

A Study on Dynamical Queuing Systems

Dr. S. Chandrasekaran¹, P. Velusamy²

¹KhadirMohideen College, Adirampattinam

²Research Scholar, Department of Mathematics, Khadir Mohideen College, Adirampattinam

Abstract: *In this article, modeling in queuing systems with heavy traffic customer flows is reviewed. Key areas include their limiting distributions, asymptotic behaviors, modeling issues and applications. Heavy traffic flows are features of queuing in modern communications, transportation and computer systems today. Initially, we reviewed the onset of asymptotic modeling for heavy traffic single server queuing systems and then proceeded to multi server models supporting diffusion approximations developed recently. Our survey shows that queues with heavy traffic customer flows have limiting distributions and extreme value maximum. In addition, the diffusion approximation can conveniently model performance characters such as the queue length or the waiting time distributions in these systems.*

Keywords: Queue, Limiting distributions, Modeling, Extreme value maximum

1. Introduction

There are times when queueing systems behave like fluid. A good scenario is when customers of a busy bus station experience rush hour. Therein, the scenery looks highly saturated and stable or completely unstable. Either way, the system dynamics resembles a continuous fluid flow rather than discrete. Medhi's analogy in of fluid flow of people coming out of a subway or an electric train during rush hour is similar to the example above. The wide sense approximate continuity in such traffic flows is created by the heaviness of queueing traffic into the system. Broadly speaking, a heavy traffic queueing system can be defined as a queueing system whose server occupation rate is barely less than unity and this phenomenon as Boxma et al. indicates is a feature in modern communications and computer systems today. Researches have shown that early investigation in this area was carried out by Kingman on a general queue called the G/G/1 and the result is referred to as central limit theorem for 1 queueing theory, see Medhi.

Our objective in this paper is to survey works on heavy traffic queueing systems generally in the light of both mathematical and statistical realities with emphasis on those queues supporting the diffusion approximation. This includes their distributions, analyses, modeling and application. It is anticipated that a survey of this kind will provide an excellent background in heavy traffic studies especially in packets and internet traffic prevalent in computers, communications and telecommunications systems. To achieve an optimum survey process as in this case, it is essential that one bears in mind the Poisson traffic controversy now prevalent in telecommunications and computer traffic modeling not because the controversy is relevant or not, but because there are diverse opinions worthy of sharing especially as regards the new traffic models such as the self-similar model. Not only that, recent studies have shown that the Poisson based models are equally relevant and could be used to describe the internet traffic central in this controversy. For instance, Lee and Kim of late have shown that at small time scales, inter-arrival times of computer protocols such as the Simple Mail Transfer Protocol (SMTP) sessions traffic for the internet are exponentially distributed and independent of each other

which makes it possible to model this kind of traffic session arrivals as a Poisson process. Before then, Karagiannis et al. have indicated that the observed long-range dependence in the internet traffic does not make the Poisson based models obsolete. That at a sub-second time scales, backbone traffic appears to be well described by Poisson packet arrivals and went further to provide evidence that the ongoing pattern of the internet evolution may potentially affect the future characteristics of its traffic. Similarly, Kos and Bester wrote that for a traffic model to be suitable, it should be able to represent traffic with few parameters devoid of complications and intricacies. This is the point the Poisson based models exceed all other models.

2. G/G/1 Model

Heavy traffic approximation started with the work of Kingman on a one-server model with a general arrival and service time distributions called the G/G/1. Kingman proved that for the G/G/1 queue under heavy traffic, the waiting time distribution could be approximated by an exponential distribution. The result is called the central limit theorem for heavy traffic queueing systems given below:

Theorem 2.1

Suppose the traffic intensity $\rho \leq 1$. Let $W(t)$ denotes the steady state waiting time distribution in a G/G/1 queue. Then $W(t)$ could be approximated by the exponential distribution

$$W(t) = 1 - \exp\left(\frac{-2(1-\rho)}{\lambda(\sigma_u^2 + \sigma_v^2)}\right)t - 1$$

where λ is the arrival rate, σ_u^2 and σ_v^2 are the variances of the inter-arrival and service time distributions of customers in the system.

Medhi noted that in 1964, Kingman made a conjecture for the seemingly more significant G/G/c queue under heavy traffic arising from developments and insights on the G/M/c model. He conjectured that the waiting time distribution in a G/G/c queue equally could be approximated similar to the G/G/1 model. Ten years later in 1974, Kollerstrom proved the conjecture affirming that the heavy traffic waiting time distribution is of the form

$$W(t) = 1 - \exp\left(\frac{-2\left(\frac{1}{\lambda}\right)(1-\rho)}{\left(\sigma_u^2 + \frac{\sigma_v^2}{c^2}\right)}\right) t \quad -2$$

Here, ρ is the occupation rate and c is the number of servers.

The two equations above represent remarkable achievements in heavy traffic analysis of modern systems with light tail or short range dependencies. Moreover, they signify outstanding developments and breakthroughs in heavy traffic approximation via classical analysis of Laplace transforms of relevant distributions as far as operations research is concerned. Apart from operational significance, the works leading to the referenced equations boasted similar works notably on the subject of convergence and behaviors of similar systems under different conditions and assumptions. Convergence here refers to convergence in distribution of sequence of stochastic queuing processes such as the arrival or service process, the waiting time or queue length process, etc, see Whitt. On convergence over the years in this area, a lot has been written on queues under heavy traffic. The bulk of these works applies the diffusion approximation technique which will be discussed later in this survey on heavy traffic queuing systems. For saturated stable systems for instance, Whitt noted that Borovkov investigated the asymptotic behavior of a single phase case with Poisson arrival and service time distributions working independently of the service process in heavy traffic. Weak convergence limits and asymptotics for heavy traffic queues are well presented also in Borovkov. Similarly, Abate and Whitt studied a multi-channel queuing system and approximated the asymptotic decay rates of the queue length and the customer service distribution in form of tail probabilities under heavy traffic. The result shows that both the queue length and the service time distributions depend on the first 3 moments of their distributions. The M/G/1 queue with priority classes is an essential model giving its numerous applications. Priorities normally aroused sequel to the realistic nature of services in systems. Abate and Whitt derived limit theorems for the case when the priorities are preemptive and non-preemptive with resumption tendency. They proved that in the low-priority case, the limiting waiting-time distribution is a geometric random sum of independent and identically distributed random variables similar to the M/G/1 first come first served (fcfs) waiting-time distribution. On the asymptotic behavior of tail probabilities for this model, Abate and Whitt added that there is routinely a region such that the tail probabilities have non-exponential asymptotics even if the service time distribution is exponential. In addition, the asymptotics formed tend to be determined by the non-exponential asymptotics for the high-priority busy-period distribution.

Essentially, heavy traffic approximations in queues under the classical procedures are difficult especially for multi-channel queues that functions as integrated systems. These queues formed the bulk found in present day computers, communications and telecommunication systems. As Whitt posited, it is not easy to work with triple or quadruple transforms and this makes it hard to obtain knitted results. Consequently, we observed two implications out of this limitation. First, attention was like shifted to obtaining other forms of approximations among others; the diffusion

approximation which describes a queuing process as a Brownian motion and appears suitable for describing heavy traffic systems, see Medhi. Secondly, asymptotics of similar models for instance the M/G/c and the M/M/c that seemed more realistic were attempted and derived. On the latter for instance, Boxma et al. derived the asymptotics for the heterogeneous server M/G/2 with an exponential and a general server of regular variation and a cumulative service time distribution denoted by $B(t)$. By regular variation, we mean a distribution whose complement can be approximated by a slowly varying process at infinity index. The M/G/2 model of Boxma et al. is simply the trivial prototype of a discrete channel system with two distinct service processes arising from two servers such that the regularly varying component keeps changing, making its complimentary distribution fatter as time grows large. Boxma et al. have shown that such a model under light traffic is asymptotically similar to the Kingman's solution distribution (waiting time distribution is semi-exponential distribution). However, under heavy traffic the regularly varying nature of the service time distribution of the general server will have a long tail effect on the complementary waiting time distribution of customers in the system. The Boxma et al. asymptotic result for the model in question under heavy traffic is summarized in the theorem below:

Theorem 2.2

Suppose $\lambda > \mu$ and $B(t)$ is the service distribution of customers served by the regularly varying general server at index $-\nu$ with mean β . If $L(t)$ is a slowly varying function on some neighborhood such that $1-B(t) \sim t^{-\nu} L(t)$, $t \rightarrow \infty$, $\nu \in (m, m+1)$, $\forall m \in \mathbb{N}$

Then the complementary waiting time distribution denoted by $1 - W(t)$ is given by;

$$1 - W(t) \sim \frac{1 - Q_0 - Q_1}{(v-1)\beta(1-\lambda\beta + \mu\beta)} \left(\frac{(\lambda - \mu)}{\lambda}\right)^{v-1} t^{1-\nu} L(t), t \rightarrow \infty$$

where Q_i is the probability that there are exactly i -customers in the system at a steady time η . From the above result, a sufficient condition for defining heavy tail phenomena in the M/G/c model generally under heavy traffic is evident. It is summarized in the lemma below.

Lemma 2.1

A sufficient condition for heavy tail phenomena in a heavy traffic M/G/2 model is that, either/both the arrival or service process of customers in the system is/are significantly regularly varying at a known index as time grows large.

Remark 2.1

It is trivial since regularly varying distributions are subclass of heavy tail or more precisely, sub-exponential distributions. The light proof (by simple argument) of the lemma follows.

Proof

Given that server-1 is exponential and server 2 is general with a regularly varying distribution $B(t)$. Then $1-B(t) \sim t^{-\nu} L(t)$, $t \rightarrow \infty$, where $L(t)$ is a slowly varying function such that

$$\lim_{t \rightarrow \infty} \frac{L(tx)}{L(t)} = 1, x \rightarrow \infty$$

If $L(t)$ is undominatedly non-decreasing, we have $\limsup_{t \rightarrow \infty} L(t) = \infty$. So that the open set of service times on server 2 is trivially sub exponential. Similarly if $L(t)$ is dominatedly non-decreasing and $v \in (-\infty, 0]$, then $1 - B(t)$ is open in \mathbb{R} . Though, finite for a given supreme point, may be long-tailed if the supreme value is closely to infinity. This suffices. The other case follows with similar argument.

Significance here is statistical and implies relativeness with the other component. On the consequence of the former limitation, Medhi indicated that researchers were motivated to seek other approximation techniques. This quest gave birth to the diffusion approximation prevalent in queuing system modeling today. Essentially, the motivation leading to the above results and many others came from the works of Kingman and others on the asymptotic behaviors of queuing systems under heavy traffic.

3. Gumbel Distribution

As indicated earlier, recent research focus in heavy traffic queuing systems is in the study of tail behaviors of queues in form of limiting and extreme value distributions for various models. For instance, Glynn and Whitt proved the extreme value limit theorem for heavy traffic queues with general arrival and service time distributions called the G1/G/1. Using strong approximations under regularity conditions, the extreme waiting time among n -sized customers was derived. The number of customers in the system is assumed increasing as time grows large. It was shown that, when the traffic intensity ρ approaches 1 from the left and n approaches infinity at a suitable rate, the normalized maximum wait among n -customers converges to the Gumbel extreme-value distribution. Also, Glynn and Whitt added that the normalization depends only on the means and the variances of the inter arrival and service time distributions. On the contrary, if ρ is a fixed point, then the maximum waiting time fails to converge to the Gumbel distribution. The General Gumbel probability distribution for a continuous random variable $W(t)$ is given by

$$P(W \leq w) = \exp(-\exp(w)), -\infty < w < \infty$$

In addition, the Gumbel extreme value distribution holds for the queue length distribution. The lemma below summarizes the ρ -region of convenience for the Gumbel distribution in the G1/G/1 model.

Lemma 3.1

In a heavy traffic G1/G/1 queue under regularity conditions, if ρ approaches one from the left origin and n steadily increases to ∞ then, the Gumbel distribution sufficiently model the extreme value limit of the waiting time or the queue length distribution of customers in the system. Similarly, Szczotka and Woyczynski studied the G/G/1 queues with service and/or the inter arrival times of heavy tailed probability distributions. Szczotka and Woyczynski obtained that the waiting time distribution is exponential if the tail of the distribution of inter arrival times is heavier than that of the service times and is non-exponential in the opposite case. In other words, if the service times have a

heavy-tailed distribution heavier than that of the inter arrival times in the domain of attraction of a Levy process then, the limiting distribution is a Mittag-Leffler distribution. In addition, Szczotka and Woyczynski emphasized that under these modeling conditions, the queue length distribution can be analyzed. Limic studied the heavy traffic behavior of a G/G/1 Last-in-First-Out (LIFO) preemptive resume queue and derived a diffusion approximation for the model. Limic showed that the queue length process exhibits perhaps an unexpected heavy traffic behavior. In addition, the diffusion limit depends on the type of arrivals and services in a fairly intricate way which is related to the Wiener-Hopf factorization for random walks. Earlier in 1999, in their heavy traffic analysis of the G1/G/1 queue with heavy tailed service or arrival distribution of the regularly varying type, Boxma and Cohen have shown that if the traffic intensity of the G/G/1 system approaches unity and the tail of the service distribution is heavier than that of the arrival distribution, the stationary actual waiting time distribution together with a contraption factor is a function of the traffic load and converges to the Kovalenko distribution. In contrast, if the reversed is the case and all other factors kept constant, the stationary actual waiting time distribution will still depend on the traffic load but converge to the negative exponential distribution. Hence, limiting distributions of queuing systems to a large extend depend on the model constructed. Moreover, a slight variation of significant parameters may shift distribution of systems.

4. Dynamic Network Queue

Heavy traffic analysis of network queues especially, multi class networks are gaining grounds recently. This is not unconnected with its numerous importance. Kimura pointed out that multi-dimensional extension of server stations analysis is a natural diffusion model for a network of queues in computer systems. Similarly, Bertsekas and Gallager wrote that multi class network queues are used in analyzing problems of congestions and delays in computer systems, communication and complex productive systems. Their importance cannot be over emphasized. Unfortunately, several studies for instance have shown that not all multi class networks especially those with feed backs under heavy traffic can be approximated using the reflected Brownian motion see Williams. However, the open multi class type under heavy traffic supports the diffusion approximation. Already, Reiser and Kobayashi in the 70's have proved that network measurements via the diffusion approximation are quite adequate. Williams studied a multi class open queuing network using the semi martingale Brownian motion process and provides sufficient conditions for which heavy traffic limit theorem holds for such queues, see Williams for details. Similarly, Dai and Dai studied an open queuing network with finite buffers consisting of d -finite server stations. Given that a server stops working when the downstream buffer is full and all customers served at a station are homogeneous in terms of service requirements and routing. They proved that the normalized d -dimensional queue length process converges in distribution to a semi martingale reflecting Brownian motion in a d -dimensional box under a heavy traffic condition. Pekoz and Joglekar considered a $J/G/k$ finite buffer queue with a stationary ergodic arrival process and a general service with delayed

feedbacks and obtained that under certain mild conditions, the feedback flow of the class of customers re-entering the queue converges to the Poisson distribution when the delay waiting time distributions is scaled up. Similar studies on the network queues followed these developments especially via the martingale representations for limiting distributions of many server systems. Already, researches have shown that the reflected Ornstein-Uhlenbeck, the geometric Brownian motion, the reflected Levy and the reflected affine diffusion processes could be used to model successfully queuing systems with noisy processes such as reneging, balking and shunting process which are in effect measurable noises⁴, see Ward and Glynn and Guodong et al. worked on the Palm model and a finite capacity $M/M/N/M_n + M$ model with reneging via the martingale diffusion approach and provided limit proofs for the heavy traffic approximations of the queue length distribution. The martingale approach applied on the queue length process of customers involves random time changes and random thinings of the stochastic queue length process. They established a key central limit theorem and a key functional weak law of large numbers for the Palm model and the finite capacity $M/M/C$ model respectively. The result shows that the limiting queue length distribution in both models is a reflected Ornstein-Uhlenbeck diffusion process, an adapted stochastic process in which the stochastic variable changes more with time in addition to a finite variation component of the process. It is worthy to note that the shape of the reflected OU-graph to a greater extent depends on the initial customer size at time zero.

5. Modeling G/G/1

Modeling in heavy data traffic queuing systems is challenged by two important factors; the nature of data traffic itself and model selection. The former creates a controversy that saddles on the later. Researches on internet and telecommunications traffic processes revealed that data traffic is characterized by properties such as regular variation, long-range dependencies, self-similarity and heavy tail distributions see Leland et al., Park et al. and Stralka et al. Long-range dependency and self-similarity in essence are associated with heavy tail distributions. The combination of the two defines how burst a traffic system is. A heavy data traffic process may be bursty or not depending on the time scales it is considered. Though, Medhi wrote that a self-similar process arising from long-range dependent process like the internet traffic decays much slower than the exponential distribution in addition to a hyperbolic decay of autocorrelations, this does not render the Poisson traffic models that is non-bursty inadequate or better still out of the internet domain as some researches tried to portray.

The Poisson Arguments:

Since the pioneering works of Leland et al., Willinger and Paxson and many others on the new network traffic description that appeared alternatives to the Poisson models, it appears the Poisson distribution has lost its place as a suitable distribution for describing the internet traffic today. Amidst its advantage, history and effectiveness, a lot has been published about its inadequacy without attending to scaling issues in both time and space. For instance in 1994, Leland et al., using long, high resolution traces of Ethernet packets indicates that arrival rates of the Internet Protocol

(IP) packets on a Local Area Network (LAN) exhibit not Poisson but self-similar behavior. However, similar studies have shown that within a given time scale, the two traffic networks coexist together. For instance, Boxma and Cohen in observing the plots of Ethernet traffic measurements on LAN of Willinger et al., WAN of Paxson and Floyd and VBR (Variable-Bit Rate) of Beran et al., Boxma and Cohen watched that bursty sub periods are alternated by less bursty sub periods in each of these traffic processes indicating the coexistence of the Poisson traffic and self-similar traffic processes. Similarly, Karagianis et al., have shown that there exist a Poisson process and a long range dependence in heavy Internet backbone traffic. Furthermore, Karagianis et al., indicated that the User Datagram Protocol/ Transmission Control Protocol (TCP/UDP) packets obey a Poisson process at sub-second time scales while they are long range dependent at large time scales. This suggests that relatively simple statistical theories of the Poisson process can still be applicable to the design and optimization of the internet. Similarly, Lee and Kim proved that at small time scales, inter-arrival times of protocols such as the simple mail transfer protocol (smtp) is a Poisson process. Finally, Karagiannis et al. have shown that the mighty internet traffic in the center of this controversy at sub-second time scales appears well described by Poisson packet arrivals and evident that, the ongoing pattern of Internet evolution may eventually renewed its Poisson tendencies even for the super second time scale. Time scaling factor appears to be a decisive factor in defining suitability in this modeling case. Consequently, the below lemma follows:

Lemma 5.1

The necessary and sufficient condition for the internet traffic or a similar process to fit in the Poisson process is that, the time scaling is sub-second.

Consequently, modeling in the heavy traffic sense may well be done with the Poisson model at sub-second time interval and at large time scales, modern day data traffic models such as the self-similar model that can capture important features such as traffic burstiness with long tail distributions are effective.

6. Applications

The role heavy traffic analysis plays in the development and advancement of service systems especially computing and telecommunications systems cannot be over emphasized. Initially, even the work of Kingman on the asymptotic waiting time distribution in a G/G/1 queue is a modeling of delay time distribution in a general service system just about to reach its service capacity. In heavy traffic analysis, any limit theorem derived is to provide understanding and approximation of distributions and tail behavior of measures for bettering performance of service systems. The general motive is advancement of corresponding service systems and application depends on the reality of model. For instance, heavy traffic analysis of networks queues of various priorities is for problem analysis of modern systems. As Bertsekas and Gallagher indicated; such models are used in analyzing problems of congestions and delays in computer systems, communications and complex productive

systems. Priority queuing analyses and those with service interruptions may be classical models of computer systems. In the context of queuing synthesis, as Kimura indicated; diffusion models are for reliability and control problems in computer and telecommunication systems. Without such analysis, problems such as those mentioned could have down our systems. The processor sharing discipline queues are used in modeling time-sharing protocols in computer systems. High speed wireless networks carrying multimedia applications under long range dependence and heavy tailed properties are troubled by excess usage. Buche et al. indicated that heavy traffic analysis of long range dependence in wireless internet traffic provides relief to troubles such as large file sizes downloads from the internet and multimedia applications for instance streaming a video. The processor sharing models such as those studied in Kleinrock and Ritchie and Thompson in the 70's and more recently, the limited processor model of Zhang and Zwart has been widely used in the analysis of computer systems, network servers and data transmission over the internet. Kruk et al. indicated that in the last decade, substantial attention has been paid to queuing systems in which customers have deadlines for service in heavy traffic (EDF queues). These types of queuing models feature in telecommunication systems carrying digitized voice or video traffic, tracking systems and real-time control systems. In the case of voice or video traffic, the packet information must be received, processed and displayed within stringent timing bounds so that the integrity of the transmission is maintained. Similarly, there are processing requirements for tracking systems that guarantee that a track can be successfully followed. Real-time control systems for instance, those associated with modern avionics systems, manufacturing plants or automobiles also gather data that must be processed within stringent timing requirements in order for the system to maintain stability or react to changes in the operating environment. Another class of heavy traffic queuing models of varied applications is the polling models. Levy and Sidi indicated that these models were first introduced in 1970 precisely when the cyclic queues were used in modeling time allocations in computer systems. In these models, queues are visited by a single server in a cyclic discipline. Such models are applied in the performance analysis of communication systems such as token rings and packet switches, where a single server resource is shared among many traffic stream demands on the resource, see Coffman et al. In addition, Levy and Sidi provided more areas of applications of polling queues in heavy traffic to include random access protocols of computer systems, robotics and manufacturing systems. In other fields such as transportation, where heavy traffic analysis has several applications, the work of David et al., on the diffusion modeling of an airport queue is an excellent use of modeling via diffusion approximation in the transportation sector. Other relevant areas of application of the heavy traffic modeling in approximations queuing systems include the repairman problem etc. Already, the works of Iglehart in the 60's has derived a diffusion limit approximations for several server case.

7. Conclusion

In this article, mathematical modeling in heavy traffic queuing systems is generally surveyed. Initially, the onset of modeling in heavy traffic queues and asymptotic behaviors for different models were reviewed and distributions uncovered. We also looked at the diffusion approximation as a remedy to queuing analysis and approximations. Modeling both in network and heavy traffic data systems and matters arising from the internet and telecommunication traffic modeling via Poisson models and assumptions were studied and a justification on the suitability of the Poisson arrival process in addition to the new network traffic in capturing the internet traffic was supported. Finally, we provide real areas of application of heavy traffic models developed for the benefits of service systems.

References

- [1] Abate, J. and Whitt, W. (1965) Asymptotics for M/G/1 Low-Priority Waiting-Time Tail Probabilities. *Queueing Systems*, 25, 173-233. <http://dx.doi.org/10.1023/A:1019104402024>
- [2] Abate, J. and Whitt, W. (1994) A Heavy-Traffic Expansion for Asymptotic Decay Rates of Tail Probabilities in Multi-Channel Queues. *Operational Research Letters*, 15, 223-230. [http://dx.doi.org/10.1016/0167-6377\(94\)90081-7](http://dx.doi.org/10.1016/0167-6377(94)90081-7)
- [3] Atar, R., Mandelbaum, A. and Shaikhet, G. (2006) Queueing Systems with Many Servers: Null Controllability in Heavy Traffic. *The Annals of Applied Probability*, 16, 1764-1804. <http://dx.doi.org/10.1214/105051606000000358>
- [4] Borovkov, A. (1980) *Asymptotic Methods in Theory of Queues*. Nauka, Moscow in Russian.
- [5] Boxma, O.J., Deng, Q. and Zwart, A.P. (2003) Waiting Time Asymptotics for the M/G/2 Queue with Heterogenous Servers. *Queueing Systems*, 40, 5-31. <http://dx.doi.org/10.1023/A:1017913826973>
- [6] Gaver Jr., D.P. (1968) Diffusion Approximations and Modes for Certain Congestion Problems. *Journal of Applied Probability*, 5, 607-623. <http://dx.doi.org/10.2307/3211925>
- [7] Glynn, P.W. and Whitt, W. (1995) Heavy-Traffic Extreme-Value Limits for Queues. *Operational Research Letters*, 18, 107-111. [http://dx.doi.org/10.1016/0167-6377\(95\)00048-8](http://dx.doi.org/10.1016/0167-6377(95)00048-8)
- [8] Guodong, P., Tareja, R. and Whitt, W. (2007) Martingale Proofs for Many-Server Heavy-Traffic Limits for Markovian Queues. *Probability Surveys*, 4, 193-267. <http://dx.doi.org/10.1214/06-PS091>
- [9] Iglehart, D.L. (1965) Limit Diffusion Approximations for the Many Server Queues and the Repairmen Problem. *Journal of Applied Probability*, 2, 429-441. <http://dx.doi.org/10.2307/3212203>
- [10] Iglehart, D.L. and Whitt, W. (1970) Multiple Channel Queue in Heavy Traffic I and II. *Journal of Applied Probability*, 2, 150-177, 355-369.
- [11] Karagiannis, T., Molle, M., Faloutsos, M. and Broido, A. (2008) A Nonstationary View on Poisson Internet Traffic. *Proceedings of IEEE INFOCOM*, 84-89.

- [12] Kimura, T. (2004) Diffusion Models for Computer/Communication Systems. Economic Journal of Hokkaido University, 33, 37-52.
- [13] Kingman, J.F.C. (1961) The Single Server Queue in Heavy Traffic. Mathematical Proceedings of the Cambridge Philosophical Society, 57, 902-904. <http://dx.doi.org/10.1017/S0305004100036094>
- [14] Kollerstrom, J. (1974) Heavy Traffic Theory for Queues with Several Servers 1. Journal of Applied Probability, 11, 544-552. <http://dx.doi.org/10.2307/3212698>.
- [15] Kos, A. and Bester, J. (2003) Poisson Packet Traffic Generation Based on Empirical Data. Systems, Cybernetics and Informatics, 1, 80-83.
- [16] Lee, C. and Weerasinghe, A. (2011) Convergence of a Queueing System in Heavy Traffic with General Patience-Time Distributions. Stochastic Processes and Their Applications, 121, 2507-2552. <http://dx.doi.org/10.1016/j.spa.2011.07.003>
- [17] Lee, Y. and Kim, J.S. (2008) Characterization of Large-Scale SMTP Traffic: The Coexistence of the Poisson Process and Self-Similarity. Proceedings of Mascot, 143-152.
- [18] Limic, V. (1999) On the Behavior of LIFO Preemptive Resume Queues in Heavy Traffic. Electronic Communications in Probability, 4, 13-27.
- [19] Medhi, J. (2003) Stochastic Models in Queueing theory. 2nd Edition, Academic Press, California, USA.
- [20] Newell, G.F. (1982) Applications of Queueing Theory. 2nd Edition, Chapman and Hall, London. <http://dx.doi.org/10.1007/978-94-009-5970-5>
- [21] Reiser, M. and Kobayashi, H. (1974) Accuracy of the Diffusion Approximation for Some Queueing Systems. IBM Journal of Research and Development, 18, 110-124. <http://dx.doi.org/10.1147/rd.182.0110>
- [22] Szczotka, W. and Woyczynski, W. (2004) Heavy-Tailed Dependent Queues in Heavy Traffic. Probability and Mathematical Statistics, 24, 67-96.
- [23] Wagenmakers, E.J., Raoul, P.P.P., Grassman, C.M. and Peter, M. (2005) On the Relationship between the Mean and the Variance of a Diffusion Model of Response Time Distribution. Journal of Mathematical Psychology, 49, 195-204. <http://dx.doi.org/10.1016/j.jmp.2005.02.003>
- [24] Whitt, W. (1974) Heavy Traffic Limit Theorems for Queues: A Survey. Springer-Verlag, Berlin, Heidelberg, New York, 1-46.