Long-run performance analysis of a multi-scale TCP traffic model

N.X. Liu and J.S. Baras

Abstract: The long-run queueing performance of a multi-scale TCP traffic model, the HOO model, is analysed. Since the model links the multi-scale behaviour with practical traffic elements and approximates TCP traffic very well, the analysis is expected to provide insights into the physical interpretation of multi-scale traffic and to give useful results for performance prediction. To derive a meaningful solution and avoid the extreme difficulty of an exact analysis, the authors adopt several techniques to track the problem, among which are techniques to establish equivalent processes to the traffic process or the queue content process in two cases, the case of fast flows and the case of slow flows. Quantitative results for the queue tails are obtained in both cases, and a unified form is derived. It indicates that three levels of traffic elements in different time scales, i.e. the connection, the burst and the packet, all affect the asymptotic queueing performance. It shows quantitatively how the connection determines the index of the queue tail, and the burst and the packet contribute to the tail with their averages. Used with simple statistical inferences, the analytical result is shown to predict the queueing performance of real traffic well.

1 Introduction

Since the seminal work of Leland *et al.* [1] on self-similar Internet traffic, scaling phenomena have been widely recognised as dominant characteristics of Internet traffic. This was enhanced later by finding of multifractality [2], which indicates that real traffic has scaling structures in multiple time scales. It has been shown that the multi-scale property may lead to a much worse network performance than does traditional exponential traffic [3–5]. Advances in modelling and performance evaluation of multi-scale traffic may impact network control, planning and operations significantly.

Traffic models proposed with regard to scaling behaviours largely fall into two categories. One large family of models are based on the on–off model and view the traffic as the aggregation of many on–off flows [1, 6, 7]. In the other category are more or less mathematically driven models that are intended to approximate the statistics of the traffic rather than mimic the structure of it. They include fractional Brownian motion [7], the M/G/∞ process [3], random cascade [9], MWM [10] and heavy-tailed renewal reward processes [11], and some others. While both categories of models can explain the real traffic quite satisfactorily in some aspects, they have certain shortcomings. Dating back to early packet-switched telephone networks [12], the on–off models may be too simplified to capture TCP protocol-related traffic textures. The mathematical models do not help much in the physical interpretation of the traffic since practical network and protocol parameters are generally not involved.

We have made an attempt to avoid the above shortcomings by presenting a model that can on the one hand capture the structure of the traffic in fair detail, and on the other hand can have the multi-scale property inherently [13]. The model is justified with protocol and real traffic data analyses and examinations of the multi-scale property. Queueing analysis of the model is important at least in two aspects. First, it would help us understand the effects of different levels/scales of traffic elements on network performance. Secondly, it would provide a basis for performance predication of practical multi-scale traffic, and thus improve network dimensioning and resource allocation in such traffic environment. However, the analysis is not an easy task and no result has been given so far. This paper will be dedicated to the problem and give useful results on the queueing performance of the model. Our primary interest is a meaningful solution for long-run performance from the engineering point of view rather than a thorough, theoretical treatment of the hard problem. The major contributions of this paper are that we adopt various techniques to make the difficult problem tractable, and get a satisfactory result to fulfill our goal.

2 Multi-scale traffic and the hierarchical on–off model

Multi-scale phenomena indicate the complex scaling behaviour of Internet traffic in multiple time scales. Generally speaking, in large scales the traffic shows long-range dependence or self-similarity [10], and in small time scales the traffic is multifractal [2]. In fact, some research claims that the long-range dependence may only exist in finite time scales, and in very long time scales the traffic is likely to be Gaussian (see [14]). This is an easier case, and we will not address it in this paper. Researchers have tried to understand why this happens and what factors contribute to the phenomena. The hierarchical on–off model (HOO) [13] is a structural model trying to link physical traffic elements to multi-scale behaviour. It is simply a natural

IEE Proceedings online no. 20040298

doi:10.1049/ip-com:20040298

Paper received 10th April 2003. Originally published online 1st June 2004

The authors are with the Institute for Systems Research, University of Maryland, College Park, MD 20742, USA

profile of the hierarchical structure of typical traffic elements. It characterises the structure from three levels: the connection level (level I), the intraconnection level (level II) and the packet level (level III). Level I describes TCP sessions. Level II describes the bursts in a connection. Level III describes the packets within bursts. In [13] it is shown that the existence of level II is an inherent property of TCP protocol. It essentially embodies the effect of the TCP window-based congestion control [15] on the traffic. In more detail, TCP makes packets be sent in bulk from the source to the destination. In a practical network setting, the change of TCP window size in response to network status makes the bulk length random. So on an intermediate link, a bulk appears to be a burst of random length.

The model is illustrated in Fig. 1. The three levels of the traffic structure are shown. Level I depicts the arrivals and durations of different TCP connections ('on' periods). We say the traffic element in this level is a connection. Details within a traffic element are given in lower levels, i.e. level II and level III. Assume a connection has a duration of τ, and the average connection length is $\mu = E[\tau]$. Level II depicts the internal structure of a connection: a series of bursts. The traffic element is thus a burst called a 'packet cluster' or a bulk. We see that between clusters are silence periods in which no packet is available. Use random variables ω_{on} and ω_{off} to represent the cluster and the silence periods. The average occupancy within a connection observed at this level is $\gamma_1 = E[\omega_{on}]/E(\omega_{on}] + E[\omega_{off}])$. Level III pictures the internal structure of a packet cluster. The traffic element at this level is a packet. We represent the duration of a packet and the interpacket interval with random variables π_{on} and π_{off}, respectively. The average occupancy within a cluster is thus $\gamma_2 = E[\pi_{on}]/(E[\pi_{on}] + E[\pi_{off}])$. Let us assume the bit rate within an 'on' period at this level is r_{on}. The whole model is then completely set.

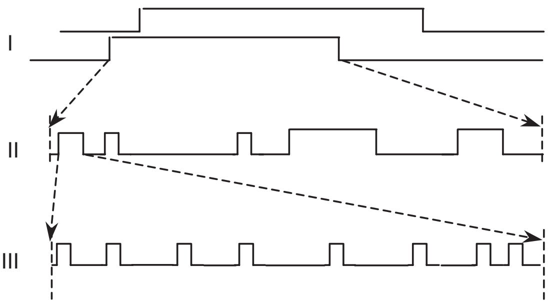

Fig. 1 *Hierarchical on–off model for TCP traffic*

This hierarchical model has some basic features that make it intuitively suitable for multi-scale traffic. Obviously the traffic elements at the three levels are in different time scales. Consequently, traffic characteristics can be expected to change with time scales.

Hierarchical descriptions of traffic structure have been seen on different occasions [16, 17], but there they are not used as a model for study of multi-scale phenomena or for performance evaluation. In fact, to our knowledge, no queueing analysis is available for this hierarchical traffic model.

3 Queueing analysis

3.1 Methodology

Mathematically, a HOO process is a very complex compound renewal process. It is extremely difficult to analyse the queue content distribution for it. The primary challenge we face is to make the problem tractable. To do this, we carefully decide the following tactics. First, the problem is separated into two cases for analysis, i.e. the case of *fast flows*, in which the average rate of a flow is higher than the queue service rate, and the case of *slow flows*, in which the average rate of a flow is lower than the queue service rate. This allows the flexibility to treat them in different ways, which turns out to be very important for this problem. Secondly, instead of analysing the queue distribution in the whole line, we study the tail behaviour over selected sample points. Sample points are chosen to be the start times of activity periods, which we will explain later. The points have a 'memoryless' feature in the sense that the intervals between the points are of an exponential distribution. Hopefully the result would give a good approximation of the tail in the whole line. Thirdly, equivalent processes are established for the HOO process or its queue content process so that we can take advantages of the 'easy' nature of the former to proceed. These techniques prove to be very helpful, and with them we are able to track the problem quite smoothly and obtain a simple form of solution that well satisfies our goal.

To facilitate the analysis, we simplify the HOO model in this way; we extend level II and level III to the whole line and view the HOO process as the product of the three levels. The resulting HOO process then has an appearance of an on–off process, but the transmission rates in different 'on' periods may be different. It is easy to see the overall occupancy of the process is $\gamma = \gamma_1\gamma_2$. We assume connections arrive in a Poisson process with average rate λ, and the connection length τ has a Pareto distribution with parameters b and α.

3.2 Queueing analysis for fast flows

Probability distributions frequently used in this paper are defined as follows:

Definition 1: A random variable X is Pareto-distributed with parameters b and α if the distribution function $F(x)$ satisfies

$$F(x) = 1 - \left(\frac{x}{b}\right)^{-\alpha}, \; x \to \infty \qquad (1)$$

where $\alpha > 0$, $b > 0$ and $x \geq b$. A Pareto distribution is denoted as $F \in P$ and a Pareto random variable is represented by (b, α).

Definition 2: A random variable X has a heavy-tailed distribution (or X is regularly varying), denoted as $F \in V$, if

$$F(x) = 1 - x^{-\alpha}L(x), \; x \to \infty \qquad (2)$$

where $\alpha > 0$ and $L(x)$ is a slowly varying function satisfying

$$\lim_{x \to \infty} \frac{L(kx)}{L(x)} = 1, \; \forall k > 0$$

Obviously, $P \subset V$.

We will establish an equivalent process for the HOO process through the $M/G/\infty$ process [14] in this case. An $M/G/\infty$ process is defined as follows:

Definition 3: An $M/G/\infty$ process is the busy server process of an $M/G/\infty$ queue.

In an $M/G/\infty$ process there are silence periods during which no servers are active. Between silence periods are activity periods during which the system is in service. We denote the start time of the ith activity period, which is exponentially distributed as T_i, $-\infty < i < \infty$.

It is not difficult to see that a HOO process and an $M/G/\infty$ process have the following basic relation:

Proposition 1: Level I of a HOO process is an $M/G/\infty$ process.

Proof: It is straightforward by comparing the level I of the HOO process with definition 3. □

It should be pointed out that in previous literature about scaling traffic modelling the $M/G/\infty$ model is primarily used as the limiting process of the aggregation of on–off flows [6, 8]. Proposition 1 indicates that it applies in a more natural sense as a profile of the connection level structure.

Assume all connections have the same average rate $\bar{r}$. Then $\bar{r} = r_{on}\gamma = r_{on}\gamma_1\gamma_2$. Based on this relation, we can define an $M/G/\infty$ companion process for a HOO process.

Definition 4: An $M/G/\infty$ *companion process* of a HOO process is a copy of the level I process of the latter except that it has a constant transmission rate $r' = \bar{r}$ for each connection.

Denote a HOO process as ξ and its $M/G/\infty$ companion process as ξ'. The following relations follow from definition 4:

(i) ξ and ξ' have the same average traffic rate for one connection.

(ii) Let vectors S and E represent sequences of start times and durations of all connections in ξ, respectively, and S' and E' for those in ξ'. Then $S = S'$ and $E = E'$ hold.

(iii) Let D_i be the arriving workload of ξ during $[T_i, T_{i+1})$ and D_i' that of ξ'. Then $D_i = D_i'$ holds.

Therefore, we can use T_i of ξ' to mark corresponding points of ξ. T_i will be referred to as the start of the ith activity period of either process hereafter.

Now we start to derive the queueing performance of a single server FIFO queue fed by a HOO process. We will first give two lemmas.

Lemma 1: For an $M/G/\infty$ process, if $r \geq c$, the following relation holds:

$$Q_{i+1} = (Q_i + D_i - C_i)^+ \tag{3}$$

where the operator $(q)^+$ means the maximum of q and 0, Q_i is the queue content at time T_i, D_i is the arriving workload during $[T_i, T_{i+1})$ and C_i is the leaving workload during $[T_i, T_{i+1})$.

Proof: See the Appendix (Section 8.1). □

We call (3) the generalised Lindley's equation.

Equivalently, (3) can be written as

$$Q_{i+1} = (Q_i + D_i - \Pi_i - \Lambda_i)^+ \tag{4}$$

where Π_i is the leaving workload in the ith activity period and Λ_i is the leaving workload in the ith silence period (the silence period immediately after the ith activity period).

In an on–off process, an 'on' period and its successive 'off' period form an on–off cycle. For level II, denote the 'on' and the 'off' period in the ith cycle as $\omega_{on}^{(i)}$ and $\omega_{off}^{(i)}$, $-\infty < i < \infty$. For level III, denote them as $\pi_{on}^{(i)}$ and $\pi_{off}^{(i)}$. So the occupancies of the ith cycle at level II and level III are $\gamma_1^{(i)} = \omega_{on}^{(i)}/(\omega_{on}^{(i)} + \omega_{off}^{(i)})$ and $\gamma_2^{(i)} = \pi_{on}^{(i)}/(\pi_{on}^{(i)} + \pi_{off}^{(i)})$, respectively. Let $\gamma_{1,min} = min\{\gamma_1^{(i)} \; -\infty < i < \infty\}|$ and $\gamma_{2,min} = min\{\gamma_2^{(i)} \; -\infty < i < \infty\}$. We can establish the following relation between a HOO process and its companion $M/G/\infty$ process with regard to their queueing content processes.

Lemma 2: If $r_{on}\gamma_{1,min}\gamma_{2,min} \geq c$, the queue contents for a HOO process at the start and the end of an activity period are equal to those for its $M/G/\infty$ companion process.

Proof: See the Appendix (Section 8.2). □

The condition $r_{on}\gamma_{1,\,min}\gamma_{2,\,min} \geq c$, denoted as Ω, seems to be very strict. However, to loosen the condition will greatly complicate the analysis, which does not make much sense for achieving our main goal. Instead of examining a more general condition, we simply make the following probability statement: there exists a more general condition $\wp \supset \Omega$ under which lemma 2 holds with high probability (→ 1). Proof is available in [18]. For simplicity we proceed with condition Ω. However, the following theorem holds with high probability under condition $\wp$.

Theorem 1: For a HOO process, if $\lambda\mu < 1$, $0 < \gamma_1 < 1$, $0 < \gamma_2 < 1$, $r_{on}\gamma_{1,min}\gamma_{2,min} \geq c$, and the distribution of τ is $F \in P$ with parameters $b > 0$ and $\alpha > 1$, then its queue content process at T_i, Q_s, has the following tail:

$$P[Q_s > x] \sim \frac{\lambda}{\alpha - 1} \frac{b^{\alpha}(\lambda\mu r_{on}\gamma_1\gamma_2 + r_{on}\gamma_1\gamma_2 - c)^{\alpha}}{c - \lambda\mu r_{on}\gamma_1\gamma_2} x^{-(\alpha-1)}, \quad x \to \infty \tag{5}$$

Proof: From lemma 2 and definition 4, the queue tail of a HOO process is the same as that of its $M/G/\infty$ companion process with $r = r_{on}\gamma_1\gamma_2$. For a queue fed by an $M/G/\infty$ process, if $\lambda\mu < 1$, $r > c$, and the distribution of τ is $F \in V$, the following relation holds [6]:

$$\lim_{x\to\infty} \frac{P[Q_s > x]}{\int_{x/(\lambda\mu r + r - c)}^{\infty} (1 - F(u))du} = \lambda\left(\frac{r}{c - \lambda\mu r} - 1\right)$$

With the Pareto distribution function $F(x)$ given in (1), we obtain (5). □

3.3 Queueing analysis for slow flows

As far as we know, the queue tail for an $M/G/\infty$ process with heavy-tailed service time in the case of slow flows is an unsolved problem. So we cannot use the approach in Section 3.2. Instead, we will approximate the queueing process by means of the $M/G/1$ queue. The approach has been used in [19, 20].

We first define two processes. One is related to the queue for an $M/G/\infty$ process

$$V_{i+1} = (V_i + D_i - \Pi_i - \Lambda_i)^+ \tag{6}$$

The other is for an $M/G/1$ queue

$$W_{i+1} = (W_i + r\tau_i - \Lambda_i)^+ \tag{7}$$

Process V looks similar to the queue content process at T_i. In the case of fast flows, they are equivalent. But in the case of slow flows, they are not, because the queue may be empty at some instants in an activity period and thus lemma 2 does not hold. Process W is exactly the queue content process of an $M/G/1$ queue at customer arrival points. The workload brought by customer i is $r\tau_i$, the interarrival time between customer i and $i+1$ is Λ_i and the queue service rate is 1. It can also be understood as the waiting time process of the $M/G/1$ queue. This understanding will be used in theorem 2.

Next, we make the following assumption:

Heavy traffic assumption: If $\lambda\mu r \to c$, $P[Q = 0] \to 0$.

This is reasonable and has been shown to be true for extensive cases [19]. It enables a probabilistic equivalent relationship between V and W, and is a key step towards the final result for this case.

Lemma 3: If $r < c$ and $\lambda\mu r \to c$, process V resembles the actual queue content at T_i with probability 1, and it has the same tail as process W.

Proof: Assuming heavy traffic, for any time unit within an activity period, $P[q_j + d_j - c \geq 0] \to 1$. So $P[Q_{i+1} = (Q_i + D_i - C_i)^+] \to 1$. This means the queue content process Q_s is equivalent to process V with probability 1. That V and W have the same tail has been shown in [20]. This completes the proof. □

Now we are ready to present the queue tail for heavy traffic of slow flows.

Theorem 2: For a queue fed by a HOO process, if $r_{on}\gamma_1\gamma_2 < c$, $\lambda\mu r_{on}\gamma_1\gamma_2 \to c$, and the distribution of τ is $F \in P$ with parameters $b > 0$ and $\alpha > 1$, then the queue content process at T_i has the following tail with probability 1:

$$P[Q_s > x] \sim \frac{\lambda}{\alpha - 1} \frac{b^{\alpha}(r_{on}\gamma_1\gamma_2)^{\alpha}}{c - \lambda\mu r_{on}\gamma_1\gamma_2} x^{-(\alpha-1)}, \; x \to \infty \qquad (8)$$

Proof: For both the HOO process and its $M/G/\infty$ companion process, assuming heavy traffic, at any time unit within an activity period $P[q_j + d_j - c \geq 0] \to 1$. So $P[Q_{i+1} = (Q_i + D_i - C_i)^{+}] \to 1$. Thus the queue content for the HOO process at time T_{i+1} is equivalent to process V with probability 1. With lemma 3, process V has the same tail as W. As defined above, W is the waiting process of an $M/G/1$ queue with arrival rate λ and service time $s = r\tau$. It is easy to get $F(s) = 1 - r^{\alpha}(s/b)^{-\alpha}$ and $E[s] = rE[\tau] = r\mu$. It is known that the following relation holds for W if the distribution of s is $F \in V$ with parameter $\alpha > 1$ and $\rho = \lambda \cdot E[s] < 1$ [20]:

$$P[s > x] \sim (\alpha - 1)\left(\frac{x}{E[s]}\right)^{-\alpha} L(x), x \to \infty$$

$$\Leftrightarrow P[W > x] \sim \frac{\rho}{1 - \rho}\left(\frac{x}{E[s]}\right)^{1-\alpha} L(x), \; x \to \infty$$

Then we can write down

$$P[W > x] \sim \frac{\rho}{(1 - \rho)E[s]} \frac{b^{\alpha} r^{\alpha}}{\alpha - 1} x^{-(\alpha-1)}, \; x \to \infty$$

Given parameters for the HOO process, i.e. $r = r_{on}\gamma_1\gamma_2$ and $\rho = \lambda\mu r_{on}\gamma_1\gamma_2/c$, we obtain (8). □

3.4 Summary

The combination of the results for the case of fast flows and the case of slow flows gives the full solution. We notice that (5) is exactly the same as (8) when $\lambda\mu r \to c$. Thus we immediately get the following corollary.

Corollary 1: Under heavy traffic, i.e. when $\lambda\mu r \to c$, both fast flows and slow flows have the same Q_s.

In more general conditions, the two cases also share important commonalities; the queue tails both follow a power law and their indices are the same. So the results can be generally presented in a simple form

$$G(x) = P[Q_s > x] \sim \beta x^{-(\alpha-1)}, \; x \to \infty \qquad (9)$$

where factor β represents the coefficient part on the right-hand side of either (5) or (8). The similarity of this result with those from the on–off model and the $M/G/\infty$ model [6] suggests the persistence of the power-law queueing behaviour in various traffic environments. Compared with the latter two, this result displays the roles of different traffic elements. First, the connection length determines the index of the queue tail while the traffic elements at level II and level III have no effects on it in the conditions when either the overall traffic load or the individual flow rate is high. Secondly, the traffic elements at level II and level III contribute the queue tail by their averages. This can be seen from the fact that β is a function of

$$\gamma_1\gamma_2 = \frac{E[\omega_{on}]E[\pi_{on}]}{(E[\omega_{on}] + E[\omega_{off}])(E[\pi_{on}] + E[\pi_{off}])}$$

4 Result evaluation and performance prediction

In this Section, we will compare the analytical tail with simulation results. Furnished with simple statistical inferences, the analytical result can be demonstrated to predicate the queueing performance quite well.

4.1 Comparisons of analytical tails with simulation results

Figure 2 compares the analytical results (5) and (8) with simulation results for different traffic loads. The connection length in level I of the HOO process is set as $\tau = (20, 1.5)$. Level II parameters are $\omega_{on} = (1, 2.0)$ and $\omega_{off} = (4, 2.0)$, which results in $\gamma_1 = 0.2$. Level III parameters are chosen as $\pi_{on} = 1$ and $\pi_{off} = 0$, meaning the packet size is fixed and the interpacket intervals are too tiny to be seen. This maintains the simulated data series in a reasonable length. Then $\gamma_2 = 1$. The traffic load is indicated by the queue system utility$\rho = \lambda\mu r_{on}\gamma_1\gamma_2/c$. Let $c = 1$. We change λ to get different values of ρ. Figures 2*a* and 2*b* are for fast flows where $\bar{r} = 1$ ($r_{on} = 5$), and Fig. 2*c* is for slow flows, where $\bar{r} = 0.2$ ($r_{on} = 1$). λ for the Figures are 1/300, 1/200 and 1/15, respectively, and corresponding traffic loads are 20%, 30% and 80%.

We observe that the analytical tail does not always match the simulation data well, especially when the load is high. A key fact, nevertheless, is the identity of tail indices between the analysis and the simulation in all situations. This is clearly indicated by the parallelism between the tail parts of the curves for the simulated and the analytical results. It suggests that the index of the tail in the analysis is accurate while the coefficient part is not. This is not unusual for asymptotic results. In consequence, it is not sensible to estimate every parameter in (5) or (8) to do statistical inferences. Instead, the coefficient may be estimated as a whole. This leads to a simple and efficient statistical inference method.

4.2 Statistical interference of the queue tail

The simple method is to estimate the tail based on (9) instead of (5) and (8). In simulations, α is a given parameter. In real networks, it can be estimated either from connection length data or directly from queue size data. Several good estimators are available [21, 22]. β can be estimated with empirical queue tail probabilities. Suppose $\hat{G}(x_i)$, $i = 1, 2, .., m$, are a set of samples of empirical tail probabilities for different x_i. A simple estimation of β is then as follows

$$\hat{\beta} = \frac{1}{m}\sum_{i=1}^{m}[\hat{G}(x_i)/x_i^{-(\hat{\alpha}-1)}] \qquad (10)$$

where $\hat{\alpha}$ is the estimation of α. To make sure the samples are in the tail part, x_i should be chosen to be considerably big.

Used with the above inference method, (9) can serve as an estimation-based performance predictor. Figure 3 compares the inferred queue tails with real ones from simulations. We can see they match very well in all cases. The gaps between the analytical tail and the real data are greatly reduced. This indicates that (9), as a long-run performance predictor, is fairly good.

4.3 Matching queueing performance of real traces

We apply the estimation-based predictor to real traffic traces and compare the predictions with real queueing performance. Two traces are used, named LBL-TCP and SAT-TCP. LBL-TCP is a frequently used trace from LBL [23]. It includes 1.8 million packets, and was measured

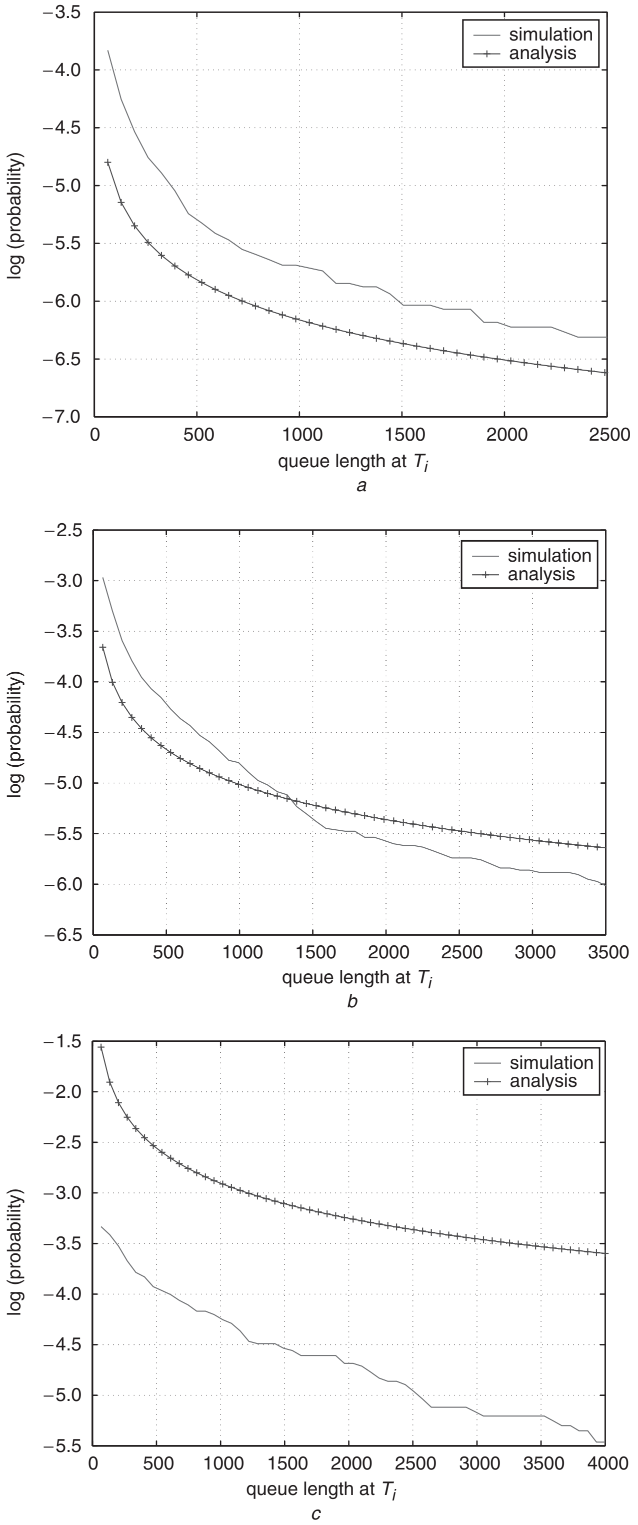

Fig. 2 *Comparisons of analytical and simulated queue tails in different traffic loads*
a Fast flows, $\rho = 20\%$, $\bar{r} = 1$
b Fast flows, $\rho = 30\%$, $\bar{r} = 1$
c Slow flows, $\rho = 80\%$, $\bar{r} = 0.2$

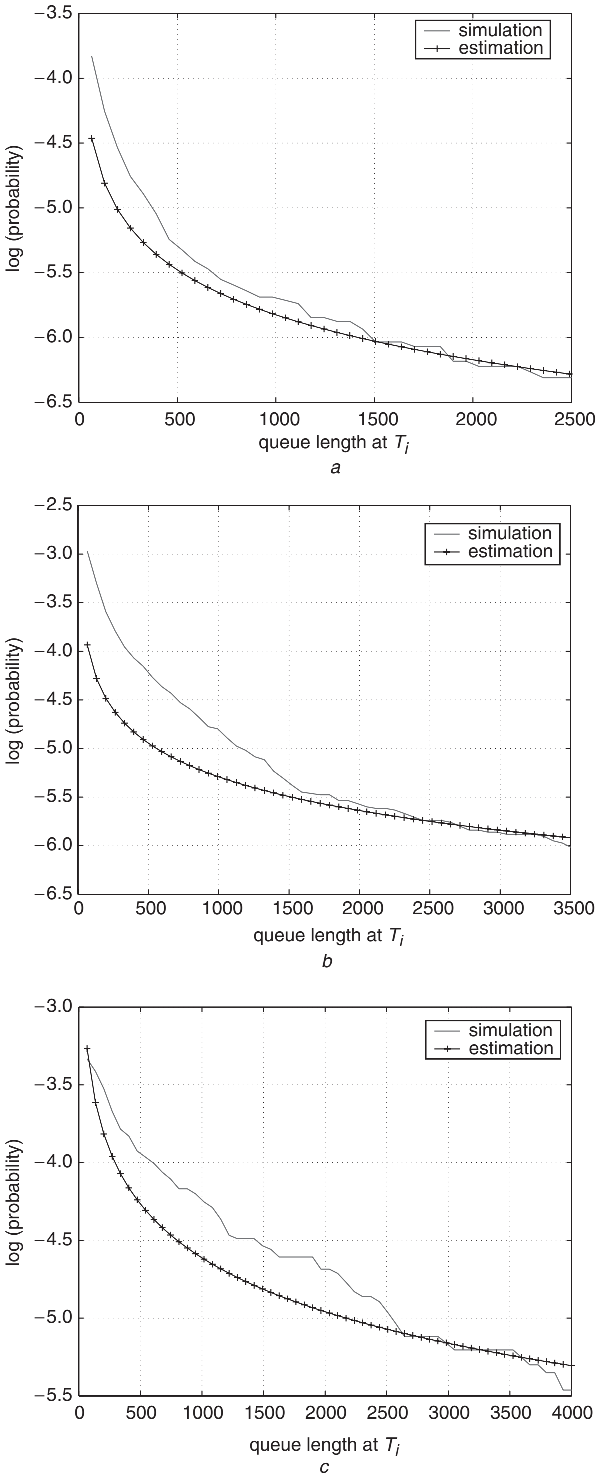

Fig. 3 *Comparisons of inferred and simulated queue tails in different traffic loads*
Simulation settings are the same as those in Fig. 2
a $\rho = 20\%$, $\bar{r} = 1$
b $\rho = 30\%$, $\bar{r} = 1$
c $\rho = 80\%$, $\bar{r} = 0.2$

from 14:10 to 16:10 on 20 January 1994 on an Ethernet of LBL. Its average rate is 282.12 kbit/s. We choose the queue service rate as 500 kbit/s, which means $\rho = 56.42\%$. SAT-TCP is collected from an Internet–satellite network gateway of NASA. It includes 4.8 million HTTP packets, and was measured in a period of 7396 s between 17:00 and 18:00 on 12 May 1999. The average rate of it is 4.39 Mbit/s. The queue service rate is set as 6.4 Mbit/s and $\rho = 68.59\%$.

Figure 4 shows the results. The unit for the queue length is a byte. We see the predicted tails match the real ones quite well. The queue length data used for the real traces are from the whole line (every time unit), not just at time T_i. So the predictor, though based on analytical results at T_i, provides a good approximation to the overall tail distribution. Considering the real traces include both fast and slow flows, and the transmission rates of connections are very diverse, the results also suggest that the predictor is quite robust.

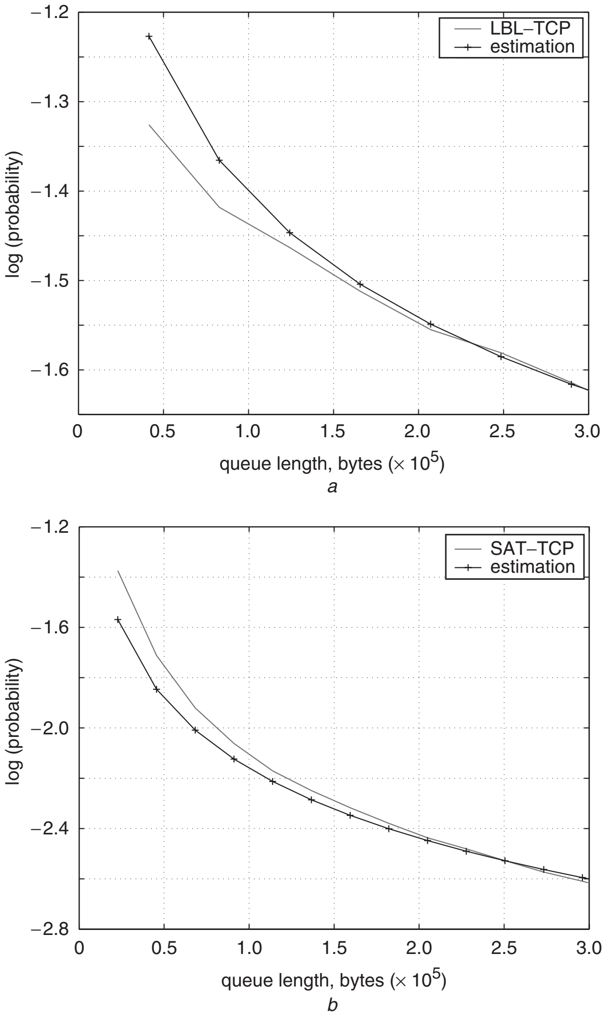

Fig. 4 *Comparisons of inferred and actual queue tails for real Internet traffic traces*
a LBL-TCP
b SAT-TCP

5 Conclusions

In this paper, we evaluate the queueing performance of a complex multi-scale traffic model, the HOO model. To track the problem we employ several techniques to avoid difficulty. In particular, an equivalent process to the traffic model is established in the case of fast flows and an equivalent process to the queue content is established in the case of slow flows. Quantitative solutions for both cases are obtained and a unified form is derived. The result indicates that the connection level of the traffic determines the index of the queue tail, while the burst and the packet levels affect the asymptotic performance with their averages. The analytical formula is further tested against simulations under different traffic loads. Using simple statistical inferences, a performance predictor based on the analysis is demonstrated to match the queueing performance of real traffic quite well.

6 Acknowledgments

This work was supported in part by DARPA under contract No. N66001-00-C-8063.

7 References

1 Leland, W.E., Taqqu, M.S., Willinger, W., and Wilson, D.V.: 'On the self-similar nature of Ethernet traffic', *IEEE/ACM Trans. Netw.*, 1994, **2**, pp. 1–15

2 Riedi, R.H., Levy Vehel, J.: 'TCP traffic is multifractal: a numerical study'. Preprint, 1997

3 Erramilli, A., Narayan, O., and Willinger, W.: 'Experimental studies on the performance impacts of long range dependence', *IEEE/ACM Trans. Netw.*, 1996, **4**, p. 209

4 Park, K., and Willinger, W.: 'Self-similar network traffic: an overview', in Park, K. and Willinger, W. (Eds.): 'Self-similar network traffic and performance evaluation' (John Wiley & Sons, 2000)

5 Ribeiro, V.J., Riedi, R.H., Crouse, M.S., and Baraniuk, R.G.: 'Multiscale queuing analysis of long-range-dependent network traffic'. Proc. INFOCOM, Tel. Aviv, Israel, March 2000, pp. 1026–1035

6 Jelenkovic, P.R., and Lazar, A.A.: 'Asymptotic results for multiplexing subexponential on-off processes', *Adv. Appl. Probab.*, 1999, **31**, pp. 394–421

7 Taqqu, M.S., Willinger, W., and Sherman, R.: 'Proof of a fundamental result in self-similar traffic modeling', *Comput. Commun. Rev.*, 1997, **27**, pp. 5–23

8 Krunz, M., and Makowski, A.: 'Modeling video traffic using M/G/∞ input processes: a compromise between Markovian and LRD models', *IEEE J. Sel. Areas Commun.*, 1998, **16**, (5), pp. 733–748

9 Feldmann, A., Gilbert, A.C., and Willinger, W.: 'Data networks as cascades: investigating the multifractal nature of Internet WAN traffic', *Comput. Commun. Rev.*, 1998, **28**, pp. 42–55

10 Riedi, R.H., Crouse, M.S., Ribeiro, V.J., and Baraniuk, R.G.: 'Multifractal wavelet model with application to TCP network traffic', *IEEE Trans. Inf. Theory*, 1999, **45**, (3)

11 Levy, J.B., and Taqqu, M.S.: 'Renewal reward processes with heavy-tailed interrenewal times and heavy-tailed rewards', *Bernoulli*, 2000, **6**, (1), pp. 23–44

12 Anick, D., Mitra, D.S., and Sondhi, M.M.: 'Stochastic theory of a data-handling system with multiple sources', *Bell Syst. Tech. J.*, 1982, **61**, (8), pp. 1871–1893

13 Liu, N., and Baras, J.: 'Understanding multi-scaling network traffic: a structural TCP traffic model' Submitted to *IEEE/ACM Trans. Netw.*

14 Figueiredo, D.R., Liu, B., Misra, V., and Towsley, D.: 'On the autocorrelation structure of TCP traffic', *Comput. Netw.*, 2002, **40**, pp. 339–361

15 Stevens, W.R.: 'TCP/IP illustrated, Volume 1: The protocols' (Addison–Wesley, 1994)

16 Misra, V., and Gong, W.B.: 'A hierarchical model for teletraffic'. Proc. 37th IEEE Conf. on Decision and Control, 1998, Vol. 2

17 Willinger, W., Taqqu, M., Sherman, R., and Wilson, D.: 'Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level', *IEEE/ACM Trans. Netw.*, 1997, **5**, (1), pp. 71–86

18 Liu, N., and Baras, J.: 'Long-run performance analysis of a multi-scale TCP traffic model'. Technical Report, Institute for Systems Research, University of Maryland, USA, Aug. 2002

19 Boxma, O.J., and Cohen, J.W.: 'The single server queue: heavy tails and heavy traffic', in Park, K. and Willinger, W. (Eds.) 'Self-similar network traffic and performance evaluation' (John Wiley & Sons, 2000)

20 Cohen, J.W.: 'The M/G/1 fluid model with heavy-tailed message length distributions'. Technical Report No. PNA-R9714;CWI, 1997

21 Taqqu, M.S., Teverovsky, V., and Willinger, W.: 'Estimators for long-range dependence: an empirical study', *Fractals*, 1995, **3**, (4), pp. 785–788

22 Veitch, D., and Abry, P.: 'A wavelet based joint estimator for the parameters of LRD', *IEEE Trans. Inf. Theory*, 1999, **45**, (3)

23 http://ita.ee.lbl.gov/html/traces.html

8 Appendix

8.1 Proof of lemma 1

Proof: Assume $[T_i, T_{i-1})$ includes N_i time units, i.e. $N_i = T_{i+1} - T_i$. First M_i $(< N_i)$ units form an activity period while the rest, $N_i - M_i$ units have no traffic. Let q_j be the queue content at the start of the *j*th time unit in this interval, d_j is the arriving workload in the time unit and c is the service rate. So

$$D_i = \sum_{j=1}^{M_i} d_j$$

Let $q_1 = Q_i$. Because $d_j \geq r \geq c$ for $j = 1, 2, \ldots M_i$, with Lindley's equation we can write down

$$q_{M_i+1} = q_1 + \sum_{j=1}^{M_i} (d_j - c) = Q_i + D_i - M_i c \qquad (11)$$

For $j = M_{i+1}, \ldots, N_i$, available queue content merely drains out. So

$$\begin{aligned} Q_{i+1} &= (q_{M_i+1} - (N_i - M_i)c)^+ \\ &= (Q_i + D_i - C_i)^+ \qquad (12) \end{aligned}$$

□

8.2 Proof of lemma 2

Proof: Recall that the *i*th activity period of a HOO process has the same workload D_i and duration M_i with its $M/G/\infty$ companion process. Inside an activity period are many 'on' and 'off' periods. These include on–off cycles within clusters and intercluster 'off' periods. A cluster always begins with an 'on' period. Because $r_{on}r_{2,min} > r_{on}\gamma_{1,min}\gamma_{2,min} \geq c$, we have

$$r_{on}\pi_{on}^{(j)} - c(\pi_{on}^{(j)} + \pi_{off}^{(j)}) \geq 0 \qquad (13)$$

for any on–off cycle *j* in a cluster *k*. Obviously, the average rate of cluster *k*, r_k, satisfies $r_k \geq r_{on}\gamma_{2,min}$. Thus $r_k\gamma_{1,min} \geq c$. Then

$$r_k\varpi_{on}^{(k)} - c(\varpi_{on}^{(k)} + \varpi_{off}^{(k)}) \geq 0 \qquad (14)$$

for any on–off cycle *k* at level II (including cluster *k* and the intercluster silence period immediately after it). Then, with Lindley's equation, it is straightforward to write down the same equation as (11) for the *i*th activity period. An equation the same as (12) for the interval $[T_i, T_{i+1})$ follows. This completes the proof. □