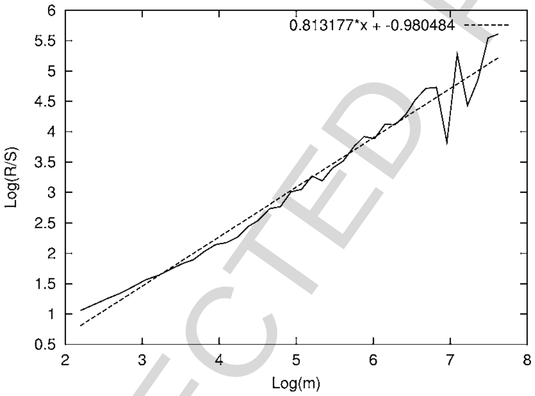

*Figure 25.* Estimation of H-parameter for main D/B server (H = 0.87).

back-end server in B2B space is also found to be more bursty than the front-end traffic. This contradicts previous assumptions about burstiness at the back-end servers in Web environment.

***4.4.2. Disk accesses*** The B2C site has four disks for the main DB system. Disk accesses are used for the study instead of disk utilization. Reliable data could not be obtained for the disk utilization due to the presence of a cluster of four disks.

Figure 28 shows the distribution of the file request rate at the main DB server. This shows the arrival rate of file requests seen by the four hard disks. Figure 29 shows the average queue length seen by the hard disks at the main DB server. The average queue length is found to be self-similar in nature with H = 0.77. This would result in a heavy-tailed behavior in the average response-time of the hard disk. The reason for the burstiness in the queue length can be attributed to the arrival of file transfers at the hard disk. This rate is also found to be bursty in nature with H = 0.83. The buffer cache does not

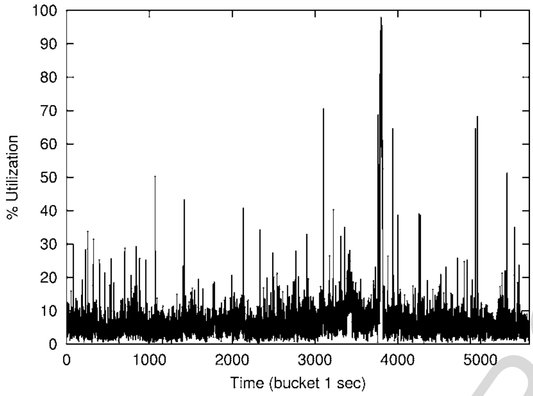

*Figure 26.* Utilization of the B2B back-end server.

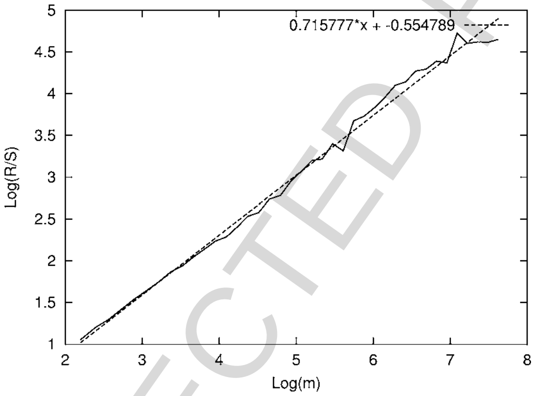

*Figure 27.* H-parameter for the B2B database server (H = 0.72).

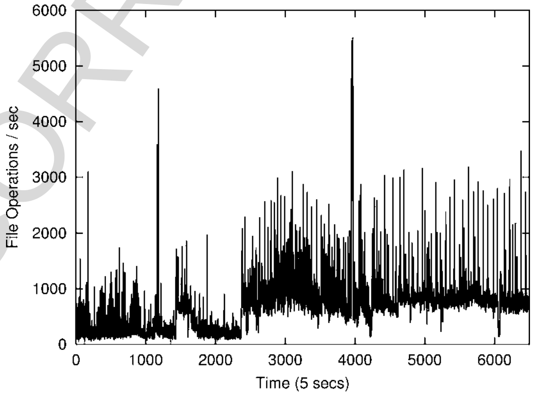

*Figure 28.* File operations per second from main DB server (5 sec).

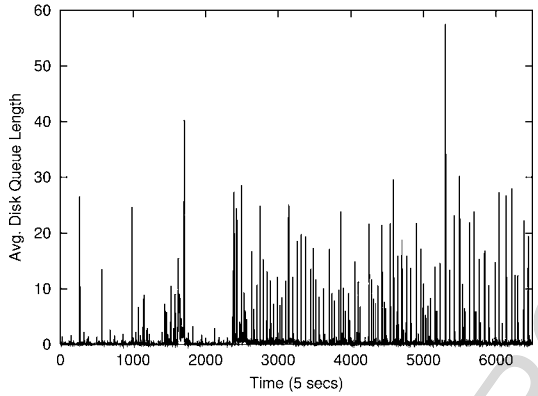

*Figure 29.* Disk queue length at main DB server (5 sec).

appear to be effective since the hard disk is experiencing requests at this level of burstiness.

In the previous section, the response-time at the front-end is found to be heavy-tailed in nature even though the request and response size did not follow this distribution. The burstiness in the service time at the back-end was attributed to this behavior. Here it can be seen that the heavy-tailed distribution of response-time at the back-end is due to the bursty arrival process to the hard disks, causing the queue length to be bursty. This high burstiness in queue length will remove the effect file sizes may have on the transfer times. This conclusion also supports the previous speculation that file-sizes were not a good representation of response-times in e-commerce environment.

## 5. Conclusion

Aggregated traffic arriving at an e-commerce servers is characterized in this paper. Live traffic was collected from two different e-commerce sites: a B2B site and a B2C site. The data were collected at three different levels. Access logs from the Web servers is collected for application level information, Microsoft performance logs were collected for system level information and processor counters were collected for architectural information like cache hit ratio etc. Information from this data was used to understand the load behavior of the traffic for a normal weekday. Only a specific set of parameters (arrival process, utilization, response-time, transfer sizes, etc.) which would impact the system to the maximum extent were used for characterization of the workload.

Self-similar nature of the traffic was established using Hurst-parameter as a measure of degree of self-similarity. Two different tests were used for measuring the Hurst-parameter: AV-estimator and the R/S plot. It was observed that the load behavior of the two sites was complimentary in nature with traffic load shifting from one type of e-commerce site to the other during the later part of the day. Unlike previous speculation, the back-end server was found more bursty than the front-end server, this was attributed to the fractal nature of the service time at the back-end.

In both the sites, the response-times were found to be heavy-tailed in nature, complying to the results found in Web environment. But in the B2C environment, highly bursty arrival of file requests was seen at the disks. It was found that this arrival process is causing high queuing delays at the disk reducing the impact of disk transfer time as compared to the queuing time. This increased the burstiness in the overall response-time seen at the front-end server.

This work provides an understanding of the complexity of the traffic arriving at e-commerce sites while providing a preliminary workload characterization.

## Acknowledgment

This work was supported in part by the National Science Foundation under Grant No. CCR-9988179.

## References

[1] Arlitt, M.F. and C. L. Williamson. (1996). "Web Server Workload Characterization: The Search for Invariants." In *ACM SIGMETRICS'96*, May 1996.

[2] Bradford, P. and M. Crovella. (1998). "Generating Representative Web Workloads for Network and Server Performance Evaluation." In *Proc. 1998 ACM/SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, Madison, July 1998.

[3] Chen, X., H. Chen, and P. Mohapatra. (2001). "An Admission Control Scheme for Predictable Server Response Time for Web Accesses." In *Proc. 10th World Wide Web Conference*.

[4] "Cisco and Microsoft E-Commerce Framework Architecture." `http://www.microsoft.com/technet/ecommerce/ciscomef.asp`.

[5] Crovella, M. and A. Bestavros. (1997). "Self-Similarity in World-Wide Traffic: Evidence and Possible Causes." *IEEE/ACM Transactions on Networking* 5, 835–846.

[6] Diffie, W. (1998). "E-Commerce and Security." Technical Report No. 3, Sun Microsystems, Palo Alto, CA, September 1998.

[7] "E-Commerce Statistics." `http://www.zdnet.com/ecommerce/stories/main/0,10475,2636088,00.html`.

[8] Floyd, R.A. (1986). "Short-Term File Reference Patterns in a Unix Environment." Technical Report, Computer Science Department, University of Rochester.

[9] Glushko, R.J., J.M. Tenenbaum, and B. Meltzer. (1999). "An XML Framework for Agent-Based E-Commerce." *Communications of ACM* 42, 106.

[10] Kant, K. *Introduction to Computer System Performance Evaluation*. (1992). New York, NY: McGraw-Hill.

[11] Kant, K. and Y. Won. (1999). "Server Capacity Planning for Web Traffic Workload." *IEEE Transactions on Knowledge and Data Engineering*, 731–747.

[12] Krishnamurthy, D. and J. Rolia. (1998). "Predicting the Performance of an E-Commerce Server: Those Mean Percentiles." In *Proc. First Workshop on Internet Server Performance, ACM SIGMETRICS'98*, June 1998.

[13] Ma, M. (1999). "Agents in E-Commerce." *Communications of ACM* 42, 78–80.

[14] Menasce, D.A. and A.A.F. Almeida. (1998). *Capacity Planning for Web Performance: Metrics, Models and Methods*. Upper Saddle River, NJ: Prentice-Hall.

[15] Menasce, D.A., A.A.F. Almeida, R. Fonseca, and M.A. Mendes. (1999). "Resource Management Policies for E-Commerce Servers." In *2nd Workshop on Internet Server Performance*, May 1999.

[16] "Microsoft Management Console: Performance." `http://www.microsoft.com/windows2000/techinfo/howitworks/management/mmcover.asp`.

[17] Ousterhout, J.K., H.D. Costa, D. Harrison, J.A. Kunze, M. Kupfer, and J.G. Thompson. (1985). "A Trace Driven Analysis of the Unix 4.2BSD File System." Technical Report, Department of Computer Science, University of California at Berkley.

[18] "Real Numbers Behind 'Net Profits." (2000). ActivMedia, June.

[19] Sahinoglu, Z. and S. Tekinay. (1999). "On Multimedia Networks: Self-Similar Traffic and Network Performance." *IEEE Communications Magazine*, January.

[20] Veitch, D. and P. Abry. (1999). "A Wavelet-Based Joint Estimator for the Parameters of LRD." Special Issue on Multiscale Statistical Signal Analysis and its Applications. *IEEE Transactions on Informatics Theory* 45.