Trends in the Mathematics of Queuing Systems

Sulaiman Sani, Onkabetse A. Daman

Abstract— **In this article, we study the trends in queuing system mathematics (mathematical study of waiting lines) from its inception in 1909 to date. The ai m is to educate on how advances in system engineering a nd operations research are transforming study trends in terms of scholarly contributions (historical evolution), pr oblems formulation, analytic techniques, modeling and r esults. To achieve this objective, articles on this field of o perations research are studied and general trends uncovered and made easily understandable for educational purposes. In the end, we came out with deductions that trends in the mathematics of queuing systems depend to a large extend on developments in operation systems and engineering. What makes this paper most interesting is the understanding that queuing problems are fast becoming pure stochastic (diffusion) problems. This understanding is made more elaborated and easily understandable for a wide variety of audience.**

2000 Mathematics Subject Classification: 60K25

Index Terms— **Queue, occupation rate, the G/G/C queuing model, regular variation**

I. INTRODUCTION

There are basic words an interested reader of a piece covering the mathematics of queuing systems should initially understand[1]. Beginning with the word **queue** which derives its meaning from the Latin word ‘’cauda” meaning tail. Literally, to queue is to tail or wait of course for a reason which may be to receive service[2]. On the other hand, **queuing** is a process more precisely, a diffusion process and **queuing theory** studies such diffusions involving the manner which inputs, arrivals, packets or customers move from a concentrated area (waiting line) to an isolated one (service area) in somewhat macroscopic, visibly continuous or semi continuous process[3]. Operationally, Medhi [33] defines queuing theory as the mathematical study of waiting lines formed whenever the demand for service exceeds the capacity to provide it. In its pure mathematical sense, it refers to the theory of formation and behavior of queues (transient and limiting)[4] involving problems connected with traffic congestions and storage systems. These definitions extend the relevance of queuing theory to a wide variety of contentious situations such as how customers checkout line forms (arrival process), how it can be minimized (queuing analysis), how many calls a telephone switch can handle (service capacity), how long a customer can wait for a service (waiting time analysis) and so on. Generally, the object of queuing theory is relieving problems in business settings primarily in operational management and operations research via the well-established theorems of mathematics.

1.1 Queue Mathematics: Significance:

To date, the mathematics of queuing systems enjoys tremendous attention resulting from the ever increasing, multi-complexity of service systems[5]. From service systems that are homogenous to those with reasonable variations[6] (deterministic and stochastic) for operational relevance, marketing, system development or advancement. This makes the mathematics a continuum with dynamical behaviors and trends. Service systems are advancing and advances are transforming service spheres necessitating changes in trends and study dimensions. Intuitively, understanding these trends will lead to a better understanding of the future of service systems where queuing is evident. On the role of studying historical trends to knowledge advancements and motivation for instance, Man-Keung and Tzanakis [32] has pointed out that historical behaviors enhances learning and teaching, an appetizer or a dessert which caters respectively to motivation, content or enrichment. More so, trends of the mathematics of queues can uncover hidden realities vital for understanding not only what Man-Keung and Tzanakis [32] indicated, but a gateway to the future of systems. This is because system process always precedes system product. Thus, studying the mathematical behaviors of queuing systems will not only enrich us with tales but aid our understanding of complicated scenarios we define for operational systems. The work is organized in (7) sections as follows; in section two, we present challenges in queuing mathematics and analysis necessitating behavioral shift generally. In section three, evolutionary (historical behavior) trends in queuing mathematics from inception to the present were identified, in each case, essential scholarly contributions are identified and stated. Section four discusses the trends of queuing systems mathematics today and section five identifies the accompanying methodological behavior presented in form of summarized and easily understandable limit theorems for selected queues. In section six, we investigate recent trends in queue mathematics with the emergence of data traffic phenomenon in telecommunications and computer systems today, fractal queuing theory and effect of long-range dependencies in queuing performance vis-a-vis contributions of eminent scholars and mathematicians. The article is

Manuscript received Dec 31, 2014

Sulaiman Sani, Department of Mathematics, University of Botswana-Gaborone

Onkabetse A. Daman, Department of Mathematics, University of Botswana-Gaborone

[2] In shops, malls, telecommunications and computer business centers, etc

[3] Continuity here denotes widely approximate continuity.

[4] Instantaneous and long-time behavior of queues

[5] Computer systems, telecommunication systems and complex productive systems.

[6] Heterogenous server systems.

concluded in section seven with summary of trends in queuing theory.

II. Why Changing Trends

Mathematizing queuing systems (queuing theory or modeling) is a challenging task. It faces serious challenges depending on the nature and reality of the queue in question. The challenges generally emanated from the inter relationship between system engineering, system design and queuing theory. A question of interest to the reader is that; how does advancements in system design and engineering create problems to queuing theory mathematics and modeling? Simply put, advances come with newness which defines volatile scenarios [7] for queues forcing challenging transformations in study trends, behaviors and dimensions. These volatile scenarios among others include; queuing systems realities, analytic technique suitability, modeling, conditioning and adaptations, etc.

On this basis, queuing systems can be seen as either stable or unstable[8] (noisy) systems; that is **the nature of queuing systems.** In early queuing period, stable queuing models that can be analyzed classically using Laplace techniques are more often constructed and modeled see Whitt [48]. Whereas modeling looks easier in stable queues, the converse is true in unstable or noisy queues. This is because the later possesses more randomness that transforms its distribution from normal to somewhat non-normal. Also, unstable queues are known to exhibit long-range dependencies, a long-term memory problem in certain queues[9] that makes decay slower than the exponential random variable, see Stralka et al [41]. What is challenging is that, noisy systems with unstable queues that seemed **difficult in analysis via the classical approaches are the bulk found in today's systems[10].** Strzalka et al [41] indicated that using classical models in this stage of network traffic modeling for instance can lead to mistaken performance predictions and inadequate network design. More generally, if we fail to represent processes in queuing systems accurately, that will lead to under estimations or otherwise of performance. This necessitates changing trends in queuing theory mathematics generally. But even for stable queues, modeling is pretty hard especially in the context of transient solution and analysis. On this difficulty, Medhi [33] indicated that, the transient-state distributions of simple models are even difficult to handle and to date, that of the M/G/1 queue as simple as it looks is unknown. Whitt [48] pointed out that classical modeling in queuing theory is under the Laplacian curtain and complex systems analysis via Laplace transform is uneasy and challenging. **Existing techniques are really difficult even for simple models not to talk of complex models which are properties of integrated systems.** Consequently, the need to shift study trends.

Another challenging aspect of modeling queuing systems necessitating changes in trends and transformation is **the physical realities of queues**. Queuing systems are generally unconstrained [11] and modeling unconstrained problems is difficult. As Sulaiman et al [43] puts it, exact solution for an unconstrained problem is merely an ideal case and practically is unrealistic. Queuing systems especially those with general arrival and service distributions (the G/G/C's) are in essence, unconstrained in nature, exact solutions are generally too ideal for such unconstrained problems. **Consequently, numerical approximations which can provide room for error analysis are more realistic, functional and operative**. Thus, the need to enact the numerical trend that proffers easy solution however, meaningful. Numerical approximations and simulations of queuing parameters nowadays are necessary though not sufficient to provide relief. Similarly, it is extremely difficult to control realities such as queuing shocks; making modeling assumes lots of stability conditions. The balking process, the shunting process and the reneging process etc for instance are realistic noises[12] on any queuing system. Controlling such noises for optimality poses a serious challenge to modeling the same way Brownian noise shocks the financial markets. This is because, the randomness in the two systems is similar and the calculus is the same. The calculus to date is not understood by very many queuing theorist. It is a different form of calculus and its definition of system estimators is really tricky and problem posing. Moments such as the variance of a stochastic waiting time may be challenging to compute not to talk of joint distributions of multiple queuing systems in a connected topological space which are properties of integrated networks. **For such queuing systems, a shift in mathematics from the classical use of Laplace techniques to a more vibrant use of the diffusion approximations is evident.**

The nature of queuing in some systems such as the internet challenges modeling in system engineering and data traffic science. Today, service systems are so complex that queuing features such as queue openness, queue security, queue scaling, failure handling rates and concurrency are understated, see Strzalka et al [41]. This occurs due to system complexities leading to degradation and parameter collapse which undermine the sanctity of results. **The absence of a unified queuing model to solve problems in queuing systems creates an intrinsic problem of multiplicity of models which solves similar queuing problems differently**. Though, most mathematical approximations via different analytic techniques seem to agree, the choice in practice and applications is difficult and often not in a usable form. As Whitt [48] pointed out, that the limitations of queuing theory are obviously due in part to the inability of obtaining queuing results (models) in a usable form. At present, queuing theory remains under the Laplacian curtain and analyzing complex systems via the Laplace transform is really uneasy and challenging. Data traffic in heavy traffic and diffusive queuing systems such as those found in telecommunications are gathering weight, [see Whitt [48], Christian et al [9], Ward and Glynn [47], David et al [12],…][13]. This indicates that all

[7] Complex scenarios often containing lots of intricacies and complications.

[8] Stability for a given queue depends on both the arrival and service processes.

[9] Heavy traffic queues for instance.

[10] Whitt [45] posited that, it is not easy to handle double, triple or quadruple transforms.

[11] Queuing problems are often open boundary problems. Applying constraints to such systems is a daunting challenge.

[12] Processes that generate continuous time unsteadiness in a queuing system.

[13] In this modeling era, the number of models constructed from the time of A.K. Erlang to date are enormous and the coverage is wide.

hope is not lost in advancing and improving queuing systems through mathematics; that is, in terms of study dimensioning and analytic techniques. In fact, we can argue with reasons that the challenges ended up strengthening queuing studies and modeling. This is evident in the quantum of models developed over time by queuing theorists of repute to understand system performance and behaviors. More so, we see the evolution of new behaviors and trends in queuing theory for instance fractal queuing theory, heavy traffic approximation, etc to capture every bit of challenge posed by system engineering and its accompanying developments.

III. Evolutionary Trends of Queuing Theory Mathematics

The historical evolution of queuing theory mathematics is interesting as the theory itself. Medhi [33] dated the origin of queuing mathematcis as far back as 1909 when the Danish mathematician Agner Krarup Erlang published his fundamental paper on congestion in telephone traffic. Erlang, in addition to formulating analytic practical problems and solutions laid a solid foundation to queuing theory in terms of basic assumptions and techniques of analysis. Interestingly, these techniques are being used to date even in the wider areas of modern communications and computer systems. For instance, using Erlang basic assumptions and techniques, Ericsson telecom developed a programming language called Erlang used in programming concurrent processes and verifications such as the conditional term rewriting systems (CTRS) see Thomas and Giesl [45]. Erlang could not live long to see how his works transformed telecommunications engineering. Interestingly, his contribution was recognized in 1946 when the International Unit of Telephone Traffic (CCITT) was named Erlangs in honor of Agner Krarup Erlang, see Brockmeyer and Halstrom [6]. His works contributed immensely to the development of queuing models vital for analyzing lost and delay behaviors in queuing systems mathematically. For instance, the Erlang-B and Erlang-C mathematical models developed by A.K. Erlang are used in telephone and telecommunications analysis to define probabilities that an incoming call[14] is rejected or delayed. Also, the Erlang-C model gives the probability that an incoming call has to wait[15] before service, see Kleinrock [28]. The developed Erlang models are deterministic and assume Poissonian arrival process and exponentially distributed service times. These models can be computed statistically to measure both probabilities. **What is essential here is that the initial trend in queuing theory mathematics is statistically deterministic**.

Though, Erlang pioneered queuing theory mathematics especially its applications to operations research, the pioneer of this type of mathematics from the perspective of stochastic processes was D. G. Kendal. In 1951, Kendal developed and introduced certain notations which to date are adopted to denote queuing systems. The Kendal's A/B/C notation[16] specifies three basic characteristics in a given queuing system namely; the arrival process (A), the service distribution (B) and the number of servers in a system. Kendal's integral equation relating the Laplace-Stieljes transformations of the busy period and that of the arrival process is a remarkable achievement and breakthrough in the field of queuing theory. For, it depicts the continuous nature and behavior of queuing systems in a more advanced manner than the Erlangian models. The Kendal's integral equation simplifies lots of queuing problems in single server systems with priority customer classes. More so, it reflects the reality of queuing systems at rush and peak times which are continuous streams of homogenous events. To date lots of priority models covering peak and rush hour (steady state models) behaviors are computed using the Kendal's equation. Also, numerical approximations of models which simplify difficulties in the analysis of transforms are computed using the integral equation see Bejan [4]. The Kendal's era in queuing theory mathematics marked the beginning of modeling queuing systems as stochastic process. From trends perspective, we see that advancements and complications in queuing systems drive the need for changing focus, analytic techniques and approaches referenced to the Erlang-Kendal shift.

In 1961, D.C. Little came up with a fundamental relationship between the averages of three quantities in a queuing system in what is known in the queuing theory mathematics as Little's formula. The formula relates the behavior of the average number of customers in the system or in the queue to the average sojourn or waiting time. To date, the formula is applied in many areas of manufacturing and service systems as well as in decision making to quantify expected behaviors of these parameters for better service delivery.

After these breakthroughs, a lot of sub areas of interest in queuing theory mathematics continue to emerge especially in the 40's. Mathematicians had understood that queuing problems can be seen in the light of both statistical and stochastic behaviors. For instance, Franken et al [15] indicated that in the early 50's, mathematicians where faced with the challenge of developing appropriate mathematical tools to describe the behavior of sequence of arrivals in a given system. The first stage of this development was taken by Conny Palm[17] in 1943 and was made mathematically precise and well expanded by Aleksandr Khinchine[18]. In 1955 precisely, Khinchine studied point processes on the positive real line which he addressed as stream of homogenous events. This development opened further examination of similar areas among others including that of insensitivity of queuing systems[19]. Here, existence and continuity statements and relationships between time and stationary quantities with special inputs were emphasized. This led to the emergence of a new class of random processes connected with point processes which seemed most suitable for describing queuing systems. The new class is termed; random processes with

[14] An arrival occurring at an arbitrary time t.

[15] For a time t strickly finite.

[16] Kendall's A/B/C notation over the years have been modified to include necessary parameters such as size of waiting rooms which may be finite or infinite, noisy processes, etc.

[17] Born in Sweden, lived between1907-1951; first paper on queuing theory in 1936.

[18] Khinchine contributions among others include; the development of an analytic technique for obtaining the steady state solution of the M/G/1. The Pollaczec-Khinchine formular is remarkable in queuing analysis of Poisson arrival uni server systems to date.

[19] Queues without waiting sense. For instance, the palm model or any finite capacity queue with expected arrivals equal to the number of servers.

embedded marked point process (MPP). The development of this class of processes was attributed to Khinchine and Kendal in 1976.

After 1976, queuing theory mathematics, problems, models and techniques have developed firm mathematical analysis with ripe continuity statements, relationships and modeling. This advancement in trends could be argued to be the driving force behind the discovery of the mobile telephones and the internet. As IFIP TC6 Working Group 6.3 [21] puts it; *That the history of communication systems research is tightly coupled with the performance evaluation. Brilliant examples range from A.K. Erlang queuing system studies applied to dimensioning the telephone network, to Kleinrock works applying queuing theory to investigate the packet switching technology which is the foundation of the internet.*

Trends in Queuing Theory Mathematics Today:

The last four decades to date symbolizes an era of model development as Medhi [33] implies. Queuing theory mathematicians today seemed more interested in model development. This era is of modeling and lots of queuing models are developed. What we witnessed of late is a shift in paradigm that centers on creating models to capture every bit of system development, advancement and conjecture, [see next paragraph for examples]. Models may be classified under two categories; deterministic models (stable models) and stochastic models (unstable and heavy traffic). The methodology today is to pose a problem and model its transient or limiting behavior. However, it requires the mathematical analysis of existence of solution so that the constructed model is realistic.

The central limit theorems and the maximum principle form the basis for proofing existence of queuing solutions today, see Whitt [48]. Limit theorems have been developed for various queuing models (stable and unstable queues) to show limiting behavior. In the rest of this article, we present eminent contributions by mathematicians as citations in various modeling conditions. For instance, the series of deterministic results obtained over the years on several models are enormous. Federgruen and Tijms [14] obtained the stationary distribution of the queue length[20] for the M/G/1. Hoksad [19] and [20] worked on a more general queue called the M/G/m in terms of its limiting state behavior and its specific case. Smith [40] specified the system performance of a finite capacity queue called the M/G/C/K. Also, Tijms et al [46] approximated the steady state probabilities for the M/G/C queue. In the area of heterogeneous queuing systems, Boxma et al [8] derived the waiting time asymptotics for the heterogeneous server M/G/2 with one exponential and one general server. Sulaiman et al [42] and [43] generalized the steady state behavior and decay approximations for (C-1)-exponential servers and a general server of regular variation modeled of Boxma et al [8]. On the other hand, the series of papers written by **Abate, Whitt, Mandelbrot, Glynn, Ward, Zwart, Zheng, Boxma, Krishnamoorthy, Mandelbaum etc** among other scholars of queuing theory [see; Christian et al [9]] provide a new and advanced trend in studying queuing systems with noise. Today, queuing theory mathematics has advanced in approach that modeling takes into account some forms of randomize noises shocking the queue in a measurable sense. This trend advances the deterministic approach and requires few stability conditions such as the scaling and dependency conditions. Unfortunately, it requires some forms of analytic calculus distinct from the conventional Lesbeque-Riemann calculus, a unique calculus that adapts the deterministic and the variation components of queuing systems. Queuing system modeling uses stochastic calculus nowadays to provide meaning to queuing system behaviors. The initial fear was that queuing systems diffuse so slowly that giving them a stochastic outlook will be an over emphasis. However, recent studies on diffusion behavior of queuing systems have shown that approximations exist and are tractable; see David et al [12] and Christian et al [9]. Equally, 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 behaviors such as reneging, balking and shunting processes which are in effect measurable noises. However, this shift is not without a price. Certain challenges in form of relevance, model adaptability, and convenience shall be overcome to categorically continue queuing theory mathematics vis-a-via developments in systems engineering and operations research. In conclusion to this section, even with these developments, advancements and trends today, the queuing theory mathematics will continue to change in this dynamic world of uncertainties and advancements

[20] See also Jain and Sigman [24], Boucherie and Boxma [5].

IV. Methodology Trends Today

From methodology point of view, queuing theory mathematics has seen transformations in trends right from the time of A.K. Erlang to date. The reason is due to changes in continuum of queuing traffic, occupation rates, distributions and available servers. These changes are reflected in the limit theorems (central and extreme valued) for different models. Whitt [48] classifies queuing processes as either standard (light traffic) or growing (heavy traffic) and summarizes limit theorems for each. In addition, Whitt [48] added that the initial behavior in methodology was to compute the steady state solutions via balanced equations of standard models for server occupation rate below the critical value. This trend was followed by the computation of time-dependent behaviors in models though, with some difficulties initially, but later done with ease. Presently, the behavior today is on growing processes (heavy traffic) with emphasis on asymptotic behaviors of models in form of tail probabilities with limit theorems at the center stage. Limit theorems have been put to use since 1909 and are fundamental in analyzing both the extreme and central behaviors of queuing systems. Whitt [48] classifies limit theorems for queuing models as classical or functional from the perspective of resulting convergence. The classical limit theorems consist of the central limit theorem, the laws of large numbers, the laws of iterated logarithms and the extreme value theorems. The functional limit theorems consist of the Donsker theorem for iterated random processes and the continuous mapping theorems. Basically, limit theorems inform us on the behavior of a queuing variable by revealing its statistical regularity under macroscopic view of uncertainty. The statement of the classical limit theorems involves obtaining the transient or asymptotic behavior of some n-dimensional queuing vectors as n grows large or its associated extreme value sequence. Our survey shows that

limit theorems (central and extreme) exist for lots of models. For instance, Glynn and Whitt [17] proved a limit theorem for heavy traffic queues with general arrival and service time distributions called the G/G/1. The proof uses strong approximations under regularity conditions to derive the maximum waiting time distribution under heavy traffic. Glynn and Whitt [17] added that the normalization depends only on the means and the variances of the inter arrival and service time distribution. Unfortunately, for a fixed value of the intensity parameter, the maximum waiting time fails to converge to the Gumbel distribution. In addition, Glynn and Whitt [17] have shown that the waiting time limit for the queue in question exist even if the waiting time function fails to converge for a fixed value of the intensity parameter. Similarly, for the heterogeneous-server M/G/2 queuing model, Boxma et al [8] proved the asymptotic limits for the waiting time distribution for varying size of the arrival rate compared to the exponential service rate. The modeling conditions require that the service time distribution of customers served by the general server has regularly varying behavior at a known index. Boxma et al [8] have shown that if the arrival rate is strictly greater than that of the exponential server, the tail waiting time distribution will have a regularly varying behavior at infinity at a complementary index of the service time distribution. For the classical M/G/1 queue with two priority classes and non preemptive and preemptive-resume disciplines, Abate and Whitt [1] derived these limit theorems. Abate and Whitt [1] proved that the low-priority limiting waiting-time behavior is a geometric random sum of independent and identically distributed random variables like the M/G/1 first come first served (fcfs) waiting-time distribution. On the asymptotic behavior of tail probabilities, Abate and Whitt also indicated that there is routinely a region such that the tail probabilities have non-exponential asymptotic behavior even if the service time distributions are exponential. In addition, the asymptotic behavior of the tail probabilities tends to be determined by the non-exponential asymptotic behavior for the high-priority busy-period distribution. On the functional limit theorems for queuing systems the work of Whitt [48] is worth mentioning. Whitt [48] summarized functional limit theorems for both noisy and non-noisy single server queues. Using the open mapping theorem, Whitt [48] indicated that just as stochastic functions converges to reflected Brownian motion (Donsker theorem), a discrete-time queuing model with cumulative net-input process of stationary increments and jumps of infinite variance or mean, the central limiting behavior is a reflected stable process. This limiting distribution can be computed by numerically inverting its Laplace transform. However, for a sequence of models (multi systems), the queue need not be in heavy traffic. The limiting behavior is a reflected Levy process. In addition, if the jumps are positive increasing then, the steady-state behavior of the reflected Levy process can be computed by numerically inverting its Laplace transform also. Finally, Whitt [48] established that the functional central limit theorem for the customers in the queue when the input process is a superposition of many independent processes with complex dependence, the limiting input behavior is a Gaussian process. Similarly, for multi server queues, limit theorems have equally been proved. For instance, Guodong et al [18] proved the heavy traffic limit approximations for the queue length distribution in a multi-server model with Poisson arrival behavior. Using the martingale approach involving random time changes and random thinings of the stochastic queue length process, a key central limit theorem and a key functional weak law of large numbers for the popular Palm model and the finite capacity M/M/C model are respectively established. Interested readers should refer to the above reference for details. Other central limit theorems for relevant queuing models under heavy traffic include Abate [2],..etc.

V. Recent Trends in Queuing Theory Mathematics

The 1980's and the last decade witnessed two significant developments in telecommunications engineering; the invention of the facsimile machine and the internet. These developments change the nature of queuing theory mathematics completely. A somewhat new form of traffic process that exhibits different statistical behavior with the Poisson process prevalent in early telecommunications modeling now emerged. The emerging process has a long-term memory which the Poisson process can not statistically conserve. Recent day traffic termed data traffic for instance the internet traffic seems continuous in its arrival behavior in contrast to the discrete voice calls behavior of the telephone age. Also, in Medhi's [33] description, data traffic do not come in steady rate like the telephone traffic rather, it has starts and fits with lulls in between. It possesses long-range dependencies and regularly varying capacities due to heavy traffic in addition to having large variability in contrast to the voice traffic which has small distributional variance. Consequently, the Poisson process becomes less realistic and limited on this process. This motivates the need for a realistic trend to tackle the present nature and behavior of the network traffic which eventually gave rise to the recent trend in queuing studies. As Florin and Jens [10] observed, the limitation of the Poisson process motivated the development of alternative trends in queuing theory over the past three decades with the emergence of rapid growth of high-speed data networks. Equally, techniques of analysis that seemed successful in queuing modeling became almost vague recently as a result; see Strzalka et al [41]. The relevance of the emerging trends to tackle the recent day traffic behavior especially for the internet community has become evident with the discovery that the internet traffic is fundamentally different [21] from Poisson traffic. It is this necessity to overcome the Poisson assumption limitation that produced the heavy traffic and diffusion approximation recently in queuing theory. Though, approximating a discrete-time stochastic process by a diffusion process is not new, application into queuing theory is of recent origin (beginning of the 70's). The support for this behavior and trend is the work of Kingman [27] on a general queue called the G/G/1. The result is called the central limit theorem for queuing theory, see Medhi [33].

[21] The Poisson suitability argument for the internet and similar processes still holds good. Boxma and Cohen in [6] observed that, in both LAN and WAN traffic, bursty sub-periods are alternated by less bursty sub-periods, indicating the coexistence of the Poisson traffic and self similar traffic processes see also Karagianis et al [24]. Most importantly, the argument gave rise to another significant model of analysis.

5.1. Data Traffic Trends and Queuing Theory Mathematics:

Data traffic analysis is synonymous to heavy traffic analysis. Heavy traffic approximations in queuing theory mathematics started with the work of Kingman [27] on a general arriving and service time queuing model called G/G/1 queue. Kingman [27] proved that for the G/G/1 queue under heavy traffic, the waiting time distribution could be approximated by an exponential distribution. Kingman made a conjecture for the seemingly more interesting multi-server G/G/C queue. He conjectured that the waiting time distribution could similarly be approximated by an exponential distribution. In 1974, Kollerstrom [29] proved the conjecture to affirm that the waiting time distribution for such queue is an exponential distribution.

On the other hand, the diffusion approximation for heavy traffic queuing systems came to light in the works of Iglehart [22], Gaver [16] and Newell [35]. It involves approximating the limit of a sequence of stochastic queuing variables (heavy traffic) as a Brownian process (diffusion). Guadong et al [18] indicated that Iglehart established the first limit theorem for the palm model via diffusion approximations and Gaver [16] considered the technique for certain congestion problems in 1968, see Medhi [33]. In 1970, Iglehart and Whitt [23] justified the suitability of the diffusion approximation for queuing variables by establishing a limit theorem for the G/G/C queue. The theorem proved that both the queue length and the waiting time distributions could be approximated by a Brownian motion process. In 1974, Reiser and Kobayashi [39] studied the accuracy of the diffusion approximation on some networks of queuing systems. The accuracy was considered for a wide class of distributional forms of inters arrival and service times for various models. Reiser and Kobayashi [39] concluded similar to Iglehart and Whitt [23] that the diffusion approximation is quite adequate in most cases, more adequate than the exponential server model prevalent in computer system modeling. Since then, several models have been developed to approximate performance of systems in the form of tail probabilities, extreme behavior, moments and distributions using the diffusion approximation. For instance, see Glynn and Whitt [17]. Similarly, Abate and Whitt [2] approximated the asymptotic decay rates of the queue length and customer service distribution in form of tail probabilities for a multi-channel queue 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. Data traffic in telecommunications queuing systems possesses long-range dependencies and self-similarity, see Medhi [33]. Recently, measuring, analysis and modeling of self-similar behavior has been one of the main research challenges. In the last couple of years, several studies have been carried out, see Yu et al [49]. More recently, Nakashima [34] worked on the queue length behavior on restricted link under busty self-similar Transmission Control Traffic (TCP). It was shown that the queue length distribution is long tailed. Christian et al [9] derived the diffusion limits for queues under shortest remaining processing service time distribution, see [9]. Another aspect of network traffic behavior has also appeared. In this case, the network traffic behavior is researched from application or data source point of view with focus on statistics of file sizes and inter-arrival times between files, see Park et al [37]. These research works are very important for describing the relation between packet network traffic on lower ISO/OSI layers[22] and data source network traffic on higher layers of ISO/OSI model. Based on the research of the World Wide Web network traffic, Crovella and Lipsky [11] have shown that file sizes of such traffic are best described by Pareto distribution with a unit shape parameter. Also, for the FTP traffic, the shape parameter lies in the set [0.9, 1.1], see Paxon and Floyd [38]. Finally, Nuzman et al. [36] have s hown that interarrival time of transmission control protocol (TCP) connections are self similar in behavior which can be described by Weibull heavily tailed distribution. In 2012, quite a number of researches covering heavy traffic and diffusive behaviors of queuing systems have been published notably; Stralka et al [41] on queue performance in the presence of long- range dependencies and Florens and Jens [10] observed network calculus under no free lunch and concluded that the future still holds good value for network calculus and finally, David et al [12] derived the diffusion approximation to a single server queue in an airport. The phenomenon of self similarity in heavy traffic and diffusive queuing system together with that of long range dependencies have been a subject of recent studies as Medhi [33] indicates. The pioneer of this trend is Kolmogorov and Mandelbrot and Co, see Mandelbrot [31]. Self similar traffic processes possess fractal features in both time and space scales and as Erramilli et al [13] pointed out, there is a considerable scope for future research in this area of fractal queuing theory[23]. This is the most recent behavior in this field and without doubt will be significant in addressing teletraffic issues in the future. Other emerging trends in queue mathematics include the analysis of new queue disciplines. The pioneering work of Krishnamoorthy [25] provides a base for constructing and analyzing these queue disciplines. Similarly, new works on the mathematics of queue schedules are gaining grounds. For a full discussion on this kind of mathematics see Alexander et al. [3], Sivasamy et al. [44] , etc. for details. What this entails is that this kind of mathematics will continue to grow and be relevant for operational purposes. This is generally the purpose of operations research, itself a new branch of mathematics.

CONCLUSIONS

In this article, trends in queuing mathematics are studied. From evolutionary trends to challenges necessitating changing trends were surveyed from inception in 1909 to date to reveal how advances in system engineering or operations research transforms study dimensions and behaviors in terms

[22] OSI means Open Systems Interconnection. It is a standard description on how messages should be transmitted between any two points in a telecom network. Its purpose is to guide product implementors so that their products will consistently work with other products. ISO means International Standards Organization, a traditional model for representing communications protocols.

[23] The fractional queuing theory dimension mostly shows the convergence of queueing processes such as the arrival process to fractional Brownian motion. It is really interesting to see queueing studies in this light.

of problems formulations, technique of analysis, results and modeling. Initially, we looked at areas of challenges necessitating trend transformations. Trends in Methodology were also identified and recent areas of interest arising from system developments for instance, data traffic science was discussed. Finally, the most recent trend of fractal queuing theory mathematics especially in teletraffic and communications engineering was discussed. Finally, an emerging trend that emphases the analysis of queue discipline for heterogeneous server systems was introduced. We conclude this article by emphasizing the need for mixing trends in analyzing complex behaviors with diffusion approximation, fractal queue mathematics at the center. Approaching queue mathematics in a mixed mode will simplify lots of challenges in this interesting field of mathematics of operations research.

ACKNOWLEDGEMENTS

The authors do acknowledge Prof. Sivasamy, R., of the Statistics Department, University of Botswana for sparing time to correct this exposition to this stage. Also is to all the sources of literature used in this piece and to the anonymous reviewers of this work.

REFERENCES

[1] [1] Abate, A. and Whitt, W., 1997. Asymptotics for M/G/1 low-priority waiting-time tail probabilities. Queueing Systems, 25, 173-233.

[2] Abate, A. and Whitt, W., 1994. A Heavy-Traffic Expansion for Asymptotic Decay Rates of Tail Probabilities in Multi-Channel Queues. Operational Research Letters.

[3] Alexander, J. B., Marcus, R., & Cristobal, M. (2014). Flow shop scheduling with heterogeneous workers. European Journal of Operational Research, 237(2), 713–720. SciVerse ScienceDirect, www.elsevier.com/locate/ejor.

[4] Bejan, A.L., 2006. Numerical treatment of the Kendal equation in the analysis of priority queueing systems. Buletinul Academiei De Stiinte, A Republich Moldova Mathematica, 0(0), 1-12.

[5] Boucherie, R.J., and Boxma, O.J., 1996. The workload in the M/G/1 with work removal. Probability in Engineering and Informational Science, 10, 261-277.

[6] Brockmeyer, E., and Halstrom, H.L., The Life of A.K Erlang, page 21.

[7] Boxma, O.J., and Cohen, J.W., 1999. Heavy Traffic Analysis of the G1/G/1 Queue with Heavy Tailed Distributions. Queueing Systems, 33, 177-204.

[8] Boxma, O.J., Deng, Q., and Zwart, A.P., 2002. Waiting time asymtotics for the M/G/2 queue with heterogenous servers. Queueing Systems, 40, 5–31.

[9] Christian, H., Kruk, G.L., and Puha, A.L., 2011. Diffusion Limits for Shortest Remaining Processing Time Queues. Stochastic Systems, 1(1), 1-16.

[10] Ciucu, F., and Schmitt, J., 2012. Perspectives on Network Calculus No Free Lunch, but Still Good Value. SIGCOMM12., Helsinki, Finland.

[11] Crovella, M.E., and Lipsky, L., 1997. Long-lasting transient conditions in simulations with heavy-tailed workloads. Proc. 1997 Winter Simulation Conference Atlanta, GA, USA and Edmonton, Canada.

[12] David, J.L, Kleoniki, V., Tarek, R., and Alexander, B., 2012. A diffusion approximations to a single airport queue. Transport. Res. Part C.

[13] Erramili, A., Narayan, O., and Willinger, W., 1997. Fractal Queueing Models in Frontiers in Queueing:Models and Applications in Science and Engineering. CRC Press, Boca Raton, FL, 245-269.

[14] Federgruen, A., and Tijms, H.C., 1980. Computation of the Stationary Distribution of the Queue Size in M/G/1 with variable Service rate. Journal of Applied Probability, 17, 515–522.

[15] Franken, P., Koonig, D., Arndt, U., and Schmidt, V., 1982. Historical development of Point Procesess as Stochastic Procesess. Queues and Point Procesess.

[16] Gaver, D.P.,Jr., 1968. Diffusion approximations and modes for certain congestion problems. J. Appl. Prob., 5, 607-623.

[17] Glynn, P.W., and Whitt, W., 1995. Heavy-Traffic Extreme-Value Limits for Queues, Operations Research Letters.

[18] Guodong, P., Tareja, R., andWhitt, W., 2007. Martingale proofs for many-server heavy-traffic limits for Markovian queues. Probability Surveys, 4, 193-267.

[19] Hoksad, P., 1978. Approximation for the M/G/m Queue. Journal of Operation Research, 26, 511–523.

[20] Hoksad, P., 1979. On the steady state solution of the M/G/2 queue. Advanced applied probability, 11, 240–255.

[21] I.F.I.P. TC6Working Group 6.3, 1993. Fifth International Conference on Data Communication Systems and their Performance. Raleigh, NC, USA, 26-28.

[22] Iglehart, D.L., 1965. Limit diffusion approximations for the many server queues and the repairmen problem. J. Appl. Prob., 2, 429-441.

[23] Iglehart, D.L., and Whitt, W., 1970a, 1970b. Multiple channel Queue in heavy traffic 1 and 11. J. Appl. Prob., 2, 150-177, and 355-369.

[24] Jain, G., and Sigman, K., 1996. A Pollaczek-Khinchine formular for the M/G/1 queues with Disasters. Journal of Applied Proability, 33, 1191-1200.

[25] Krishnamoorthy, B. (1962). On Poisson queue with two heterogeneous servers. Operations Research, 2(3), 321–330.

[26] Karagiannis, T., Molle, M., Faloutsos, M., and Broido, A., 2004. Department of Computer Science and Engineering, University of California, Riverside and San Diego. ftkarag,mart,michalisg@cs.ucr.edu, broido@caida.org.

[27] Kingman, J.F.C., 1961. The single server queue in heavy traffic. Proc. Camb. Phill. Soc., 57, 902-904.

[28] Kleinrock, L., 1975. Queuing Theory. Queuing Systems,1: (1975), pp.103.

[29] Kollerstrom, J., 1974. Heavy traffic theory for queues with several servers 1. J. Appl. Prob.,11, 544-552.

[30] Leland, W.E., Taqqu, M.S., Willinger, W., and Wilson, D.V., 1994. On the self-similar nature of Ethernet traffic (Extended version). IEEE/ACM Transactions on Networking, 2, 1-15.

[31] Mandelbrot, B.B., 1965. Self-similar error clusters in communication systems and the concept of conditional stationarity, IEEE Trans. Comm. Tech, COM-13, 71-90.

[32] Man-Keung, S., and Tzanakis, C., 2004. History of Mathematics in Classroom Teaching, Appetizer? Main Course? Or Dessert?, Excert from the Mediterranean Journal for Research in Mathematics Education, 3, v-x, 1-2.

[33] Medhi, J., 2003. Stochastic models in queueing theory. Stochastic Models in Queueing Theory,(2003).

[34] Nakashima, T., 2009. Queue length Behaviour on Restricted Link Under Bursty Self-Similar TCP Traffic. Advanced Information Networking and Application Workshops. WAINA 09. Proc. of International Conference, 452-457.

[35] Newell, G.F., 1982. Applications of Queueing Theory. 2nd ed., Chapman and Hall, London.

[36] Nuzman, C., Saniee, I., Sweldens W., and Weiss, A., 2002. A compound model for TCP connection arrivals for LAN and WAN applications. Computer Networks: The International Journal of Computer and Telecommunications Networking, 40(3), 319-337.

[37] Park, K., Kim, G., and Crovella,M.E., 1996. On the Relationship Between File Sizes Transport Protocols, and Self-Similar Network Traffic. International Conference on Network Protocols,171-180.

[38] Paxon, V., and Floyd, S., 1995. Wide area traffic: the failure of Poisson modeling. IEEE/ACM Transactions on Networking, 3(3): 226-244.

[39] Reiser, M., and Kobayashi, H., 1974. Accuracy of the Diffusion Approximation for Some Queueing Systems. IBM. J. RES. DEVELOP., 110-124.

[40] Smith, J.M., 2003. M/G/C/K blocking probability models and system performance. www.computerscience.com powered by science@direct, 52, 237–267.

[41] Strzalka, B., Mazurek, M., and Strzalka, D., 2012. Queue Performance in the Presence of Long- Range Dependencies-an Empirical Study. International Journal of Information Science, 2(4), 47-53.

[42] Sulaiman, S., Daman, O.A., Olabode B.T., and Basimanebotlhe, O.S., 2011. Generalized Steady State Probabilities for the M/G/C Queue with Heterogenous Servers: Implication for Quality Assurance in System Operations. International Journal of Mathematical Science, 3(1),67-73.

[43] Sulaiman, S., Daman, O.A., Olabode B.T., 2011. Customer-Decay Approximations for the M/G/C Queue with Heterogeneous Servers. International Journal of Physical Science, 3(5),125-132.

[44] Sivasamy, R., Daman, O.A. and Sulaiman, S. An M/G/2 Queue subject to a minimum violation of the FCFS queue discipline, European Journal of Operational Research, (2014), Available at http://www.sciencedirect.com/science/article/pii/S0377221714005529;DOI: 10.1016/j.ejor.2014.06.048.

[45] Thomas, A., and Giesl, J., 1999. Applying Rewriting Technique to the Verification of the Erlang Processes. Proceedings of the Annual Conference of the European Association for

[46] Computer Science Logic (CSL '99), Madrid, Spain. Lecture notes in Computer Science 1683,Springer-Verlag, 96-110.

[47] Tijms, H.C., Vaan Hoorn, M.H., and Federgruen, A., 1981. Approximation for the steady state probabilities in the M/G/C queue. Advances in Applied Probability, 13(1): 186-206.

[48] Ward, A., and Peter, W.G., 2003. Properties of the Ornstein-Uhlenbeck Process. Queuing Systems, 44, 109-123. Whitt, W., 1974. Heavy Traffic Limit Theorems for Queues; A survey, Springer-Verlaq, 1-46. Whitt, W., 2000. An overview of Brownian and non-Brownian FCLTs for the single-server queue. Queuing Systems, 36, 39-70.

[49] Yu, Y., Liu, D., Li, J., and Shen, C., 2006. Traffic Identification and Overlay Measurement of Skype. Proc. International Conference on Computational Intelligence and Security, (2), 1043-1048.