Teletraffic Models

Michael D. Logothetis
University of Patras, Greece.

Ioannis D. Moscholios
University of Peloponnese, Greece.

Keywords: *traffic-load, system capacity, quality-of-service, blocking probability, call, Markov*

Abstract: First, we present the key features of teletraffic models while emphasizing on the classification of their parameters to arrive at a possible categorization. Second, we briefly discuss the relationship between teletraffic models and the Internet, and third, we present three representative multirate teletraffic loss models.

1 Introduction

Teletraffic models[1] are mathematical formulas that combine three parameters of a communication system: the system *capacity*, the *traffic-load*[2][1] and the *quality of service (QoS)*; any two of them are inputs (usually the first two) and the rest is the output (usually the third). Teletraffic models are key tools in teletraffic engineering[3] for network planning or QoS assessment/guarantee.

Human analogy: For a sales store, its size (system capacity), expressed by the number of cashiers, the store and parking space, and so on, is a defining factor of the number of products (traffic-load) available for sale per day (QoS). Similarly, the offered traffic is a defining factor of the capacity of a communication system. The sales store example is even more important than for understanding purposes, since the same teletraffic model can be used to estimate the traffic offered either to a communication system or to the sales store. In many such examples, the size of an installation is directly related to the throughput of the installation. Thus, the applicability of teletraffic models can be extended to other systems (e.g., smart grids or banking).

Right or wrong: Since traffic-load comes from calls, the number of calls varies randomly as calls start and end randomly. By expressing the traffic-load with a single number, which is the average traffic-load, the resulting model could be right on "average" only. Teletraffic models are created based on basic assumptions, whereby we describe a communication system[1]. Since it is difficult to find one-to-one correspondence between the components of a communication system and a teletraffic model, in that sense a teletraffic model is wrong. On the other hand, if the assumptions are valid, then the resulting teletraffic model can be an accurate model or an approximate one (due to purely mathematical approximations, e.g., rounding).

Statistics Reference Online, © 2014–2022 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat08379

Usefulness: Teletraffic models have been an inseparable part of the telecommunications infrastructure and Information & Communication Technology since the beginning of their existence. Regardless of the changes that new networking technologies can bring from generation to generation, the essential task of teletraffic models remains the same: to determine and evaluate the relationship between (i) the QoS parameters (e.g., call blocking probability), (ii) the parameters that determine the intensity of connection requests and the demanded resources (traffic-load), and (iii) the parameters that describe the available network resources (capacity). They are favorably applied to connection-oriented networks or services, where capacity is a key parameter. Teletraffic models are particularly helpful in controlling the access of different services to network resources and the bandwidth distribution between service classes (Call Admission Control – CAC). This has been widely recognized as a necessary solution for the QoS guarantee in both the existing and the future networks. Call-level multi-rate teletraffic loss models aim at assessing the call-level QoS of networks with resource reservation capabilities, as well as of the emerging and future all-optical core networks. In short, teletraffic models are useful because they help us design a system or evaluate the basic performance metrics and predict its behavior even under strange conditions. The more complex the system, the more useful the model. Having created a teletraffic model, you can implement it into your computer as a small program (tool) and have a robust way to study a communication system in a short time, using small computer memory. Thus, one can make it much easier to study the system and come to safe conclusions.

Loss and queueing models: In telecommunications, call service systems (networks) are treated either as loss systems (where calls are cleared when they cannot be served immediately), or as queueing systems (where calls are queued when they cannot be served immediately – *see stat02990*). Two main parts are distinguished: call servers and incoming calls requesting service upon arrival. The number of servers reveals the bandwidth capacity of the system and corresponds to bandwidth units (b.u.). If a call starts service, it occupies the required bandwidth for as long as necessary. Although teletraffic models could include both loss and queueing systems, they refer mainly to loss systems[4]. This is due to the Erlang-B formula[5], the useful and famous mathematical formula of the past used to represent the term teletraffic model. The Erlang-B model applies to a system accommodating a single service-class only, which is a serious limitation in modern communication networks such as the Internet, where many services are simultaneously conveyed[6].

Multirate models: In the multidimensional traffic environment of contemporary communication networks, new teletraffic models consider multiple service-classes and thus are characterized as multirate models. The motivation for developing efficient multirate teletraffic models is manifold: The accuracy of network dimensioning and optimization depends to a large extent on the accuracy of the incorporated teletraffic model, which in turn depends on the precise modeling of the traffic categories (service-classes). Dimensioning is an endless, ongoing process of analyzing and designing network performance. To achieve this effectively, it is necessary to work out models that incorporate the network parameters in a reliable way.

Efficient models: Efficiency means an effective computer implementation of the teletraffic model achieved by a recursive formula (*see stat06426*). In the past, useful teletraffic models were available through tables or charts containing their values (e.g., Erlang B and Engset tables). The recursive feature however, along with the fact that computers are ubiquitous and used daily, makes such tables obsolete today.

1.2 Classification of the Parameters of Teletraffic Loss Models

The global network[7] of either 4G or 5G (or whatever in the future) consisting of many interacting heterogeneous systems supports widely used mobile devices and cloud computing that have given rise not only to a tremendous growth of network traffic but also to a high variety of traffic flows. The latter more

DOI: 10.1002/9781118445112.stat08379

than ever necessitates the development of specialized teletraffic models according to the input traffic. Fortunately, teletraffic models do not refer to specific technologies, but are abstracted from the technologies and receive a high degree of independence. Teletraffic models can be distinguished according to the:

(i) call arrival process (i.e., input traffic),
(ii) call bandwidth requirements upon arrival, (i.e., service-classes), and
(iii) call behavior under service (regarding the number of occupied b.u. per call over time).

The combination of the call characteristics of (i), (ii) and (iii) lead to different teletraffic models. However, not all combinations are realistic. In what follows we describe, in more detail, several call attributes, each of which leads to a different teletraffic model.

- ***According to the arrival process, calls are classified into:***

(i) Random calls – Random traffic (infinite number of traffic sources).
(ii) Quasi-random calls – Quasi-random traffic (finite number of traffic sources – *see stat02329*).
(iii) Batch Poisson arrivals (infinite number of traffic sources), with calls from different service-classes arriving in batches and batches arriving randomly following a Poisson process.

- ***According to the bandwidth requirements upon call arrival, calls are classified into:***

(i) Calls with fixed bandwidth requirements.
(ii) Calls with several alternative, contingency, and fixed bandwidth requirements, called elastic bandwidth requirements (Figure 1).

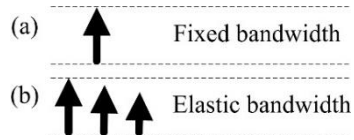

Figure 1 Visualization of (a) fixed and (b) elastic bandwidth requirements

- ***According to their behavior when calls are in service, calls are classified into:***

(i) Calls with fixed bandwidth allocation (stream traffic – *see stat08310*).
(ii) Calls tolerant to bandwidth compression or expansion (elastic traffic / bandwidth).
(iii) Calls that alternate between transmission periods of fixed bandwidth (ON) and no transmission periods (OFF) (ON-OFF traffic which is a simple representation of bursty traffic) (Figure 2).

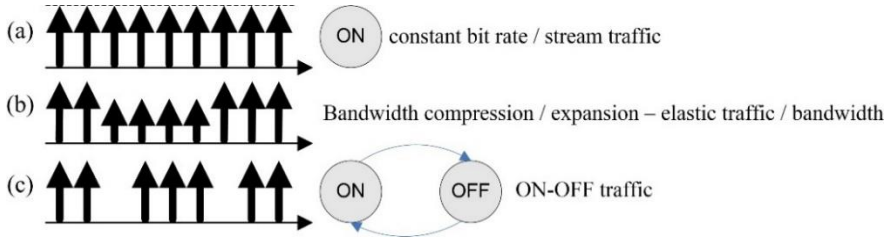

Figure 2 Visualization of (a) stream, (b) elastic, and (c) ON-OFF traffic

DOI: 10.1002/9781118445112.stat08379

1.3 Three Major Categories of Teletraffic Loss Models

Realistic combinations of the call arrival process, the bandwidth requirement upon call arrival and the in-service behavior of a call, lead to the following three major categories of teletraffic models.

- ***Teletraffic Models of Random Input***

  (I) With fixed or elastic bandwidth requirements and fixed bandwidth allocation during service.
  (II) With fixed or elastic bandwidth requirements and elastic bandwidth during service.
  (III) With fixed or elastic bandwidth requirements and ON-OFF traffic behavior during service.

- ***Teletraffic Models of Quasi-Random Input***

  (I) With fixed or elastic bandwidth requirements and fixed bandwidth allocation during service.
  (II) With fixed bandwidth requirements and elastic bandwidth during service.
  (III) With fixed bandwidth requirements and ON-OFF traffic behavior during service.

- ***Teletraffic Models of Batched Poisson Input***

  (I) With fixed bandwidth requirements and fixed bandwidth allocation during service.
  (II) With fixed bandwidth requirements and elastic bandwidth during service.

Before presenting some basic teletraffic models of the above major categories it is important to discuss the relation of teletraffic models with the Internet.

2 Teletraffic Models and the Internet

The need for a teletraffic model on the Internet arises from the necessity to guarantee QoS of the various network services. Over-provisioning (e.g., over-dimensioning of transmission link bandwidth) is an insufficient solution, because the network is not able to ensure low latency for packets (e.g., of real-time services) while maintaining sufficiently high throughput. According to forecasts[9], the Internet traffic-load will be so high that it is questionable whether we could provide ample bandwidth and over-dimensioning. Fortunately, the so-called best-effort Internet (i.e., without QoS guarantee) can be enhanced by two basic resource (bandwidth) allocation strategies that can ensure QoS[10]: the Integrated Services (IntServ) and the Differentiated Services (DiffServ). Which strategy is preferable depends on various conditions, such as the specific QoS requirement[11].

Although Internet traffic is very complicated to be modeled using traditional techniques (developed for telephone networks or computer systems), conventional teletraffic models are applicable to the Internet and provide handy tools for performance evaluation (*see stat00413*), as long as Internet traffic is considered at a flow level[6]. Three service-classes can be distinguished for flows: stream traffic (i.e., flow streams generated under the User Datagram Protocol (UDP)[12]), elastic traffic (i.e., elastic flows generated under the Transport Control Protocol (TCP)[13]) and ON-OFF traffic (i.e., flows of distinct active and passive periods – bursts of steaming traffic). Regarding the traffic in the Internet, although it varies a lot during the day, a busy period can be identified, where the traffic-load can be considered constant and crucial for QoS assessment[1].

DOI: 10.1002/9781118445112.stat08379

3 The Erlang Multirate Loss Model – EMLM

Let us start with a teletraffic model of random input and fixed bandwidth requirements both upon call arrival and under service, considering a transmission link of capacity C that accommodates calls of K different service-classes. In the context of teletraffic models, a service-class (also called traffic stream by the ITU-T – International Telecommunication Union –Telecommunication Standardization Sector) consists of calls that require the same number of b.u. Each call of service-class k (k = 1,…,K) arrives in the system following a Poisson process with mean rate λ_k and requires b_k b.u. for service. If the required bandwidth is available, then a call is accepted in the system and remains under service for an exponentially distributed service time, with mean μ_k^{-1} (*see stat00969*). Otherwise, the call is blocked and lost (QoS index: Call Blocking Probability – CBP), that is, a blocked call is not allowed to retry. After service completion, the b_k b.u. become available to new arriving calls. The service-classes equally share the system capacity. This sharing policy is called Complete Sharing (CS), and the model is called Erlang Multirate Loss Model (EMLM).

Teletraffic theory proves that when two services (one of high-speed calls and the other of low-speed calls) share the system bandwidth capacity equally, the high-speed service-class (i.e., the one with the highest b.u. per call) always receives the worst QoS (highest CBP). For a fair share, the following CAC, called Bandwidth Reservation (BR) policy, should apply: A new service-class k call is accepted in the system, if, after acceptance, the system has at least t_k b.u. available to service calls of other service-classes. We can achieve CBP equalization among service-classes, by choosing the BR parameters t_k so that $b_1 + t_1 = b_2 + t_2 = \ldots = b_k + t_k = \ldots = b_K$ (assuming that $b_K > \ldots > b_k > \ldots > b_2 > b_1$); that is, $t_K = 0$, since it is reasonable not to reserve bandwidth against the service-class which requires the maximum bandwidth per call.

In the EMLM/BR, the determination of CBP of service-class k, B_k, is given by the following formula:

$$B_k = \sum_{j=C-b_k-t_k+1}^{C} \frac{q(j)}{G}$$

where the ratio $q(j)/G$ is the *link occupancy distribution*, that is, the probability that j out of C b.u. are occupied. The system is described by a Markov Chain (*see stat00360*). The unnormalized values of *link occupancy distribution* are given by the $q(j)$'s, according to the following approximate but recurrent formula[1]:

$$q(j) = \begin{cases} 1 & \text{for } j = 0 \\ \dfrac{1}{j}\displaystyle\sum_{k=1}^{K} a_k D_k(j-b_k)\, q(j-b_k) & \text{for } j = 1,\ldots,C \\ 0 & \text{otherwise} \end{cases}$$

where $\alpha_k = \lambda_k \mu_k^{-1}$ is the offered traffic-load in erlang (*erl* – the unit of traffic load in honor of A. K. Erlang (Danish)),

$$\text{and } D_k(j-b_k) = \begin{cases} b_k & \text{for } j \le C - t_k \\ 0 & \text{for } j > C - t_k \end{cases}$$

The $q(j)$'s are becoming probabilities after the division by the normalization constant $G = \sum_{j=0}^{C} q(j)$.

By setting $t_k = 0$ for all service-classes k, the EMLM/BR becomes the EMLM (under the CS policy). In the EMLM, the calculation of $q(j)$'s is accurate (Kaufman–Roberts recursion)[14]. Figure 3 depicts a helpful visualization regarding the CBP calculation. In the case of only one service-class in the system, the EMLM provides the same CBP with the Erlang-B formula (*see stat00968*). This fact justifies the name EMLM.

DOI: 10.1002/9781118445112.stat08379

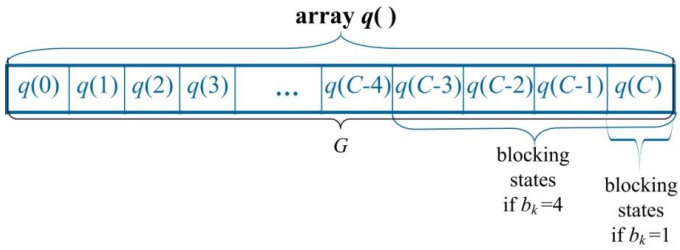

Figure 3 Visualization of CBP calculation

Example: For C = 5 b.u., K = 2 service-classes, $\alpha_1 = \alpha_2 = 1$ erl, b_1 = 1 b.u., b_2 = 2 b.u., t_1 = 1 b.u. and t_2 = 0 b.u., we obtain:

Table 1 Results of the example.

	$q(0)$	$q(1)$	$q(2)$	$q(3)$	$q(4)$	$q(5)$	G	B_1	B_2
EMLM/BR	1.0	1.0	1.5	1.1667	1.0417	0.4667	6.1751	24.43%	24.43%
EMLM (t_1=0)	1.0	1.0	1.5	1.1667	1.0417	0.6750	6.3834	10.57%	26.89%

4 The Connection-Dependent Threshold Model – CDTM

We continue with a teletraffic model of random input and elastic bandwidth requirements on call arrival but fixed bandwidth allocation during service. Consider a service system (Figure 4), where the requested b.u. (upon a call arrival) and the corresponding service time of a new call are related to the value of the occupied link bandwidth j. When j is lower or equal to a threshold J_{k_0}, a new arriving call of service-class k is accepted in the system with its initial requirements (b_k, μ_k^{-1}). If $j > J_{k_0}$, the call tries to be connected in the system with a reduced bandwidth and increased service time so that the product *bandwidth requirement* by *service time* remains constant. Thanks to thresholds (*see stat07783*), a call does not need to be blocked to retry with lower bandwidth requirements. Each service-class k may have $S(k)$ thresholds (different thresholds among service-classes). Thus, there may exist $S(k) + 1$ bandwidth and service-time requirements; the initial requirements and $S(k)$ more requirements with values $(b_{kc_s}, \mu_{kc_s}^{-1})$, where $s = 1,\ldots, S(k)$ and $b_{kc_{S(k)}} < \ldots < b_{kc_1} < b_k \equiv b_{kc_0}$ and $\mu_{kc_{S(k)}}^{-1} > \ldots > \mu_{kc_1}^{-1} > \mu_k^{-1} \equiv \mu_{kc_0}^{-1}$. The pair $(b_{kc_s}, \mu_{kc_s}^{-1})$ is used when $J_{k_{s-1}} < j \leq J_{k_s}$, where $J_{k_{s-1}}$ and J_{k_s} are two successive thresholds of service-class k, while $J_{k_{S(k)+1}} = C$, while the highest possible (other than C) threshold is $J_{k_{S(k)}} = C - b_{kc_{S(k)}}$. If the $b_{kc_{S(k)}}$ b.u. are not available, the call is blocked and lost.

DOI: 10.1002/9781118445112.stat08379

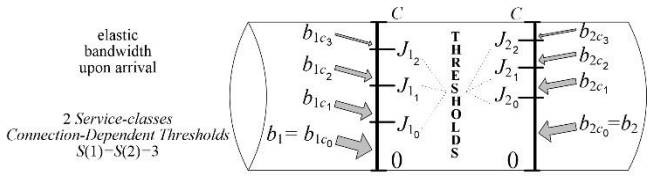

Figure 4 Principle of CDTM service system

The determination of CBP of service-class k, $B_{kc_{S(k)}}$, is given by the following approximate formula[1]:

$$B_{kc_{S(k)}} = \sum_{j=C-b_{kc_{S(k)}}+1}^{C} \frac{q(j)}{G}$$

$$q(j) = \begin{cases} 1 & \text{if} \quad j = 0 \\ \frac{1}{j}\left(\sum_{k=1}^{K} a_k b_k \delta_k(j) q(j-b_k) + \sum_{k=1}^{K}\sum_{s=1}^{S(k)} a_{kc_s} b_{kc_s} \delta_{kc_s}(j) q(j-b_{kc_s})\right) & j = 1,\ldots,C \\ 0 & \text{otherwise} \end{cases}$$

where $\delta_k(j) = \begin{cases} 1 \text{ (if } 1 \le j \le J_{k_0} + b_k \text{ and } b_{kc_s} > 0) \text{ or (if } 1 \le j \le C \text{ and } b_{kc_s} = 0) \\ 0 \quad \text{otherwise} \end{cases}$,

and $\delta_{kc_s}(j) = \begin{cases} 1 & \text{if } J_{k_s} + b_{kc_s} \ge j > J_{k_{s-1}} + b_{kc_s} \text{ and } b_{kc_s} > 0 \\ 0 & \text{otherwise} \end{cases}$ and $a_{kc_s} = \lambda_k \mu_{kc_s}^{-1}$.

Example: Let C = 580, K = 4, traffic characteristics (initially): $(\lambda_1, \mu_1^{-1}, b_1) = (20, 1, 1)$, $(\lambda_2, \mu_2^{-1}, b_2) = (12, 1, 6)$, $(\lambda_3, \mu_3^{-1}, b_3) = (28, 1, 12)$, $(\lambda_4, \mu_4^{-1}, b_4) = (6, 1, 20)$. Calls of service-classes 2, 3, and 4 can reduce their bandwidth requirements one, two and three times respectively: $b_{2c} = 4$, $\mu_{2c}^{-1} = 1.5$, $b_{3c_1} = 8$, $\mu_{3c_1}^{-1} = 1.5$, $b_{3c_2} = 4$, $\mu_{3c_2}^{-1} = 3$, $b_{4c_1} = 16$, $\mu_{4c_1}^{-1} = 1.25$, $b_{4c_2} = 12$, $\mu_{4c_2}^{-1} = 1.667$, $b_{4c_3} = 8$, $\mu_{4c_3}^{-1} = 2.5$. The following thresholds are used: $J_{2_0} = 540$, $J_{3_0} = 520$, $J_{3_1} = 524$, $J_{4_0} = 500$, $J_{4_1} = 504$, $J_{4_2} = 508$ (Table 2).

Table 2 Results of the example together with simulation results to reveal the accuracy of CDTM.

B_1	B_{2c}	B_{3c_2}	B_{4c_3}	B_1 (simul.)	B_{2c} (simul.)	B_{3c_2} (simul.)	B_{4c_3} (simul.)
0.95%	4.01%	4.01%	8.21%	(0.51 ± 0.05)%	(2.15 ± 0.13)%	(2.12 ± 0.12)%	(4.28 ± 0.24)%
				95% confidence interval (see stat00130 or stat00165)			

5 The Elastic Multi-Retry Model – E-MRM

We proceed to a teletraffic model of random input with elastic bandwidth requirements and elastic bandwidth allocation during service. Calls may retry several times upon arrival (requiring less bandwidth each time) to be accepted for service. If call admission is not possible with the last (minimum) bandwidth requirement, then bandwidth compression is attempted. Consider a link of capacity C b.u. that accommodates calls of K service-classes. The arriving calls follow a Poisson process with mean rate λ_k, and have a peak-bandwidth requirement of b_k b..u. and an exponentially distributed service time with mean μ_k^{-1}.

Statistics Reference Online, © 2014–2022 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat08379

A blocked call of service-class k can have more than one retry parameters $(b_{kr_s}, \mu_{kr_s}^{-1})$, where $s = 1,\ldots,S(k)$, and $b_{kr_{S(k)}} <\ldots< b_{kr_1} < b_k$ and $\mu_{kr_{S(k)}}^{-1} >\ldots> \mu_{kr_1}^{-1} > \mu_k^{-1}$. To introduce bandwidth compression, we permit the occupied link bandwidth j to virtually exceed C up to a limit of T. Let j be the occupied link bandwidth, $j = 0,1,\ldots,T$, when a new service-class k call arrives in the link. To simplify the description of CAC, assume that a service-class k call may retry two times to be connected in the system. The first time with $b_{kr_1} < b_k$ and the second time (if blocked with b_{kr_1}) with $b_{kr_2} < b_{kr_1}$. We consider the following cases for CAC:

- If $j + b_k \leq C$, no bandwidth compression takes place and the call is accepted in the link with b_k b.u.
- If $j + b_k > C$, then the call is blocked with b_k and retries immediately to be connected with $b_{kr_1} < b_k$. If $j + b_{kr_1} \leq C$, no bandwidth compression occurs and the retry call is accepted with b_{kr_1} and $\mu_{kr_1}^{-1} > \mu_k^{-1}$.
- If $j + b_{kr_1} > C$ the retry call is blocked with b_{kr_1} and immediately retries with $b_{kr_2} < b_{kr_1}$. If $j + b_{kr_2} \leq C$, no bandwidth compression occurs and the retry call is accepted with b_{kr_2} and $\mu_{kr_2}^{-1} > \mu_{kr_1}^{-1} > \mu_k^{-1}$. If $j + b_{kr_2} > T$ the retry call is blocked and lost, while if $T \geq j + b_{kr_2} > C$ the retry call is accepted by compressing its bandwidth requirement b_{kr_2} together with the bandwidth of all in-service calls of all service-classes. In that case, the compressed bandwidth of the retry call becomes $b'_{kr_2} = rb_{kr_2} = \frac{C}{j + b_{kr_2}} b_{kr_2}$ where r is the compression factor, common to all service-classes. Similarly, all in-service calls, which have been accepted in the link with b_k (or b_{kr_1} or b_{kr_2}), compress their bandwidth to $b'_k = rb_k$ (or $b'_{kr_1} = rb_{kr_1}$ or $b'_{kr_2} = rb_{kr_2}$) for $k = 1,\ldots,K$. After the compression of all calls the link state is $j = C$. The minimum value of the compression factor is $r_{\min} = C / T$.

When a service-class k call, with bandwidth b'_k (or b'_{kr}), departs from the system, the remaining in-service calls of each service-class i ($i = 1,\ldots,K$), expand their bandwidth in proportion to their initially assigned bandwidth b_i (or b_{ir}). After bandwidth compression/expansion, the elastic service-class calls increase/decrease their service time so that the product *service time* by *bandwidth* remains constant.

The determination of CBP of the service-class k, $B_{kr_{S(k)}}$, is given by the following approximate formula[1]:

$$B_{kc_{S(k)}} = \sum_{j=C-b_{kr_{S(k)}}+1}^{C} \frac{q(j)}{G}$$

$$q(j) = \begin{cases} 1 & \text{if } j = 0 \\ \dfrac{1}{min(j,C)} \left(\displaystyle\sum_{k=1}^{K} a_k b_k \gamma_k(j) q(j - b_k) + \sum_{k=1}^{K} \sum_{s=1}^{S(k)} a_{kr_s} b_{kr_s} \gamma_{kc_s}(j) q(j - b_{kr_s}) \right) & j = 1,\ldots,T \\ 0 & \text{otherwise} \end{cases}$$

where: $a_{kr_s} = \lambda_k \mu_{kr_s}^{-1}$, $\gamma_k(j) = \begin{cases} 1 & \text{if } 1 \leq j \leq C \text{ and } b_{kr} > 0 \\ 1 & \text{if } 1 \leq j \leq T \text{ and } b_{kr} = 0 \text{ and} \\ 0 & \text{otherwise} \end{cases}$

$$\gamma_{kr_s}(j) = \begin{cases} 1 & \text{if } C - b_{kr_{s-1}} + b_{kr_s} < j \leq C \text{ and } s \neq S(k) \\ 1 & \text{if } C - b_{kr_{s-1}} + b_{kr_s} < j \leq T \text{ and } s = S(k). \\ 0 & \text{otherwise} \end{cases}$$

Example: For $C = 3$, $T = 4$, $K = 2$, $b_1 = 1$, $b_2 = 3$, $b_{2r_1} = 2$, $b_{2r_2} = 1$, $\lambda_1 = \lambda_2 = \mu_1^{-1} = \mu_2^{-1} = 1$, $\mu_{2r_1}^{-1} = 2.0$, and $\mu_{2r_2}^{-1} = 4.0$, we get the results shown in Table 3. Since blocking occurs when both service-classes request the same b.u., it is expected that $B_1 = B_{2r_2}$.

Statistics Reference Online, © 2014–2022 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat08379

Table 3 Results of the E-MRM example.

$q(0)$	$q(1)$	$q(2)$	$q(3)$	$q(4)$	G	B_1	B_{2r_2}
1.0	1.0	0.5	3.1667	5.27778	10.94445	48.22% (exact 47.10%)	48.22% (exact 47.10%)

References

[1] Moscholios, I., and Logothetis, M. (2019) *Efficient Multirate Teletraffic Loss Models Beyond Erlang*, John Wiley & IEEE Press, Hoboken, NJ, USA.

[2] Akimaru, H., and Kawashima, K. (1999) *Teletraffic: Theory and Application*, Springer, London, UK.

[3] Iversen, V. (2010) *Teletraffic Engineering Handbook,* Dept. Photonics Engineering, Technical Univ. Denmark. Lyngby, Denmark, https://orbit.dtu.dk/en/publications/teletraffic-engineeringandnetwork-planning.

[4] Ross, K. (1995) *Multiservice Loss Models for Broadband Telecommunication Networks*, Springer, Berlin.

[5] Brockmeyer, E., Halstrom, H., and Jensen, A. (1948) The life and works of A.K. Erlang, *Transactions of the Danish Academy of Technical Science*, 2, Kopenhagen, Denmark.

[6] Roberts, J. (2001) Traffic Theory and the Internet. *IEEE Communications Magazine*, **39**, 94–99.

[7] Stasiak, M., Głąbowski, M., Wisniewski, A., and Zwierzykowski, P. (2011) *Modeling and Dimensioning of Mobile Networks: From GSM to LTE*, John Wiley, Hoboken, NJ, USA.

[8] Bonald, T., and Roberts, J. (2012) Internet and the Erlang Formula, *ACM SIGCOMM Computer Communication Review*, **42**, 23-30.

[9] Cisco®: The Zettabyte Era: Trends and Analysis (White Paper) https://webobjects.cdw.com/webobjects/media/pdf/Solutions/Networking/White-Paper-Cisco-The-Zettabyte-Era-Trends-and-Analysis.pdf (accessed 24 October 2021).

[10] Daniel Reid and Michael Katchabaw, (2004) Internet QoS: Past, Present, and Future, *Technical Report* Department of Computer Science, The University of Western Ontario https://www.csd.uwo.ca/~mkatchab/pubs/tr_internetqos.pdf (accessed 24 October 2021).

[11] Shioda, S. (2014) Fundamental Trade-offs between Resource Separation and Resource Share for Quality of Service Guarantees, *IET Networks*, **3**, 4-15.

[12] Postel, J. (1980) User Datagram Protocol, RFC 768, IETF, Fremont, CA, USA.

[13] Berger, A., and Kogan, Y. (2000) Dimensioning Bandwidth for Elastic Traffic in High-speed Data Networks, *IEEE/ACM Trans. on Networking*, **8**, 643–654.

[14] Kaufman, J. (1981) Blocking in a shared resource environment, *IEEE Trans. Commun.* **29** (10):1474–1481.

DOI: 10.1002/9781118445112.stat08379