RC23960 (W0605-105) May 17, 2006
Physics

IBM Research Report

Biased Diffusion and Universality in Model Queues

G. Grinstein, R. Linsker
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Biased Diffusion and Universality in Model Queues

G. Grinstein and R. Linsker
IBM T.J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598
(Dated: May 16, 2006)

We study the structure and robustness of universality classes for queueing, deriving analytic results for priority-based models with continuous-valued priorities. By mapping one model onto the problem of biased diffusion, we show that its distribution of waiting times, $P(\tau)$, decreases for large times τ as $P(\tau) \sim \tau^{-3/2}$ or as $P(\tau) \sim \tau^{-5/2} \exp(-\tau/\tau_0)$ in different parameter regimes. In a second model, introducing a cost for switching between different classes of tasks substantially changes the asymptotic behavior of $P(\tau)$.

PACS numbers: 02.50.-r, 02.50.Ey, 89.20.-a

I. Introduction

The management of queues is a pervasive feature of modern life, from the operation of hospital emergency rooms, to highway congestion, to computer jobs awaiting an available processor. Most of the intensive study of the statistics of model queues[1–3] has been devoted to situations wherein the distribution, $P(\tau)$, of waiting times, τ, in the queue falls off rapidly – typically exponentially – with time. A series of interesting recent papers[4–6] by Barabási and co-workers has focused attention on waiting-time distributions with longer tails, by analyzing data on such activities as the exchange of letters and e-mail messages, web browsing, and library use. These activities were reported to have waiting time distributions with heavy tails consistent with power laws, $P(\tau) \sim \tau^{-\alpha}$, over some range of τ. The reported values of α were close to 3/2 for written correspondence and close to 1 for the other activities. Refs. [4–6] also described two queueing models devised to try to account for this behavior. The first, a fixed-length queue, has tails with $\alpha = 1$[4, 7], while the second, a variable-length queue, was reported numerically to have α near 3/2.

A full explanation of heavy tails in the waiting-time statistics of human activities will obviously require both larger data sets[8] and progressively more realistic models. Here we are concerned with the latter issue, studying power-laws and universality in queueing in two different models. First, we map the variable-length, continuous-priority-queue model studied in refs. [4–6] onto the familar model of biased diffusion[9]. In this way we derive analytic, asymptotic expressions for $P(\tau)$, thereby explaining the origin of the numerical result $P(\tau) \sim \tau^{-3/2}$ for $\lambda \geq \mu$, where λ and μ are the respective rates of task arrival and of task execution. We show further that for $\lambda < \mu$, $P(\tau) \sim e^{-\tau/\tau_0}\tau^{-5/2}$ for asymptotically large τ, i.e., for $\tau >> \tau_0$, where τ_0 is a characteristic time that diverges as $1/(\mu - \lambda)^2$ as $\lambda \to \mu$[5].

Second, we generalize the fixed-length model queue[4] to contain tasks of two or more different classes, with a "start-up cost" for switching from one class to another, thereby representing schematically the management of jobs of different types – e-mail messages to be answered and household chores to be performed, for example. We show numerically that this seemingly modest modification produces a substantial change in behavior: For intermediate waiting times τ and moderate values of the switching cost, $P(\tau)$ still exhibits power-law behavior, but with an exponent α of approximately 3/2, rather than 1. Beyond a characteristic time that grows with the length L of the fixed-length queue, the decay of $P(\tau)$ becomes exponential. We explain these results, and why they differ from those for single-class models.

II. Models

Model A: One-Class, Continuous-Priority Queue: We start by considering Model A, defined as follows. At each discrete time step: (1) With probability λ, a new task with priority x ($0 \leq x \leq 1$), chosen from the probability distribution $\rho(x)$, arrives in the queue. (2) Then, with probability μ, the highest priority task in the queue is executed. The execution is assumed to occur instantaneously. The case $\lambda = \mu = 1$ is analyzed in ref. [4]; numerical results and scaling arguments for other (λ, μ) are given in refs. [4]-[7].

For any μ and λ, the transformation from the original priority variable x to a new variable $y \equiv \int_0^x \rho(z)dz$ with a uniform distribution, $\tilde{\rho}(y) = 1$, over the interval $0 \leq y \leq 1$, satisfies $\tilde{\rho}(y)dy = \rho(x)dx$, and so produces a model equivalent to the original[7]. Thus we take $\rho(x) = 1$ here.

The case $\lambda = \mu = 1$ of model A is special in that the queue length remains strictly constant, and the distribution of priorities of tasks in the queue approaches $\delta(x)$ in the long-time limit. The highest value, x_M, of x in the queue after each complete time step is a nonincreasing function of time that approaches 0 as the number of time steps, T, becomes large. Thus, as $T \to \infty$, the probability of newly arrived tasks having priorities $x > x_M$ and so being executed immediately approaches unity. The result[4], $P(\tau) \sim 1/\tau$, applies to the remaining tasks that have $x < x_M$ on arrival[10].

We now analyze Model A for λ and μ less than unity. Eq. (1) expresses the overall probability, $P(\tau)$, that a given task sits in the queue for a time τ before being ex-

ecuted, in terms of two quantities[3]: (a) the probability, $G(n,x,\tau)$, that a given task of priority x, which arrives in the queue at time $t=0$ with exactly n items of higher priority (i.e., larger x) already in the queue, gets executed at precisely time $t=\tau$; and (b) the probability, $\tilde{Q}(n,x)$, of there being exactly n items in the queue with priority greater than x, once a steady state has been achieved.

$$P(\tau) = \Sigma_{n=0}^{\infty} \int_0^1 dx\, \tilde{Q}(n,x)\, G(n,x,\tau)\ . \qquad (1)$$

Let $Q(m,x,t)$ be the probability that at time t there are precisely m tasks with priority greater than x in the queue. Then $Q(m,x,t)$ satisfies the discrete master equations, valid for $m>0$ and $m=0$, respectively:

$$\begin{aligned} Q(m,x,t+1) &= a(x)Q(m+1,x,t) + b(x)Q(m-1,x,t) \\ &\quad +(1-a(x)-b(x))Q(m,x,t)\ ; \\ Q(0,x,t+1) &= a(x)Q(1,x,t) + (1-b(x))Q(0,x,t); \qquad (2) \end{aligned}$$

here $a(x)=\mu(1-q(x))$ and $b(x)=q(x)(1-\mu)$ are the respective probabilities of the number of tasks with priorities $>x$ in the queue decreasing and increasing by 1 in a given time step, where $q(x)=\lambda\int_x^1 \rho(z)dz = \lambda(1-x)$ is the probability of a task with priority $>x$ arriving in the queue during phase (1) of a given time step.

We first consider $\lambda<\mu<1$. In steady state, $Q(m,x,t+1)=Q(m,x,t)$, which yields the normalized steady-state distribution

$$\tilde{Q}(m,x) = [1-b(x)/a(x)][b(x)/a(x)]^m. \qquad (3)$$

As λ approaches μ from below, $b(0)$ approaches $a(0)$, and the distribution $\tilde{Q}(m,0)$ becomes uniform in m. The mean number of tasks in the queue in steady state, $\langle m(x=0)\rangle$, thus diverges as $1/(\mu-\lambda)$. In this strict sense, the steady-state distribution is ill-defined for $\lambda=\mu$[1–3]. However, the mean number of tasks having priorities greater than x, $\langle m(x)\rangle$, remains finite for any $x>0$ when $\lambda=\mu$, behaving as $1/x$ as $x\to 0$. Owing to this fact, the queue does have well-defined steady-state properties, as we shall see.

Next we compute $G(n,x,t)$ of Eq. (1) by deriving an estimate for $Q(m,x,t)$, starting from the initial condition in which exactly n tasks have priority exceeding x in the queue at $t=0$. This is most easily accomplished through study of the continuum limit of Eq. (2) in both the variable m [11] and the time, t, viz.:

$$\partial Q(y,x,t)/\partial t = c(x)\partial^2 Q/\partial y^2 + d(x)\partial Q/\partial y. \qquad (4)$$

Here the discrete number of tasks m has been replaced by the continuum variable y, t is now a continuous time variable, $c(x)\equiv ra(x)$ and $d(x)\equiv r[a(x)-b(x)]$, where r is an arbitrary time constant that sets the time scale for the biased-diffusion equation (4). Eq. (4), with the initial condition $Q(y,x,t=0)=\delta(y-n)$ corresponding to there being n tasks in the queue initially, and the absorbing boundary condition $Q(y=0,x,t)=0$, has the solution[9] (with the x-dependence of a, b, c, and d suppressed):

$$Q(y,x,t) = \frac{1}{\sqrt{4\pi ct}}[e^{(-(y+dt-n)^2/4ct} - e^{dn/c}e^{(-(y+dt+n)^2/4ct}]. \qquad (5)$$

The probability of there being a positive number of tasks having priority greater than x in the queue, at time t, is $R(n,x,t)=\int_0^\infty Q(y,x,t)dy$. The probability that the queue of tasks with priorities greater than x empties at precisely time t (i.e., the first-passage probability) is $G(n,x,t)=-\partial R/\partial t$, yielding, from (5)[9]:

$$G(n,x,t) = \frac{n}{\sqrt{4\pi c}\, t^{3/2}} e^{-(dt-n)^2/4ct}. \qquad (6)$$

Given expressions (6) and (3), Eq. (1) for $P(\tau)$ is

$$P(\tau) = \Sigma_{n=0}^{\infty} \int_0^1 dx\, g(n,x,\tau)\, e^{[-\frac{(d\tau-n)^2}{4c\tau} + n\log(b/a)]}\ , \qquad (7)$$

where $g(n,x,t)\equiv \frac{n}{2\sqrt{\pi c}\, t^{3/2}}(1-b/a)$. Rescaling n via $n=l\tau$, where $l=0,1/\tau,2/\tau,...$ yields

$$P(\tau) = \tau^{-1/2}\Sigma_{l=0,1/\tau,2/\tau,...} \int_0^1 dx\, h(l,x)\, e^{-\tau j(l,x)}\ , \qquad (8)$$

with $h\equiv l(1-b/a)/2\sqrt{\pi c}$ and $j\equiv -l\ \log(b/a) + \frac{(d-l)^2}{4c}$. For large τ, the right side of (8) is dominated by the smallest value of $j(l,x)$, which can be shown to occur at $l=x=0$. The behavior of $P(\tau)$ for asymptotically large τ is derived by expanding the functions h and j around this point and extending the integral over x to infinity. We consider three cases, distinguished by the relative arrival and execution rates of tasks.

Case (1): $\lambda=\mu<1$. Here $h(l,x)$ and $j(l,x)$ are quadratic in l and x for small l and x, and $j(l,x)$ is never negative. The rescaling $(x,l)=\tau^{-1/2}(\tilde{x},\tilde{l})$ then removes the τ dependence from the integrand, whereupon, for large τ, the sum over $\tilde{l}$ can be replaced by an integral, $\Sigma_{\tilde{l}=0,\tau^{-1/2},2\tau^{-1/2},...} \to \tau^{1/2}\int_0^\infty d\tilde{l}$. This leaves $P(\tau)$ proportional to $\tau^{-3/2}$ times a convergent double integral over $\tilde{x}$ and $\tilde{l}$; i.e., $P(\tau)\sim\tau^{-\alpha}$ with $\alpha=3/2$.

Case (2): $\lambda<\mu<1$. Again, $j(l,x)$ is never negative. For small l and x, $j(l,x)$ has a term, $1/\tau_0\equiv r(\mu-\lambda)^2/4\mu(1-\lambda)$, independent of l and x, and terms both linear and quadratic in l and x; $h(l,x)$ has terms of O(l) and of O(lx). 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 the rescaling $(x,l)=\tau^{-1}(\tilde{x},\tilde{l})$ removes the τ dependence from the integrand, 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 $P(\tau)\approx e^{-\tau/\tau_0}\tau^{-3/2}\approx\tau^{-3/2}$. These behaviors are confirmed numerically in Fig. 1.

Case (3): $\mu<\lambda<1$. Here tasks arrive faster than they are executed on average, producing a queue that grows in time t like $(\lambda-\mu)t$. A fraction $x^\star\equiv(\lambda-\mu)/\lambda$ of

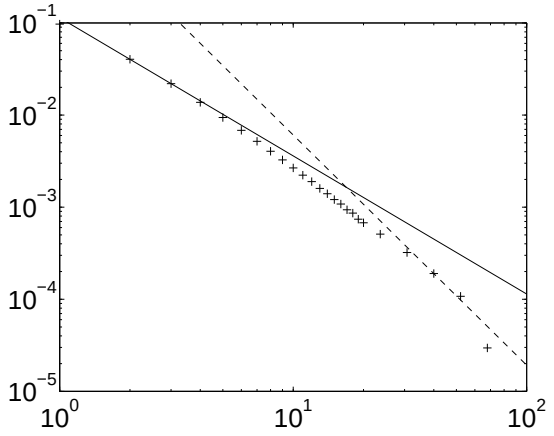

FIG. 1: $P(\tau)\exp(\tau/\tau_0)$ vs. τ, for Model A with $T = 10^8$, $\lambda = 0.6$, $\mu = 0.9$, and $\tau_0 = 7.33$. This τ_0 was obtained as the best-fit constant value of $A(\tau,\alpha) \equiv -\tau/[\log(P(\tau)) + \alpha\log\tau]$ at large τ for $\alpha = 5/2$; $A(\tau,\alpha)$ is not constant for $\alpha = 3/2$. Guide lines are $\propto \tau^{-3/2}$ and $\tau^{-5/2}$, confirming that $P(\tau) \approx \tau^{-3/2}\exp(-\tau/\tau_0)$ for $\tau \ll \tau_0$, and that $P(\tau) \sim \tau^{-5/2}\exp(-\tau/\tau_0)$ for $\tau \gg \tau_0$.

arriving tasks waits in the queue forever without getting executed. The analysis of the tasks that do get executed, however, can be cast in a form identical to Case (1). To see this, note that for $x > x^\star$, the rate, $\lambda(1-x)$, of arrival of tasks with priority greater than x is less than the rate, μ, at which such tasks are executed. Thus for $x > x^\star$, our earlier analysis of the master equation (2) remains valid. In particular, the steady-state distribution of Eq. (3) holds for $x > x^\star$; the mean number of particles having priority greater than x, $\langle m(x)\rangle = \Sigma_{m=0}^{\infty} m\,\tilde{Q}(m,x)$, diverges as $x \to x^\star$ from above. Changing the variable x to the new variable y defined by $x = x^\star + (1-x^\star)y$ maps the range $x^\star < x < 1$ onto the range $0 < y < 1$, and transforms the quantities $a(x)$ and $b(x)$ in Eq. (2) to $\tilde{a}(y) = \tilde{\mu}(1-\tilde{q}(y))$ and $\tilde{b}(y) = \tilde{q}(y)(1-\tilde{\mu})$, where $\tilde{q}(y) \equiv \tilde{\lambda}(1-y)$, $\tilde{\mu} \equiv \mu$, and $\tilde{\lambda} \equiv \lambda(1-x^\star)$. Thus, apart from the change from $(\mu,\ \lambda)$ to $(\tilde{\mu},\ \tilde{\lambda})$, the functions $\tilde{a}(y)$ and $\tilde{b}(y)$ are the same as the original functions $a(y)$ and $b(y)$, respectively. Given the definition of $x^\star$, moreover, $\tilde{\mu} = \tilde{\lambda}$, so the problem maps precisely onto Case (1) above, with the asymptotic result $P(\tau) \sim \tau^{-3/2}$. Concerning those tasks having $x < x^\star$: Because the total number of tasks in the queue with $x > x^\star$ grows without bound as time progresses, the probability of executing a task with $x < x^\star$ approaches 0 in the long-time limit. Asymptotically, all such tasks thus remain in the queue in perpetuity.

Model B: Multi-Class, Fixed-Length Queue with Switching Cost: We now modify Model A in the fixed-length-queue limit $\lambda = \mu = 1$, by assigning to each task a class label as well as a priority. We consider the case of two classes. At time $t = 0$ the queue contains L tasks and one of the classes (say class I) is arbitrarily designated the 'active' class. At each subsequent time step, a task of either class (chosen with probability $\frac{1}{2}$), having priority $0 < x < 1$ chosen from a uniform distribution, is added to the queue. If the highest priority of all the tasks of the inactive class exceeds that of the active class by at least a fixed amount c (or if the active class has no remaining tasks in the queue), then the inactive queue becomes active and the active queue inactive. The highest priority task of the active queue is then executed. Taking $c > 0$ simulates the inertia, or start-up cost, of shifting from one type of activity to another. Model A with $\lambda = \mu = 1$ is recovered for $c = 0$.

Figure 2a shows simulation results for the waiting-time distribution $P(\tau)$ of tasks for Model B as a function of τ, on a log-log plot, for queue lengths L from 2 to 1000. For $\tau \geq 10$ but not too large, $P(\tau)$ decays as $\tau^{-\alpha}$, with α close to 3/2. Beyond a characteristic time that increases with L, however, $P(\tau)$ falls off from the power-law curve and decreases much more rapidly, consistent with exponential rather than algebraic decay. Figure 2b is a semilog plot of $P(\tau)\tau^{3/2}$ vs. τ/L^β for $\beta = 2.25$. The roughly linear behavior (apart from large-τ statistical fluctuations) shows that $P(\tau) \sim \tau^{-3/2}\exp[-\tau/\tau_0(L)]$. The fact that the curves for widely differing L approximately collapse onto one another for $\beta = 2.25$ shows that $\tau_0(L)$ is approximately proportional to L^β for this β. In contrast, if the Fig. 1b curves are replotted using $\beta \leq 2.125$ or ≥ 2.375 (not shown), they remain approximately linear but have quite different slopes from one another. [The characteristic time at which $P(\tau)$ departs from the power-law curve also scales approximately as L^β with the same β.]

To understand this behavior heuristically, first note that even though the queue length L is fixed, the model generates, for $c > 0$, a distribution of tasks in which the highest priority value of the tasks in the queue does *not* tend to zero as the number of time steps T becomes large. In this respect, Model B is more like Model A with a variable-length queue (i.e., λ and $\mu < 1$), than it is like the fixed-length version of Model A with $\lambda = \mu = 1$. The inactive class is constantly replenished by the addition of tasks that cannot be executed until that class becomes active; thus the sizes of the individual classes fluctuate while L stays fixed. The rules for the addition and removal of tasks of a given class (I, say), look very similar to the rules for the single-class Model A, with $\lambda = \mu = \frac{1}{2}$. This is because the probability of executing a class-I task is $(1 + x_M + c)/2$ or $(1 - x_M - c)/2$ when the active class is class I or class II, respectively, where x_M is the highest priority of all tasks in the queue. Thus the average probability of executing a class-I task is $\frac{1}{2}$, as is the probability of adding a class-I task to the queue. One would therefore expect $P(\tau) \sim \tau^{-3/2}$ for modest τ values.

This argument, however, ignores the fixed length, L, of the queue, which allows all tasks of a given class to be eliminated in at most $L + 1$ time steps with nonzero probability. To see this, consider the example where all L tasks in the queue belong to class I, which is the active

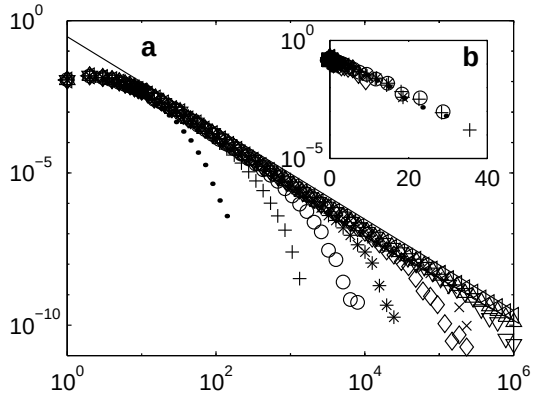

FIG. 2: (a) Waiting-time distribution $P(\tau)$ vs. τ (log-log plot), from Model B numerical simulations with 'switching cost' 0.2, and $L = 2, 5, 10, 20, 50, 100, 200, 500, 1000$ (denoted by: $\cdot, +, \bigcirc, *, \Diamond, \times, \nabla, \triangle$, and left-pointing triangle). Guide line is $\propto \tau^{-3/2}$. Total number of time steps $T = 10^6$ for $L \leq 100$ and 4×10^6 for $L \geq 200$. Bins for computing $P(\tau)$ are of unit width for $1 < \tau \leq 10$, and proportional to τ (log-binned) for $\tau > 10$. (b) Semilog plot of $P(\tau)\tau^{3/2}$ vs. τ/L^β for $\beta = 2.25$, showing power-law times exponential falloff of $P(\tau)$ at large τ and scaling of the coefficient of τ in the exponential as $L^{-\beta}$ (see text). Bins are as in (a). For clarity, only points having both $\tau > 10$ and $P(\tau) > 10^{-9}$ are plotted in (b).

class. If on each of the subsequent L time steps, the task added to the system belongs to class II and has priority $x < c$, then class I will remain the active class, and all the class-I tasks will be executed. Thus $(c/2)^L$ is a loose lower bound for the probability of the class-I queue being eliminated in L time steps, in this case. Similar arguments for arbitrary initial conditions of the queue show that the class-I tasks can always be eliminated in $L+1$ time steps with probability $q_L = (c/2)^{L+1}$. Thus on average the class-I queue will take no longer than $(L+1)/q_L$ time steps to empty. One concludes that the waiting-time distribution must decay – presumably exponentially – with a characteristic time τ_L bounded above by $(L+1)/q_L$. This is consistent with the numerical results of Fig. 2.

Thus the essential difference between Model A with $\lambda = \mu < 1$ and Model B is that the number of tasks in the queue of Model A performs a random walk and thus can increase without bound. It is these large excursions of the queue length that allow[5] the long waiting times necessary to produce the asymptotic power-law behavior of $P(\tau)$ in Model A with $\lambda = \mu < 1$, rather than more rapid, e.g., exponential, decay.

III. Discussion

By mapping the continuous-priority queueing model (here called Model A) onto the biased-diffusion model, we have shown analytically that the exponent α characterizing the asymptotic decay of $P(\tau)$ has the value 3/2 for $\lambda \geq \mu$, in agreement with existing numerical results[4–6]. For $\lambda < \mu$, our result, $P(\tau) \sim e^{-\tau/\tau_0}\tau^{-3/2}$ for $\tau \lesssim \tau_0$, is consistent with our simulation results (Fig. 1) and those of ref. [5], while the result $P(\tau) \sim e^{-\tau/\tau_0}\tau^{-5/2}$ for $\tau \gg \tau_0$ is also consistent with Fig. 1.

Our Model B attempts to make the fixed-length-queue model[4] more realistic, by incorporating schematically the cost of switching execution between different classes of tasks. We showed that this produces a notable change in behavior, making the fixed-length queue model look much like the variable-length Model A, with $P(\tau)$ decreasing as $\tau^{-3/2}$ for $1 \ll \tau \lesssim \tau_L$, where $\tau_L \sim L^\beta$. Empirically, we found $\beta \approx 2.25$ for the range of L's studied. This is close to, but distinct from, two – the value one might expect asymptotically[5], since the diffusion time over distance L behaves like L^2. For $\tau \gtrsim \tau_L$, $P(\tau)$ decays exponentially, since with nonzero probability the queue empties of all tasks of a particular class.

The marked difference in behavior between the fixed-length queue models with one and two classes of tasks suggests that the study of simple models like the ones treated here may be useful in identifying efficient prioritizing strategies. More generally, the analysis of such models is essential for understanding the intriguing phenomenology[1]-[6] of human activities that involve waiting times.

We thank Sid Redner for providing helpful information about first-passage probabilities.

[1] D. R. Cox and W. L. Smith, *Queues* (Methuen & Co., London, 1961).

[2] D. Gross and C.M. Harris, *Fundamentals of Queueing Theory*, 3rd ed. (Wiley, New York, 1998).

[3] J. Abate and W. Whitt, Queueing Systems **25**, 173 (1997), and references therein.

[4] A.-L. Barabási, Nature **207**, 435 (2005), and Supplementary Online Information.

[5] A. Vázquez et al., Phys. Rev. E**73**, 036127 (2006).

[6] J.G. Oliveira and A.-L. Barabási, Nature **437**, 1251 (2005).

[7] A. Vázquez, Phys. Rev. Lett. **95**, 248701 (2005), showed that the $P(\tau) \sim 1/\tau$ result also holds, for $\tau < \tau_p \equiv 1/\ln(\frac{2}{1+p})$, in a stochastic version of the fixed-length-queue model, wherein the highest priority task is executed with probability p.

[8] Some researchers have argued that $P(\tau)$ for the e-mail data is actually log-normal. See comment by D.B. Stouffer, R.D. Malmgren, and L.A.N. Amaral, physics/0510216; and reply by A.-L. Barabási, K.-I. Goh, and A. Vázquez, physics/0511186.

[9] S. Redner, *A Guide to First-Passage Processes* (Cambridge University Press, NY, 2001).

[10] See the exchange in ref. [8], and also ref. [7] for further comments on this point.

[11] The asymptotic results for $P(\tau)$ given here can be derived without approximating the discrete variable m in Eq. (2) by the continuous variable y of Eq. (4). We present the

continuum calculation here because of its relative simplicity.