Empirical Model of WWW Document Arrivals at Access Link

Shuang Deng
GTE Laboratories, Inc.
40 Sylvan Road
Waltham, MA 02254, USA
(sdeng@gte.com, +1 617 466 2165)

Abstract

The cable and telephone industries have already begun constructing the information super highway. The network capacity planning for data services, however, cannot start until a realistic traffic model is developed based on actual traffic data. Since the future data services are perceived to resemble today's World Wide Web (WWW) browsers, we propose in this paper a traffic model for the access network based on today's WWW traffic. The model is an ON-OFF two-state model with the ON period consisting of a sequence of document transmission requests from an individual subscriber. Actual traffic data with over 20,000 data points was used to fit distributions to the model. The ON and OFF periods are found to be of heavy-tailed Weibull and Pareto distributions, respectively. The inter-arrival times of requests within the ON periods can be described with another Weibull distribution. This empirical model can be used as a realistic basis for network capacity planning for the future access networks.

1. Introduction

The construction of the information super highway has begun with the cable and telephone companies investing billions of dollar into new network infrastructure. Network architecture includes hybrid fiber-optic (HFC) switched digital video (SDV) or fiber to the curb (FTTC), and asymmetric digital subscriber line (ADSL) or very high-speed asymmetric digital subscriber line (VDSL) [1-4].

The future network will provide not only video services, but also such data services as customized news, on-line search and on-line shopping and so on. It is widely believed that future data services will resemble today's World Wide Web (WWW) browsers such as Mosaic and Netscape Navigator. The network capacity planning for data services, however, has yet to evolve beyond preliminary estimates.

The access traffic characteristics must first be understood before the bandwidth demands can be projected, and bandwidth capacity be planned for future networks. The access link refers to the network segment from individual users, or subscribers, to the local switch (i.e., central offices or head-ends). This paper concerns itself with the data traffic characteristics of individual subscribers on the future network based on today's WWW traffic, and proposes a model for the arrival process of document transmissions at the access link.

Paxson and Floyd discovered that user-initiated TCP sessions arrive at a WAN according as a Poisson process [5]. Their findings are consistent with other studies of human-initiated process such as telephone calls. The Poisson process, however, may not be applicable to WWW document arrivals, because WWW document transmissions are not entirely initiated by the user. A WWW page usually contains several in-line images (e.g., logos, icons and buttons). When a user requests a page, the browser program automatically generates a series of additional requests to download these in-line images. Instead of a simple Poisson arrival, an on-off process can be used to model this process with several requests during the active, or ON, period followed by an inactive, or OFF, period that is significantly longer than the inter-arrival time during the on period.

Crovella and Bestavros performed extensive analysis of WWW traffic on an Ethernet LAN in [6], and found document sizes to be Pareto distribution. They noticed that the inter-arrival time of document requests is not always Pareto, but did not provide a model for it. A model will be provided in this paper.

This study is based on two sets of empirical data with over 20,000 document requests by 293 active users during a two-day period. A model is obtained in this paper to describe completely the stochastic behavior of the document arrival process at the access link. The remainder of this paper is organized as follows. The empirical data collection is described in the next section. The traffic model is presented in Section 3. The application of the model and its relationship with self-similarity are discussed in Section 4. This paper is concluded in Section 5 with a summary and discussion of future research work.

2. Traffic Data Collection

We monitored all WWW traffic on the access link between GTE Laboratories and the Internet (Figure 1) in two sessions. The first session lasted about two hours from 2pm to 4pm on August 2, 1995, and the second session was from 9:30am to 12:00pm on August 3, 1995. The users include GTE Laboratories employees at Waltham, Massachusetts, as well as users at remote GTE sites that connect to the Internet via GTE Laboratories. There are over a thousand IP nodes on the LAN, four of which are WWW servers that are externally accessible.

During the first session (August 2), a total of 2,964 nodes were active with approximately 10% nodes local, and 90% remote. Nearly one-third of all nodes were involved with WWW activities which accounted for about 30% of the inbound traffic. During the second session, 27% of all the 5,328 active nodes were using the WWW, and generated about 37% of all inbound traffic.

The inbound WWW requests to four local servers are not included in the data sets for three reasons. First, we are interested in the traffic process of individual subscribers, not the aggregate behavior. Secondly, many of the data services under planning now for the future networks will offer asymmetric bandwidth services that are intended for information consumption, as opposed to large-scale provision. It is, therefore, unlikely that the access link

0-7803-3250-4/96$5.00©1996 IEEE

will be used to connect to servers. Servers probably will use traditional symmetric links. Thirdly, traffic patterns can be influenced by the organization of pages on the server. For example, a server may be created with few or no in-line images, resulting in very short ON period, or another server may have higher than typical amount of in-line images. The amount of document requests received by the four local servers are roughly 10% of the outbound requests from local usrs to remote servers. We excluded the local server traffic from our data sets to avoid statistical bias introduced by the particular local server.

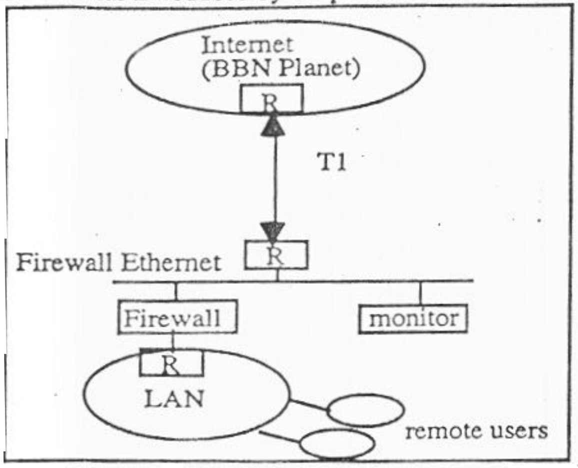


Figure 1. Network Under Observation

Each data set was then processed to separate individual streams by the WWW client's IP address. The first data set contains 8,592 document requests from 157 local users to 775 remote servers. The second set contains 11,590 requests from 185 local users to 1,242 remote servers. Some users and servers were active in both sessions. Because of the large number of remote servers present in our data sets, we felt confident that the analysis is representative of the typical network activities.

3. Traffic Model

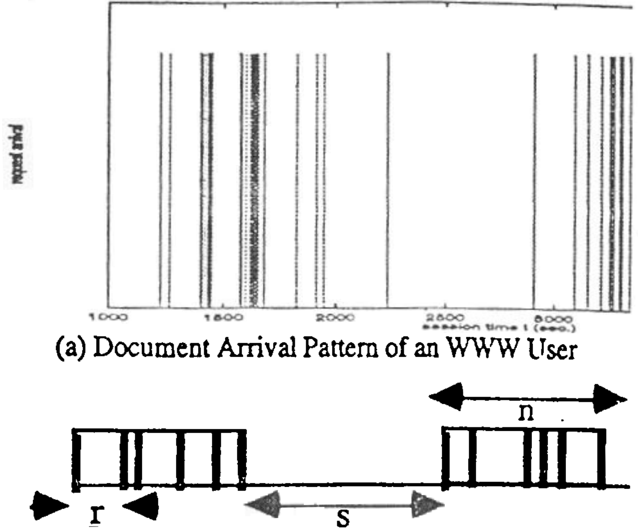
We took the approach of creating a model for the physical process of WWW document request arrivals, and then finding the distribution and parameters from empirical data sets. Combining this model with the Pareto distribution of document sizes by Crovella and Bestavros [6], one can get a complete description of the data traffic on the access link.

A. WWW Request Arrival Model

When a WWW user clicks on a hypertext link, several URL requests may follow. The first one transmits the user's direct request to the server. During the execution of the user's request, other requests may be automatically generated by the client program. For example, each in-line image requires a separate request be sent automatically by the client program during the download of a page. These requests each open a TCP new connection, and the TCP connections are either overlapping or back-to-back.

After the user's request and its associated requests are completed, the user typically will take time to absorb the information just received before initiating the next request. This physical process is evident in the arrival pattern of a random user in Figure 2-a. Hence, the traffic can be modeled with an inactive, or OFF, period following an active, or ON, period that consists of a series of requests as depicted in Figure 2-b.

The first request in an ON period is initiated directly by the user. The new few requests may be generated automatically by the client program. An ON period may include more than one direct request by the user, for the user may make another request before all the current requests are completed.



(a) Document Arrival Pattern of an WWW User

(b). ON-OFF Mode

Figure 2. Model of WWW Document Requests

The model can be described by the distributions of three random variables

r: the inter-arrival time of requests during an ON period

s: the duration of an OFF period and

n: the duration of an ON period

These three distributions are examined in the remainder of this section.

B. ON Period Distribution

While the traffic of page requests clearly shows an ON OFF pattern, the exact boundaries of an ON and OFF periods are dependent on the choice of threshold. We follow the method used in [7] paper to heuristically choose a specific value, and then to test the insensitivity of the distribution to the specific choice of threshold value within a large range.

In this study we use a threshold value of 60 seconds. A sequence of document requests with inter-arrival times less than 60 seconds are considered to form an ON period, and the occurrence of a request more than 60 seconds after the previous request indicates an OFF period. Two consecutive OFF periods separated by a single request are combined into one.

Both sets of actual data appear to belong to Weibull distribution. A Weibul distribution is given by probability density function

$$p(x) = \frac{k}{\theta}\left(\frac{x}{\theta}\right)^{k-1} e^{-(x/\theta)^k} \quad \ldots(1)$$

and probability distribution function

$1-e^{-(\frac{x}{\theta})^{k}}$

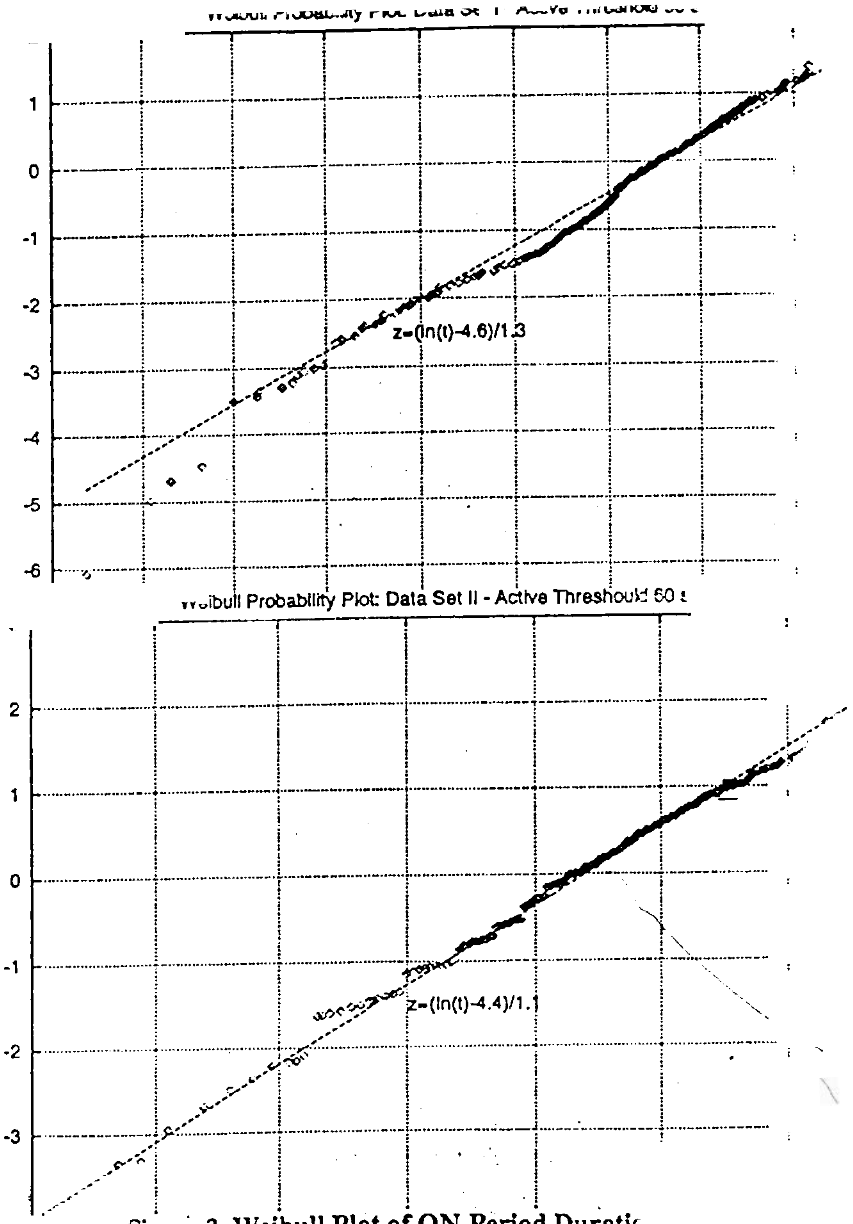
$$F(x)=\text{Prob.}(t\leq x)=1-e^{-(\frac{x}{\theta})^{k}} \quad (2)$$

The parameter k determines the shape of the distribution. The distribution function is positively skewed when $k>3.6$ and negatively skewed when $k\leq 3.6$. Moreover, the distribution is light tailed when $k>1$, and heavy tailed when $k<1$. A Weibull distribution becomes negative exponential distribution when $k=1$

Testing for Weibull distribution is performed by detecting a straight line on a probability plot of y=ln(x) versus $z=ln(-ln(1-F(x))$ [8]. From a straight line fit $z=(y-a)/b$, one obtains $k=1/b$, and $\theta=exp(a)$ for Eq. (1).

Figure 3 shows the Weibull probability plotting for ON period duration with threshold 60 seconds for Data Set I and II. The results clearly suggest a Weibull distribution with k–0.91 to 0.77, and θ–exp(4.4) to exp(4.6). Based on the average of these data , an empirical model could be devised for ON period probability density as

$$f(x)=\frac{0.88}{x^{0.12}}e \quad (3)$$



Y

Figure 3. Weibull Plot of ON-Period Duratic

To demonstrate the distribution's insensitivity to the particular choice of 60 seconds for the threshold valu Figure 4 shows the probability distribution's Weibu plotting with threshold values 30 and 120 seconds. In both cases, the shape value maintains at k–0.91, and the scale parameter θ increases or descreases accordingly as data are added to or remoted from the t

C. OFF Period Distributio

The OFF period represents the thinking time of the us and typically indicates the existence of significant pau in the communication activities. Figure 5 shows the results of fitting the complementary probabili distribution function 1-F(x) with a power function, whe R is the residual value of the fitti The fitting w applied to all data points in the first data set, and to only the period shorter than 6,000 seconds in the second four-hour data set. Requests longer than 6,000 seconds apart can be viewed as two separate sessions.

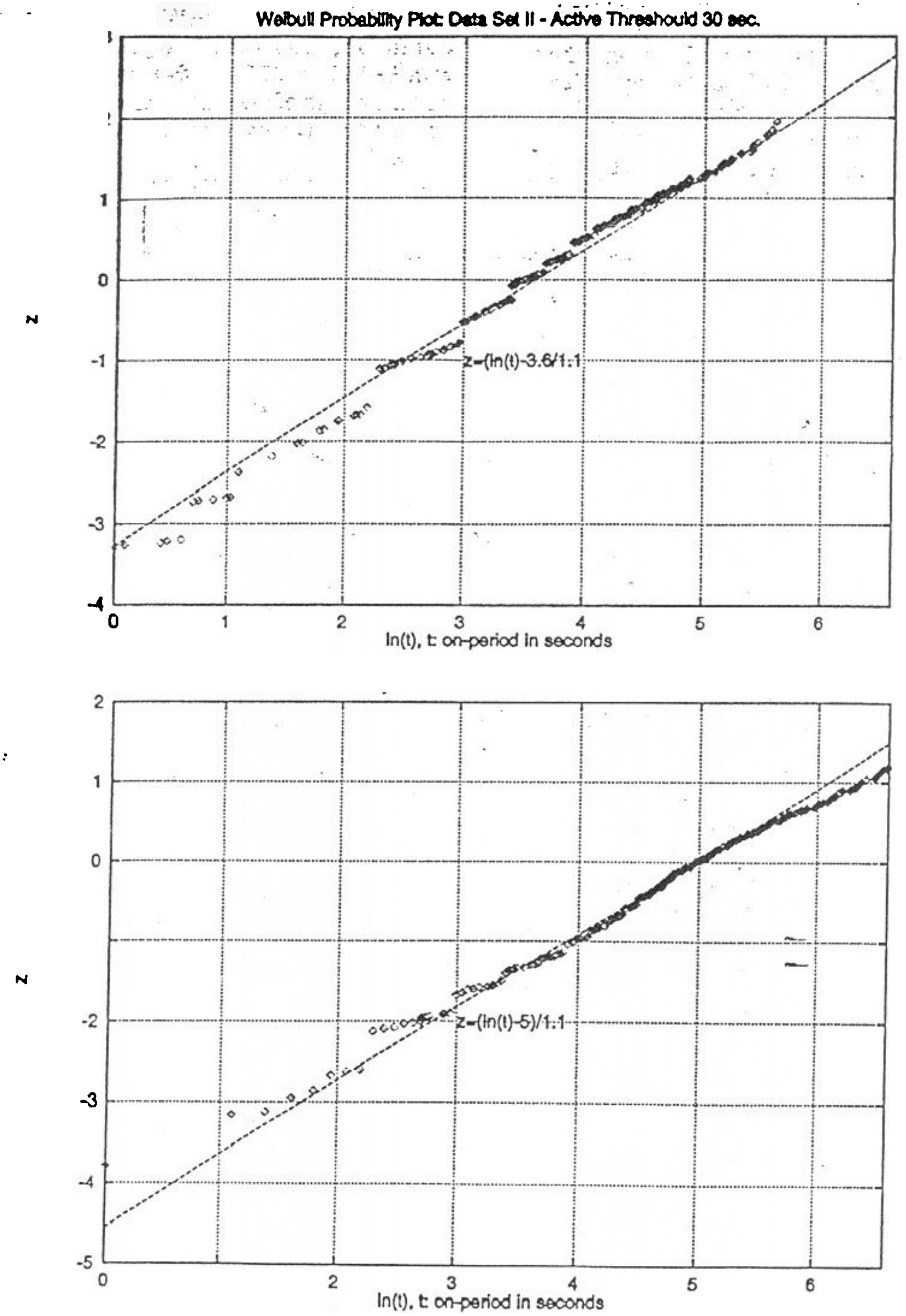


Figure 4. Insensitivity Test with Different Threshold Values

The results suggest that the duration of the OFF period belongs to a Pareto distribution with probability distribution function:

$$F(x) = \Pr ob.(t \le x) = 1 - (k / x)^{\alpha} \quad (4)$$

or the density function

$$p(x) = \alpha k^{\alpha} / x^{\alpha+1} \quad (5)$$

where k represents the smallest value.

A Pareto distribution has infinite mean if $\alpha \le 1$, and infinite variance if $\alpha \le 2$. Our WWW traffic data suggest an α value of 0.9 and 0.58, respectively. In both case, the distribution has infinite mean and variance. Using $\alpha = 0.5$, one would get the following probability density function as an empirical model for the OFF period length:

$$p(x) = \frac{\sqrt{60}}{2x\sqrt{x}} \qquad (6)$$

Table 1: Moments of Inter-Arrival Times During ON Period

Data Set	Mean	Variance
I	6.64	148.16
II	6.79	144.47

D. Interarrival Distribution During ON Period

The first data set produces 8,592 samples of inter-arrival time, and the second 11,590. There appears to be more WWW activities in the afternoon session (Data Set I) than the morning session (Data Set II) with a mean request rate of 0.48 versus 0.24 requests per user per minute. The statistics of interarrival times, nevertheless, are remarkably similar, as suggested by the first and second moment values in Table 1, and the Weibull probability plots in Figure 6.

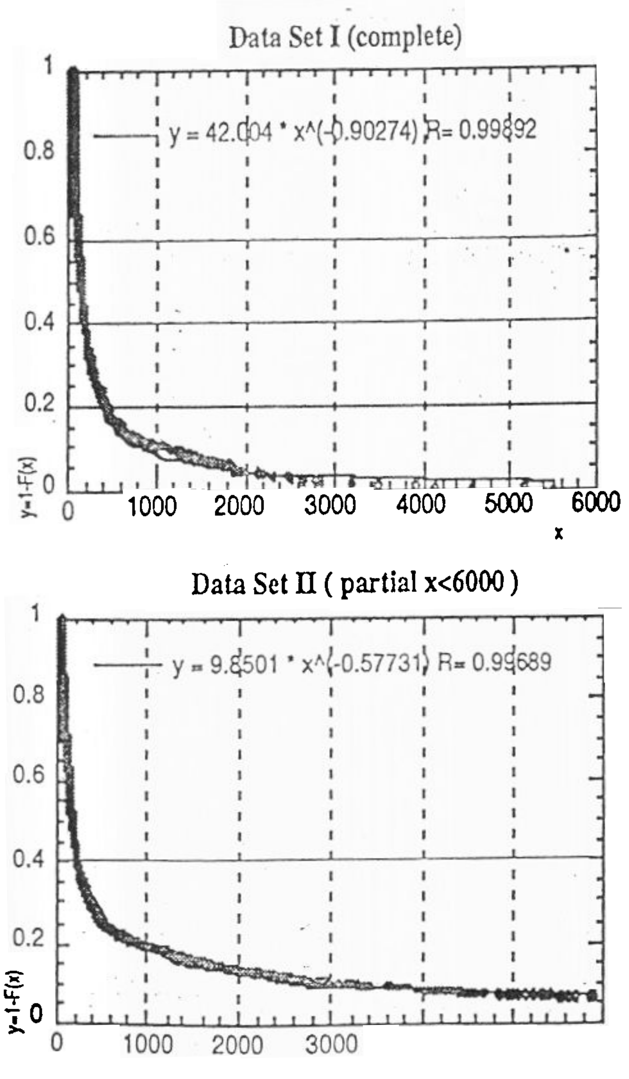


Figure 5. Fitting Pareto Distribution to Empirical Data

$\theta = e^{1.5}$

The results indicate a same distribution function for both data sets. Instead of attempting to find the exact values, we suggest an approximate model of $k=0.5$ and $\theta=1.5$ for Weibull distribution, i.e., probability density function:

$$p(t) = \frac{e^{\sqrt{t/e^{1.5}}}}{2\sqrt{e^{1.5}t}} \qquad (7)$$

Although they may not be the optimal fitting, these two values appear to be very close by visual inspection of Figure 6, and provide the advantage of simplicity.

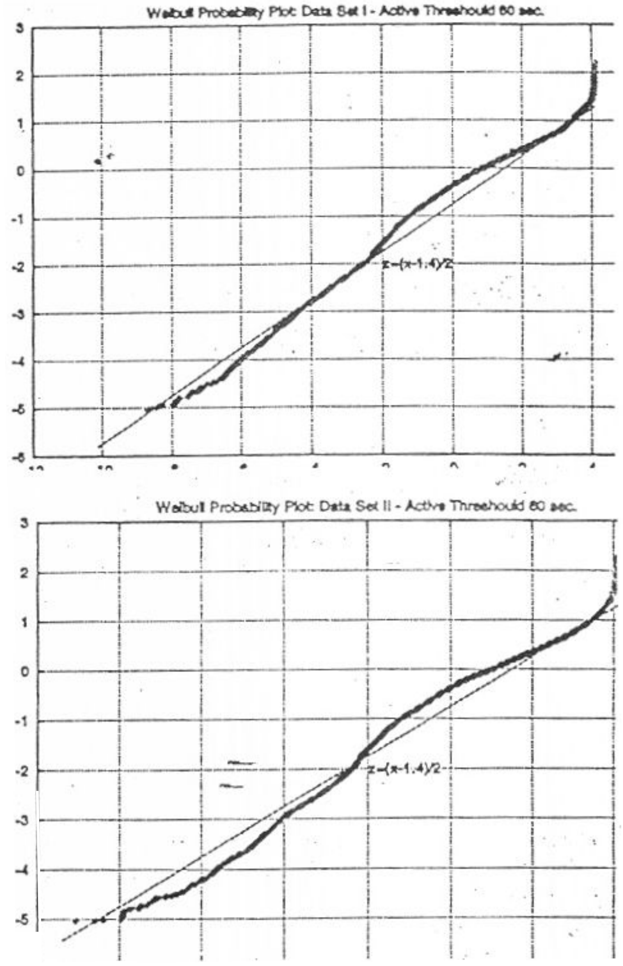


Figure 6. Weibull Probability Plot of Inter-Arrival Times

4. Applying the Model to Network Planning

Because the model was developed from actual traffic data, it can provide an accurate and realistic foundation to network planning and traffic engineering. The first part of this section describes the applications of this model to network planning problems, and the second part relates the findings to the traffic self-similarity issue.

A. Applications

We use the problem of estimating the bandwidth requirement of a central office (CO) with S subscribers for different active ratios A. A subscriber is active if he or she logs on the data services. It can be modeled as a queue with SA i.i.d. sources. Each source generates alternates between the ON and OFF states with distributions defined by Eq. (3) and Eq. (6), respectively. Within the ON process, the source generates documents with random inter-arrival time a as defined by probability density function Eq. (7). Finally, the length of each document is given by a Pareto distribution as described in [6]. This model can be used directly in a simulation to obtain measurements of such performance metrics as queue length, packet loss rate, packet delay and trunk utilization.

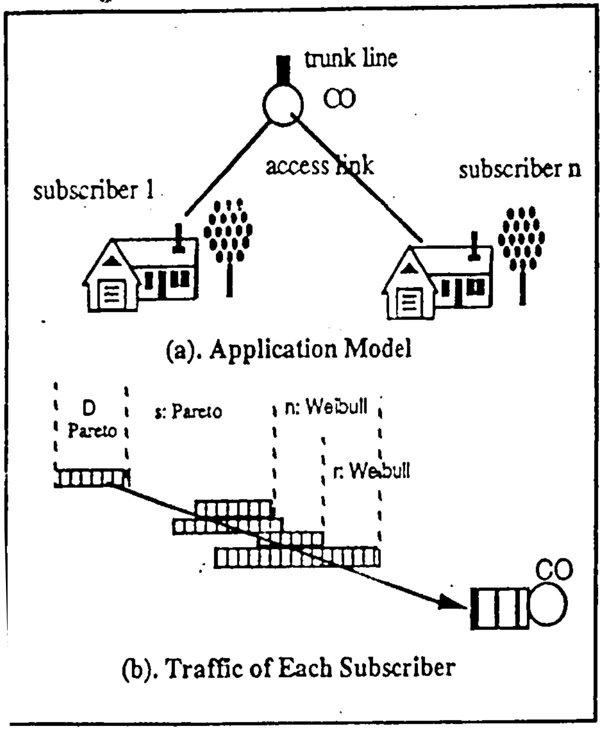


Figure 7. Application in Network Planning

B. Relating to Self-Similar Traffic

The OFF period is indeed heavy tailed with a Pareto distribution of $\alpha \leq 2$. Both the duration and the inter-arrival time during the ON period are heavy tailed with a Weibull distribution of $k<1$. We suspect the *packets* arrival of each source as depicted in Figure 7-(b) is still heavy tailed. The packet arrival process appears to be long range dependent, or self-similar [9] with Hurst parameter $H{-}0.9$, as indicated by the R/S plot based on our data. One likely explanation for self-similarity is believed to be the heavy tail distributions of individual ON-OFF sources [7]. This can be related to the heavy-tail distribution in the actual data and our model.

5. Conclusion

A empirical model was developed in this paper for the WWW document request arrivals at the access link. This model can be used for capacity planning for data services with HFC, FTTC/SDV or ADSL/VDSL networks. The distributions of ON and OFF periods were found from the actual data to be Weibull and Pareto with parameters $(k{-}0.91{\sim}0.77,\ \theta{-}e^{4.4}{\sim}\ e^{4.6})$ and $(\alpha{-}0.9 \sim 0.58)$, respectively. A series of requests are generated during the ON period, usually a user-initiated request leading a series of computer-generated requests. The inter-arrival time of requests within an ON period is governed by another Weibull distribution with parameters $k=0.5$ and $\theta=1.5$. The document size distribution was previously found in [6] to be Pareto distribution of $\alpha<1$. These four distributions completely describe the data source of individual users in a WWW network or a future data service network. $\theta = e^{1.5}$

The present model can be easily used in simulation study to model multiple data subscribers. Results can be obtained on various network engineering metrics such as CO switch buffer size, quality of service, and trunk dimensioning.

The analytical solution, however, is not immediately obvious, and is a possible future research topic. Additional future work may include verifying the model with data from a variety of sources.

Acknowledgment

The author thanks Dr. Robert Olshansky for suggesting and supporting this study, and Dr. Les Servi and Dr. David Krinsky for many insightful discussions.

References

[1] A. Paff, "Hybrid Fiber/Coax in the Public Telecommunications Infrastructure," *IEEE Communications Magazine*, vol. 33, no. 4, pp. 40-45, 1995.

[2] B. W. Phillips, "Broadband in the Local Loop," *Telecommunications (American Edition)*, vol. 28, no. 11, pp. 37-42, 1994.

[3] D. Veeneman and R. Olshansky, "ADSL for Video and Data Services," in the proceeding of IEEE ICC'95, pp. 837-841, 1995.

[4] F. V. d. Putten (ed), " Part 8: Lower Layer Protocols and Physical Interfaces," in *DAVIC 1.0 Specifications*, Digital Audio-Visual Council 1995.

[5] V. Paxon and S. Floyd, "Wide Area Traffic: The Failure of Poisson Modeling," *IEEE/ACM Transactions on Networking*, vol. 3, no. 3, pp. 226-244, 1995.

[6] M. E. Crovella and A. Bestavros, "Explaining World Wide Web Traffic Self-Similarity," Computer Science Dept., Boston University, Technical Report TR-95-015, August 29 1995.

[7] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson, "Self-Similarity Through High-Variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level," in the proceeding of ACM SIGCOMM'95, pp. 100-113, 1995.

[8] R. B. D'Agostino and M. A. Stephens, *Goodness-of-fit Techniques*: Marcel Dekker, 1986.

[9] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the Self-Similar Nature of Ethernet Traffic," in the proceeding of ACM SIGCOMM'93, pp. 183-193, 1993.