Journal of the Korean Physical Society, Vol. 52, February 2008, pp. S171∼S175

Universal Heavy-Tailed Behavior in Waiting Time Distribution in a Queue

Jin Seop KIM, Byungnam KAHNG* and Doochul KIM
FPRD, Department of Physics and Astronomy, Seoul National University, Seoul 151-747

We study the waiting time distribution of queuing models operating in first-in-first-out (FIFO) and priority-based protocols by mapping the dynamics onto random-walk problems. The number of tasks in the queue in the long time limit gives the initial condition of a random walker, and the waiting time of a task is related with its first passage time (FPT). The formalism for the FIFO protocol is established first, successfully reproducing the exponential waiting time distribution, and then the priority-based case by Grinstein and Linsker [1] is reviewed with minor corrections, yielding the power-law waiting time distributions. We also discuss the universality in random systems, comparing the queuing model with other systems from the viewpoint of the ubiquitous exponent 1.5 in the FPT distributions.

PACS numbers: 89.75.-k, 05.40.Fb
Keywords: Queuing theory, Priority-based protocol, Random walk, Power law, Universality

I. INTRODUCTION

One outstanding feature of modern society is massiveness. From mass manufacture and consumption of products to congestion of urban transfer systems to lines of people awaiting services, the excessive requirement for services beyond the limitation of available resources inevitably calls for management of queues. On that account, queuing theory has been developed to support effective allocation of limited resources [2,3].

In spite of extensive studies, however, the term "queue" is often used without any decisive definition. Here, we adopt a rough definition of queue (or queuing system) as a system which consists of objects (*tasks*) awaiting execution by a processing unit in a sequence. The word *sequence* in the definition suggests an important ingredient of a queuing system — the *protocol*. The protocol states the way that the order of task execution is determined. For most queuing systems, the sequence of execution is chosen to be the same as the order of tasks' arrival. In this case, the queuing system is said to have FIFO protocol. All the examples noted above are typical queuing systems with FIFO protocol. Less frequently, there are occasions when the order of execution is taken to be the reverse of the tasks' arrival sequence. One of the data structures used in computer science, called *stack*, works in this manner. The protocol of such queuing systems is called last-in-first-out (LIFO) [4].

Besides these protocols used for practical systems, one can also consider the priority-based protocol. This was recently brought to attention by Barabási and following workers [1,5–7]. They first took notice of the fact that human activity can also be considered as a dynamic process of a complex system, and named this *human dynamics*. A human virtually can do only a single task at a time. That is to say, if he chooses to do one thing, the other tasks should be postponed. The choice of what to do next is determined by the order of importance or emergency of the tasks, *i.e.*, the *priority*. In this regard, task management by a human was modelized into a priority-based queuing system where the human takes the role of the processing unit. The model successfully emulated the ubiquitous phenomenon called *burstiness* [5]. Burstiness is a pattern of activity found in diverse human dynamics, which can be summarized by short periods of high activity followed by long periods of inactivity. The model produces a heavy-tail behavior in the distribution of the tasks' waiting time in the queue, which is consistent with the heterogeneous activity pattern.

In this work we study queuing models operating in such conditions. First, a generic model operating in the priority-based protocol with arbitrary task processing rate is introduced, which generalizes previously studied models. The model is compared to various systems from the viewpoint of universality. To better understand such a model, we pose it onto a random-walk problem [8]. As a preliminary study, the traditional queuing model with FIFO protocol and constant task processing rate is considered first, and then the priority-based model by Barabási and following workers [1,5–7] is reviewed. The

*E-mail: kahng@phya.snu.ac.kr

primary concern is the distribution, $P(\tau)$, of the time τ that a task resides in the queue from its arrival to execution and removal (the waiting time). We find that the waiting time distribution is Poissonian for the FIFO case and follows a power law $P(\tau) \sim \tau^{-\delta}$ for the priority-based case.

II. THE MODEL

We first introduce a generic model as given below. At each discrete time step, (1) m tasks are delivered and added to the queue with probability λ_m, depending on m. Each task has its own priority drawn from the flat distribution $[0, 1]$. (2) Simultaneously, the top n highest-priority tasks are executed and removed from the queue with probability μ_n, which depends on n. All the processes are assumed to occur instantaneously, and the capacity of the queue is assumed to be infinite, so as to stock all the tasks delivered.

The model above can be specified into individual systems of concern by proper choice of the probability distributions λ_m and μ_n. For example, the model in Ref. [1] is reproduced when λ_m and μ_n are given as

$$\lambda_m = \begin{cases} 1-\lambda & (m=0) \\ \lambda & (m=1) \\ 0 & (m>1) \end{cases}, \quad \mu_n = \begin{cases} 1-\mu & (n=0) \\ \mu & (n=1) \\ 0 & (n>1) \end{cases}. \tag{1}$$

With this choice of probability distribution, one task is delivered with probability λ and executed with probability μ in unit time, and λ and μ themselves are the task arrival and execution rates, respectively. The magnitudes of λ and μ lead to different behaviors in the waiting time distribution $P(\tau)$, as we shall see in the next section.

There seem to be three different universality classes according to λ and μ. When $\lambda = \mu = 1$, it was shown that $P(\tau) \sim \tau^{-1}$ first by Barabási [5] *via* a scaling argument and then by Vázquez [7] in an exact analytic solution. The scaling argument of Barabási is introduced in the next section. For the cases of $\mu \leq \lambda < 1$ and $\lambda < \mu < 1$, Grinstein and Linsker [1] showed that $P(\tau) \sim \tau^{-3/2}$ and $P(\tau) \sim \tau^{-5/2}$, respectively.

The dynamics of the $\lambda = \mu = 1$ case is special in that L, the number of tasks in the queue (queue length), is strictly fixed. If we assume FIFO protocol for the queuing dynamics in this case, it is easy to see that $P(\tau) = \delta(\tau - L)$. On the other hand, the dynamics for the priority-based queue is changed dramatically. It is shown that the case $L = 2$ already exhibits the $P(\tau) \sim \tau^{-1}$ distribution. This means that a task with sufficiently low priority is not selected to be executed for a very long time, while other tasks come in and go out repeatedly. Vázquez obtained an exact solution for the case of a stochastic version of the model with $L = 2$: the queuing is started at $t = 0$ by adding L new tasks to the queue. At each time step $t > 0$, the task with higher priority is selected with probability α, or one task is selected randomly between the two with probability $1 - \alpha$. The selected task is executed and removed from the queue, and a new task is added. Then, the waiting time distribution is

$$P(\tau) \sim \frac{1-\alpha^2}{4} \frac{1}{\tau} \exp\left(-\frac{\tau}{\tau_0}\right), \tag{2}$$

where $\tau_0 = 1/\ln\big(2/(1+\alpha)\big)$. In the limit of $\alpha \to 1$ where the deterministic version is reproduced, the power-law behavior $P(\tau) \sim 1/\tau$ becomes dominant. As one can see, the origin of the $\delta = 1$ exponent is not of the kind of $1/f$ noise, nor any previously known dynamics. Thus, additional attention and further study are needed for this new kind of universality class.

On the other hand, the $\delta = 1.5$ exponent for the cases $\lambda > \mu$ and $\lambda = \mu < 1$ is quite ubiquitous behavior. As appears clearly in the next section, the $\delta = 3/2$ exponent originates from the FPT distribution (FPTD) of the random walker. It is robust among the dynamics of random systems that the FPTD is given by a power law, $T(\tau) \sim \tau^{-3/2}$. For example, the FPTD of Lévy's flight random-walk problem asymptotically decays as $\sim \tau^{-3/2}$, irrespective of the Lévy exponent and the initial condition [9, 10]. Moreover, the Sparre Andersen theorem claims that for any discrete-time random-walk process starting at $x_0 \neq 0$ with each step chosen from a continuous, symmetric but otherwise arbitrary distribution, the FPTD asymptotically decays as $\sim \tau^{-3/2}$ [11, 12]. Whilst such results suggest the robustness of the exponent $\delta = 3/2$, whether the waiting time is fully understood solely by FPTD is still in question and should be resolved in future study.

III. QUEUING AND RANDOM WALK

Single task with FIFO.— Let us first consider the case of a single task processing queue operating in FIFO protocol Figure (1). More precisely, a queuing model operating in FIFO protocol whose task arrival and execution rate is given by Eq. (1) is concerned. We are interested in the waiting time distribution, $P(\tau)$, of the model. In a given time step, the queue length, x, can be (a) increased by 1 when a task arrives and no task is executed, (b) decreased by 1 when no task arrives and a task is executed, or (c) the same as for the previous time step, when no task arrives and no task is executed or a task arrives and a task is executed. Notice here that if we regard x as position variable of a random walker and the three ways that the queue length is changed as its transition processes, the dynamics of the queue length is exactly mapped onto a biased random-walk problem. The master equation of such a system for the probability, $Q(x, t)$, that there are x tasks in the queue at time t, or equivalently that the random walker is at position x at time t,

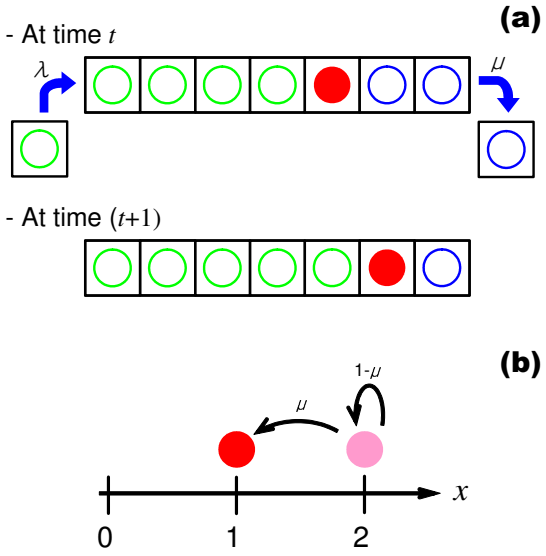

Fig. 1. (Color online) Schematic diagram of the dynamic process in a single-task FIFO queue. In (a), the change in total number of tasks is described. A new task is delivered to the left end, and the oldest task in the queue is executed and removed from the right end. The event that both arrival and execution do occur in a time step, as in the figure, appears with probability $\lambda\mu$. If we focus on the task depicted with filled circles (), the number of preceding tasks is relevant for its execution. Such a situation is shown in (b), from the viewpoint of random walk. The walker is initially at the position of (number of preceding tasks +1) and moves toward the origin with probability μ. The event that the walker reaches the origin corresponds to the execution of the task.

is given by

$$\begin{aligned} Q(x,t+1) &= \big[\lambda(1-\mu)\big]Q(x-1,t) \\ &\quad +\big[\lambda\mu+(1-\lambda)(1-\mu)\big]Q(x,t) \\ &\quad +\big[(1-\lambda)\mu\big]Q(x+1,t), \\ Q(0,t+1) &= \big[1-\lambda(1-\mu)\big]Q(0,t) \\ &\quad +\big[(1-\lambda)\mu\big]Q(1,t) \end{aligned} \tag{3}$$

for $x > 0$ and $x = 0$, respectively. When $\lambda < \mu$, a stationary-state solution exists and, letting $Q(x,t+1) = Q(x,t)$, we obtain

$$\tilde{Q}(x) = \big[(\mu-\lambda)/(1-\lambda)\mu\big]\big[\lambda(1-\mu)/(1-\lambda)\mu\big]^x. \tag{4}$$

Next, consider a task which arrives in the system when the queue is stocked with x_0 tasks in front of it. For the task to be executed, the preceding x_0 tasks should be executed first and then the task itself should also be executed. Consequently, the waiting time of the task is the time required for all the (x_0+1) tasks to be executed. Since the execution of each task is a stochastic process occurring with probability μ, $T(\tau,x_0)$, the probability that the waiting time will be given by τ when there exist x_0 preceding tasks in the queue, has to be found. Meanwhile, from the definition of the model, an execution process occurs simultaneously with the arrival of a task in a given time step, and thus the waiting time should be counted from zero. For example, when the queue is empty, a task can be delivered and then immediately executed and the waiting time of the task is measured to be zero. The dynamic process, again, can be understood by another random-walk problem: let (x_0+1) be the initial position of a random walker and x be the position variable. The execution of a task corresponds to the step (-1) of the random walker, and the execution of that specific task corresponds to the random walker's reaching the origin, $x = 0$, for the first time. In this picture, $T(\tau,x_0)$ is the probability that a unidirectional random walker which moves only to $(-x)$ direction with probability μ passes through the origin at time τ when the initial condition is given by $x = x_0+1$, *i.e.*, FPTD of the random walker. If the random walker should pass the origin at the precise time τ $(\geq x_0)$, it should move left during (x_0+1) time steps and should keep its position during the remaining $(\tau - x_0)$ time steps. (Note the difference between the number of time steps and the time required.) On considering the number of ways to do so, $T(\tau,x_0)$ is obtained as

$$T(\tau,x_0) = \binom{\tau}{x_0}\mu^{x_0+1}(1-\mu)^{\tau-x_0} \quad (\tau \geq x_o \geq 0). \tag{5}$$

The distribution of waiting time, $P(\tau)$, can now be calculated from the two results above. It is easy to see that

$$P(\tau) = \sum_{x=0}^{\infty}\tilde{Q}(x)T(\tau,x), \tag{6}$$

which is valid when $\lambda < \mu$. From Eqs. (4) and (5), we find

$$P(\tau) = \frac{\mu-\lambda}{1-\lambda}\left(\frac{1-\mu}{1-\lambda}\right)^{\tau} \quad (\lambda < \mu). \tag{7}$$

When $\lambda \geq \mu$, the distribution $P(\tau)$ generally does not converge but shows a scattered pattern. On the other hand, if $\lambda = \mu = 1$, it is easy to see that $P(\tau) = \delta(\tau)$. Since the arrival and execution occur deterministically, each task is removed as soon as it is delivered to the queue. Of course, if the queue is initially stocked with L tasks, the distribution will then be $P(\tau) = \delta(\tau - L)$.

Single task with priority.— The model by Barabási and following workers has priority-based protocol and stochastic task arrival and execution rates given by Eq. (1) [1]. In this case, the number of tasks that should be removed from the queue before a given task is to be executed increases if a task with higher priority is delivered to the queue. Neither the total number of tasks in the queue nor the number of pre-existing tasks at the moment of the arrival of the task is relevant any more. Instead, the probability distribution $Q(x,t)$ and the FPTD $T(\tau,x)$ should be redefined to include the factors due to priority. Let $Q(x,p,t)$ be the probability that there are x tasks with priority higher than p in the queue at time

t. Then, $Q(x,p,t)$ satisfies the master equations valid for $x>0$ and $x=0$, respectively:

$$\begin{aligned} Q(x,p,t+1) &= a^\dagger(p)Q(x-1,p,t) \\ &\quad +a(p)Q(x+1,p,t) \\ &\quad +\big[1-a^\dagger(p)-a(p)\big]Q(x,p,t), \\ Q(0,p,t+1) &= \big[1-a^\dagger(p)\big]Q(0,p,t) \\ &\quad +a(p)Q(1,p,t), \end{aligned} \tag{8}$$

where $a(p)=\mu\big[1-\lambda(1-p)\big]$ and $a^\dagger(p)=\lambda(1-p)(1-\mu)$ are the respective probabilities that the number of tasks with priorities larger than p in the queue is decreased or increased by 1 in a given time step. As one can easily see, Eq. (8) is in exactly the same form as Eq. (3) for the FIFO-protocol case, provided that λ in Eq. (3) is replaced by $\lambda(1-p)$. At this stage, it is more convenient to analyze Eq. (8) by using its continuum version, where x and t are treated as continuous variables. The continuum equation is

$$\frac{\partial}{\partial t}Q(x,p,t)=C(p)\frac{\partial^2 Q}{\partial x^2}+D(p)\frac{\partial Q}{\partial x}, \tag{9}$$

where $C(p)\equiv\big[a(p)+a^\dagger(p)\big]/2$ and $D(p)\equiv a(p)-a^\dagger(p)$. The continuum equation with the initial condition $Q(x,p,t=0)=\delta(x-x_0)$ and absorbing boundary condition $Q(x=0,p,t)=0$ has the solution:

$$\begin{aligned} Q(x,p,t) &= \frac{1}{\sqrt{4\pi ct}}\Big[\exp\big[-(x+Dt-x_0)^2/4Ct\big] \\ &\quad -\exp(Dx_0/C)\exp\big[-(x+Dt+x_0)^2/4Ct\big]\Big]. \end{aligned} \tag{10}$$

On the other hand, the steady-state solution of Eq. (8) is given as

$$\begin{aligned} \tilde{Q}(x,p) &\equiv Q(x,p,t\to\infty) \\ &= \big[1-a^\dagger(p)/a(p)\big]\big[a^\dagger(p)/a(p)\big]^x. \end{aligned} \tag{11}$$

In this queuing model, the waiting time of a task with priority p is not solely decided by the initial condition, but is derived from Eq. (8). Let $T(\tau,x_0,p)$ be the waiting time (τ) probability of a given task with priority p which arrived in the queue with x_0 higher-priority tasks having already been in the queue. By using the random-walk picture once again, $T(\tau,x_0,p)$ is the FPTD of a random walker, whose motion is described by the probability distribution function $Q(x,p,t)$ with initial condition given by $Q(x,p,t=0)=\delta(x-x_0)$. Then, from the definition of the FPT and Eq. (10),

$$\begin{aligned} T(\tau,x_0,p) &= -\frac{\partial}{\partial\tau}\int_0^\infty Q(x,p,\tau)dx \\ &= \frac{x_0}{\sqrt{4\pi C}}\tau^{-3/2}\exp\big[-(D\tau-x_0)^2/4C\tau\big]. \end{aligned} \tag{12}$$

The waiting time distribution corresponding to Eq. (6) is written as

$$P(\tau)=\int_0^\infty dx\int_0^1 dp\tilde{Q}(x,p)T(\tau,x,p). \tag{13}$$

On plugging Eq. (11) and Eq. (12) into Eq. (13),

$$P(\tau)=\int_0^\infty dx\int_0^1 dpxH(p)\tau^{-3/2}\exp\big[-\tau J(\tau,x,p)\big]. \tag{14}$$

with $H(p)\equiv(1-a^\dagger/a)/\sqrt{2\pi(a+a^\dagger)}$ and $J(\tau,x,p)\equiv\big[(a-a^\dagger)\tau-x\big]^2/2(a+a^\dagger)\tau^2-x/\tau\log\big(a^\dagger/a\big)$. The dependencies on p of $a(p)$ and $a^\dagger(p)$ are dropped for simplicity. The asymptotic form is determined by the magnitudes of λ and μ, yielding four cases as stated below.

Case (1): $\lambda=\mu=1$. Here, a task is always delivered and the highest-priority task is always executed. Consequently, the number of tasks in the queue remains strictly constant. In this case, the formulae derived above fail and another argument is needed. For the moment, let us extend the model to the case where the highest-priority task is chosen stochastically. The probability of choosing a task with priority p for execution is $\Pi(p)\sim p^\eta$, where η is a parameter that determines the randomness of the model. When $\eta=0$, $\Pi(p)$ is independent of p and the task selection is random. If $\eta=\infty$, however, the highest-priority task is always selected. The probability that a task with priority p is executed at time t is $f(p,t)=\big(1-\Pi(p)\big)^{t-1}\Pi(p)$. The average waiting time of a task with priority p is

$$\tau(p)=\sum_{t=1}^\infty tf(p,t)=\frac{1}{\Pi(p)}\approx\frac{1}{p^\eta}, \tag{15}$$

that is, the higher a task's priority, the shorter the waiting time. If we let the flat distribution of the priority as $\rho(p)$ and note that $\rho(p)dp=P(\tau)d\tau$,

$$P(\tau)=\frac{dp}{d\tau}\approx\tau^{-1-1/\eta}. \tag{16}$$

In the $\eta\to\infty$ limit, $P(\tau)\sim\tau^{-1}$ is obtained [5].

Case (2): $\lambda=\mu<1$. Here, the number of tasks in the queue is constant only in the average sense. The integral over p yields the τ factor from J, which removes τ dependency in the exponential function. As a result, $P(\tau)\sim\tau^{-3/2}$ is obtained.

Case (3): $\lambda<\mu<1$. The execution of tasks is faster than their arrival. The number of tasks in the queue goes to zero frequently. For small x and p, J has a term, $1/\tau_0\equiv(\mu-\lambda)^2/4\mu(1-\lambda)$, independent of x and p. Additionally, both linear and quadratic dependence in x and p come from the expansion of J. The $1/\tau_0$ term produces the exponential factor $e^{-\tau/\tau_0}$ in $P(\tau)$. For $\tau\gg\tau_0$, the linear terms dominate and τ dependence in the integrand is removed, yielding $P(\tau)\sim e^{-\tau/\tau_0}\tau^{-5/2}$. For $1\ll\tau\ll\tau_0$, the quadratic terms dominate and case (1) is recovered, yielding $P(\tau)\approx e^{-\tau/\tau_0}\tau^{-3/2}\sim\tau^{-3/2}$.

Case (4): $\mu<\lambda<1$. Here, the average rate of task arrival is greater than that of task execution. The number of tasks grows as $(\lambda-\mu)t$. A fraction, $(\lambda-\mu)/\lambda$, of arriving tasks fall into a deadlock, never being executed.

However, the other tasks do get executed and follow identical statistics as for case (1), with the asymptotic result $P(\tau) \sim \tau^{-3/2}$.

IV. CONCLUSION

We studied the waiting time distribution, $P(\tau)$, in queuing models from the viewpoint of random-walk problems. We first set up a general queuing model and discussed the special case where the task delivery and execution rates are given by Eq. (1). The case of FIFO protocol with the same task processing rate was studied first, *via* the formalism utilizing the random-walk idea. We obtained the Poissonian waiting time distribution when the system is in a stationary state, $\lambda < \mu$, which is consistent with the known results from various other solutions[2, 3]. Next, the solution for the priority-based queuing system in Refs. [1] and [5] is reproduced with minor corrections in the derivation which do not affect the asymptotic behaviors. The asymptotic behaviors change, depending on the magnitudes of λ and μ. In contrast to the case of FIFO protocol, a solution also exists when the system is not in a stationary state, $\mu \leq \lambda < 1$. The exponent in that case is found to be $\delta = 3/2$, which is consistent with the robust exponents of the FPTD in various systems. When $\lambda < \mu$, $P(\tau)$ decays as $\sim e^{-\tau/\tau_0}\tau^{-5/2}$ in the asymptotic region $\tau \gg \tau_0$, but $\sim e^{-\tau/\tau_0}\tau^{-3/2}$ for $1 \ll \tau \ll \tau_0$. Finally, when $\lambda = \mu = 1$, the exponent is found to be $\delta = 1$, which is not usual for previously known systems. Notice that the Sparre Andersen theorem does not apply here, since the corresponding random walks are asymmetric. Further studies on the origin of this new behavior with exponent $\delta = 1$ and a new class of queuing system which can interpolate the two classes above are required. The case where λ_m follows a power law will be a feasible candidate.

ACKNOWLEDGMENTS

This work was supported by a KOSEF grant funded by MOST (No. R17-2007-073-01001-0).

REFERENCES

[1] G. Grinstein and R. Linsker, Phys. Rev. Lett. **97**, 130201 (2006).
[2] D. R. Cox and W. L. Smith, *Queues* (Methuen, London, 1961).
[3] D. Gross and C. M. Harris, *Fundamentals of Queueing Theory, 3rd ed* (Wiley, New York, 1998).
[4] A. V. Aho, J. D. Ullman and J. E. Hopcroft, *Data Structures and Algorithms* (Addison-Wesley, Boston, 1982).
[5] A.-L. Barabási, Nature **435**, 207 (2005).
[6] A. Vázquez, J. G. Oliveira, Z. Dezsö, K.-I. Goh, I. Kondor, A.-L. Barabási, Phys. Rev. E **73**, 036127 (2006).
[7] A. Vázquez, Phys. Rev. Lett. **95**, 248701 (2005).
[8] See, *e.g.*, S.-H. Lee, J. H. Park, T.-S. Chon and H. K. Pak, J. Korean Phys. Soc. **48**, 236 (2006); H. S. Song and J. M. Kim, J. Korean Phys. Soc. **48**, 245 (2006); S. Lee and Y. Kim, J. Korean Phys. Soc. **48**, 249 (2006).
[9] A. V. Chechkin, V. Y. Gonchar, J. Klafter and R. Metzler, Adv. Chem. Phys. **133(B)**, 439 (2006).
[10] A. V. Chechkin, R. Metzler, V. Y. Gonchar, J. Klafter and L. V. Tanatarov, J. Phys. A: Math. Gen. **36**, L537 (2003).
[11] E. Sparre Andersen, Math. Scand. **1**, 263 (1953).
[12] S. Redner, *A Guide to First-Passage Processes* (Cambridge University Press, New York, 2001).