Queueing Systems Vol. 100
Special Issue: 100 views on queues

Asymptotics in priority queues: from finite to infinite capacities

Joris Walraevens

February 14, 2022

1 Introduction

The stationary system content $s(n)$ of any lower-priority class in a priority queue with infinite capacities for all classes is known to potentially exhibit non-exponential (exact) asymptotics for $n \to \infty$. More precisely, asymptotics are of the form $c \cdot R^{-n}$ (exponential), $c \cdot n^{-3/2} R^{-n}$ (which I will coin 'non-exponential' in this letter) or $c \cdot n^{-1/2} R^{-n}$ ('bordercase'), with c, R real numbers with $c > 0$ and $R > 1$[1]. In the remainder, we do not discuss the border case. Also, to focus the discussion, we keep the following particular model in mind throughout, namely a discrete-time model with two priority classes, single-slot service times and independent arrivals from slot to slot (cf. [9]). However, the discussion is also valid for other models (more priority classes [8], continuous-time models [1,6], multi-server models [5], ...). It also applies to other low-priority distributions, such as the waiting time and unfinished work distributions [1].

These different types of asymptotics have *physical* as well as *functional* explanations. The exponential tail asymptotics occur when the low-priority arrival rate is sufficiently large. High low-priority system contents typically occur in this case when many customers have arrived. The non-exponential tail asymptotics occur when the low-priority arrival rate is small. In that case, the (number of low-priority arrivals during) *busy periods* of the high priority class play a pivotal role; high low-priority contents occur because of excessively long busy periods of class-1. This can also be concluded from a functional perspective. In terms of Probability Generating Functions (PGFs) of this model, the non-exponential asymptotics are caused by the branchpoint of a *implicitly* defined function that is a solution of the kernel [9]. This exhibits a clear relation with (arrivals during) busy periods, whose transform functions are also characterized by

Joris Walraevens
Department of Telecommunications and Information Processing, Ghent University, Belgium. E-mail: Joris.Walraevens@UGent.be

[1] Note, that we assume the input processes (arrival process, service process, ...) to be sufficiently regular so that they do not impact the asymptotics in a direct way (for instance, in case of power-law service times, the asymptotics of the system content inherit these power laws).

implicit functions. In case of exponential asymptotics, the branchpoint is also present, but is dominated by a regular pole. In other words, the 'busy period' effect on the low-priority system content distribution is only of second order in that case.

In models where the high-priority queue has *finite* capacity, the story is different. In that case, the asymptotics of the low-priority system content are always (purely) exponential. The functional explanation is that the busy period of a finite-capacity queue is no longer determined by an implicit function and therefore has exponential asymptotics. The *physical* distinction that one can make in the infinite-capacity case is also no longer straightforward (or even possible?) in the finite-capacity case.

2 Problem Statement

Although both the finite- as the infinite-capacity systems are well-studied, the convergence of the asymptotics from the finite- to the infinite-capacity case is ill understood. Related questions are:

- How is the branchcut of the PGF in the infinite capacity case 'formed' for $N \to \infty$ if there is convergence? Are they formed by regular poles that 'clump' together? However, the number of poles is countable (even for $N \to \infty$?) while a branchcut can be regarded as a continuum of singularities. How does that match?
- Of all poles (on the positive axis) of the PGF in case of finite capacity, can we identify the one that converges to the regular pole in the infinite case? A candidate in case this regular pole is dominant in the infinite case is the smallest (dominant) one. When the branchpoint is dominant in the infinite case it is even more unclear (there is no regular pole; or is it part of the branchcut?).

These questions are formulated as *functional* questions, as we feel that the key is in functional analysis. However, when answered they would also lead to explanations on a physical level. The other way around is true as well: if physical explanations could be given (perhaps through large-deviations theory?) it would help answering the functional questions too.

3 Discussion

We have done some research related to the problem statement for the discrete-time model described in the beginning [2]. We have calculated the PGF of the low-priority content in a queue with finite capacity N for the high-priority content. We then calculated the poles of this PGF numerically in case of specific distributions of the number of arrivals per slot. In Fig. 1, the dominant pole and branchpoint for the infinite-capacity system are shown, as well as all the poles of the finite-capacity system for several values of N. Some convergence of the poles to the dominant pole and to the branchcut is observed. For the special case that is shown in Fig. 1, we actually proved (unpublished) that there are exactly N poles on the positive real axis and that they all shift to the left when N increases (extra poles are added to the right). In the case of a maximum of two

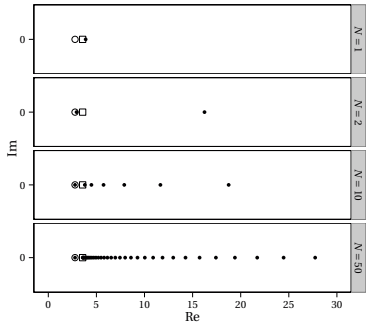

Fig. 1 Singularities on the positive real axis for several finite capacities N. The circle is the dominant pole in the infinite capacity case, while the square is the branchpoint in that case. Taken from [2], p. 4-9.

high-priority arrivals, we furthermore showed that the PGF of the finite case could be written in terms of *two* implicitly defined functions (the one from the infinite-capacity model and a related one), that the branchcuts of both cancel eachother out for all finite N (with isolated poles as only singularities as result) and where the second one disappears when $N \to \infty$. Extending these results could be a step in solving this problem.

Links can be made to alternative models and methods. Quasi-Birth-and-Death (QBD) processes for instance also typically exist of a component of infinite dimension (the level) and one of finite dimension (the phase). The finite-capacity priority queue is therefore only one example of such QBD-type processes. The matrix-geometric approach exploits the geometric decay through the finite $R - matrix$. When the number of phases is infinite, the story is again very different (R is an infinite matrix) [3]. A generalization are random walks in the quarter plane (infinite-capacity generalization) and in a strip in the quarter plane (the finite-capacity case) [4]. Finally, large deviations principle could be an alternative methodology that leads to answers [7].

References

1. J. Abate and W. Whitt. Asymptotics for M/G/1 low-priority waiting-time tail probabilities. *Queueing Systems*, 25(1-4):173–233, 1997.
2. T. Demoor. *Priority queues with limited capacity*. PhD thesis, Ghent University, 2014.
3. L. Haque, Y. Zhao, and L. Liu. Sufficient conditions for a geometric tail in a QBD process with many countable levels and phases. *Stochastic Models*, 21(1):77–99, 2005.
4. S. Kapodistria and Z. Palmowski. Matrix geometric approach for random walks: Stability condition and equilibrium distribution. *Stochastic Models*, 33(4):572–597, 2017.
5. K. Laevens and H. Bruneel. Discrete-time multiserver queues with priorities. *Performance Evaluation*, 33(4):249–275, 1998.
6. H. Li and Y. Zhao. Exact tail asymptotics in a priority queue - characterizations of the non-preemptive model. *Queueing Systems*, 68(2):165–192, 2011.
7. M. Mandjes and M. Van Uitert. Sample-path large deviations for tandem and priority queues with gaussian inputs. *Queueing Systems*, 54(2):85–97, 2005.
8. J. Walraevens, H. Bruneel, T. Maertens, D. Fiems, and S. Wittevrongel. Delay analysis of multiclass queues with correlated train arrivals and a hybrid priority/fifo scheduling discipline. *Applied Mathematical Modelling*, 45:823–839, 2017.
9. J. Walraevens, B. Steyaert, and H. Bruneel. Performance analysis of a single-server ATM queue with a priority scheduling. *Computers & Operations Research*, 30(12):1807–1829, 2003.