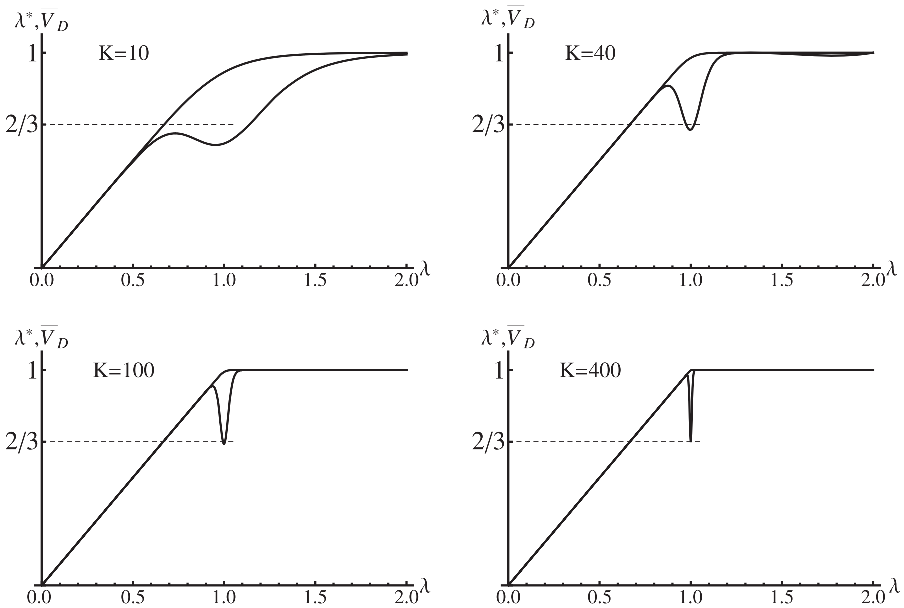

Fig. 1 M/M/1/K: λ^* (*top curve*) and $\overline{V}_{\mathcal{D}}$ (*bottom curve*) as a function of λ when $\mu = 1$ for various buffer sizes

output process associated with the queue: $D(0) = 0$ and $\mathcal{D}$ increases by 1 when $\mathcal{Q}$ decreases. It can be shown that the expectation and the variance functions of $\mathcal{D}$ are $O(t)$ (cf. formulas (9), (10) and accompanying discussion and references) and may thus be described by the *flow rate*, λ^* and *asymptotic variance rate*, $\overline{V}_{\mathcal{D}}$:

$$\mathbb{E}\,[D(t)] = \lambda^* t \tag{1}$$

$$\mathrm{Var}(D(t)) = \overline{V}_{\mathcal{D}} t + o(t) \tag{2}$$

Evaluation of $\overline{V}_{\mathcal{D}}$ is important in manufacturing type settings. When the system operates for a long duration, T, the variance of the number of items produced is approximately $\overline{V}_{\mathcal{D}} T$. Several studies have investigated computational procedures that evaluate this quantity for the output of a series of queues (cf. [14, 16, 17, 21, 29, 30]). In this paper we concentrate on the seemingly simpler case of a one-pass single class queueing system with losses. In general, output processes of one-pass single class systems and their second moments have been studied extensively (cf. the surveys [9, 12, 27]). For finite state space loss systems, the overflow process has received a considerable amount of attention (cf. [3, 4, 7, 24, 26, 31, 36]). Fewer papers have considered the output process of loss systems (cf. [2, 10, 23]) and to the best of our knowledge none have analyzed the asymptotic variance rate of the outputs.

Figures 1 and 2 display $\overline{V}_{\mathcal{D}}$ for different parameter values of the M/M/1/K queue with arrival rate λ and service rate μ. The plots may be partially understood as follows: For $\lambda \ll \mu$ the finite queue is hardly ever full and it behaves almost like an M/M/1 queue. In the M/M/1 queue, reversibility arguments imply that $\mathcal{D}$ is a Poisson process (cf. [19]), and thus for M/M/1/K we expect $\overline{V}_{\mathcal{D}} \approx \lambda^* \approx \lambda$ when $\lambda \ll \mu$. For

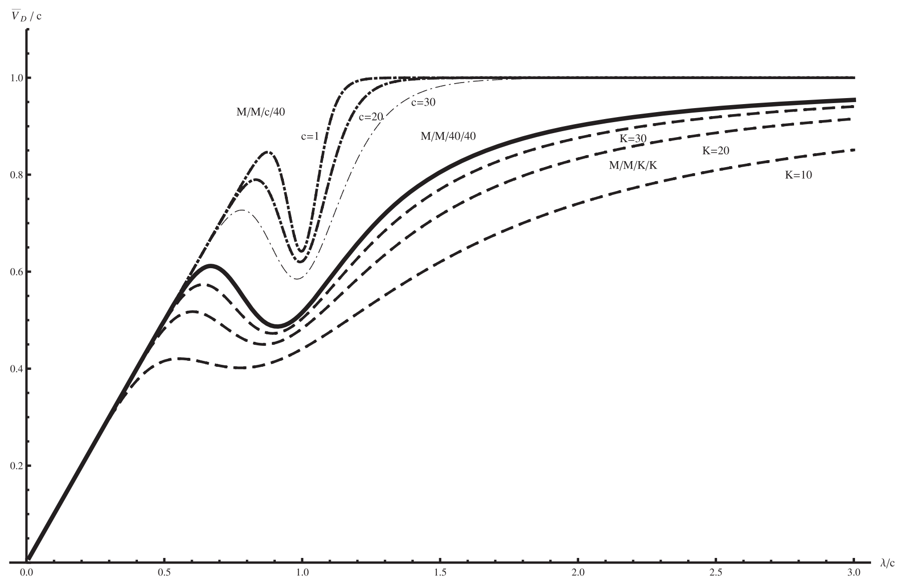

Fig. 7 M/M/c/K: $\frac{\overline{V}_{\mathcal{D}}}{c}$ as a function of $\frac{\lambda}{c}$ when $\mu = 1$

transitions on the edge states, 0 and K are λ and μ respectively. Observing the steady state distribution, (20), we see that when $\lambda = \mu$ the system spends very little time on the edge states and thus the "modulation" between rates $\lambda + \mu$ and λ or μ is minimal. On the contrary when $\lambda \neq \mu$ the system often switches between an edge state and a non-edge state and thus there is substantial "modulation" in the transition rates and as a result the variance of the transition process is greater.

This intuition does not immediately carry over to more complex systems but the BRAVO effect does. We now show some examples.

5.1 M/M/c/K

The M/M/c/K queue with $1 \le c \le K$ is an example a of a birth-death queue with monotone rates (the birth rates are constant and the death rates are increasing). While Theorem 3.1 is applicable to this system, the calculation of the normalization constant of the stationary distribution does not simplify and thus we are not able to obtain simple a formula for $\overline{V}_{\mathcal{D}}$ except for the case $c = 1$ (Sect. 4). Nevertheless, the computation of $\overline{V}_{\mathcal{D}}$ using the formula of (3.1) is simpler and more efficient than using the matrix formula (11).

Figure 7 shows that the BRAVO effect appears in the M/M/c/K queue: in this case "**b**alancing" implies setting $\lambda = c\mu$. The thick curve is for the Erlang loss system $(c = K)$ with $K = 40$ to which we compare other systems. It is apparent that as the number of servers decreases, the asymptotic variance rate normalized by the number of servers increases. Alternatively, keeping the number of servers equal to the buffer size and decreasing the number of servers causes a decrease in the asymptotic