Modeling human activity in the spirit of Barabasi's queueing systems

Ph. Blanchard
Fakultad für Physik and BiBoS, Universität Bielefeld, D-33615 Bielefeld, Germany

M.-O. Hongler
Sciences et Techniques de l'Ingénieur/IPR/LPM, EPFL, CH-1015 Lausanne, Switzerland

(Received 7 August 2006; published 9 February 2007)

Barabasi has shown that the priority-based scheduling rules in single-stage queuing systems (QS) generate fat tail behavior for the task waiting time distributions (WTD). These fat tails are induced by the waiting times of very low priority tasks that stay unserved almost forever as the task priority indices are "frozen in time" (i.e., a task priority is assigned once for all to each incoming task). Here, we study the new dynamic behavior expected when the priority of each incoming task is time-dependent (i.e., "aging mechanisms" are allowed). For two classes of models, namely a population-type model with an age structure and a QS with deadlines assigned to the incoming tasks, which is operated under the "earliest-deadline-first" policy, we are able to extract analytically some relevant characteristics of the task waiting time distribution. As the aging mechanism ultimately assigns high priority to any long waiting tasks, fat tails in the WTD cannot find their origin in the scheduling rule alone, thus showing a fundamental difference between our approach and Barabasi's class of models.

DOI: 10.1103/PhysRevE.75.026102 PACS number(s): 89.75.Da

I. INTRODUCTION

In recent contributions Barabasi [1] and Vázquez *et al.* [2,3] propose a simplified model of the human activity dynamics. These authors view the human activity as a decision-based queueing system (QS) in which tasks to be executed arrive (randomly) and accumulate before a server $\mathcal{S}$—here $\mathcal{S}$ stands for the processing action of the human operator. The time required to process a task (i.e., *the service time*) is generally drawn from a probability distribution. In addition to the usual features inherent to any QS, each incoming task is endowed with a priority index (PI) expressing the urgency to process the job. In this setting, [1–3] study the dynamics arising when the service policy is not restricted to the usual first-come-first-served (FCFS) rule but follows scheduling policies based on PI's. Under such priority-based scheduling rules, it is shown that the timing of the tasks follows fat tail probability distributions (i.e., the activity of the server exhibits bursts separated by long idle periods). This "burst" character has to be contrasted with the ubiquitous Poisson behavior, which arises when tasks are executed according to FCFS or to purely random order scheduling rules. In this general context, we shall distinguish between two types of dynamics:

(i) *Service policies based on fixed* (*i.e., frozen in time*) *priority indices*. This case, which is considered in [1–3], assumes that the value a of the PI is fixed once for all. Therefore, very low priority jobs are likely to never be served. To circumvent this difficulty [1–3], introduce an ad-hoc random mechanism $0 \leq p \leq 1$ in terms of which the limit $p \to 1$ corresponds to a deterministic scheduling strictly based on the PI's while in the other limit $p \to 0$ the purely random scheduling is working. In this setting, the *waiting time distribution* (WTD) of the tasks before service exhibits asymptotically a fat tail behavior. The main point of Barabasi's contribution is to show that *PI-based scheduling rules alone can generate fat tails in the WTD of unprocessed jobs*.

(ii) *Service policies based on time-dependent priority indices*. Here the priority index is *time-dependent*. This typically models situations in which *the urgency to process a task increases with time* and $a(t)$ will hence be represented by increasing time functions. Clearly, scheduling rules based on such time-dependent PI do offer new specific dynamical features. They are directly relevant in several contexts, such as the following

(a) *Flexible manufacturing systems with limited resources*. Here a single server is conceived to process different types of jobs but only a single type can be produced at a given time t (i.e., this is the limited resource constraint). The basic problem is therefore to schedule, in real time, the production for matching the random demand arrivals for each type of item. This can be optimally achieved by using time-dependent priority indices (*Gittins' indices*), which specify in real time the type of production to be engaged [4]. Problems of this type belong to a wider class referred to as the *multi-armed bandit problems* in operations research.

(b) *Tasks with deadlines*. This situation can be idealized by a queueing system in which each incoming item has a deadline before which it definitely must be processed [5–7]. In this case, to be discussed later in the present paper, we can explicitly derive the lead-time profile of the waiting jobs obtained under several scheduling rules, including the (optimal) time-dependent priority rule known as the *earliest-deadline-first* policy.

(c) *Waiting time-dependent feedback queueing systems*. In queueing networks, priority indices based on the waiting times can be used to schedule the routing through the network. For networks with loops, such scheduling policies are able to generate generically stable oscillations of the populations contained in the waiting room of the queues [8].

In the context of QS, the waiting time probability distribution (WTD) (i.e., the time the tasks spend in the queue before being processed) is a central quantity characterizing the dynamics. It strongly depends on the arrival and service stochastic processes—in particular to the distributions of the *inter-arrival* and *service* time intervals. The first moments of these distributions enable us to define the traffic load $\rho := \frac{\lambda}{\mu} \geq 0$ (i.e., the ratio between the mean service time $\frac{1}{\mu}$ and the mean arrival time $\frac{1}{\lambda}$) and clearly the stability of elementary QS is ensured when $0 \leq \rho < 1$. Focusing on the WTD [1–3] emphasized that heavy tails in the WTD can have several origins, three of which are listed below:

(i) The *heavy traffic load of the server*, which induces large "bursty" fluctuations in both the WTD and the busy period (BP) of the QS. For QS with feedback control driving the dynamics to heavy traffic loads, this allows us to generate self-organized critical (SOC) dynamics [9], and the resulting fat tails distribution exhibits a power-law decay with exponent $-3/2$.

(ii) The presence of *fat tails in the service time distribution* produces fat tails of the WTD, a property that here is independent of the scheduling rule [10]. For completeness, we summarize these recent results in Appendix A.

(iii) Priority index scheduling rules as discussed in [1–3].

This paper focuses on case (iii), but contrary to [1–3], we shall consider here the dynamics in the presence of *age-dependent priority indices*. As might have been expected, these aging mechanisms generate new behaviors that will be explicitly discussed for two classes of models.

II. SCHEDULING BASED ON TIME-DEPENDENT PRIORITY INDICES

The most naive approach to discuss the dynamics of QS with scheduling based on *time-dependent priority indices* is to think of a population model in which the members suffer aging mechanisms that ultimately will kill them. Hence, we may consider the population of a city in which members are either born in the city or immigrate into it at a certain age and finally die in the city. Assuming that the death probability *depends on each individual age*, the study of the age structure of the population exhibits some of the salient features of our original QS. This is the class of models to be discussed in Sec. II A. Later in Sec. II B, we shall return to the original model of Barabási and consider a simple QS in which each task waiting to be processed carries a deadline (playing the role of a PI), and as time flows these deadlines steadily reduce—this implies a *time dependence of the PI*. At each given time, the scheduling of the tasks follows the "earliest-deadline-first" (EDF) policy, and given a queue length configuration, we shall discuss the lead-time (lead time is deadline - current time) profile of the tasks waiting to be served.

A. Task population dynamics with time-dependent priority indices

Consider a population of tasks waiting to be processed by $\mathcal{S}$ and satisfying the following assumptions:

(i) An inflow of new tasks steadily enters into the queueing system. Each task is endowed with a priority index (PI) a, which indicates its degree of urgency to be processed. In general, the tasks are heterogenous as the PI are different. During the time interval $[t, t+\Delta t]$, the number of incoming jobs exhibiting an initial PI in the interval $[a, a+\Delta a]$ is expressed by $G(a,t)\Delta t \Delta a$.

(ii) Contrary to the situations discussed in [1], an "aging" mechanism affects directly the urgency to process a given task. In other words, the priority index a is not frozen in time, but $a=a(t)$ monotonically increases with time t. For an infinitesimal time increase Δt, in the simplest case we shall have $a(t+\Delta t)=a(t)+\Delta t$. Here we slightly generalize this and allow inhomogeneous aging rates $p(a)>0$ such that $a(t+\Delta t)=a(t)+p(a)\Delta t$.

(iii) The scheduling rule depends on the PI of the tasks in the queue, and we will focus on the natural policy "*process the highest PI first.*"

(iv) At time t, $M(a,t)$ counts the number of waiting tasks with priority index a. Hence $M(a,t)\Delta a$ is the number with PI $\in [a, a+\Delta a]$ from which it follows that the total workload facing the human server $\mathcal{S}$ at time t is given by

$$L(t) = \int_0^\infty M(a,t) da. \quad (1)$$

(v) In the time interval $[t, t+\Delta t]$, the server $\mathcal{S}$ processes tasks with an a-dependent rate $\mu(a)\Delta t$. Typically $\mu(a)$ could be a monotonically increasing function of a. The service rate $\mu(a)$ depending explicitly on the PI a plays, therefore, an effective role of service discipline.

The previous elementary hypotheses imply

$$M(a + p(a)\Delta t, t + \Delta t)\Delta a \approx M(a,t)\Delta a - \mu(a)M(a,t)\Delta a \Delta t + G(a,t)\Delta t \Delta a.$$

Dividing by $\Delta a \Delta t$, we obtain, in the limits $\Delta a \to 0$ and $\Delta t \to 0$, the linear equation

$$\frac{\partial}{\partial t}M(a,t) + p(a)\frac{\partial}{\partial a}M(a,t) + \mu(a)M(a,t) = G(a,t). \quad (2)$$

It is worthwhile to remark that the dynamics given by Eq. (2) is closely related to the famous McKendrick's age structured population dynamics [11].

Assuming stationarity for the incoming flow of tasks [i.e., $G(a,t)=G_s(a)$], from the linearity of Eq. (2) we obtain its stationary solution,

$$M(a) = \pi(a)\left[C + \int_0^a \frac{G_s(z)}{p(z)\pi(z)} dz\right], \quad (3)$$

where

$$\pi(z) = \exp\left\{-\int^z \frac{\mu(y)}{p(y)} dy\right\}, \quad (4)$$

with an integration constant $C < \infty$ still to be determined. Assume that the PI attached to the incoming jobs does not exceed a limiting value T, namely

$$G(a,t) = \mathbb{I}(a < T)\hat{G}(a,t) \quad \Rightarrow \quad G_s(a) = \mathbb{I}(a < T)\hat{G}_s(a), \tag{5}$$

where $\mathbb{I}(a<T)$ is the indicator function. In other words, Eq. (5) indicates that the newcoming jobs do not exhibit arbitrarily high PI's.

This enables us to define

$$\Psi(T) := \int_0^\infty \frac{G_s(z)}{p(z)\pi(z)} dz = \int_0^T \frac{\hat{G}_s(z)}{p(z)\pi(z)} dz \tag{6}$$

and Eq. (3) reads

$$M(a) = \begin{cases} \pi(a)\left[C + \displaystyle\int_0^a \frac{\hat{G}_s(z)}{p(z)\pi(z)} dz \right] & \text{if } a \leq T \\ \pi(a)[C + \psi(T)] & \text{if } a > T. \end{cases} \tag{7}$$

The asymptotic behavior of $M(a)$ for $a \to \infty$ depends only on $\pi(a)$ (the square bracket terms are bounded by constants) and therefore Eqs. (4) and (7) imply

$$M(a) \approx \pi(a) \approx \begin{cases} e^{\frac{-k}{q}a^q} & \text{when } \dfrac{\mu(a)}{p(a)} \propto k a^{q-1} \quad \text{with } q > 0, \\ \dfrac{1}{a^k} & \text{when } \dfrac{\mu(a)}{p(a)} \propto \dfrac{k}{a}. \end{cases} \tag{8}$$

Equation (8) exhibits the following alternatives:

(a) For $q<0$ in Eq. (8), the integral $\int_0^\infty M(z)dz$ does not exist. In this case, an ever-growing population of tasks accumulates in front of the server and the queueing process is exploding.

(b) For $q>0$, a stationary regime exists and in this case the constant $C<\infty$ in Eq. (7) can be determined by solving

$$\int_0^\infty G_s(z)dz = \int_0^\infty M(z)\mu(z)dz, \tag{9}$$

which expresses a global balance between the stationary incoming and outgoing flows of tasks.

(c) For $q=0$, which implies that $\frac{\mu(a)}{p(a)} \propto \frac{k}{a}$, Eq. (8) *produces a power law with exponent k for the distribution M*(a), *the number of waiting tasks with PI a* in the system. For $T<\infty$ and $a \to \infty$, the fat tail of $M(a)$ takes into account the long waiting tasks, i.e., those having waited more than $a-T$ inside the system before being served. In the limiting case, for which $\mu(a)=\mu=\text{const}$ and $p(a)=a$ (i.e., aging directly proportional to time), which leads to $q=0$ in Eq. (8), the density $M(a)$ coincides with the WTD for $a \to \infty$.

This population model shares several features with Barabási's model, namely (a) when a stationary regime exists, the function $\hat{G}_s(a)$, which here plays the role of the initial PI distribution in [1–3], does not affect the tail behavior given by Eq. (8); (b) the scheduling rule here is implicitly governed by the service rate $\mu(a)$, which itself depends on time as the PI $a=a(t)$ are time-dependent. Note that $\mu(a)$ directly influences the asymptotic behavior of Eq. (8). In particular, for case (c), the tail exponent explicitly depends on $\mu(a)$.

Besides the similarities, we now emphasize the important differences between the present population model and the model discussed in [1–3]:

(a) The service is not restricted to a single task at a given time (i.e., the service resource is not limited). Indeed $\mu(a)$ describes an average flow of service and hence several tasks can be processed simultaneously (in the city population model, the service corresponds to death and several individuals may die simultaneously).

(b) While the fat tail in [1–3] is entirely due to the scheduling rule and therefore occurs even for QS far from traffic saturation, this is not the case in the population model. Indeed in this last case, fat tails are due to heavy traffic loads for which the incoming flow of tasks nearly saturates the server capacity [this is implied by $q=0$ in Eq. (8)]—for lower traffic loads arising when $q>0$, the fat tail in Eq. (8) disappears.

B. Stochastic dynamics: Real-time queueing dynamics

In this section, we will use the results of the real-time queueing theory (RTQS), pioneered in [5], to explore situations in which the incoming jobs have a deadline—this problem was already suggested in [1]. Based on [5–7] and [12], let us first recall the basic hypotheses and the relevant results of RTQS's. Consider a general single-server QS with arrival and service being described by independent renewal processes with average $\frac{1}{\lambda}$ and $\frac{1}{\mu}$, respectively, and *finite variances* for both renewal processes. Each incoming task arrives with a random hard time relative deadline $\mathcal{D}$ drawn from a PDF $G(x)$ with a density $g(x)$,

$$\text{Prob}\{0 \leq \mathcal{D} \leq x\} = G(x),$$

with average $\langle \mathcal{D} \rangle$,

$$\langle \mathcal{D} \rangle = \int_0^\infty [1 - G(x)]dx = \int_0^\infty x g(x) dx.$$

At a given time t, we define the lead time $\mathcal{L}$ by

$$\mathcal{L} = \mathcal{D} - t. \tag{10}$$

Assume now that the lead time $\mathcal{L}$ plays the role of a priority index and the service is delivered by using the *earliest-deadline-first* (EDF) rule with preemption (i.e., the server always processes the job with the shortest lead time $\mathcal{L}$). Preemption implies that whenever an incoming job exhibits a shorter $\mathcal{L}$ than the one currently in service, this incoming job is processed before (i.e., *preempts*) the currently engaged task, which postpones service. The EDF rule directly corresponds to the deterministic policy (i.e., $p=0, \gamma=0$ in the original Barabási's contribution [1]).

At a given time, one can define a probability distribution corresponding to the *lead-time profile* (LTP), $F(x) := \text{Prob}\{-\infty \leq \mathcal{L} \leq x\}$, of the jobs waiting in the QS. The LTP specifies the repartition of tasks having a given $\mathcal{L}$ at time t. Knowing the queueing population Q at a given time,

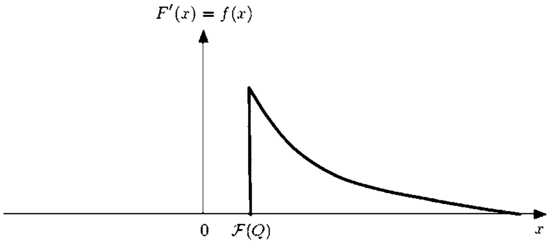

FIG. 1. Qualitative sketch of the probability density of the lead time profile $f(x)=\frac{dF(x)}{dx}$ when $\mathcal{F}(Q)>0$.

it is shown in [7] that for heavy traffic regimes, the LTP can, in a first-order approximation scheme, be expressed by a simple analytical form. Indeed, following [7], let us define a frontier $\hat{\mathcal{F}}(Q)>0$ as the unique solution of the equation

$$\frac{Q}{\lambda} = \int_{\hat{\mathcal{F}}(Q)}^{\infty} [1 - G(x)]dx, \quad (x \in [l, \infty) \subset \mathbb{R}^{+}). \tag{11}$$

Let now define the frontier $\mathcal{F}(Q)$ as

$$\mathcal{F}(Q) = \begin{cases} \hat{\mathcal{F}}(Q) & \text{when } Q \le Q^{*}, \\ \left(\langle \mathcal{D} \rangle - \dfrac{Q}{\lambda}\right) \le l & \text{when } Q > Q^{*}, \end{cases} \tag{12}$$

with Q^{*} being defined by $\hat{\mathcal{F}}(Q^{*})=l$, with l defined in Eq. (11). Two regimes can occur.

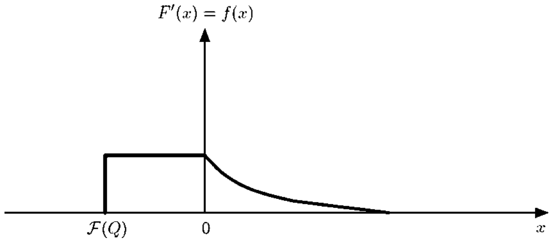

FIG. 2. Qualitative sketch of the probability density of the lead time profile $f(x)=\frac{dF(x)}{dx}$ when $\mathcal{F}(Q)<0$.

(a) *Jobs served before deadline*. When $\langle \mathcal{D} \rangle > \frac{Q}{\lambda}$ and therefore $\hat{\mathcal{F}}(Q)=\mathcal{F}(Q)$, the LTP cumulative distribution $F(x)$ takes the form (see Fig. 1)

$$F(x) = \begin{cases} 0 & \text{when } x < \mathcal{F}(Q), \\ 1 - \dfrac{\lambda}{Q}\left(\displaystyle\int_{x}^{\infty} [1 - G(\eta)]d\eta\right) & \text{when } 0 < \mathcal{F}(Q) \le x. \end{cases} \tag{13}$$

(b) *Jobs served after deadline*. When $\mathcal{F}(Q)=\left(\langle \mathcal{D} \rangle - \frac{Q}{\lambda}\right) \le 0$, the LTP cumulative distribution $F(x)$ takes the form (see Fig. 2)

$$F(x) = \begin{cases} 0 & \text{when } x < \langle \mathcal{D} \rangle - \dfrac{Q}{\lambda} < 0, \\ \dfrac{1 - \dfrac{\lambda \langle \mathcal{D} \rangle}{Q}}{l - \langle \mathcal{D} \rangle + \dfrac{Q}{\lambda}}\left[x - \langle \mathcal{D} \rangle + \dfrac{Q}{\lambda}\right] & \text{when } \langle \mathcal{D} \rangle - \dfrac{Q}{\lambda} \le x < l, \\ 1 - \dfrac{\lambda}{Q}\left\{\displaystyle\int_{x}^{\infty} [1 - G(\eta)]d\eta\right\} & \text{when } l \le x. \end{cases} \tag{14}$$

Remark. The alternative regimes given by Eqs. (13) and (14) can be heuristically understood by invoking the *Little law*, which connects the average queue length $\langle Q \rangle$ with the average waiting time $\langle W \rangle$ [13],

$$\langle Q \rangle = \lambda \langle W \rangle, \tag{15}$$

a result independent of the scheduling policy. In view of Eqs. (11) and (15), one obviously suspects that the *LTP* strongly depends on the sign of the difference $\langle \mathcal{D} \rangle - \frac{\langle Q \rangle}{\lambda} = \langle \mathcal{D} \rangle - \langle W \rangle$. Intuitively, when $\langle W \rangle$ exceeds $\langle D \rangle$, we expect, in the average, that processed jobs will be delivered too late, and conversely when $\langle W \rangle < \langle D \rangle$ jobs will be processed before their deadline. While the above heuristic arguments is strictly valid only for the averages, [5–7] shows that in heavy traffic regimes, it holds also for the LTP given in Eqs. (13) and (14).

Assuming that the arriving tasks have positive deadlines, the LTP given by Eqs. (13) and (14) imply the following. (a) If the left-hand support of the LTP is negative, then a job entering into service is already late [case of Eq. (14)]; see Fig. 2. (b) If the left-hand support of the LTP is positive then a job enters into service with a positive lead time [case of Eq. (13)]; see Fig. 1. Accordingly, it is likely that the tasks will be completed before the deadline expires. (c) The critical value $Q^{*}=\lambda\langle \mathcal{D} \rangle$, for which $\mathcal{F}(Q^{*})=l$, corresponds to a queue length for which customers are likely to become late. Choosing $Q=Q^*$, we cannot expect lateness to disappear completely, but for $Q<Q^*$ lateness will be strongly reduced, a behavior clearly confirmed by simulation experiments [7] and [12]. (d) For deadline distributions $G(x)$ with fat tails, it follows from Eqs. (13) and (14) that the LTP also exhibits a fat tail.

1. "First come first served" (FCFS) scheduling policies

Choosing the deadline probability density as $g(x)=\delta(x)$ (i.e., zero deadline), the EDF scheduling policy directly coincides with the FCFS rule. For this case we have $Q^*=0$, and Eq. (12) implies

$$\mathcal{F}(Q)=\begin{cases}\hat{\mathcal{F}}(Q)=0 & \text{when } Q\le 0,\\ -\dfrac{Q}{\lambda} & \text{when } Q>0.\end{cases} \tag{16}$$

Hence the LTP density given by Eq. (13) is merely the uniform probability density $U[-\frac{Q}{\lambda},0]$ ($[-\frac{Q}{\lambda},0]$ being its support). This expresses the fact that in the heavy traffic regime $\rho=\lambda/\mu\approx 1$, the waiting time behaves as $Q\times(\frac{1}{\mu})\approx Q\times(\frac{1}{\lambda})$ leading to a LPT linearly growing with Q. For general $G(x)$, the LTP associated with a FCFS scheduling rule will be given by the convolution of the deadline distribution $G(x)$ with the uniform distribution $U[-\frac{Q}{\lambda},0]$. Indeed, adding the task deadlines with the time spent in the queue, we recover the task lead time. Therefore, in the heavy traffic regime and for a given queue length Q, one explicitly knows the LTP's for both the EDF and the FCFS scheduling policies, thus enabling us to explicitly appreciate their respective characteristics. In particular, using Eqs. (13) and (14), one can conclude that for a given queue length Q, with the FCFS scheduling rule and the associated LTP $F(x)$ being theconvolution of $G(x)$ with the $U[-\frac{Q}{\lambda},0]$, we obtain

$$F(x)=\begin{cases}0 & \text{when } x<-\dfrac{Q}{\lambda},\\ \dfrac{\lambda}{Q}\displaystyle\int_{-(Q/\lambda)}^{x}\left[G\left(\xi+\frac{Q}{\lambda}\right)\right]d\xi & \text{when } -\dfrac{Q}{\lambda}\le x<0,\\ \kappa+\dfrac{\lambda}{Q}\left\{\displaystyle\int_{0}^{x}\left[G\left(\xi+\frac{Q}{\lambda}\right)-G(\xi)\right]d\xi\right\} & \text{when } x\ge 0,\end{cases} \tag{17}$$

where the constant κ is given by

$$\kappa=\frac{\lambda}{Q}\int_{-(Q/\lambda)}^{0}\left[G\left(\xi+\frac{Q}{\lambda}\right)\right]d\xi.$$

Equation (17) allows us to emphasize the following features:

(i) When the left-hand support of the deadline distribution $G(x)$ is larger than $\frac{Q}{\lambda}$, the left boundary of the support of $F(x)$ is larger than 0 and therefore the jobs experience no delay when entering into service.

(ii) If the left-hand support of $G(x)$ is smaller than $\frac{Q}{\lambda}$, then it may happen that the LTP exhibits a negative left-hand support under the FCFS policy and a positive left-hand support under the EDF scheduling rule. Hence in this last situation, the FCFS policy would deliver tasks late while the EDF tasks will be processed in due time. This fact explicitly confirms the intuition that EDF is indeed an efficient policy. It has been shown that the EDF scheduling rule is optimal for minimizing the number of jobs processed after the deadline [14].

(iii) If $G(x)$ exhibits a fat tail for $x\to\infty$, that will also be the case for the LTP regardless of the scheduling rule used. This can be directly verified from Eq. (17) by studying the LTP density $f(x)=\frac{dF(x)}{dx}$ for $x\to\infty$. We have

$$f(x)=\frac{\lambda}{Q}\left[G\left(x+\frac{Q}{\lambda}\right)-G(x)\right]\quad\text{for } x\to\infty,$$

which for $G(x)\sim 1-x^{-q}$ and for $\frac{Q}{\lambda}<$const implies that

$$f(x)\sim x^{-(q+1)}\quad\text{for } x\to\infty. \tag{18}$$

Hence, the LTP inherits the fat tail property of $G(x)$ even when using the optimal EDF scheduling rule—a fully explicit illustration involving the Pareto probability distribution is given in Appendix B.

The results obtained for the LTP enable us to get asymptotic properties of the waiting time distribution (WTD). Indeed, assume a heavy traffic regime with the EDF scheduling policy. Assume further that for a given queue length, some jobs are served too late (i.e., the left boundary of the LTP is negative). As under the EDF rule the more urgent jobs are always served first, the waiting times of the last jobs in the queue necessarily exceed their deadlines. Therefore, when the deadline distribution exhibits a fat tail, so will the WTD distribution. Note that while the EDF policy decreases, compared with the FCFS rule, the number of jobs served after their deadline, it cannot get rid of the fat tail of the WTD, which is due to the fat tail of $G(x)$. This result is fundamentally different from the situation that is valid for the frozen in time PI models discussed in [1–3], where the fat tail behavior does not depend on $G(x)$ itself. This can be heuristically understood as, in [1–3], the fat tail is mainly due to the low priority jobs, which, as no aging mechanism occurs, are likely to never be served. Note that in [1–3], stable queueing models (i.e., those for which the traffic $\rho < 1$) and fat tails of the WTD disappear under a FCFS scheduling rule. Indeed without priority scheduling, the WTD always follows an exponential asymptotic decaying behavior. In the presence of time-dependent PI, all tasks do finally acquire a high priority and this aging mechanism precludes the formation of a fat tail solely due to the scheduling rule. Accordingly, in the presence of aging PI, the appearance of WTD with fat tails is due to $G(x)$.

2. Some remarks about human resources and work organization

The results for the LTP derived in the preceding section can be directly measured on actual queueing systems (QS). Consider the queue content of a single-stage QS. Assume that at a given time, Q is the observed queue content, and at this instant take a snapshot of the lead time associated with each waiting item and construct the associated LTP (i.e., the histogram of the observed lead times). In heavy traffic regimes (i.e., typically $0.95 \leq \rho < 1$ leading to stationary average queue lengths $\frac{\rho}{1-\rho}$) and under the EDF scheduling policy, the LTP will approximately be given by Eqs. (13) and (14). Actual simulation experiments are reported in [5–7] and [12], where an excellent agreement between measured data and theory is observed.

From the human activity viewpoint, the explicit expressions of the LTP obtained both for the FIFO and EDS policies show clearly that organizing the work scheduling is extremely important. As an illustration, consider a situation in which the deadline distribution $G(x)$ follows an exponential law:

$$G(x) = 1 - e^{-\alpha x} \Rightarrow \langle \mathcal{D} \rangle = \frac{1}{\alpha}. \tag{19}$$

For this situation, we compare two different organization policies.

a. EDF scheduling policy. Introducing Eq. (19) into Eq. (11), we obtain $\hat{\mathcal{F}}(Q) = \frac{-1}{\alpha}\log\left(\frac{\alpha Q}{\lambda}\right)$ and with $l=0$ we have $Q^* = \frac{\lambda}{\alpha}$. This enables us to write Eq. (12) as

$$\mathcal{F}(Q) = \begin{cases} \dfrac{-1}{\alpha}\log\left(\dfrac{\alpha Q}{\lambda}\right) & \text{when } Q \leq \dfrac{\lambda}{\alpha}, \\ \dfrac{1}{\alpha} - \dfrac{Q}{\lambda} & \text{when } Q > \dfrac{\lambda}{\alpha}. \end{cases} \tag{20}$$

When $\mathcal{F}(Q) > 0$ [i.e., the upper line in Eq. (20)], Eq. (13) implies

$$F(x) = \begin{cases} 0 & \text{when } x \leq \mathcal{F}(Q), \\ 1 - \dfrac{\lambda}{\alpha Q} e^{-\alpha x} & \text{when } x > \mathcal{F}(Q) \end{cases} \tag{21}$$

and when $\mathcal{F}(Q) < 0$ [i.e., the lower line in Eq. (20)], Eq. (14) yields

$$F(x) = \begin{cases} 0 & \text{when } x < \mathcal{F}(Q), \\ 1 - \dfrac{\lambda}{\alpha Q}(1 - \alpha x) & \text{when } \mathcal{F}(Q) \leq x < 0, \\ 1 - \dfrac{\lambda}{\alpha Q} e^{-\alpha x} & \text{when } 0 \leq x. \end{cases} \tag{22}$$

b. FIFO scheduling policy. With $G(x)$ given by Eq. (19), the result given in Eq. (17) reads

$$F(x) = \begin{cases} 0 & \text{when } x < -\dfrac{Q}{\lambda}, \\ 1 + \dfrac{\lambda x}{Q} - \dfrac{\lambda}{\alpha Q}\left[1 - e^{-\alpha\left(x + \frac{Q}{\lambda}\right)}\right] & \text{when } -\dfrac{Q}{\lambda} \leq x < 0, \\ 1 - \dfrac{\lambda\left[1 - e^{-\frac{\alpha Q}{\lambda}}\right]}{\alpha Q} e^{-\alpha x} & \text{when } x \geq 0. \end{cases} \tag{23}$$

Comparing Eqs. (21) and (23), we conclude that in a heavy traffic regime, for a given work load Q, the use of EDF enables us to process tasks in due time with a high probability while the naive FIFO policy generates large delays. Specifically, when $Q < \frac{\lambda}{\alpha}$, the EDF policy guarantees that most jobs enter into service before the deadline [see Eq. (20)] and will therefore be served before deadline, with a high probability. On the contrary, the FIFO policy result given in Eq. (23) [i.e., obtained for $x=0$ in the last line of Eq. (23)] shows that a proportion of $1 - (\lambda[1 - e^{-\alpha Q/\lambda}]/\alpha Q)$ jobs enter the service with delays and will therefore be late.

As far as human resources are concerned, this simple model enables us to quantify the importance of adopting performant scheduling policies to respond to the "burnout"-generating challenge: *deliver more in less time with fewer resources*. Along the same lines, one of the key rules to avoid burnout is to *learn to say no* to new incoming tasks if the queue length exceeds a threshold. In our modeling framework, the critical threshold does depend closely on the level Q^*, above which lately served tasks (and hence complaints) are unavoidable.

III. CONCLUSION AND SUMMARY

There are several possibilities to discuss analytically the scheduling of tasks in QS with time-dependent priority indices and to infer the existence of fat tails for the asymptotic behavior of the resulting WTD. In this paper, we propose two distinct models in which an explicit analysis can be developed. Our first model is directly inspired by the study of age classes in population dynamics for which the mortality rate increases with the age of the individuals. In this context, identifying the service of the QS with the death of an individual, this dynamics is closely related to the scheduling based on PI, the indices here being the age of the individuals, and the immigration with different ages plays the role of incoming tasks with different priorities. For this class of dynamics, it is straightforward to show that fat tails of the WTD can develop on the onset of stability of the population model. As in the original Barabási model, the tail behavior of the WTD does not depend on the nature of the PI but only on the scheduling rule (corresponding in the population model to the mortality rate). In our second modeling ansatz, which is closer to Barabási's original idea, we consider a classic QS in which the scheduling rule is based on the deadlines attached to each incoming task. As time flows, the deadlines reduce and hence the waiting tasks acquire a higher priority to be processed. In the heavy traffic limit, i.e., for regimes where the law of large numbers dominates, it is possible to derive analytically the lead-time profile (lead time equals deadline minus the time elapsed in queueing) of the waiting tasks and from this to get information on the asymptotic behavior of the associated WTD. In this case, and contrary to the conclusions made in [1–3], the scheduling rule alone cannot generate fat tails in the WTD. Fat tail in [1–3] are due to low-priority jobs that are likely to never be served. This possibility disappears if time-dependent PI are introduced as, due to aging, initially low priority tasks do acquire, with time, high priorities and hence will not stay unprocessed forever. This precludes the formation of fat tails in the WTD. We finally observe that, in this second class of models, the only possibility to generate fat tails is to feed the QS with task deadlines drawn from a fat tail distribution.

ACKNOWLEDGMENTS

It is our pleasure to acknowledge support from the Portuguese Fundaçao para Ciencia e Technologica (Bolsa FEDER/POCTI-SFA-1-219) and the Fonds National Suisse pour la Recherche Scientifique (FNRS-200021-109191/1). M.O.H. thanks O. Gallay and R. Filliger for numerous discussions.

APPENDIX A: WAITING TIME DISTRIBUTIONS FOR QS WITH FAT TAIL SERVICE TIMES

Let us reproduce here a result recently obtained by [10].

Theorem 1. Assume that the (random) service time in an $M/G/1$ QS is drawn from a PDF with a regularly varying tail at infinity with index $\nu \in (-1,-2)$ [regularly varying with index $\nu \in (-1,-2) \Rightarrow$ fat tail with index $\nu \in (-1,-2)$]. For this range of asymptotic behaviors of the PDF, the first moment β of the service exists. Assume further that the service is delivered according to a random order discipline. Then the waiting time distribution W_{ROS} exhibits a fat tail with index $(1-\nu) \in (-1,0)$ and, more precisely, we can write

$$\mathrm{Prob}(W_{\mathrm{ROS}} > x) \propto C \frac{\rho}{1-\rho} \frac{h(\rho,\nu)}{\beta\Gamma(2-\nu)} x^{1-\nu}\mathcal{L}(x), \qquad \text{(A1)}$$

where $\rho<1$ is the traffic intensity, β is the average service time, $\mathcal{L}(x)$ is a slowly varying function, and

$$h(\rho,\nu) := \int_0^1 f(u,\rho,\nu)du,$$

with

$$f(u,\rho,\nu) := \frac{\rho}{1-\rho}\left(\frac{\rho u}{1-\rho}\right)^{\nu-1}(1-u)^{1/(1-\rho)} + \left(1+\frac{\rho u}{1-\rho}\right)^{\nu}(1-u)^{[1/(1-\rho)]-1}.$$

The fat tail behavior given in Eq. (A1) is therefore entirely inherited from the fat tail behavior of the service and is not affected by any reduction of the traffic intensity ρ. Note also that a change of the scheduling rule cannot get rid of this fat tail behavior. This point can be explicitly observed in [13,15], where it is shown that for the previous $M/G/1$ QS with a random order service (ROS) service discipline, one can prove that

$$\mathrm{Prob}(W_{\mathrm{ROS}} > x) \propto h(\rho,\nu)\mathrm{Prob}(W_{\mathrm{FCFS}} > x) \quad \text{for } x \to \infty, \qquad \text{(A2)}$$

from which we directly observe that *the fat tail in the asymptotic behavior is not altered by a change of the scheduling rule*.

Note finally that for the $M/M/1$ QS (i.e., exponential service distributions and hence no fat tail), [16] shows that the random order service scheduling rule yield

$$\mathrm{Prob}(W_{\mathrm{ROS}} > x) \propto C_\rho x^{-5/6} e^{-\gamma x - \delta x^{1/3}} \quad \text{for } x \to \infty, \qquad \text{(A3)}$$

with

$$C(\rho) = 2^{2/3}3^{-1/2}\pi^{5/6}\rho^{17/12}\frac{1+\rho^{1/2}}{[1-\rho^{1/2}]^3}\exp\left\{\frac{1+\rho^{1/2}}{1-\rho^{1/2}}\right\},$$

$$\gamma = (\rho^{-1/2}-1)^2 \quad \text{and} \quad \delta = 3\left[\frac{\pi}{2}\right]^{2/3}\rho^{-1/6},$$

which has to be compared with the FCFS scheduling discipline, which for the same $M/M/1$ QS reads [13]

$$\mathrm{Prob}(W_{\mathrm{FCFS}} > x) = \frac{1}{\beta}(1-\rho)e^{-(1/\beta)(1-\rho)x}. \qquad \text{(A4)}$$

While the detailed behaviors given by Eqs. (A3) and (A4) clearly differ, they both share, in accord with [1], an exponential decay.

APPENDIX B: DEADLINE DRAWN FROM PARETO DISTRIBUTION

Here, we focus on

$$G(x) = \begin{cases} 0 & \text{when } x < B, \\ 1-\left(\frac{B}{x}\right)^{(\omega-1)} & \text{when } x \geq B, \quad \omega > 1, \end{cases} \qquad \text{(B1)}$$

which has no moment of order $\geq\omega-1$. For $\omega>2$, we have $\langle\mathcal{D}\rangle=[\frac{\omega-1}{\omega-2}]B$. Using Eq. (12) with $l=B$, which implies $Q^* = \frac{\lambda B}{\omega-2}$, we have

$$\mathcal{F}(Q)=\begin{cases}\hat{\mathcal{F}}=B\left(\dfrac{B\lambda}{Q(\omega-2)}\right)^{[1/(\omega-2)]} & \text{when } Q\leq\dfrac{\lambda B}{\omega-2},\\[2ex] \left[\dfrac{\omega-1}{\omega-2}\right]B-\dfrac{Q}{\lambda} & \text{when } Q>\dfrac{\lambda B}{\omega-2}.\end{cases}\tag{B2}$$

Using Eqs. (13) and (14), the LTP distribution reads

$$Q\geq\frac{\lambda B}{\omega-2}\Rightarrow F(x)$$
$$=\begin{cases}0 & \text{when } x\leq\mathcal{F}(Q),\\[1ex] 1-\dfrac{\lambda}{Q}\left(\dfrac{\omega-1}{\omega-2}B-x\right) & \text{when } \mathcal{F}(Q)\leq x<B,\\[2ex] 1-\dfrac{B\lambda}{Q(\omega-2)}\left(\dfrac{B}{x}\right)^{\omega-2} & \text{when } x\geq B,\end{cases}\tag{B3}$$

$$Q<\frac{\lambda B}{\omega-2}\Rightarrow F(x)$$
$$=\begin{cases}0 & \text{when } x\leq\mathcal{F}(Q),\\[1ex] 1-\dfrac{B\lambda}{Q(\omega-2)}\left(\dfrac{B}{x}\right)^{\omega-2} & \text{when } x>\mathcal{F}(Q).\end{cases}\tag{B4}$$

Equations (B3) and (B4) exhibit a fat tail with power $\omega-2$. Note that Eq. (B4) implies that for $\omega>2$ and for $\frac{Q}{\lambda}<\frac{B}{(\omega-2)}$, the EDF scheduling policy part of the tasks enters into the service before the due date expired. Finally, note also that for $\omega\leq2$, no moments exist for the deadline distribution, hence the theory [7] cannot be applied directly. We conjecture that for these regimes, no scheduling rule will be able to deliver tasks in due time.

[1] A.-L. Barabási, Nature (London) **435**, 207 (2005).

[2] A. Vázquez, Phys. Rev. Lett. **95**, 248701 (2005).

[3] A. Vázquez, J. G. Oliveira, Z. Dezsö, K.-I. Goh, I. Kondor, and A.-L. Barabási, Phys. Rev. E **73**, 036127 (2006).

[4] F. Dusonchet and M.-O. Hongler, IEEE Trans. Rob. Autom. **19**, 997 (2003).

[5] J. Lehozcky, *Proceedings of the 17th IEEE Real-Time Systems Symposium, New York*, (IEEE, Piscataway, NJ, 1996), p. 186.

[6] J. Lehozcky, Perform. Eval. **25**, 158 (1997).

[7] B. Doytchinov, J. Lehozcky, and S. Shreve, Ann. Appl. Probab. **11**, 332 (2001).

[8] R. Filliger and M.-O. Hongler, Europhys. Lett. **70**, 285 (2005).

[9] Ph. Blanchard and M.-O. Hongler, Phys. Lett. A **323**, 63 (2004).

[10] O. J. Boxma, S. G. Foss, J. M. Lasgouttes, and R. Nùñez Queija, Queueing Syst. **46**, 35 (2004).

[11] F. Brauer and C. Castillo-Chávez, *Mathematical Models in Population Biology and Epidemiology*, Text in Applied Mathematics Vol. 40 (Springer, Berlin, 2001).

[12] R. O. Baldwin, N. J. Davis, J. E. Kobza, and S. F. Mikdiff, Queueing Syst. **35**, 1 (2000).

[13] J. Cohen, J. Appl. Probab. **10**, 343 (1973).

[14] S. Panwar, D. Townsley, and J. K. Wolf, J. ACM **35**, 832 (1988).

[15] A. G. Pakes, J. Appl. Probab. **12**, 555 (1975).

[16] L. Flatto, Ann. Probab. **7**, 382 (1997).