Accounting for characteristics of session workloads: A study based on partly-open queue

Nikola Janevski and Katerina Goseva-Popstojanova
Lane Department of Computer Science and Electrical Engineering
West Virginia University, Morgantown, WV 26506, USA
E-mail: njanevsk@mix.wvu.edu, katerina.goseva@mail.wvu.edu

Abstract—**Many systems, including Web and Software as a Service (SaaS) are best characterized with session-based workloads. Empirical studies have shown that Web session arrivals exhibit long range dependence and that the number of request in a session is well modeled with skewed or heavy-tailed distributions. However, models that account for session workloads characterized by empirically observed phenomena and studies of their impact on performance metrics are lacking. In this paper, we use partly-open queue to account for session-based workloads in a physically meaningful way and use simulation to analyze the behavior of the Web system under Long Range Dependent (LRD) session arrival process and skewed distribution for the number of requests in a session. Our results show that the percentage of dropped sessions, mean queue length, mean waiting time, and the useful server utilization are all affected by the LRD session arrivals and the statistics of the number of requests within a session. The impact is higher in the case of more prominent long-range dependence. Interestingly, both request arrival process and request departure process are long-range dependent, even in the case when session arrivals are Poisson.**

I. Introduction

Many businesses are using Web technologies to build new communication channels with customers around the globe. Therefore, it is of crucial importance to be able to assess Web system performance realistically and assure the quality of service. Traditionally, evaluation of Web server performance accounted for request-based workloads and it was focused on assessment and prediction of request-based metrics (e.g., throughput in number of completed requests, percentage of dropped requests, and so on). Web workload, however, is in a form of sessions, each consisting of multiple individual requests originated from the same user. For example, placing an order on an e-commerce Web site involves requests relating to selecting a product, providing payment and shipping information, and receiving a confirmation. So, for a customer trying to place an order or a retailer trying to make a sale, the real measure of a Web server performance is its ability to complete the entire sequence of requests within a session [4].

Using either closed or open queuing system with request-based workload does not account for session characteristics and therefore does not result in a realistic model of a Web system. Instead, in this paper we use so called partly-open queue which accounts for session-based workloads (see Figure 1). In this queue sessions arrive as in open queue, but for each request within a session the user sends a new request only after receiving a response on the previous request. In other words, partly-open queue allows us to model a varying number of users at the site over time as in open system (rather than a fixed number of users N as in closed system). On the other side, it behaves as a closed system for the requests within each session. Upon completing all requests in a session, the user leaves the system, again as in open system.

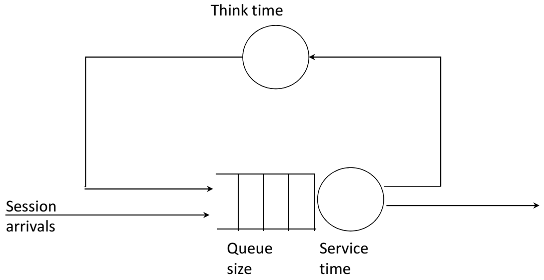


Fig. 1. Partly-open queueing system

In addition, we consider a queue with a finite size, since our interest is the throughput in successfully completed sessions, that is, the percentage of sessions being dropped due to the queue being full. When a server works under high utilization the queue length tends to grow to the point when the queue becomes full, resulting in dropping the incoming request. For a server that runs session-based workloads a dropped request could be anywhere in the session, and will lead to aborted, incomplete session. Obviously, the quality of service of a Web system or SaaS system, from both user's and provider's perspective, is best assessed by the number of successfully completed sessions. Considering only request-based workload does not allow for assessment of the percentage of dropped sessions (i.e., unsatisfied users) or the amount of server utilization wasted on completing requests from aborted sessions.

The realism of the assessment of performance metrics is not based only on the type of the queuing model (i.e., open, closed or partly-open), but also on the models and distributions used for the associated random variables and the values of the corresponding parameters. In choosing these models, distributions, and parameters our work is motivated by recent empirical results. For example, recent studies on session Web workloads have shown that the arrival of Web sessions, for systems under moderate to high load, is a Long Range Dependent (LRD) process (i.e., asymptotically second order self-similar process) [23], which means that sessions arrive in bursts over many time scales. In addition, the number of requests in a session follows a skewed or heavy-tail distribution [9], [16]. Motivated by these empirical results, in our partly-open queueing system we use LRD process for session arrivals and then distributions with different coefficients of variation for the number of requests in each session.

In what follows we briefly discuss the related work on performance assessment of Web systems. We first address the closed queueing models. In [17] the authors showed that having an autocorrelated service time in a closed queueing network propagates the autocorrelation to all tiers. An Approximate Mean Value Analysis (AMVA) algorithm for analysis of closed queueing networks with service time modeled as a MAP process was proposed in [2]. Modeling mean response time and throughput in a closed queueing network model of multi-tier Web system was done in [21]. All these models assumed fixed number of users (typical for closed queueing systems), request-based workloads, and infinite queues.

Open queueing models have also been used for modeling Web systems performance. One of the earliest papers that considered the request loss probability [13] used single open queue with Poisson arrivals and exponentially distributed service time. In [12] authors presented approximate analytical results for the queue length, request loss, and waiting time, for an open queue with fractional Brownian motion (fBm) request arrival process and long-tailed service time. Even though some of the open queue models considered finite queues and/or LRD arrival process, all of them accounted for request-based workloads only.

It appears that the only studies that took into account session-based workload are [3], [4], and [20]. In [3], authors proposed an overload control mechanism for closed queues with geometrically distributed numbers of requests in a session. Session-based admission control mechanism was proposed in [4], where the number of requests in a session was modeled with an exponential distribution, while the session arrival process was not explicitly specified. The goal of [20] was to study the difference between open and closed queueing models and to explore the use of a partly-open queue as a model for systems with session workloads. The session arrivals in [20] were modeled with a Poisson process, the number of requests per session was assumed to follow a geometric distribution, and the queue size was infinite.

The main contributions of our work are as follows:

- We use a partly-open queue to account for session-based workloads. So far, the only paper in the literature that used partly-open queues is [20]. However, [20] considered infinite queue size, Poisson session arrivals, and geometric distribution for the number of requests in a session. Furthermore, the mean response time was the only output metric explored in [20].
- Based on the empirical findings in [23], we model session arrivals with a long-range dependent (LRD) process. In addition, we use Poisson session arrival process, which is obtained by reshuffling the LRD process to have independent arrivals, allowing for a fair comparison with models that assume Poisson session arrivals, such as [20].
- The number of requests within a session is modeled with discrete lognormal distribution with different means and coefficients of variation to explore the impact of these statistics on several performance metrics. In the related work which considered session-based workloads, the number of requests in a session was modeled with an exponential [4] or geometric distribution [20]. Both of these distributions have a tail that decays exponentially which is not the case with real Web workloads [9], [16].
- We use several performance metrics, such as the percentage of dropped sessions, mean queue length, mean waiting time, and useful utilization. In addition, we explore the nature of the request arrival process and request departure process, both of which are dependent not only on the session arrival process and the distribution of the number of requests per session, but also on the service and think time. These processes have not been studied in the related work on session-based workloads.

Our results show that the percentage of dropped sessions, mean queue length, mean waiting time, and the useful server utilization are all affected by the LRD session arrivals and the statistics of the number of requests within a session. The impact is higher in the case of more prominent long-range dependence. Interestingly, both request arrival process and request departure process are long-range dependent, even in the case when session arrivals are Poisson. Our findings have strong practical implications on the performance assessment of Web systems, as well as on developing scheduling policies and admission control policies.

II. Approach

Not much work exists on analytical solution of partly open queues. Even more, using LRD session arrival process, with skewed distributions for the number of request per session, and finite queue size impose using simulations to solve the partly-open queuing system. The models/ distributions and corresponding parameters used for the random variables associated with the partly-open queue are given in Table I and are briefly described next.

Motivated by the empirical findings for Web servers working under moderate and heavy workloads [23], we use LRD process to model *session arrivals*. Web traffic has dual nature [15], that is, both the number of sessions per second (i.e., the count) and the inter-arrival time can be LRD. In order to achieve the dual nature of the session arrivals, we use the method of inverse transformation proposed in [11]. A LRD process, $\{Y_k\}$, with a marginal cumulative distribution function (CDF), $F_Y(y)$, can be generated from another LRD process, $\{X_k\}$, with CDF $F_X(x)$ using the transformation:

$$\{Y_k\} = F_Y^{-1}(F_X(X_k)), \quad k = 1, 2, ... \tag{1}$$

where, F_Y^{-1} is the inverse CDF of $\{Y_k\}$. This transformation actually first transforms the sequence $\{X_k\}$ into a uniformly distributed random variable ($F_X(X_k) \sim U(0,1)$) and then generates the sequence $\{Y_k\}$ using the inverse CDF, F_Y^{-1}, of the desired marginal CDF, F_Y [19]. In addition to generating

TABLE I
MODELS/ DISTRIBUTIONS AND THE CORRESPONDING PARAMETERS

Random variable	Model	Parameters
Session arrivals	LRD process	$H_{sessions} = \{0.6, 0.8\}$
	Poisson	reshuffled LRD process
Number of requests in a session	Discrete lognormal distribution	$mean = \{2, 6, 11\}, C = \{0.5, 1.5\}$
Service time	Pareto distribution	$\mu_s = 57ms, \alpha_s = 1.6$
Think time	Pareto distribution	$\mu_t = 5000ms, \alpha_t = 1.6$

a sequence with desired marginal distribution, the LRD is also preserved [11]. For our simulation $\{X_k\}$ is the Fractional Gaussian Noise (FGN), which we simulated using the FFT method proposed by Paxson [18] because it provides fast way to simulate FGN and still preserves all the relevant statistical properties [8]. For $F_X(x)$ we use normal distribution because FGN has a normal marginal distribution. Finally, for $F_Y(y)$ we use the exponential distribution. As a result, $\{Y_k\}$ defined with equation (1) is a LRD process with exponentially distributed inter-arrival times. This enables us to obtain Poisson process for session arrivals by reshuffling of Y_k and thus carry on a fair comparison with the LRD arrival process.

A predominant way to quantify the long-range dependence is through the Hurst exponent, H, which is also a parameter of the FGN. For a LRD process $0.5 < H < 1.0$. We use two values $\{0.6, 0.8\}$ for the Hurst exponent $H_{session}$ of the LRD session arrival process. Thus, $H_{session} = 0.6$ represents low level of LRD. Considering values of $H_{session}$ higher than 0.8 is not of practical interest because such values have not been observed in the empirical research.

For the *number of requests in a session* we use the discrete lognormal distribution [6] whose probability mass function $P(X = r)$ is given by

$$P_r(\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \frac{1}{r!} \int_0^\infty e^{-\lambda} \lambda^{r-1} \exp\left(-\frac{(\ln\lambda - \mu)^2}{2\sigma^2}\right) d\lambda \tag{2}$$

where μ is the location parameter and σ is the shape parameter. (For the relationship between these parameters and the mean and standard deviation see [6].)

The discrete lognormal distribution is flexible; it can decay slower than the geometric distribution and have higher variance, or it can decay faster and have lower variance. This allows us to examine the impact that the coefficient of variation has on the performance metrics of interest. For the mean of the discrete lognormal distribution, based on the findings in [9] and [20], we use three values: $\{2, 6, 11\}$ requests. For the coefficient of variation C, defined as the ratio of the standard deviation to the mean, we use the values $\{0.5, 1.5\}$. Thus, discrete lognormal distribution with $C = 0.5$ has lower variance then the geometric (or exponential) distribution, while for $C = 1.5$ it has higher variance.

For the *service time*, motivated by [5], we use Pareto distribution. For the mean value of the service time, μ_s, we use $57ms$ which is in the range of values reported in [20], while, for the tail of the Pareto distribution we use, $\alpha_s = 1.6$, because for this value the distribution is heavy-tailed and also the simulations are stable [5]. For the *think time* we also use Pareto distribution which is in agreement with empirical and theoretical research [1]. The mean value $\mu_t = 5000ms$ was chosen as in [4] and $\alpha_t = 1.6$. For the analysis presented in this paper the distributions and parameters of the service time and think time are kept fixed because our focus is on the impact of the LRD session arrival process and characteristics of the number of request within a session distribution on the performance metrics.

Finally, for the *queue size* (i.e., the maximum number of requests in the queue) we use 511 requests, which is the default value for Apache [14]. The scheduling policy used is FCFS.

In the simulation, sessions arrive as in open system. If the session has more than one request, the next request is generated after the first request was served and an amount of think time has passed. In other words, requests belonging to a same session are processed as in closed system. Of course, at any point of time there may be multiple active sessions in the partly-open queue. If a request arrives but the queue is full then that request and the session it belongs to are dropped.

We wrote a program in R language to run the simulations. The removal of the transient warm-up and cool-down periods from the simulations was done by visual inspections of the request arrival process. The validation of the simulation was done by checking the limiting cases, i.e., for the number of requests in a session equal to one the partly-open queue becomes an open queue, while for a very high number of requests in a session it becomes closed queue [20].

III. Analysis of the main findings

We study the impact of the session workload characteristics (i.e., session arrival process and the number of requests within a session) on several performance metrics: percentage of dropped sessions, mean queue length, mean waiting time, and useful server utilization. We also study the characteristics of the request arrival process and request departure process, which in case of partly-open queue are dependent on session arrivals, the number of request per session, as well as on the service and think time. The time series of the number of session arrivals, number of request arrivals, queue length, and number of dropped sessions (for $H_{session} = 0.8, C = 1.5$ and utilization of 92%) are shown in Figure 2.

The **percentage of dropped sessions**, as shown in Figure 3, is highest for LRD session arrivals with high value of the Hurst exponent $H_{session} = 0.8$. The percentage of dropped sessions for less self-similar arrivals (i.e., $H_{session} = 0.6$) is very close to the case with Poisson session arrivals. In particular, for a high utilization values, the 10-15% more sessions are dropped for highly self-similar process (i.e., $H_{session} = 0.8$) then when session arrivals follow the Poisson distribution obtained by reshuffling the LRD process.

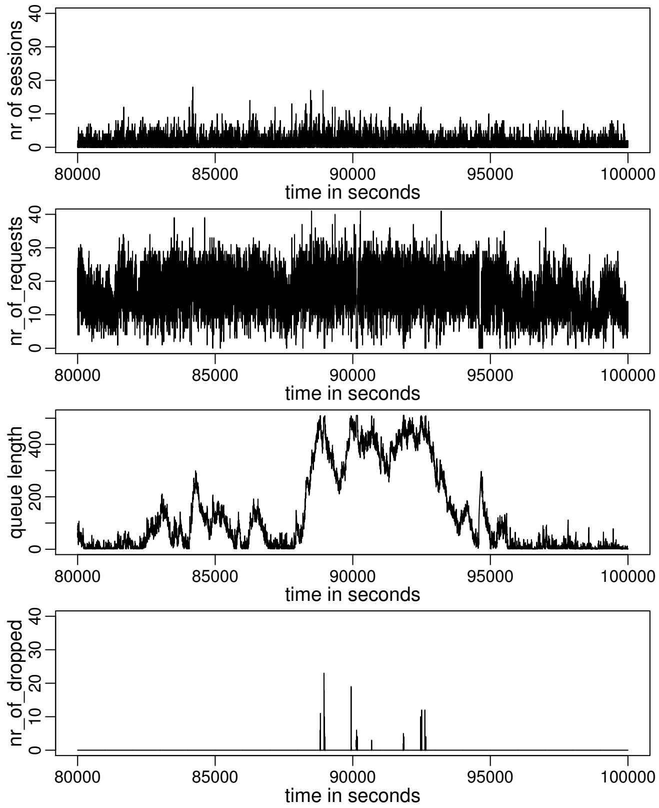


Fig. 2. Time series of performance metrics

Moreover, the percentage of dropped sessions is larger for higher mean number of requests per session (see Figures 3 (b) and (c)), with a more noticeable impact of the LRD, observable for lower utilization values than in case of mean number of requests per session equal to two requests (Figures 3 (a)).

When the number of requests per session has higher coefficient of variation (i.e., $C = 1.5$) the percentage of dropped sessions is smaller then for lower coefficient of variation ($C = 0.5$), for each value of $H_{session}$ (i.e., 0.8 and 0.6), as well as for Poisson session arrivals. This counterintuitive result is explained by the fact that for $C = 1.5$ a significant number of longer sessions are generated compared to the case when $C = 0.5$. A closer inspection of the number of requests in the dropped sessions shows that these longer session are dropped, which decreases the load in number of requests (i.e., frees the queue for new arrivals). This also shows that the server discriminates against longer sessions which typically will have less chance to complete all requests. Having in mind that in e-commerce sites sessions that include purchase are typically much longer than sessions in which users only browse the site [4], this result means that although the overall percentage of dropped sessions is smaller for higher coefficients of variation ($C = 1.5$), the e-commerce business may experience loss of revenue. In addition, the server wastes its resources on completing requests that belong to long sessions that may be dropped under high utilization.

The **mean queue length**, shown in Figures 4 (a), (b) and (c), is also affected by the LRD of the session arrival process, with the highest mean queue length for the high self-similar session arrivals ($H_{session} = 0.8$), and less significant difference between ($H_{session} = 0.6$) and Poisson arrivals. The reason for this behavior is the fact that under the LRD model sessions arrive in bursts and tend to fill in the queue fast, resulting in larger mean queue length.

Comparing Figure 4 (a) to Figures 4 (b) and (c) we observe that smaller mean number of requests per session (i.e., 2 requests) results in larger mean queue length, especially noticeable for high values of Hurst exponent ($H_{session} = 0.8$). This is due to the fact that less sessions are dropped when the mean number of requests per session is smaller (see Figure 3), which actually increases the number of requests in the queue.

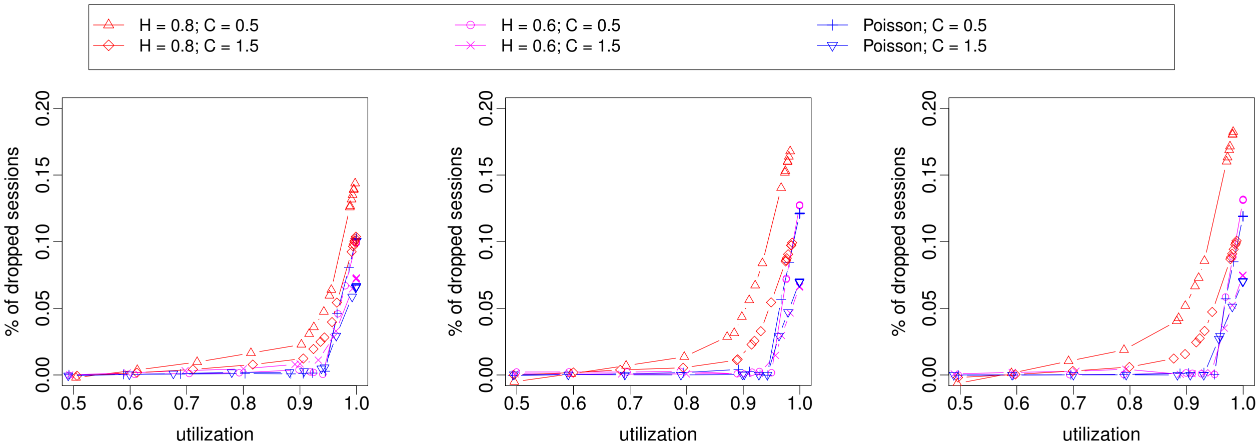
As in case of the percentage of dropped sessions, higher coefficient of variation $C = 1.5$ of the number of requests per session results in lower mean queue length then when $C = 0.5$, which again is due to the fact that for $C = 1.5$ longer sessions are generated, which tend to be dropped from the queue more often then short sessions. The coefficient of variation C, however, has smaller impact on the queue length than on percentage of dropped sessions.

The observations for the **Mean waiting time** are similar to the observations made for the mean queue length. The reason being, the mean waiting time depends on the mean queue length; the higher the mean queue length the higher the mean waiting time of requests in the queue. (The figures for the mean waiting time are not shown due to space limitations.)

The **useful request utilization** is defined as the ratio of the completed requests that belong to successfully completed sessions and the total number of completed requests (including those requests that have been completed, but belong to dropped sessions). As seen in Figures 5 (a)–(c) highly LRD session arrivals can have $5 - 8\%$ lower useful request utilization then in case of Poisson session arrivals. Also, the higher the mean number of requests in a session the lower the useful request utilization. Note that although for higher C the percentage of dropped sessions is lower, the useful request utilization is significantly affected because the server wastes resources on completing requests in longer sessions that are dropped.

Request arrival process in the case of partly-open queue depends not only on the session arrivals and the number of requests per session, but also on the service time and think time, as in closed queueing system. Therefore, we study the long-range dependence of the request arrival process. We use the Abry-Veitech method [22] to estimate the Hurst exponent of the request arrival process. (Figures plotting the values of the Hurst exponent are not shown due to space limitations.) The results show that the request arrival process in the partly-open queue is also LRD, with somewhat higher values of the Hurst exponent than the session arrival process, which is expected, having in mind that there are more request arrivals than session arrivals (see Figure 2). Similarly, the **request departure process** from the partly-open queue, which sums all the request departing from the server, is also LRD process, but it has lower value of the Hurst exponent than the request arrival process.

More interesting observation is that both the request arrival

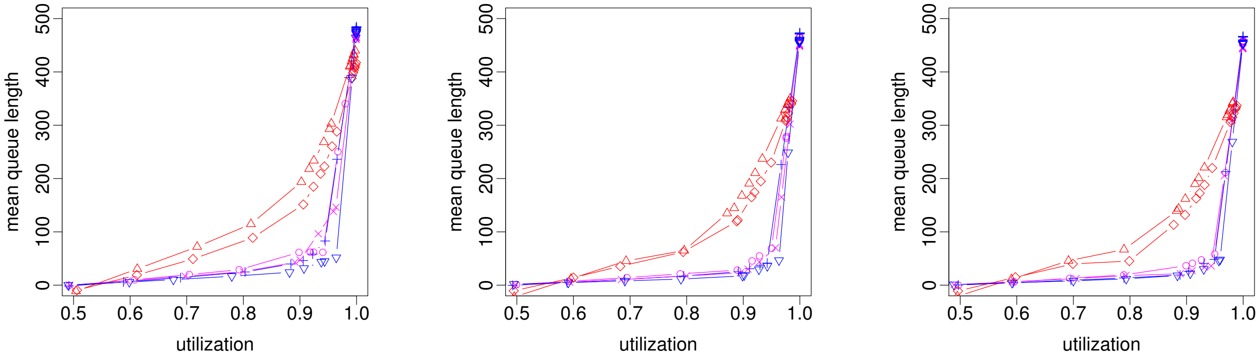


(a) mean = 2 requests

(b) mean = 6 requests

(c) mean = 11 requests

Fig. 3. Percentage of dropped sessions

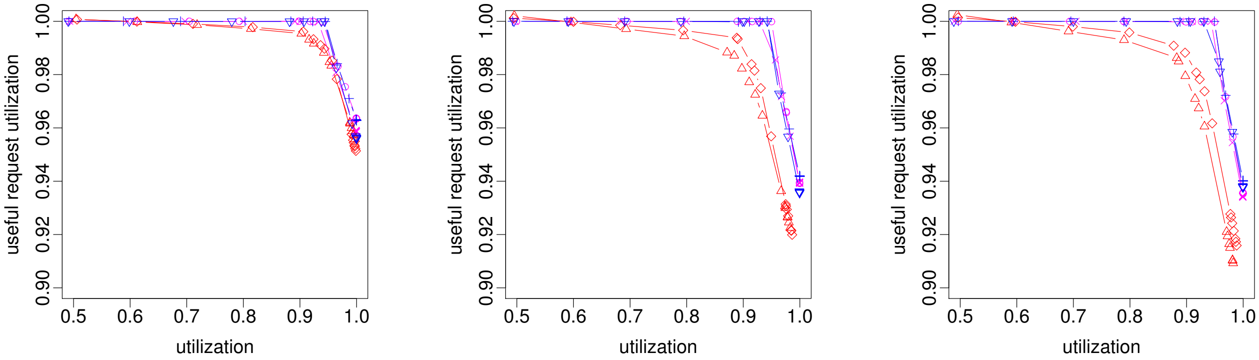


(a) mean = 2 requests

(b) mean = 6 requests

(c) mean = 11 requests

Fig. 4. Mean queue length



(a) mean = 2 requests

(b) mean = 6 requests

(c) mean = 11 requests

Fig. 5. Useful request utilization

process and request departure process are LRD even when session arrival process is Poisson, for all considered values of the mean and coefficient of variation of the number of requests per session. It should be noted that, based on empirical observations, we use heavy-tailed distributions to model both the service time and think time. The fact that heavy-tail service time is a reason for long range dependence of the output (i.e., departure) process has been established as a theoretical result for an open queue with infinite queue size in [7]. Our work generalizes this theoretical result having in mind that we consider partly-open queue with a finite queue size. Furthermore, in our case, in addition to the request departure process, the request arrival process is also LRD, even when session arrivals are Poisson.

Another work related to this result is the semi-experimental methodology used for manipulation of IP level traffic presented in [10]. That work considers IP flow arrivals, which similarly to HTTP session arrivals, were shown to be LRD. By randomizing the arrival process of the flows to become Poisson process, while maintaining the full integrity of the packet arrivals patterns within each flow, the authors showed that the LRD of the flow point process does not significantly affects the LRD at packet level. Our result based on simulation of partly-open queue is consistent with [10]. Our work, however, shows that in the context of performance assessment and prediction, the LRD nature of the session arrival process has to be considered, even though it does not significantly affect the LRD of the request arrival process, because it significantly affects (especially under high utilization) the queue length, and thus the percentage of dropped sessions and the useful utilization of the server.

IV. Conclusion

In this study we use a partly-open queue model to account for session-based workload and study the impact of its characteristics (i.e., the session arrival process and the number of request per session) on performance metrics. In particular, we first assume a point process for the session arrivals, and then another model for intra-session characteristics (in this case a distribution for the number of requests within a session). By considering LRD session arrivals (instead of Poisson), skewed distribution for the number of request within a session (instead of geometrical distribution), and a finite queue size (instead of infinite) our work generalizes the existing work on partly-open queues and prior simulation study which considered session workloads.

We explore several performance metrics: percentage of dropped sessions, average queue length, waiting time, and useful server utilization. In addition, we pay particular attention on the nature of the request arrival and request departure point processes, which has not been done in the related work.

To summarize, our results show that the LRD of the session arrival process, especially for higher values of the Hurst exponent, has bigger impact on all performance metrics (i.e., percentage of dropped sessions, mean queue length, mean waiting time, and useful utilization) when compared to the statistics of the number of requests per session. Performance metrics for lower values of the Hurst exponent (e.g., $H_{session} = 0.6$) are close to the values for Poisson session arrivals.

More interestingly, for a higher variation of the number of request per session the percentage of dropped sessions is smaller, as well as the mean queue length and mean waiting time. This is due to the fact that longer sessions have greater chance to be dropped, which will affect less users. However, this results in lower useful utilization of the server, which has been busy serving requests of the dropped sessions. Furthermore, it is likely to lead to loss of revenue in the case of e-commerce sites which typically observe more purchasing requests in the longer sessions.

When session arrivals are Poisson, performance metrics are affected only at very high utilization and the impact is less significant. More interesting observation is that in a partly-open queue with heavy-tailed service time and think time both the request arrival process and request departure process are LRD, even when the session arrivals are Poisson. Our future work is focused on further exploring the effect of the service time and think time distributions and the values of their parameters on the correlation structure of the request arrival and request departure processes.

The results presented in this paper have several strong implications for performance modeling. It is obvious that in order to build a realistic model for systems with session-based workloads, both inter-session characteristics (e.g., LRD of session count and inter-arrivals) and intra-session characteristics (e.g., distribution of the number of request per session) have to be modeled accurately. The combination of partly-open queue with finite queue size under LRD session arrivals and skewed distribution of the number of requests in a session, provides a model with physically meaningful parameters, which preserves both the session and the request characteristics. Furthermore, we show that although the session arrival model is not affecting significantly the LRD of the request arrival and request departure processes, it has to be taken into account because the LRD at session level determines the queueing delays and session losses, and thereby the quality of delivered services.

Acknowledgments

This work is funded in part by the National Science Foundation under the grants CNS-0447715 and CCF-0916284.

References

[1] P. Barford and M. Crovella, “Generating representative Web workloads for network and server performance evaluation,” *Performance Evaluation Review*, vol. 26, pp. 151–160, 1998.

[2] G. Casale and E. Smirni, “MAP-AMVA: Approximate Mean Value Analysis of bursty systems,” *IEEE/IFIP Int'l Conf. Dependable Systems Networks*, vol. 2, pp. 409–418, July 2009.

[3] H. Chen and P. Mohapatra, “Session-based overload control in QoS-aware Web servers,” *21st Int'l Conf. Computer Communications*, vol. 2, pp. 516 – 524, 2002.

[4] L. Cherkasova and P. Phaal, “Session-based admission control: A mechanism for peak load management of commercial Web sites,” *IEEE Trans. Computers*, vol. 51, pp. 669–685, 2002.

[5] M. E. Crovella and L. Lipsky, *Self-Similar Network Traffic and Performance Evaluation*, John Wiley & Sons, Inc., 2002.

[6] E. Crow and K. Shimizu, *Lognormal distributions: theory and applications*, M. Dekker, 1988.

[7] D. J. Daley, R. Vesilo, “Long range dependence of point processes, with queueing examples”, *Stochastic Processes and their Applications*, vol. 70, no.2, pp. 265–282, 1997.

[8] T. Dieker, “Simulation of fractional Brownian motion,” Master's thesis, Vrije Universiteit Amsterdam, 2002.

[9] K. Goseva-Popstojanova, F. Li, X. Wang, and A. Sangle, “A contribution towards solving the Web workload puzzle,” *IEEE/IFIP Int'l Conf. Dependable Systems and Networks*, pp. 505–516, 2006.

[10] N. Hohn, D. Veitch, P. Abry, “Cluster processes: a natural language for network traffic,” *IEEE Trans. Signal Processing*, vol.51, no.8, pp. 2229–2244, 2003.

[11] H.-D. J. Jeong, K. Pawlikowski, and D. C. McNickle, “Generation of self-similar processes for simulation studies of telecommunication networks,” *Mathematical and Computer Modelling*, vol. 38, no. 11-13, pp. 1249–1257, 2003.

[12] X. Jin and G. Min, “QoS analysis of queuing systems with self-similar traffic and heavy-tailed packet sizes,” *IEEE Int'l Conf. Communications*, pp. 100–104, 2008.

[13] M. Kaaniche, K. Kanoun, and M. Martinello, “A user-perceived availability evaluation of a Web based travel agency,” *IEEE/IFIP Int’l Conf. Dependable Systems and Networks*, pp. 709–718, 2003.

[14] B. Laurie and P. Laurie, *Apache: the definitive guide*, O’Reilly, 2002.

[15] J. C. López-Ardao, C. López-García, A. Suárez-González, M. Fernández-Veiga, and R. Rodríguez-Rubio, “On the use of self-similar processes in network simulation,” *ACM Trans. Modeling and Computer Simulation*, vol. 10, pp. 125–151, 2000.

[16] D. A. Menascé, V. A. F. Almeida, R. Fonseca, and M. A. Mendes, “A methodology for workload characterization of e-commerce sites,” *1st ACM Conf. Electronic Commerce*, pp. 119–128, 1999.

[17] N. Mi, Q. Zhang, A. Riska, E. Smirni, and E. Riedel, “Performance impacts of autocorrelated flows in multi-tiered systems,” *Performance Evaluation*, vol. 64, no. 9-12, pp. 1082–1101, 2007.

[18] V. Paxson, “Fast, approximate synthesis of fractional gaussian noise for generating self-similar network traffic,” *SIGCOMM Computer Communication Review*, vol. 27, pp. 5–18, 1997.

[19] S. Ross, *Simulation*, Academic Press, 2002.

[20] B. Schroeder, A. Wierman, and M. Harchol-Balter, “Open versus closed: A cautionary tale,” *USENIX Symp. Networked Systems Design and Implementation*, pp. 239–252, 2006.

[21] B. Urgaonkar, G. Pacifici, P. Shenoy, M. Spreitzer, and A. Tantawi, “An analytical model for multi-tier Internet services and its applications,” *ACM SIGMETRICS Int’l Conf. Measurement and Modeling of Computer Systems*, pp. 291–302, 2005.

[22] D. Veitch and P. Abry, “A wavelet-based joint estimator of the parameters of long-range dependence,” *IEEE Trans. Information Theory*, vol. 45, no. 3, pp. 878–897, 1999.

[23] X. Wang and K. Goseva-Popstojanova, “Modeling Web request and session level arrivals,” *IEEE Int’l Conf. Advanced Information Networking and Applications*, pp. 24–32, 2009.