Effect of Prioritization on the Waiting Time

Yannic Jäger[1] and Christoph Roser[2(✉)]

[1] Locom Consulting, Karlsruhe, Germany
yannic.j@gmx.de
[2] Karlsruhe University of Applied Sciences, Karlsruhe, Germany
christoph.roser@hochschule-karlsruhe.de

Abstract. In industry, it is common to prioritize some orders over others. This is done to reduce the lead time and waiting time of these prioritized orders, hence the customer will get the order earlier than otherwise. However, whenever an order is prioritized, the remaining orders are de-prioritized, and their lead time and waiting time will increase. In industry, a rule of the thumb that no more than 30% of the orders should be prioritized is often used. This paper will verify this assumption using simulations for different conditions. It will show that this rule of thumb is generally a valid approach. The paper will offer more detail on the trade-off between prioritizing some orders and hence delaying other orders.

Keywords: Prioritization · Waiting time · Lead time

1 Introduction

The behavior of single-arrival single-server systems as shown in Fig. 1 is generally well understood, and relevant to most industries [1]. If an actual system can be observed, the lead time can easily be calculated using Little's law [2].

If the system is understood in more abstract terms, the Kingman equation (also known as Kingman formula or Kingman approximation) gives an approximation of the waiting time of the orders for a single process based on its utilization and variance [3]. Other calculations and approximations exist like [4–6] or [7]. These equations are valid over a wide range of assumptions and estimate the behavior of a steady state system quite well.

In industry, it is common practice to prioritize some orders over others to reduce the lead time and waiting time of the prioritized orders at the cost of an increased lead time and waiting time of the non-prioritized orders. Examples include the food industry with its limited product lifespan [8], maintenance tasks [9] or other resources [10], as well as general throughput improvements [11]. It is important to note that as long as the average system behavior does not change, the equations in [2–6] and [7] are still valid. Even if some orders are prioritized and accelerated, the slowdown of the not-accelerated orders will cause the overall system to keep a constant average lead time and waiting time.

Take for example Little's law [2]. Little's law is "not influenced by the arrival process distribution, the service distribution, the service order, or practically anything else" [12]. As prioritization does nothing but change the "service order," prioritization

I. Moon et al. (Eds.): APMS 2018, IFIP AICT 535, pp. 21–26, 2018.
https://doi.org/10.1007/978-3-319-99704-9_3

Fig. 1. Illustration of a single-arrival single-queue single service system

has no effect on the average lead time. It does, however, affect the distribution of this lead time. As some orders are accelerated at the expense of others, the width of the distribution of the lead times and waiting times will increase, even though the mean remains unchanged.

Please also note that this paper uses orders as the item processed in the single-server single-queue system. However, the wide application of this prioritization makes the following simulations, discussions, and calculations equally valid for a system processing parts (as for example a machine), customers (for example in a supermarket, a hospital, or a call center), products (as for example a shipping warehouse), or many other applications.

2 System Outline

The system simulated is a single-arrival single-server system as a simplification of more complex production systems. The arrival times and service times are randomly distributed. For the arrival times, we used an exponential distribution, as this is the most commonly used distribution to model inter-arrival times [13, 14]. The service times are modeled using a lognormal distribution, which is also commonly used for service times [13, 15]. The exponential distribution has only one parameter, which was used to adjust the mean value. The lognormal distribution has two parameters, hence besides the mean, it is also possible to influence the standard deviation. During this analysis, the standard deviation was set to be 25% of the mean value (i.e., the coefficient of variation is 25% for the service times). This is within the range of common values in the industry.

The utilization of the system has a major influence on the waiting time. Hence different systems were simulated using different utilizations. Table 1 gives an overview of the different settings to achieve different utilizations. Please note that the units of the mean times are here only for a complete understanding of the set-up, but does not influence the results. The lot size of arrivals and processing is both one. For simplicity, we also did not model any set-up changes, breakdowns, or other interruptions. Transport times were also assumed instantaneous.

During the simulations, we measured the mean waiting time for each order type as well as the joint mean waiting time. We also measured the standard deviation of the waiting time for order types A and B individually as well as jointly. The 95% confidence interval of all of these was also determined. Each simulation experiment had a duration of 120,000 min, representing 20,000 orders processed or around one year of simulated time. Each simulation was repeated thirty-nine times to calculate the 95% confidence interval. For details on the set-up, see [16].

Table 1. Overview of the mean inter-arrival times and service times to achieve different utilizations

Utilization	Mean inter-arrival time (min)	Mean service time (min)
75%	6	4.50
80%	6	4.80
85%	6	5.10
90%	6	5.40
95%	6	5.70

2.1 Un-prioritized Baseline System

As a baseline reference, we used a system without prioritization, using a simple first-come-first-served approach for the arriving orders. The layout is shown in Fig. 1.

2.2 Prioritized System

The main part of the analysis is the prioritized system. Two different order types were simulated, order type A and B. Order type A always has priority over order type B. This is simulated by having two different waiting queues, both of which have an independent first-in-first-out logic. Orders in the B queue are only processed if there are no more orders waiting in the A queue. The service time for both order types is identical and depending on the selected utilization as shown in Table 1 (Fig. 2).

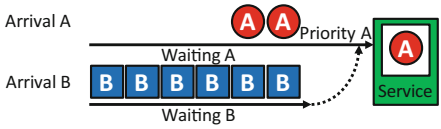

Fig. 2. Illustration of a prioritized system with a double-arrival double-queue single-service system

The percentage of prioritized orders was modified from 0.1% to 99.9% as shown in Table 2. The exponentially distributed inter-arrival times were adjusted accordingly to maintain a joint mean inter-arrival rate of 6 min between orders. Combining the 11 different percentages A with the 5 different utilizations gives a total of 55 simulation experiments in addition to the 5 utilizations of the baseline system.

3 Simulation Results

3.1 Baseline System

As expected and predicted by theory, the waiting time of the queue of the baseline system was influenced by the utilization. The exponential relation is shown in Fig. 3.

Table 2. Mean inter-arrival times for order A and B for different percentages of A

%A	Mean inter-arrival time A (min)	Mean inter-arrival time B (min)	Joint mean inter-arrival time (min)
0.1%	6000	6.006	6
10%	60	6.666	6
20%	30	7.5	6
30%	20	8.571	6
40%	15	10	6
50%	12	12	6
60%	10	15	6
70%	8.571	20	6
80%	7.5	30	6
90%	6.666	60	6
99.9%	6.006	6000	6

For a utilization of 100%, the average waiting time would approach infinity. These values serve as our baseline system.

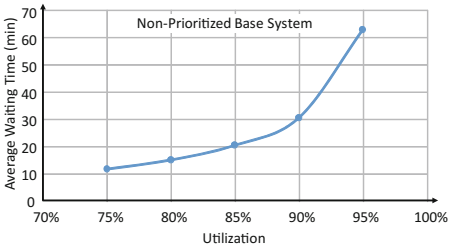

Fig. 3. Waiting times in dependence of the utilization of the un-prioritized baseline system

3.2 Prioritized System

Figure 4 shows the behavior of the systems under different utilizations and percentages of prioritized orders. For simplification, the percentage improvement of orders A and the percentage worsening of orders B compared to the baseline waiting time from Fig. 3 is shown. It is clearly visible that for low percentages of A, there is a substantial benefit for orders A with an up to 90% reduction in waiting time without much disadvantage for orders B. However, as the percentage of A increases, this benefit for A shrinks, whereas the disadvantage for orders B becomes exponential, with the waiting time being a multitude of the baseline.

Figure 5 shows the impact on the coefficient of variation of the waiting time. For both orders A and B, this increases as the percentage of prioritized orders A increases. Hence, not only does the average waiting time increase, but the range of the fluctuations also increases. The full data including the confidence intervals can be found in [16].

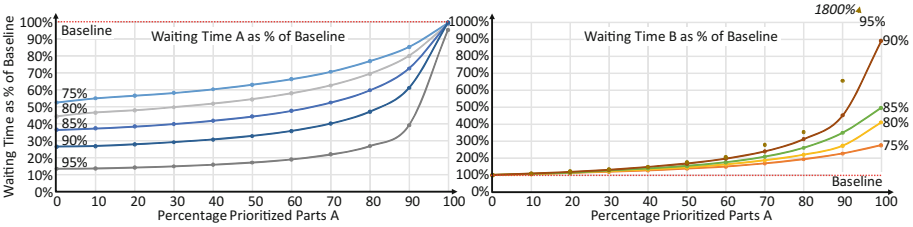

Fig. 4. Waiting times relative to the baseline in dependence of the utilization and percentage prioritized orders of the prioritized system

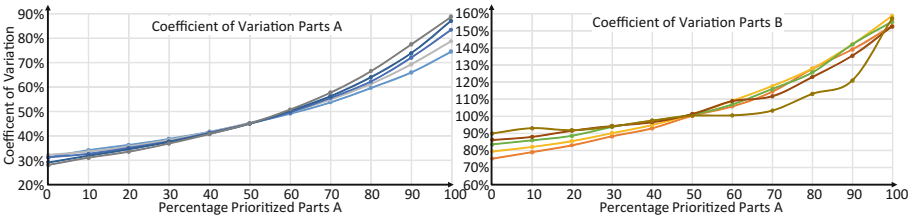

Fig. 5. Coefficient of variation of the waiting times in dependence of the utilization and percentage prioritized orders of the prioritized system

4 Conclusions

Overall, prioritizing important orders can have a significant benefit as long as it is done in moderation. In industry, often a general rule of thumb is used, recommending to prioritize no more than 30% of the workload. Since this number originates in industry, there is no academic reference for this that we are aware of. While this rule is an oversimplification, Fig. 4 shows that this is a workable assumption without having a too much negative effect on the not-prioritized orders. However, this is not a hard cut off, but rather a sliding scale, and prioritizing 35% or even 40% may also be possible, although the benefit shrinks and the disadvantages grow. It boils down to the tradeoff that has to be made between benefitting the prioritized products while disadvantaging everything else.

However, prioritizing an excessive number of orders will diminish the effect of this prioritization. The negative effect on the not-prioritized orders will multiply and become significant. Even worse, the range of the fluctuations increases faster than the mean waiting time. If the products are made to order (MTO), this means a prediction of a delivery date will become more difficult, as the actual delivery dates become more erratic. For a reasonable delivery performance, the promised delivery date now has to be much later, as this not only has to include the mean but also a substantial share of the outliers. If 95% of the true delivery dates should be within the estimate, the estimated delivery date needs to be the 95th percentile of the actual delivery dates.

Similarly, for made-to-stock (MTS) items, if the waiting and process time would always be constant, it would be sufficient to have one item in stock plus the coverage

for the customer behavior. As the fluctuations of the waiting time and delivery time increase, more stock is needed to cover for these fluctuations. Again, if a 95% delivery performance is promised, there needs to be stock covering at least 95% of the shortest waiting times and processing times in addition to the customer behavior.

Overall, prioritizing too many orders will drastically push the promised delivery dates to the future (for MTO) or require significant increases in inventory (for MTS), or have a significant negative impact on the delivery performance (for both cases). Practitioners are strongly advised to prioritize no more than 30% of their order volume!

References

1. Tom, G., Lucey, S.: A field study investigating the effect of waiting time on customer satisfaction. J. Psychol. **131**(6), 655–660 (1997)
2. Little, J.D.C.: A proof for the queuing formula: L = λW. Oper. Res. **9**(3), 383–387 (1961)
3. Kingman, J.F.C.: The single server queue in heavy traffic. Math. Proc. Camb. Philos. Soc. **57** (4), 902–904 (1961)
4. Bhat, U.N.: The general queue *G*/*G*/1 and approximations. An Introduction to Queueing Theory. SIT, pp. 169–183. Birkhäuser, Boston (2008). https://doi.org/10.1007/978-0-8176-4725-4_9
5. Adan, I.: Queueing theory: Ivo Adan and Jacques Resing. Department of Mathematics and Computing Science, Eindhoven University of Technology (2001)
6. Marchal, W.G.: An approximate formula for waiting time in single server queues. E Trans. **8** (4), 473–474 (1976)
7. Krämer W., Lagenbach-Belz, M.: Approximate formulae for the delay in the queueing system GI/G/1. In: Proceedings of the 8th International Teletraffic Congress, Melbourne, pp. 235.1–235.8 (1976)
8. Akkerman, R., van Donk, D.P.: Product prioritization in a two-stage food production system with intermediate storage. Int. J. Prod. Econ. **108**(1), 43–53 (2007)
9. Li, L., Ni, J.: Short-term decision support system for maintenance task prioritization. Int. J. Prod. Econ. **121**(1), 195–202 (2009)
10. Gupta, S., Bhattacharya, J., Barabady, J., Kumar, U.: Cost-effective importance measure: a new approach for resource prioritization in a production plant. Int. J. Qual. Reliab. Manag. **30**(4), 379–386 (2013)
11. Pascual, R., Godoy, D., Louit, D.M.: Throughput centered prioritization of machines in transfer lines. Reliab. Eng. Syst. Saf. **96**(10), 1396–1401 (2011)
12. Simchi-Levi, D., Trick, M.A.: Introduction to 'Little's Law as viewed on its 50th anniversary. Oper. Res. **59**(3), 535 (2011)
13. Law, A.M., Kelton, D.W.: Simulation Modeling & Analysis, 3rd edn. McGraw Hill, Maidenherd (2000)
14. Domschke, W., Drexl, A.: Einführung in Operations Research, 7th edn. Springer, Berlin (2007). https://doi.org/10.1007/978-3-662-48216-2
15. Kühn, W.: Digitale Fabrik: Fabriksimulation für Produktionsplaner. Carl Hanser Verlag GmbH & Co. KG, München (2006)
16. Jäger Y.: Einfluss von Priorisierung auf das Verhalten eines Produktionssystems. Master thesis, Karlsruhe University of Applied Sciences, Karlsruhe, Germany (2017)