Models of Mail Server Workloads

Laura Bertolotti, Maria Carla Calzarossa
Dipartimento di Informatica e Sistemistica
Università di Pavia
I-27100 Pavia, Italy
{bertolot, mcc}@alice.unipv.it

ABSTRACT

Electronic mail has become an integral part of our daily lives. With this trend, mail servers have to provide a fast, highly available, reliable and secure service. Hence, workload characterization and performance evaluation of mail servers are to be addressed as primary issues. This paper deals with a detailed characterization of mail server workloads. Our study is based on the analysis of a large set of measurements collected on various mail servers. We analyze SMTP and POP3 requests and we obtain models able to capture and reproduce their behavior and most relevant characteristics. These models represent the basis for the definition of the workload of SPECmail2001, a benchmark currently under development within SPEC to assess the ability of a system to act as a mail server.

1 Introduction

The large availability of smart devices and ubiquitous networking has changed the way people approach personal and professional communications. Electronic mail is part of our daily lives. It has achieved a wide spread acceptance for personal communication and it has become a fundamental prerequisite for doing any kind of business.

With this trend, Internet users require a fast, highly available, reliable and secure email service. Mail servers have to meet these requirements by efficiently handling the delivery of incoming and outgoing messages and allowing users to access and download their messages. Hence, organizations, such as, ISPs and enterprises, which provide email services, have to address performance evaluation and capacity planning of their mail servers as a primary issue.

It is known that the accuracy and the representativeness of any performance study are influenced by the definition of the workload processed by the system under test. The characterization of the workload has then to be seen as a preliminary step towards performance analysis.

Workload characterization can be addressed from different perspectives [CS93], [CMT00]. Recent studies focus on the characterization of the workload of web servers [AW97], [CB97], [BC98], [ISZ99], proxy servers [AFJ99], and ecommerce sites [MAFM99]. These studies analyze various aspects of the workload, such as, the characteristics of the requests, the reference locality, the navigational patterns of the customers, and develop distributional and graph-based models able to accurately reproduce the properties of the load.

This paper presents a detailed characterization of mail server workloads. Our approach is experimental, that is, based on the analysis of measurements collected on various mail servers. We analyze the requests associated to the open standard Internet protocols, namely, SMTP [RFC821] and POP3 [RFC1939] protocols, and we obtain static and dynamic descriptions of these workloads. The results of this characterization provide useful suggestions on how to manage and configure a mail server. Moreover, these models are used for the definition of the load of SPECmail2001, a standardized benchmark currently under development within SPEC (Standard Performance Evaluation Corporation) [SPEC], whose goal is to measure a system's ability to act as a mail server.

The paper is organized as follows. Section 2 introduces the raw data sets used in the study. The methodology adopted for the analysis of the requests received by mail servers is briefly discussed in Section 3. The results of the characterization of SMTP and POP3 workloads are presented in Sections 4 and 5. In Section 6, we summarize our work by drawing a few conclusions and outlining future research directions in the field of workload characterization of mail servers.

2 Data collection

The workload of a mail server is the superposition of requests due to its interactions with other mail servers, i.e., SMTP requests, as well as to the interactions with the users, i.e., POP3 requests. To characterize the behavior of this composite workload, we analyzed measurements collected during April and May 1999, on various mail servers, namely, the mail servers of a small ISP, two enterprises and the University of Pavia. Because of non–disclosure agreements, we will not be able to name the enterprises and the ISP.

Table 1 presents a summary of the raw data considered in our study. This data refers to the access logs of the four mail servers. Each log file records information collected over a day, that is, 24 hours. The logs contain information on both SMTP and POP3 requests.

mail server	number of log files	number of users	total SMTP requests	total POP3 requests
Enterprise A	2	unknown	35,840	–
Enterprise B	1	unknown	20,790	89,546
ISP	11	8,057	214,225	–
University	11	1,058	138,117	243,684

Table 1: Summary of the collected raw data.

As can be seen from the table, only the log files measured on the University of Pavia and Enterprise B mail servers contain POP3 data. Indeed, collecting measurements, which is always a challenging issue, becomes even more challenging in the case of POP3 workload. Many installations do not log any information about this workload because of the overhead introduced on their mail servers.

The table lists, whenever available, the number of users who have subscribed to the various mail servers. This figure provides a rough characterization of the mail server in terms of its potential load.

It is known that the information that can be collected on a mail server depends on the messaging software running on the server itself and on the logging options selected. Because of the heterogeneity of the mail servers considered in our study, we had to identify the set of information recorded by every mail server. This allows us to keep a uniform description of the workload across different mail servers. We developed a set of Perl scripts which extract from the log files the information common to all log files. These scripts are also used to sanitize the collected log files by removing all the confidential information they might contain.

For each STMP request, the information extracted from the log files are its time stamp, the message identifier, the size of the message to be sent or being received, the number of recipients of the message, the IP addresses of the sender and of the recipient(s), the status of the delivery and the duration of the request.

For each POP3 request, we extracted its time stamp, the identifier of the user who issued the request, the IP address of the client and information about mailbox statistics, such as, number and size of the messages in the mailbox.

Note that the resolution of the time stamps recorded into the log files for both types of requests is one second only. In the following sections, we will discuss how this coarse granularity might affect the accuracy of the distributions of the times between two consecutive requests.

3 Methodology

The analysis of the measurements collected on the various mail servers is based on a typical workload characterization methodology, consisting of various steps dealing with the identification of the parameters describing the workload, their exploratory analysis, the application of various statistical and numerical techniques to build the appropriate models and the validation of these models.

Due to the different nature of SMTP and POP3 requests, we considered each workload type separately and we built their models accordingly. Indeed, SMTP workload, which is managed by application software, such as, sendmail [AV95], deals with the requests coming from or addressed to other mail servers to deliver incoming and outgoing messages, whereas POP3 workload reflects to a certain extent the user behavior in that it refers to the requests issued by the users against the mail server to access their mailboxes and read their messages.

For each workload type, we identified the characterizing parameters. The choice of these parameters was driven by the objective of our study, that is, to build workload models for describing the input load of benchmark experiments. Hence, the parameters used in our analysis deal with the intensity and the functional characteristics of the requests, whereas we did not include any parameter dealing with the consumptions of the requests on the mail server resources, as it would have been required by workload models used to describe the input of analytic or simulation system models.

The exploratory analysis of the parameters extracted from the log files is based on their descriptive statistics, i.e., basic statistics and distributions. This analysis provides preliminary insights into the behavior and the characteristics of the workloads. The application of techniques, such as, clustering [Hart75], numerical fitting [LH74], and probabilistic graphs [Ferr84], leads to the identification of models which are representative of the static and dynamic behavior of the mail server workloads. A comparison of the models obtained for each data set will be used to assess their validity and representativeness.

Note that the data of each log file are analyzed separately and, unless otherwise stated, the results presented in what follows, refer to the requests received by one of the mail servers during a day, that is, 24 hours.

4 STMP Characterization

This section focuses on the characterization of the SMTP workload, that is, the requests processed by the mail server to send or receive messages. As previously stated, our analysis focuses on the various aspects of the workload and, in particular, on its intensity and functional characteristics.

The workload intensity is first analyzed by studying the arrival process of the requests. Figure 1 shows the number of arrivals per minute, as measured on the ISP mail server for one of the days considered in our study, namely, May 10, 1999. The number of requests received

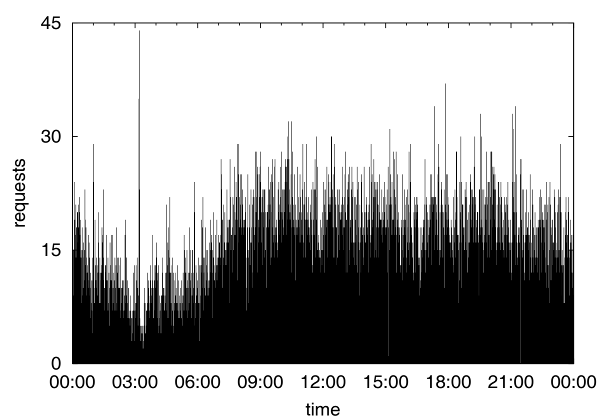

Figure 1: Arrival rate of the SMTP requests over a time period of 24 hours.

by the mail server over the 24 hours is equal to 23,880, with an average arrival rate of 16.58 requests/minute. As can be seen, the arrivals exhibit large fluctuations. The load is light in the early hours of the day, with a deep valley around 3am, then it gradually increases. Graphically, we have noticed that the burstiness of the arrivals does not change when varying the time scale, that is, the aggregation level of the arrivals. This scale-invariant or self-similar behavior is also proved by the value of the Hurst parameter, equal to 0.9186, estimated by applying standard tests, such as, the time–variance plot [LTW+94]. Values of this parameter close to one denote a high degree of self-similarity.

Another figure describing the intensity of the workload is represented by the volume of data exchanged by the messages sent and received by the mail server. Figure 2 shows the volume of data, expressed in kbytes, exchanged per minute.

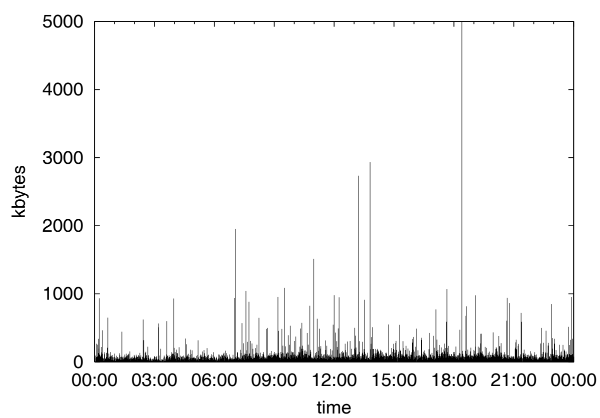

Figure 2: Volume of data exchanged per minute over a time period of 24 hours.

The total number of bytes exchanged in 24 hours is about 180 Gigabytes, with an average of 126 kbytes per minute. As can be seen, the early hours of the day are characterized by a low data volume, this volume then increases starting from the morning hours and it even exhibits spikes larger than 2000 kbytes. Despite of the arrivals, the behavior of this series is less bursty, as also denoted by the value of the Hurst parameter, which is equal to 0.573.

By looking at the IP addresses of the sender and recipients of the messages, we could classify the SMTP requests as local or remote. Local requests involve the local mail server only, as opposed to remote requests which involve the local mail server as well as other mail servers reachable over the network. We have noticed that about 82% of the messages processed by the ISP mail server are remote, whereas this percentage decreases to about 58% in the case of the enterprises and University mail servers. This classification provides useful hints on the actual load of the server. Indeed, the delivery of a remote message requires activities, such as, authentication between mail servers, which are not necessary for local messages.

The analysis of the status of the requests shows that about 95% of the messages were successfully delivered without being deferred because of network errors or unavailability of the destination mail server. This figure is valid over all the mail servers considered in our study.

A more detailed description of the quantitative aspects of the workload is obtained by analyzing the distributions of parameters, such as, the interarrival time of the requests, the size of the message and the number of its recipients. Table 2 presents the basic statistics of these parameters.

parameter	mean	std. dev.	min	max
interarrival time [s]	3.62	3.84	0	105
message size [bytes]	7,776.83	52,419.40	176	4,998,263
number of recipients	1.24	3.87	1	406

Table 2: Basic statistics of the parameters of the SMTP workload.

Figure 3 shows the interarrival time distribution. As can be seen, the distribution is highly positively skewed. We have noticed that approximately 30% of the requests are characterized by small interarrival times, that is, 0 or 1 second, and about 5% by interarrival times are in the range between 10 and 105 seconds. This means that most of the requests arrive at the mail server very close apart from each other, even though there is a non negligible fraction of requests with long interarrival times. From these results, it seems that the accuracy of the distribution has not been compromised by the coarse time stamp resolution. Of course, a finer granularity will guarantee a better accuracy.

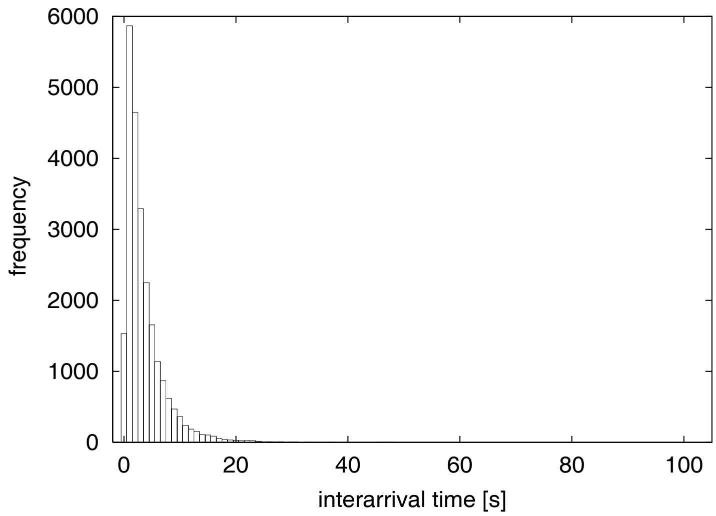

Figure 3: Interarrival time distribution.

Figure 4 shows a zoom the message size the distribution in the range 0-100 kbytes. As in the case of interarrival times, the distribution is highly positively skewed. Its 90–th percentile corresponds to 11 kbytes, and its 99–th percentile is 67 kbytes. Moreover, the distribution is characterized by a standard deviation which is more than six times larger than the corresponding mean. The size of approximately 99.35% of the messages is small, namely, less than 100 kbytes.

As per the number of recipients of the messages, we have discovered that approximately 94% of the messages have one recipient, only 0.02% of the messages have more than 20 recipients and there are messages with as many as 406 recipients.

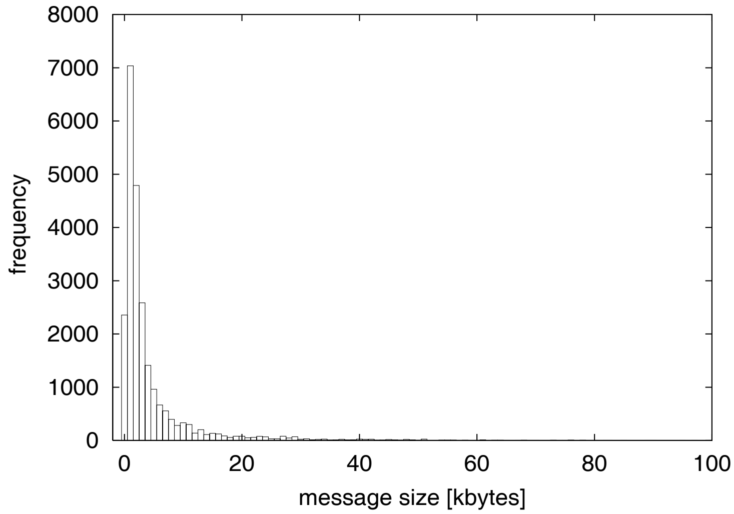

Figure 4: Message size distribution.

To further investigate the behavior of the workload and to develop the appropriate distributional models, we applied numerical fitting techniques, based on the least squares method. These techniques identify the parameters of the distribution which best fits the empirical data. The method used is based on the analysis of the properties of the empirical distribution together with the analysis of an error function associated to the fitting algorithm. For the message size distribution, a log–normal distribution best fits the measured data because most of its mass is concentrated on small values. The shape and location parameters σ and μ identified by fitting techniques are equal to 0.739 and 0.87, respectively.

In the case of the interarrival time distribution, we could not fit the experimental data with one function only because of its heavy tail. We applied right censoring techniques with the objective of finding a threshold between the body and the heavy tail of the distribution. A Weibull distribution, with shape parameter a equal to 3.027 and scale parameter b equal to 1.124, best fits the body of the empirical distribution. A Pareto distribution with shape parameter α equal to 2.459 and location parameter k equal to 3.958, best fits the tail.

Figure 5 shows the empirical interarrival time distribution together with the two distributions identified by fitting techniques. The dotted curve represents the experimental data, the solid curves represent the models. The dashed vertical line denotes the threshold identified by censoring the empirical distribution. This value is equal to 7 seconds. Hence, the Weibull and the Pareto distributions model the interarrival times whose values are below and above the identified threshold, respectively.

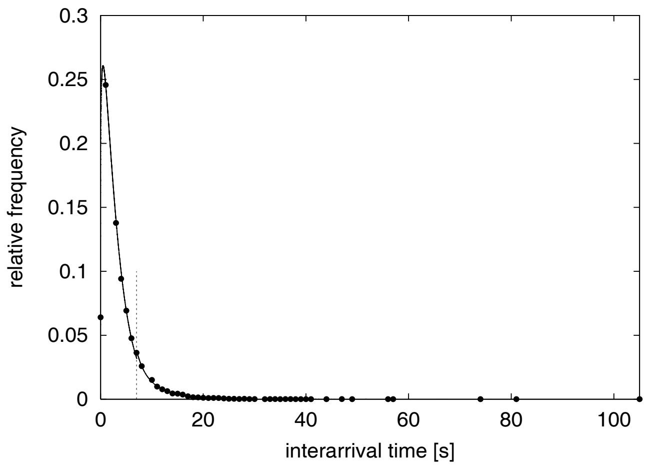

Figure 5: Empirical interarrival time distribution (dotted curve) and corresponding models (solid curves).

To study the sensitivity and the representativeness of our results, we analyzed each of the log files collected on the mail servers. By looking at the corresponding distributional models, we have noticed that the behavior of SMTP workload is quite similar across different mail servers and across different days. Note that all the log files refer to working days, we did not consider any week-end day.

As an example, Figure 6 shows the envelope obtained from all the curves corresponding to the models of the interarrival time distributions of the 25 days considered in our study. The solid curves in the envelope correspond to the centroids of the models, namely, they repre-

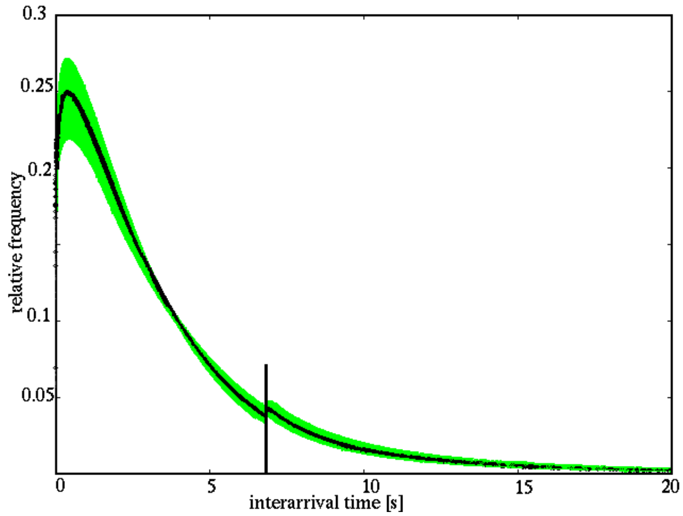

Figure 6: Envelope of the interarrival time distributions.

sent a Weibull and a Pareto distribution whose parameters are obtained as the average of the corresponding parameters of the distributions identified for the various days. As the narrow envelope denotes, the models identified for the various days are very close to each other. Similar conclusions can be drawn for the models identified for the message size distribution.

5 POP3 Characterization

The POP3 workload is the set of requests generated by the users against the mail server to have their messages delivered to their client. Our study focuses on the analysis of these requests with the objective of characterizing their behavior. The characterization of the POP3 workload is then two-folds. We analyze the accesses of POP3 users to the mail server and how the maildrops, i.e., the user mailboxes on the server, vary as a consequence of these accesses.

The parameters considered for each POP3 request are the interarrival time, the user identifier, the number and size of messages deleted from the maildrop and the number and size of the messages left in the maildrop.

Note that the options adopted in the configuration of the POP3 protocol on the client side, determine the operations performed on the server side. For example, a POP3 client can be configured with the option "download and keep", that is, messages passed from server to client are also left on the server. As an alternative, a POP3 client can be configured with the option "download and delete", that is, messages passed to client are deleted on the server.

As already pointed out, the POP3 log files were collected on two mail servers only for a total of 12 days. From the analysis of the arrival rate of each day, we have noticed that, because of the peculiarity of these installations, that is, a mail server used by faculty and staff members, and an enterprise mail server, most of the POP3 requests were issued during working hours. Both servers process approximately 90% of their daily POP3 workload from 8:00am to 7:30pm. Hence, within each day, we restricted our analysis to this time period only.

Table 3 presents the basic statistics of the POP3 requests issued during one of the days considered in our study, namely, April 28, 1999. The data refers to the mail server of our University. The number of requests processed during the working hours is equal to 21,116.

parameter	mean	std. dev.	max
interarrival time [s]	1.96	2.32	40
number of messages per maildrop	13.98	59.04	590
maildrop size [bytes]	437,643.25	1,789,244.25	12,569,464
size of deleted messages [bytes]	54,857.38	314,329.51	9,304,370

Table 3: Basic statistics of the parameters describing POP3 requests.

As can be seen, all the parameters are characterized by a large variability. Each maildrop contains on the average about 14 messages, and there are maildrops which contain as many as 590 messages. The average maildrop size is about 427 kbytes, and there are maildrops whose size is as large as 12 Mbytes.

As for SMTP workload, to describe the overall behavior of the POP3 workload, we have derived analytic distributional models of these parameters. For example, a Weibull distribution, with parameters equal to 2.184 and 1.078, best fits the interarrival time distribution.

To obtain a more detailed description of the POP3 workload, we have characterized the requests by analyzing the behavior of the users as seen on the server side. In particular, we studied what a request does in terms of maildrop manipulations. For this purpose, we described each request by two attributes (D, F). The attribute D ("Delete") denotes whether any message was deleted from the maildrop $(D = 1)$, or not $(D = 0)$. The attribute F ("Full") denotes the status of the maildrop, that is, whether it contains any message $(F = 1)$, or not $(F = 0)$. By using these attributes, the actions performed by each request can be described by one of the following four states:

- `state 00`: "empty session", that is, the maildrop is empty, hence, no deletion can take place;
- `state 01`: "download and keep", that is, all messages are left in the maildrop (which is non empty);
- `state 10`: "download and delete", that is, all messages are deleted from the maildrop which is then emptied;
- `state 11`: "delete and keep", that is, a few messages are deleted from the maildrop and a few messages are left in the maildrop.

The characterization of these four classes of requests is very important because of their different impact on the performance of a mail server. To reproduce the dynamic behavior of the requests, that is, the sequences of actions performed on user maildrops, we have introduced models based on probabilistic graphs. The nodes of these graphs, one for each user, represent the various states as defined above, and the arcs, with their associated probabilities, represent the transitions among these states. A user graph is then described by its states and the corresponding transition probability matrix $P = [p_{s_i,s_j}]$, $(s_i, s_j \in S)$, where $S = \{00,\ 01,\ 10,\ 11\}$ denotes the space state. From the transition probability matrix P, we obtain the steady state probabilities, π_{s_i}, $(s_i \in S)$, that is, the limiting state probabilities of being in a given state, by solving the system of linear equations

$$\pi_{s_j} \;=\; \sum_{s_i \in S} \pi_{s_i}\ p_{s_i,s_j} \qquad s_j \in S$$

with the condition $\sum_{s_i \in S} \pi_{s_i} \;=\; 1$.

We built such a graph for each of the 636 users who checked their mail more than once, that is, issued more than one POP3 request against the mail server. By analyzing the steady state probabilities for these graphs, we have noticed that the most popular state is the state 00, which corresponds to an empty session. On the average, the probability for a user of opening an empty POP3 session is approximately 0.71. This is a consequence of the option set by most users on the client side to automatically periodically check for mail on the server.

To characterize users with similar behavior, we have applied clustering techniques. Each graph was described by its steady state probabilities π_{s_i}. The k–means algorithm adopted for our analysis classified the graphs into an optimal partition consisting of three groups. Table 4 presents the centroids, i.e., the geometric centers, of the these groups. The first group is a big group containing about 93.87% of the user graphs. The graphs belonging to this group are characterized by a high probability of being in state 00. The graphs of the second group, which are about the 5.66% of the user graphs, are characterized by high probability of being

parameters	group 1	group 2	group 3
π_{00}	0.747	0.070	0.
π_{01}	0.014	0.855	0.361
π_{10}	0.239	0.045	0.111
π_{11}	0.	0.030	0.528

Table 4: Steady state probabilities of the centroids of the three groups of graphs.

in state 01. Finally, the graphs of the third group, which account for the remaining 0.47%, are characterized by a large value of π_{11}. Figure 7 shows an example of a graph belonging to group 1.

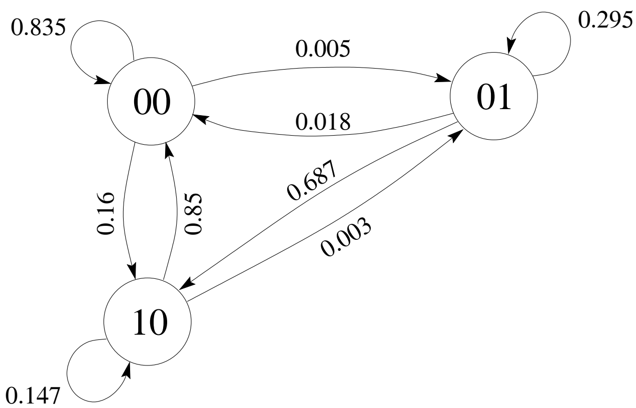

Figure 7: Probabilistic graph belonging to group 1.

To assess the validity and representativeness of these graph based models, we have applied reassignment techniques. We have found that more than 90% of the graphs built for all the users of the mail servers considered in our study fall in one of the three groups identified for the users of one day only and the percentages of graphs in each group are roughly maintained. Moreover, the remaining 10% mainly refers to graphs consisting of one node only, corresponding to either the state 00 or 11. We can then conclude that this characterization is representative of the workload of both mail servers.

6 Conclusions

Performance of mail servers is a challenging issue because of the explosive growth of subscribers and of the intensive use of email by existing as well as new subscribers. A detailed analysis of the workload of mail servers allows organizations providing email services to understand the current status of their servers and to derive useful hints on how to manage and configure these systems.

The workload models presented in this paper have been obtained by extensively analyzing a large set of measurements collected on mail servers representative of different environments. We have studied both SMTP and POP3 workloads. We have derived analytic models to describe the behavior of both types of workloads. Our models are being used as the basis for the definition of the intensity and the characteristics of the workload of SPECmail2001. Because of the confidentiality of the project, we are not able to provide any further details about the benchmark design and specifications.

Future work will be dedicated to extend these studies to the load generated on the mail server by emerging mail access protocols, such as, web mail and IMAP4 [RFC2060], and by proprietary protocols. Indeed, the wider availability of high speed connections based on new technologies, such as, coaxial cables and Digital Subscriber Loop, makes feasible the use of more sophisticated email services even for residential subscribers.

References

[AFJ99] M. Arlitt, R. Friedrich, and T. Jin. Workload Characterization of a Web Proxy in a Cable Modem Environment. Technical Report HPL-1999-48, HP Labs., 1999.

[AV95] F.M Avolio and P.A. Vixie. *Sendmail - Theory and Practice.* Digital Press, 1995.

[AW97] M.F. Arlitt and C.L. Williamson. Internet Web Servers: Workload Characterization and Performance Implication. *IEEE/ACM Transactions on Networking*, 5(5):631–645, 1997.

[BC98] P. Barford and M.E. Crovella. Generating Representative Web Workloads for Network and Server Performance Evaluation. In *Proc. ACM SIGMETRICS Conf.*, pages 151–160, 1998.

[CB97] M.E. Crovella and A. Bestavros. Self–similarity in World Wide Web Traffic: Evidence and Possible Causes. *IEEE/ACM Transactions on Networking*, 5(6):835–846, 1997.

[CMT00] M. Calzarossa, L. Massari, and D. Tessera. Workload Characterization - Issues and Methodologies. In G. Haring, C. Lindemann, and M. Reiser, editors, *Performance Evaluation – Origins and Directions*, volume 1769 of *Lecture Notes in Computer Science*, pages 459–484. Springer, 2000.

[CS93] M. Calzarossa and G. Serazzi. Workload Characterization: A Survey. *Proc. of the IEEE*, 81(8):1136–1150, 1993.

[Ferr84] D. Ferrari. On the Foundations of Artificial Workload Design. In *Proc. ACM SIGMETRICS Conf.*, pages 8–14, 1984.

[Hart75] J.A. Hartigan. *Clustering Algorithms*. Wiley, 1975.

[ISZ99] A.K. Iyengar, M.S. Squillante, and L. Zhang. Analysis and Characterization of Large-scale Web Server Access Patterns and Performance. *World Wide Web - An International Journal*, 2(1-2):85–100, 1999.

[LH74] C.L. Lawson and R.J. Hanson. *Solving Least Squares Problems*. Prentice–Hall, 1974.

[LTW+94] W.E. Leland, M.S. Taqqu, W. Willinger, and D.V. Wilson. On the Self–similar Nature of Ethernet Traffic (Extended Version). *IEEE/ACM Trans. on Networking*, 2(1):1–15, 1994.

[MAFM99] D. Menascé, V.A.F. Almeida, R. Fonseca, and M.A. Mendes. A Methodology for Workload Characterization of E–commerce Sites. In *Proc. ACM Electronic Commerce Conf.*, 1999.

[RFC821] J. Postel. Simple Mail Transfer Protocol. RFC 821, August 1982. http://www.cis.ohio-state.edu/htbin/rfc/rfc821.html.

[RFC1939] J. Myers and M. Rose. Post Office Protocol - Version 3. RFC 1939, May 1996. http://www.cis.ohio-state.edu/htbin/rfc/rfc1939.html.

[RFC2060] M. Crispin. Internet Message Access Protocol - Version 4. RFC 2060, December 1996. http://www.cis.ohio-state.edu/htbin/rfc/rfc2060.html.

[SPEC] Standard Performance Evaluation Corporation. http://www.spec.org.