Available online at www.sciencedirect.com
SCIENCE DIRECT®
PHYSICS LETTERS A
Physics Letters A 323 (2004) 63–66
www.elsevier.com/locate/pla

Self-organization of critical behavior in controlled general queueing models ☆

Ph. Blanchard [a], M.-O. Hongler [b,*,1]

[a] Universität Bielefeld, Fakultät für Physik, D-33619 Bielefeld, Germany
[b] STI/Institut de Production et Robotique/LPM, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland

Received 16 July 2003; received in revised form 16 December 2003; accepted 9 January 2004

Communicated by A.P. Fordy

Abstract

We consider general queueing models of the (G/G/1) type with service times controlled by the busy period. For feedback control mechanisms driving the system to very high traffic load, it is shown the busy period probability density exhibits a generic $-\frac{3}{2}$ power law which is a typical mean field behavior of SOC models.

PACS: 02.50.Hb; 05.90.+m

Keywords: Self-organized criticality; Queueing systems

1. Introduction

The avalanches occurring in extremal models exhibiting self-organized criticality (SOC) (the paradigmatic example being the Bak–Sneppen (BS) model) are defined from the value of a global minimal number $f_{\min}(s)$ as a function of time s. Then for any value of the auxiliary parameter f_0, an f_0 avalanche of size S is defined as a sequence of $S-1$ successive events with $f_{\min}(s) < f_0$ confined between two events having $f_{\min}(s) \geqslant f_0$. Accordingly, an avalanche is a stochastic process in which the numbers $f_i < f_0$ play the role of active particles that are randomly created or annihilated. Hence, the avalanche ends when there are no particles left in the system. Clearly, the time axis is divided into a series of avalanches. In the BS model, it exists a critical value f_c for which the creation of particles is marginally balanced by their annihilation and avalanches of all sizes can happen.

Consider now a queueing systems (QS) which consists of random arrival of customers to a server. Each customer requires a random service time and the system is equipped with a waiting room of unlimited capacity. Here also the time axis can be divided by successive cycle times (CT), a CT being the sum of a busy period (BP) (i.e., the time interval separating two successive instants where the server is starving) and an idle period (i.e., periods during which the server is starving). In QS, the role of the f_0 pa-

☆ Partially supported by FCT (Portugal) at CCM University of Madeira.

* Corresponding author.
E-mail address: max.hongler@epfl.ch (M.-O. Hongler).

1 Partially supported by the Fonds National Suisse pour la Recherche.

0375-9601/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.physleta.2004.01.015

rameter in the extremal models will now be played by the traffic load $\rho \in [0,1] \subset \mathbb{R}$. The traffic load measures the ratio between the arrival and the service rates and the limiting regime, characterized by $\rho \to 1$, is well-known to lead to a diverging population in the waiting room and BP of all sizes can be realized. In a situation where the traffic load ρ is tuned (and approaches unity from below) by the dynamics of the system itself, the basic elements for a SOC model are present. This is the construction that we adopt in the present Letter where a class of controlled QS is studied in the heavy traffic (i.e., $\rho \approx 1$) regime. The fact that QSs do play a role in SOC models should not come as a surprise. Indeed, the dynamics of the general QS is equivalent to a continuous time, generally non-Markovian, random walk with a reflecting boundary at the origin (i.e., the empty queue state). Accordingly, the dynamics can be described by master equations an approach already adopted by [1] in their study of mean-field behavior of SOC. Recently the role played by QS in the context of SOC has been pointed out in [2]. In this paper, the authors introduce the discrete time queueing model with Bernoulli arrivals and general service processes with an infinite number of parallel servers. They use this model to describe the dynamics of the avalanches in the sandpile model. Here, we shall consider the general class of continuous time QS for which the arrival and service random processes have finite two first moments but are otherwise arbitrary.

2. Basic model

Let us consider a queueing system (QS) formed by customers arriving to a server. The waiting customers are stored in a waiting room with a capacity assumed to be unlimited. The time between successive arrivals t_a and the service time t_s are independent random variables with cumulative distribution (CDF) given, respectively, by $A(x)$ and $B(x)$, i.e.,

$$\mathrm{Prob}\{0 \leqslant t_a \leqslant x\} = A(x),$$
$$\mathrm{Prob}\{0 \leqslant t_s \leqslant x\} = B(x). \tag{1}$$

We assume that the CDFs $A(x)$ and $B(x)$ admit moments to any orders and write the averages as:

$$\frac{1}{\lambda} = \int_0^\infty x\, dA(x), \qquad \frac{1}{\mu} = \int_0^\infty x\, dB(x), \qquad \rho = \frac{\lambda}{\mu},$$

with $\rho \in [0,1] \subset \mathbb{R}$ being the traffic load parameter.

Remark (*concerning the notation*). Models using QS are very common in telecommunication and production engineering. Their ubiquitous presence called for a standardized notation which was introduced in [3]. This notation characterizes the basic elements forming the "anatomy" of a QS. In the simplest setting, as the one used in this Letter, one has a single server, an infinite capacity waiting room and the stochastic processes (SP) characterizing the customer arrivals (i.e., the CDF $A(x)$ and the SP characterizing the service time (i.e., the CDF $B(x)$). For this simple setting, the standard notation will be A/B/1 (i.e., arrival CDF/service CDF/Nb of servers). Using this notation, one usually classifies QSs according to general dynamical behaviors. Accordingly, when both the $A(x)$ and $B(x)$ are exponential CDF, one usually writes M/M/1 to indicate that the underlying processes are Markovian (M stands for Markov). For $A(x)$ being an exponential CDF and for $B(x)$ a general CDF, the notation is M/G/1 to indicate that only the arrivals follow a Markov process. In the case when both $A(x)$ and $B(x)$ are general CDF, the notation will be G/G/1 (G standing here for general).

As it is common in QS theory, we now define the *busy period* (BP) to be the random variable U characterizing the time interval which begins with the arrival of a customer to the idle server and ends when the server next become idle. We will write:

$$G(x) = \mathrm{Prob}\{0 \leqslant U \leqslant x\},$$
$$g(x)\,dx = \mathrm{Prob}\{x \leqslant U \leqslant x + dx\}.$$

Consider the evolution of the QS during a time horizon T. The time interval T can be divided into successive cycle times (CT) ξ_k with $k = 1, 2, 3, \ldots, M$, with M such that $\sum_{k=1}^{M} \xi_k \leqslant T$ and $\sum_{k=1}^{M+1} \xi_k > T$ and

$$\xi_k = U_k + I_k.$$

Here I_k denotes the kth idle period starting directly after the end of the kth BP.

Construction of a controlled queueing system (CQS)

The QS is now equipped with a *self-regulating mechanism* which adapts the service rate μ_{k+1} offered during the $(k+1)$th CT by taking into account the length of the kth BP, namely, U_k. The tuning is somehow a "natural" one, that is to say: *the shorter the observed length of the kth BP, the stronger the server's availability is reduced during the* $(k+1)$*th CT*. To mathematically incorporate these qualitative features, we shall write for $k = 1, 2, \ldots, M$:

$$\mu_k = \min\{\mu_{k-1}, \lambda + \phi(U_{k-1})\},$$
$$\mu_0 = \lambda + \phi(0) > \lambda > 0, \tag{2}$$

with $\phi(x)$ being a monotonously decreasing, positive, function such that:

$$\lim_{x\to\infty} \phi(x) \to 0, \qquad \phi(0) = \text{const} > 0. \tag{3}$$

Using Eqs. (2) and (3), the traffic intensity of the QS fulfills:

$$\rho_k = \frac{\lambda}{\mu_k} = \frac{\lambda}{\min\{\mu_{k-1}, \lambda + \phi(U_{k-1})\}} \leqslant \rho_{k+1}$$
$$= \frac{\lambda}{\mu_{k+1}} = \frac{\lambda}{\min\{\mu_k, \lambda + \phi(U_k)\}} \leqslant 1. \tag{4}$$

In view of Eqs. (2), (3) and (4), the traffic load ρ_k of the CQS remains constant during the kth CT and ρ_k is increasing as k increases. With the choice given in Eq. (2), the QS modifies itself its traffic load and for T long enough, Eq. (3) implies:

$$\lim_{k\to\infty} \rho_k = 1 \quad \text{(almost surely)}. \tag{5}$$

The actualization of the service rate μ_k given by Eq. (2) does play a similar role as the tuning of the relaxation probability in the self-organized branching process (SOBP) studied by [4]. It is important to emphasize that for the CQS model the critical regime reached when the traffic load $\rho \to 1$, *does not depend on an external tuning* but rather it is controlled by *the dynamics of the system itself*. This is one of the key features governing SOC systems. In the sequel, we shall show that for asymptotic times (i.e., for $k \to \infty$), the BP probability density $g(x)$ of the CQS, exhibits the $-\frac{3}{2}$ critical exponent characterizing the mean-field behavior of SOC as it is discussed in, e.g., [1,4–6].

2.1. The M/G/1 queue

First, we focus on the M/G/1 QS for which we have Poisson arrival (i.e., $A(x) = 1 - \exp\{-\lambda x\}$) and a general CDF $B(x)$ for the service times. It is well known that for M/G/1 QS, the BP solves a functional equation (also known as the Takacs equation) [7,8]:

$$G^*(s) = B^*\big(s + \lambda - \lambda G^*(s)\big), \tag{6}$$

with $G^*(s) = \int_0^\infty e^{-sx}\, dG(x)$ and $B^*(s) = \int_0^\infty e^{-sx} \times dB(x)$.

To unveil the analogies between the queueing models and the branching processes, it is instructive to derive functional Eq. (6). Following the derivation given in [8], we condition on two events: (i) the duration of the service v of the initiating customer (call it the ancestor) and (ii) the number of new arrivals A during the service time of the ancestor. Given that $v = x$ and $A = n$, then n sub-busy periods $T_1, \ldots, T_n$ are generated by the descendants and

$$T = x + T_1 + T_2 + \cdots + T_n.$$

Since the T_i's are independent and identically distributed and are also independent of x, we have:

$$\mathbb{E}\{e^{-sT} \mid v = x, A = n\}$$
$$= \mathbb{E}\{e^{-sx}\}\mathbb{E}\{e^{-s(T_1+\cdots+T_n)}\} = e^{-sx}\big[G^*(s)\big]^n,$$

where $\mathbb{E}\{p \mid q_1, q_2\}$ stands for the conditional expectation of p given q_1 and q_2.

Now, we can write:

$$\mathbb{E}\{e^{-sT} \mid v = x\}$$
$$= \sum_{n=0}^{\infty} \mathbb{E}\{e^{-sT} \mid v = x, A = n\}\, \text{Prob}[A = n].$$

By definition of the M/G/1 QS,

$$\text{Prob}[A = n] = \frac{(\lambda x)^n}{n!} e^{-\lambda x}$$

and hence:

$$\mathbb{E}\{e^{-sT} \mid v = x\} = e^{-[s+\lambda-\lambda G^*(s)]x}.$$

Finally, the result given in Eq. (6) follows directly by noting that:

$$\mathbb{E}\{e^{-sT}\} = \int_0^\infty dB(x)\, \mathbb{E}\{e^{-sT} \mid v = x\}$$

$$= \int_0^\infty dB(x)\, e^{-[s+\lambda-\lambda G^*(s)]x}.$$

Successive differentiations of Eq. (6) yield [9]:

$$\langle U\rangle = \frac{1}{\mu(1-\rho)}, \qquad \langle U^2\rangle = \frac{b^{(2)}}{(1-\rho)^3},$$
$$\langle U^3\rangle = \frac{b^{(3)}}{(1-\rho)^4} + 3\lambda\frac{[b^{(2)}]^2}{(1-\rho)^5},$$
$$\langle U^4\rangle = \frac{b^{(4)}}{(1-\rho)^5} + 10\lambda\frac{b^{(2)}b^{(3)}}{(1-\rho)^6} + 15\lambda^2\frac{[b^{(2)}]^3}{(1-\rho)^7},$$
$$\vdots$$

where $b^{(m)}$, $m = 2, 3, \ldots$, is the mth moment of $B(x)$, (remember that $b^1 := 1/\mu$). As from Eq. (5), we have that $\lim_{k\to\infty}\rho_k \to 1$ all moments of $G(x)$ diverge in this limiting traffic regime. This is the typical signature of a SOC behavior. For the Markovian case characterized by $B(x) = 1 - \exp\{-\mu x\}$ (i.e., the M/M/1 queue), Eq. (6) can be solved in a closed form and we obtain [7]:

$$g(x)\,dx = \sqrt{\frac{\mu}{\lambda}}\,\frac{e^{-(\lambda+\mu)x} I_1(x2\sqrt{\lambda\mu})}{x}\,dx, \tag{7}$$

with I_1 being a Bessel function. Using the asymptotic expansion:

$$I_1(z) \simeq \frac{e^z}{\sqrt{2\pi z}} + O\left(\frac{1}{z}\right) \quad \text{for } z\to\infty,$$

Eq. (7) takes the form:

$$g(x) \simeq \text{const}\cdot\frac{\exp\{-\mu(1-\sqrt{\rho})^2 x\}}{x^{3/2}}.$$

Hence, for $\rho\to 1$, we observe that

$$g(x) \simeq \text{const}\cdot x^{-3/2}. \tag{8}$$

Hence, the probability density $g(x)$ of the BP exhibits the $-\frac{3}{2}$ power law typical for the mean field behavior of SOC [1,4–6]. Let us now show that this behavior also holds when general QS are considered.

2.2. The G/G/1 queue in heavy traffic regimes

Let us now consider general distributions $A(x)$ and $B(x)$ with finite two first moments (i.e., G/G/1 QS). For this type of dynamics, it has been established that in the heavy traffic regime (i.e., $\rho\to 1$), the BP probability density $g(x)$ can be written as, see Eq. (3.10) in [10]:

$$g(x)\,dx \simeq \frac{b}{\sqrt{2\pi d}}\frac{1}{\sqrt{x^3}}\exp\left\{-\frac{x(1-\rho)^2}{2d}\right\}dx \tag{9}$$

with:

$$d = C_a^2 + C_s^2,$$

$$b = \frac{1}{1-(C_a^2-1)h(\rho, C_a^2, C_s^2)} \quad \text{when } C_a^2 \neq 1,$$

with C_a^2, C_s^2 being the square of the coefficient of variation of the distributions $A(x)$, respectively, $B(x)$. The functions $h(\rho, C_a^2, C_s^2)$ read as [10]:

$$h(\rho, C_a^2, C_s^2) = \frac{1 + C_a^2 + \rho C_s^2}{1 + \rho(C_s^2 - 1) + \rho^2(4C_a^2 + C_s^2)} \quad \text{when } C_a^2 \leqslant 1,$$

$$h(\rho, C_a^2, C_s^2) = \frac{4\rho}{C_a^2 + \rho^2(4C_a^2 + C_s^2)} \quad \text{when } C_a^2 > 1.$$

Observe that for the $\rho\to 1$ limit, Eq. (9) exhibits the $-\frac{3}{2}$ power law showing that the general class of controlled G/G/1 models does exhibit the SOC mean-field exponent.

References

[1] J. de Boer, B. Derrida, H. Flyvbjerg, A.D. Jackson, T. Wettig, Phys. Rev. Lett. 75 (1994) 906.
[2] B. Tadic, V. Priezzhev, Phys. Rev. E 62 (2000) 3266.
[3] D.G. Kendall, J. R. Stat. Soc. B 13 (1951) 151.
[4] S. Zapperi, K.B. Lauritesen, H.E. Stanley, Phys. Rev. Lett. 75 (1995) 4071.
[5] H.J. Jensen, Self-Organized Critically. Emergent Complex Behavior in Physical and Biological Systems, in: Cambridge Lecture Notes in Physics, 1998.
[6] H. Flyvbjerg, P. Bak, K. Sneppen, Phys. Rev. Lett. 71 (1993) 4087.
[7] D. Gross, C.M. Harris, Fundamentals of Queueing Theory, third ed., Wiley, New York, 1998.
[8] J. Medhi, Stochastic Models in Queueing Theory, second ed., Academic Press, San Diego, 2003.
[9] H. Takagi, Queueing Analysis—A Foundation and Priority Systems, North-Holland, Amsterdam, 1991, see vol. 1, Ch. 1.
[10] J. Abate, W. Whitt, Probab. Theory Eng. Inform. Sci. 9 (1995) 581.