The Use of Queueing Models to Gain Insight into the Robustness of Supply Chains

H.L.J. Bink BSc (0739250)
Industrial and Applied Mathematics

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Industrial and Applied Mathematics

Eindhoven University of Technology
Department of Mathematics and Computer Science

26 October 2015

Supervisors
ir. M.S. van den Broek (CQM)
dr.ir. J.B.M. van Doremalen (CQM)
prof.dr. A.P. Zwart (TU/e)
Abstract

In this thesis we are interested in designing a robust supply chain by dealing with uncertainty in the demand rate. We model a supply chain as a queueing system, to cover the stochastics. In this queueing system, orders arrive as a Poisson processes and the service time are independent exponential distributed random variables. Goal is to optimize the service rate such that all orders can be served in reasonable time, without purchasing too much capacity.

We investigate several models. The first one is a model with one queue and one server. In the second model, another queue is added with its own demand and service rate. The server serves both queues according to a First Come, First Served policy without switchover times. The last model is a $k$-limited polling model in which switchover times are added to create a more realistic model.

In all models, two time periods are considered. The demand rate for the first time period is assumed to be known, while for the second time period only a discrete distribution is given. This second period represents the long-term, for which there is in general less certainty about the demand. The service rate must be chosen large enough to serve all customers in finite time. Costs are counted for the amount of capacity (service rate), the expected waiting or sojourn time and for eventually changing the capacity just before period 2 starts. The optimization in the models is done partially algebraic and partially in a numerical way, by using Mathematica.

For the first two models, the optimal solution for small instances can be found within a few seconds. For the $k$-limited polling model, at least a few minutes are needed. To create insight in how to make a supply chain robust, we analyze what happens with the optimal service rate and corresponding cost when some parameters are changed. Also, we show that our method outperforms another method, where first the demand rate is estimated and subsequently the optimal server capacity is determined.
## Contents

1 Introduction .......................................................... 4

2 One customer, one server ............................................. 6
   2.1 Model ............................................................. 6
   2.2 Notation .......................................................... 6
   2.3 Cost function ..................................................... 7
   2.4 Common approaches .............................................. 7
   2.5 Optimizing capacity in period 2 .................................. 7
   2.6 Optimizing capacity in period 1 .................................. 9
   2.7 Analysis .......................................................... 10
      2.7.1 Basic example ............................................. 10
      2.7.2 Costs for sojourn time and increasing capacity .......... 10
      2.7.3 Capacity cost ............................................. 13
      2.7.4 Sojourn cost ............................................. 14
   2.8 Average scenario ................................................ 15
      2.8.1 Some examples ........................................... 15
      2.8.2 Increasing variability ................................... 20
   2.9 Conclusions ...................................................... 22

3 Two customers, one server - First Come, First Served ............ 23
   3.1 Model ............................................................. 23
   3.2 Notation .......................................................... 23
   3.3 Cost function ..................................................... 24
   3.4 Optimizing capacity in period 2 .................................. 25
   3.5 Optimizing capacity in period 1 .................................. 26
   3.6 Analysis .......................................................... 27
      3.6.1 Basic example ............................................. 27
      3.6.2 Different ratios for the service rates ................. 28
      3.6.3 Demand rate for queue 1 in period 1 ............... 30
      3.6.4 Cost for increasing capacity .......................... 31
   3.7 Conclusions ...................................................... 31

4 Multiple customers, one server - polling model .................... 32
   4.1 Model ............................................................. 32
   4.2 Notation .......................................................... 33
   4.3 Cost function ..................................................... 33
      4.3.1 Expected waiting time .................................. 33
   4.4 Optimizing capacity .............................................. 35
      4.4.1 Two examples ........................................... 35
      4.4.2 Switchover times ....................................... 36
      4.4.3 Service time ratios .................................. 37
   4.5 Uncertain demand ............................................... 39
      4.5.1 Impact of the uncertainty ............................... 39
   4.6 Conclusions ...................................................... 41
5  General conclusions and recommendations 42

Appendix A  Mathematica program - one customer, one server 45

Appendix B  Mathematica program - two customers, one server - First Come, First Served 46

Appendix C  Mathematica program - multiple customers, one server - polling model 49
1 Introduction

Supply chains come in all shapes and sizes and often form the basis of a company. Whether a company is large or small, everything is arranged around parts of supply chains (what products, product parts or services are delivered and in what way is this achieved). In [11], Van den Broek and Van Doremalen give some properties of supply chains: supply chains are in most cases very complex with many factories, warehouses and shops all over the world. When designing a supply chain, many choices have to be made and many alternatives need to be considered to optimize the process. Even for a small supply chain, many facets can be considered [2, p. 5]. Choices are the number, locations and capacity of factories and warehouses, amongst others. Also, it is important how the production of products in a factory is organized and managed, to keep lead times small and costs low.

Besides all the choices, one must also make sure that a supply chain is robust. That is, ‘resistant to big and small shocks and changes’ [11]. Because a supply chain is designed for a longer period, the decisions for the supply chain design have a strategic-tactical character. This also means that many events that are difficult or impossible to predict will occur. For example, a machine can break down, suppliers stop delivering materials or customers do not want to buy your product anymore. However, despite of the more or less unpredictable nature of those events, the supply chain must remain operational by low costs, as best as possible. Therefore, the supply chain must be ‘risk-insensitive, flexible and adaptive’ [11]. In [10], four steps are distinguished to create a robust supply chain. In the first step, an optimal network under normal circumstances is created. In the next steps, some extreme and alternative networks are designed, compared and adjusted. Although it is often difficult, especially for large networks, it would be useful if special circumstances and uncertainties could already be taken into account in the first step.

People are interested in robust supply chains and its design because it can be of great advantage for the production flows and service rates. Our interest for the models in this thesis is also fueled by an application in the chemical industry. In this application, a company has plants and customers all over the world. Different types of products are made at the plants and transported to the customers. However, not all plants can produce all product types and different plants produce at different speeds. In [4], a method (Kleinrock’s flow deviation algorithm) to find the optimal production and transportation scheme is investigated. This is a so-called ‘flow-assignment problem’ [8].

In the context of the supply chain most often more-or-less complicated optimization problem formulations are used to assess and evaluate the set-up of the supply chain over a longer period. We are interested in the stochastic component of the equation. That is, can we incorporate the uncertainties in the modeling and analysis of our future supply chains? We investigate this on the basis of another main problem when designing a supply chain: the capacity assignment problem. In the capacity assignment problem, the optimal capacity of servers in a system needs to be determined for given flows, such that the total expected cost is minimized. In many papers that treat the capacity assignment problem, a system with multiple servers is considered and the number of servers is optimized, while the service rate for each server is known (for example [6], [3]). However, it can also be interesting to consider a single-server queue and optimize the service rate of the server, for example if one can install only one server in a plant. In the capacity assignment problems in this report, the server speed must be high enough to serve all customers and to provide sufficient speed
in delivery, but not too high because faster servers are more expensive.

In this report, we consider the capacity assignment problem for three different models. To cover robustness, we include uncertain demand in these models. In each model, there are two time periods. For the first time period (short term), the demand rate is assumed to be given, but for the second period (long term) only a distribution is known. After period 1, when the demand rate for the second period is known, it is possible to change the server speed if this is necessary. However, this gives some extra cost, so in order to keep costs low, the uncertain demand for period 2 must be taken into account when deciding for the service rate in period 1. If we can optimize such a system in a fast and intelligent way, we create a more robust system than if we consider a stochastic parameter as deterministic.

In section 2, we start with a model with one queue and one server. Then, in section 3, we extend the model with an additional queue. Both queues are served by the same server, but have their own arrival and service rate. We assume that the customers are served according to a First Come, First Served policy, where the server can switch queues costless. In section 4, we consider a model with multiple queues and add switchover times. This gives a polling model, in which one server visits the multiple queues. Finally, we give our final conclusions and recommendations in section 5.
2 One customer, one server

2.1 Model

Consider a queueing system with independent and identically distributed exponential service times, where customers are served by a First Come, First Served (FCFS) policy. We consider two time periods, for which the demand rate is known only for the first period. Customers arrive according to a Poisson process with given parameter $\lambda_1$ in period 1 and unknown parameter $\lambda_2$ in period 2. We model this unknown parameter as a random variable $\Lambda_2$ with discrete distribution $F$. If we have a continuous distribution for $\Lambda_2$, we can discretize this to get $F$. Because of the unknown parameter in period 2, we have a doubly stochastic Poisson process: both the exponential distribution and its uncertain parameter provide stochasticity.

Note that we have an $M/M/1$ queue in both periods. Our goal is to choose the service rates in both periods, so that customers do not have to wait too long until they are being served. In both period 1 and 2, costs are involved for the amount of capacity (service rate) and the number of customers in the system. The mean number of customers in the system is equal to the expected sojourn time of a customer by Little’s Law: $E(L) = \lambda E(S)$. Note that the number of customers in the system can be related to potential stock sizes for that product. Also, costs are made for switching from speed $\mu_1$ and $\mu_2$ after period 1.

2.2 Notation

In this section, we use the following notation:

- $\lambda_i$ is the arrival rate of the customers in period $i$ for $i = 1, 2$.
- $\Lambda_2$ is the random variable for the unknown demand rate $\lambda_2$ in the second period.
- $\mu_i$ is the service rate in period $i$ for $i = 1, 2$.
- The occupation rate of the server is equal to $\rho_i = \frac{\lambda_i}{\mu_i}$ in period $i$ for $i = 1, 2$.
- $c$ is the cost for sojourn time (per time unit) for each product.
- The capacity cost in period $i$ is equal to $k\mu_i$ for $i = 1, 2$.
- $h_u$ and $h_d$ are the costs for increasing and decreasing the service rate by one unit between period 1 and 2.
- $L_i$ is the mean number of customers in the system in period $i$ for $i = 1, 2$.
- $S_i$ is the expected sojourn time for a customer in period $i$ for $i = 1, 2$.

The parameters $k$ and $c$ could be different in both periods, but to simplify notation, we choose them to have the same value for both periods in our model. Further, note that in a queueing system, we can see the arrivals as customers demanding for a product or service, but we can also see them as one customer with multiple orders. When no confusion arises, we use both options interchangeably.
2.3 Cost function

We have the following function for the expected cost:

\[ T(\mu_1, \mu_2, \lambda_2) = k\mu_1 + cE(L_1) + k\mu_2 + cE(L_2) + h_u(\mu_2 - \mu_1)^+ + h_d(\mu_1 - \mu_2)^+ \quad (1) \]

\[ = k\mu_1 + c\lambda_1 E(S_1) + k\mu_2 + c\lambda_2 E(S_2) + h_u(\mu_2 - \mu_1)^+ + h_d(\mu_1 - \mu_2)^+ \quad (2) \]

\[ = k\mu_1 + c\frac{\rho_1}{1 - \rho_1} + k\mu_2 + c\frac{\rho_2}{1 - \rho_2} + h_u(\mu_2 - \mu_1)^+ + h_d(\mu_1 - \mu_2)^+, \quad (3) \]

where \( x^+ = \max\{0, x\} \).

2.4 Common approaches

An approach that is commonly used when multiple scenarios can occur, is to assume that the worst case will happen and respond to that scenario [9]. However, this is not a good method in general because in most cases a very bad scenario will happen only with a small probability. Being prepared for this scenario involves many costs and is often not profitable.

Another approach is to respond to the weighted average of the different scenarios. In some cases this works well, in other cases is does not. This depends (among other factors) on the symmetry of the parameters. We discuss this in more detail in section 2.8.

2.5 Optimizing capacity in period 2

To optimize the capacity for the first period in our model, we first compute the optimal capacity in the second period with given demand rate and capacity in the first period. After that, we use this optimization to optimize the capacity in period 1 (see section 2.6).

Intuitively, the server capacity must be increased after period 1 if the savings on sojourn cost are greater than the additional capacity cost and changing cost. On the other hand, the server capacity must be decreased after period 1 if the savings on capacity cost are greater than the additional sojourn cost and changing cost.

The situation is clarified in figure 1. In this figure, ‘OPT’ would be the optimal value if the solution does not depend on \( \mu_1 \) (so if no costs for changing capacity were charged). However, because one has to pay for changing the capacity after period 1, it is not always optimal to set the amount of capacity equal to OPT. Close to OPT, the cost for changing capacity is more than the savings on sojourn or capacity cost. In figure 1, the area ‘close to OPT’ is between OPT\(_{\text{low}}\) and OPT\(_{\text{high}}\). Therefore, if \( \mu_1 \leq \text{OPT}_{\text{low}} \), it is optimal to set \( \mu_2 \) equal to \( \text{OPT}_{\text{low}} \). If \( \mu_1 \geq \text{OPT}_{\text{high}} \), it is optimal to set \( \mu_2 \) equal to \( \text{OPT}_{\text{high}} \) and if \( \text{OPT}_{\text{low}} \mu_1 \leq \text{OPT}_{\text{high}} \), it is optimal to set \( \mu_2 \) equal to \( \mu_1 \). How to determine the values of \( \text{OPT}_{\text{low}} \) and \( \text{OPT}_{\text{high}} \) is described later in this section.
Formally, if the capacity for the first period is known, we need to consider the capacity cost for the second period, the sojourn cost for the second period and the cost for increasing/decreasing capacity between period 1 and 2. When increasing the capacity after period 1, the capacity cost for period 2 and cost for changing capacity increase, while the sojourn cost (for period 2) decreases. Because the capacity cost and increasing cost increase proportionally to the amount of capacity, while the sojourn cost decreases in a convex way, it is optimal to increase the capacity until the capacity cost and increasing cost increase faster than the sojourn cost decreases. In other words, until the derivative of $k\mu_2 + c\frac{\lambda_2}{\mu_2 - \lambda_2} + h_u(\mu_2 - \mu_1)$ with respect to $\mu_2$ is equal to zero. This gives

$$0 = k + h_u - \frac{c\lambda_2}{(\mu_2 - \lambda_2)^2}, \quad (4)$$

$$0 = k + h_u - \frac{c\lambda_2}{\mu_2^2 + \lambda_2^2 - 2\lambda_2\mu_2}, \quad (5)$$

$$0 = (k + h_u)\mu_2^2 - (2\lambda_2(k + h_u))\mu_2 + (k + h_u)\lambda_2^2 - c\lambda_2, \quad (6)$$

$$\mu_2 = \frac{2\lambda_2(k + h_u) \pm \sqrt{(2\lambda_2(k + h_u))^2 - 4(k + h_u)((k + h_u)\lambda_2^2 - c\lambda_2)}}{2(k + h_u)}, \quad (7)$$

$$\mu_2 \geq \lambda_2 \Rightarrow \lambda_2 + \sqrt{\frac{c\lambda_2}{k + h_u}}. \quad (8)$$

Hence, if $\mu_1 < \lambda_2 + \sqrt{\frac{c\lambda_2}{k + h_u}}$, then choose $\mu_2 = \lambda_2 + \sqrt{\frac{c\lambda_2}{k + h_u}}$. 

---

Table:

<table>
<thead>
<tr>
<th>Amount of capacity</th>
<th>OPT&lt;sub&gt;low&lt;/sub&gt;</th>
<th>OPT</th>
<th>OPT&lt;sub&gt;high&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Increase capacity to OPT&lt;sub&gt;low&lt;/sub&gt; to save on sojourn cost</td>
<td>Close to optimum: changing capacity is not profitable</td>
<td>Decrease capacity to OPT&lt;sub&gt;high&lt;/sub&gt; to save on capacity cost</td>
<td></td>
</tr>
</tbody>
</table>
Decreasing the capacity is profitable until the derivative of \(k\mu_2 + c\frac{\lambda_2}{\mu_2 - \lambda_2} + h_d(\mu_1 - \mu_2)\) with respect to \(\mu_2\) is equal to zero. This gives

\[
0 = k - h_d - \frac{c\lambda_2}{(\mu_2 - \lambda_2)^2},
\]

\(0 = k - h_d - \frac{c\lambda_2}{\mu_2^2 + \lambda_2^2 - 2\lambda_2\mu_2},\)

\[
0 = (k - h_d)\mu_2^2 - (2\lambda_2(k - h_d))\mu_2 + (k - h_d)\lambda_2^2 - c\lambda_2,
\]

\[
\mu_2 = \frac{2\lambda_2(k - h_d) \pm \sqrt{(2\lambda_2(k - h_d))^2 - 4(k - h_d)((k - h_d)\lambda_2^2 - c\lambda_2)}}{2(k - h_d)},
\]

\[
\mu_2 = \lambda_2 + \frac{c\lambda_2}{k - h_d}.
\]

Hence, if \(\mu_1 > \lambda_2 + \sqrt{\frac{c\lambda_2}{k-h_d}},\) then choose \(\mu_2 = \lambda_2 + \sqrt{\frac{c\lambda_2}{k-h_d}}\). This only holds for \(h_d < k\). Note that it is never optimal to decrease the capacity if \(h_d \geq k\), because then the cost for decreasing capacity is larger than savings on capacity cost.

If \(\lambda_2 + \sqrt{\frac{c\lambda_2}{k+h_u}} \leq \mu_1 \leq \lambda_2 + \sqrt{\frac{c\lambda_2}{k-h_d}}\), then choose \(\mu_2\) equal to \(\mu_1\). In this case, cost for changing capacity does not compensate for reduced capacity or sojourn cost.

Summarizing, the optimal service rate for the second period, for given demand rate and service rate for the first period, is equal to

\[
\mu_2^* := \begin{cases} 
\lambda_2 + \sqrt{\frac{c\lambda_2}{k+h_u}} & \text{if } \mu_1 < \lambda_2 + \sqrt{\frac{c\lambda_2}{k-h_d}}, \\
\lambda_2 + \sqrt{\frac{c\lambda_2}{k-h_d}} & \text{if } \mu_1 > \lambda_2 + \sqrt{\frac{c\lambda_2}{k-h_d}} \text{ and } h_d < c, \\
\mu_1 & \text{else.}
\end{cases}
\]

### 2.6 Optimizing capacity in period 1

In section 2.5, we described an expression for the optimal service rate in period 2. We can use this expression to create a function for period 1 that need to be optimized only for \(\mu_1\), instead of for both \(\mu_1\) and \(\mu_2\):

\[
T'(\mu_1, \lambda_2) = k\mu_1 + c\frac{\rho_1}{1 - \rho_1} + k\mu_2^* + c\frac{\rho_2}{1 - \rho_2} + h_u(\mu_2^* - \mu_1)^+ + h_d(\mu_1 - \mu_2^*)^+,
\]

where \(\mu_2^*\) depend on \(\mu_1\), and \(\lambda_2\) is unknown original problem. Because a probability distribution for \(\lambda_2\) is given, we can calculate the corresponding expected cost for any choice of \(\mu_1\) and for any possible scenario for \(\lambda_2\). A Mathematica script to compute the optimal capacity for a given instance can be found in appendix A.
2.7 Analysis

Since we have a script to evaluate instances of our capacity assignment problem, we can now analyze different instances and compare the results.

2.7.1 Basic example

For the analysis of different instances, we start with a basic example. By changing parameters in this basic example, we can evaluate the behavior of the optimal solution and corresponding cost.

For our basic example, we choose the following parameters:

- \( k = 2 \) (factor for capacity cost)
- \( c = 10 \) (factor for sojourn cost)
- \( \lambda_1 = 10 \)
- \( \lambda_2 = \begin{cases} 
5 & \text{with probability 0.7} \\
10 & \text{with probability 0.15} \\
15 & \text{with probability 0.15} 
\end{cases} \)
- \( h_u = 5 \)
- \( h_d = 1 \)

The optimal solution for this instance is \( \mu_1 = 17.1 \) with corresponding cost 93.5, but it is more interesting to see what happens with this solution if we change one or more parameters.

2.7.2 Costs for sojourn time and increasing capacity

In figure 2, the optimal value for \( \mu_1 \) is shown for different values of \( c \) (sojourn cost) and \( h_u \) (cost for increasing capacity).

\( c = 0 \)

For \( c = 0 \), the optimal value for \( \mu_1 \) is equal to 10 for \( h_u \leq 19 \) and equal to 15 for \( h_u \geq 19 \). This can be explained by considering the problem as a newsvendor problem. Roughly speaking, there are two reasonable options for \( \mu_1 \): 10 and 15. Because \( \lambda_1 = 10 \), \( \mu_1 \) cannot be smaller than 10 and because no sojourn cost is counted, 10 is a reasonable value for \( \mu_1 \). It can also make sense to choose \( \mu_1 = 15 \), because it is possible that \( \lambda_2 = 15 \) and increasing capacity after the first period is expensive. Before we explain why it is never optimal to choose values between 10 and 15 for \( \mu_1 \), we first show how to determine which of the two mentioned solutions (\( \mu_1 = 10 \) or \( \mu_1 = 15 \)) is best.

- If \( \mu_1 = 10 \) and \( \lambda_2 = 10 \), or if \( \mu_1 = 15 \) and \( \lambda_2 = 15 \), it is (afterwards) obvious that the right choice is made. If \( \mu_1 = 10 \) and \( \lambda_2 = 15 \), or if \( \mu_1 = 15 \) and \( \lambda_2 = 10 \), the choice for \( \mu_1 \) appears not to be optimal and compared with the optimal choice for \( \mu_1 \), some additional costs need to be paid.

- If \( \mu_1 = 10 \) and \( \lambda_2 = 15 \), we must increase the capacity by 5 units (to \( \mu_2 = 15 \)) after period 1. This gives extra cost because we need to increase the capacity, but we also save little money by having less capacity in the first period. All together, it gives an extra cost of \( h_u - k = h_u - 2 \) per unit, relative to the choice that \( \mu_1 = 15 \). This is also called the underage cost.
Figure 2: Optimal service rate for variable $c$ and $h_u$

- If $\mu_1 = 15$ and $\lambda_2 = 10$, we can decrease the service rate for period 2 to 10 with extra cost $h_d + k = 1 + 2 = 3$ per unit (relative to the choice that $\mu_2$ is equal to 10). We also should decrease the service rate if $\mu_2 = 5$, but this is inevitable, and therefore not taken into account in this analysis, since $\mu_1$ must be at least 10. The additional costs that need to be paid in this situation are called the overage cost.

To compute the optimal value of $\mu_1$, we have to take the above mentioned additional cost for wrong choices into account, but also the probability that these additional cost must be paid. If $\mu_1 = 10$, we need to increase the service rate after period 1 with probability 0.15, with additional cost $5(h_u - 2) = 5h_u - 10$ (relative to the case where $\mu_1 = 15$). On the other hand, if $\mu_1 = 15$, we should decrease the service rate after period 1 with probability 0.85, with additional cost $5 \cdot 3 = 15$ (note that we omit the cost for decreasing the service rate to 5). Hence, for least expected total costs, it is optimal to choose $\mu_1 = 10$ if $0.15(5h_u - 10) \leq 0.85 \cdot 15$, which gives $h_u \leq 19$. In the same way, it is optimal to choose $\mu_1 = 15$ if $h_u \geq 19$.

Because all costs are linear when $c = 0$, it is never optimal to choose other values than 10 and 15 for $\mu_1$. Per unit increment of $\mu_1$, the expected additional profit or loss is the same. Therefore, if for example $\mu_1 = 12$ gives lower expected cost than $\mu_1 = 10$, then $\mu_1 = 15$ gives even lower expected cost.
Newsvendor problem

In terms of a newsvendor problem, we have underage cost $c_u = h_u - 2$ and underage cost $c_o = 3$.

For newsvendor problems, it is known that the optimal solution is $F^{-1}\left(\frac{c_u}{c_o + c_u}\right)$, with $F^{-1}$ the inverse distribution function of the demand rate. For our example, we have the following distribution function for the demand rate:

$$F(x) = P(A_2 \leq x) = \begin{cases} 
0 & \text{if } x < 5, \\
0.7 & \text{if } 5 \leq x < 10, \\
0.85 & \text{if } 10 \leq x < 15, \\
1 & \text{if } x \geq 15.
\end{cases}$$

For $h_u \leq 19$, $\frac{c_u}{c_o + c_u} = \frac{h_u - 2}{h_u + 1} \leq 0.85$, so then the optimal solution is $\mu_1 = F^{-1}\left(\frac{c_u}{c_o + c_u}\right) = 10$ (or even smaller, but for period 1 we need a service rate of at least 10) and for $h_u \geq 19$, $\frac{c_u}{c_o + c_u} = \frac{h_u - 2}{h_u + 1} \geq 0.85$, so then the optimal solution is $\mu_1 = F^{-1}\left(\frac{c_u}{c_o + c_u}\right) = 15$. Note that for $h_u = 19$, both solutions $\mu_1 = 10$ and $\mu_1 = 15$ are optimal.

$c > 0$

For larger values of $c$, sojourn cost influences the optimal service rate with the result that there is a more gradual transition when $h_u$ increases. For $h_u = 0$ and a small value for $c$, the optimal service rate is a little greater than 10 because with a demand rate of 10, some extra capacity is needed to keep the sojourn cost within bounds. For larger values of $c$, even more excessive capacity is profitable because of more sojourn cost. The optimal service rate is increasing in $h_u$, because it becomes more expensive to increase the service rate after period 1. In contrast to the case where $c$ is equal to zero, we now also save sojourn cost by increasing $\mu_1$. This is the reason of the more gradual transition in contrast to the jump at $h_u = 19$ for $c = 0$.

Horizontal part of the graph

In all cases, the optimal value does not increase anymore after $h_u$ exceeds a certain value (which value depends on the fixed parameters). If increasing the service rate after period 1 is too expensive, it is set to a high value in advance such that this increase is not necessary, even if the demand rate in period 2 reaches the highest possible value. The point from where the graph is horizontal, is more to the left for larger values of $c$. This is because $h_u$ has relative less influence on the total cost for larger $c$.

Shape of the graphs

Before the graph runs horizontally, the optimal value for $\mu_1$ increases faster and faster when $h_u$ increases. Indeed, if $h_u$ increases, the expected cost for increasing the service rate after period 1 increases. By increasing $\mu_1$, not only this expected cost for increasing capacity, but also the sojourn cost, decreases. When $\mu_1$ increases further, the sojourn cost decreases slower, with the result that it is not efficient to increase $\mu_1$ too much immediately (because there is also cost charged for that). Because all (expected) costs are linear, except for the sojourn cost, it is optimal to increase $\mu_1$ until the effect of decreasing sojourn cost, combined with the effect of expected increasing cost, does not outweigh the increasing capacity cost anymore.

When $\mu_1$ increases, the sojourn cost decreases, but this will happen slower if $\mu_1$ increases further. Because the effect of $\mu_1$ on the sojourn cost increases less for high values of $\mu_1$, we can increase $\mu_1$ faster because it takes longer until the effect is too small to be profitable, and therefore the optimal
value for $\mu_1$ increases faster when $h_u$ increases further.

In general, for greater values of $c$, sojourn cost plays a more important role and more capacity is needed to minimize total cost. Therefore, for any value of $h_u$, the optimal value for $\mu_1$ is increasing in $c$.

**Kinks in the graphs**

At the left side, there is a small kink in the graphs if $c > 0$ (more clearly to see for larger values of $c$). On the left side of this point in our example, $\mu_1$ needs to be increased when $\lambda_2$ turns out to be equal to 10 after period 1, whereas at the right side it does not (because it is too expensive compared to the savings on the sojourn cost). For both scenarios, total cost and optimal $\mu_1$ increase with a different speed in $\mu_1$. If more scenarios are possible for the second period, there would be more kinks in the graphs, as can be seen in figure 3.

![Figure 3: Instance with 10 scenarios for the second period](image)

When we take a look at the total cost corresponding to the optimal choices for $\mu_1$ and $\mu_2$ in the example with two scenarios (figure 4), we see that total cost is increasing in $h_u$ (as expected). Of course, if $h_u$ is large enough and the optimal $\mu_1$ does not increase anymore, the expected cost also does not increase anymore. Also, there is a more gradual increase for larger values of $c$.

For variable $h_d$, the same analysis holds, but now $\mu_1$ is decreasing in $h_d$.

### 2.7.3 Capacity cost

As can be seen in figure 5, the optimal value for $\mu_1$ is decreasing in $k$ (capacity cost). Indeed, more cost for capacity leads to savings on the capacity. Because the service rate has to be greater than the demand rate, $\mu_1$ approaches $\lambda_1$ asymptotically for $k \to \infty$. For $k = 0$, it is optimal to set the service rate at infinity, because capacity is free and sojourn cost is saved in this way.
2.7.4 Sojourn cost

The optimal value for $\mu_1$ is increasing in $c$ (sojourn cost). Moreover, $\mu_1 \to \infty$ for $c \to \infty$, because more and more capacity is needed for increasing sojourn cost. For $c = 0$, the optimal service rate is equal to 10, as in the basic example. This is visualized in figure 6.
2.8  Average scenario

Instead of comparing all possible scenarios for the second period, as we did in the previous subsections, it is also possible to consider only one ‘average scenario’ without uncertainty. This average scenario is a weighted average of all possible scenarios. In particular, when many scenarios are possible for period 2, it can be time-saving to combine all these scenarios in one single scenario. In this subsection we investigate how well the method of only considering a weighted average scenario works.

When comparing the different methods, we use the methods only for determining the solution. To compute the corresponding costs, we always take into account all possible scenarios and their probabilities.

2.8.1  Some examples

First, we take a look at two examples to show that considering an average scenario works well in some cases, but does not in other cases.

Example 1
The first example has quite realistic and more or less symmetric parameters:

- \( k = 5 \)
- \( c = 1 \)
- \( \lambda_1 = 10 \)
- \( \lambda_2 = \begin{cases} 8 & \text{with probability } p \\ 12 & \text{with probability } 1 - p \end{cases} \)
- \( h_u = 8 \)
• $h_d = 4$

When we compare the optimal $\mu_1$ and cost (figures 7, 8, 9 and 10), we see that there is not much difference in the results between the two approaches. For multiple scenarios, some more safety capacity is optimal because increasing the service rate after period 1 is relatively expensive. When considering the average scenario, the demand rate for the second period decreases gradually and without uncertainty, so also the optimal value for $\mu_1$ decreases gradually for increasing $p$. The graph for multiple scenarios does not decrease gradually, because both scenarios need to be taken into account for all $0 < p < 1$. The kink in the graph shows the point from where the service rate needs to be increased after period 1 when $\mu_2$ appears to be equal to 12.

For all $p$, the difference in optimal $\mu_1$ is smaller than 10% and for optimal cost even smaller than 1.2%. Hence, the method of considering only the average scenario works well for this example. Note that for $p = 0$ and $p = 1$, there is no difference at all because there is no uncertainty for the second period in both cases.

![Figure 7: Optimal $\mu_1$ for both approaches (example 1)](image-url)
Example 2
Let us now consider an example with more asymmetric and less realistic parameters:

- $k = 5$
- $c = 1$
- $\lambda_1 = 10$
- $\lambda_2 = \begin{cases} 8 & \text{with probability } p \\ 58 & \text{with probability } 1 - p \end{cases}$
- $h_u = 8$
- $h_d = 4$
For this example, there is more difference in the solutions of both approaches (see figures 11, 12, 13 and 14).

First, we see that the optimal value for $\mu_1$ is 60 for $p < 0.26$ and 11 for $p > 0.4$, approximately (see section 2.7.2 for an explanation). The optimal $\mu_1$ when considering only one scenario decreases gradually for increasing $p$, so this approximation is too low for small $p$ and too high for large $p$.

For the expected cost of both approaches, the difference goes up to 20%. For $\mu_1$, the difference is even more than 200%. We can conclude that the method which considers an average scenario does not make sense in this example.

One of the reasons that the method does not work well for this example, is the fact that it does not take the asymmetry of the costs into account. It only straightens the asymmetry of the possible demand rates, but for example, if increasing capacity after period 1 is expensive, then higher demand rates should get more weight in the average. Also, we have seen in section 2.7.2, that a
gradual change for gradually changing parameters is not optimal in many cases, which seems also the case in the second example. However, the approximation always gives such a gradual change (because parameters change gradually and there is no uncertainty in the demand rate) such that results become worse.

We conjecture that the approximation works better if the parameters are more symmetric. This will be investigated in the next part of this section.
2.8.2 Increasing variability

We consider an instance of our problem with the following parameters ($0 < a \leq 1$):

- $k = 5$
- $c = 1$
- $\lambda_1 = 10$
- $\lambda_2 = \begin{cases} a\lambda_1 & \text{with probability } \frac{1}{a+1} \\ \frac{1}{a}\lambda_1 & \text{with probability } \frac{a}{a+1} \end{cases}$
- $h_u = 8$
- $h_d = 4$

In this example, the (weighted) average demand rate is equal to 10 for any $a > 0$, but for $a$ closer to zero there is more variability between the different scenarios. The results are shown in figures 15 and 16.

Because the weighted average demand rate for the second period is equal to 10 for any positive $a$, this approach always gives the same solution (independent of $a$). In figure 15, this is represented by the yellow horizontal line. When all scenarios are considered, the optimal value of $\mu_1$ depends on $a$. For $a$ close to 0, a small $\mu_1$ is optimal because $\lambda_2$ is small with high probability. For $0.5 < a < 1$, a larger $\mu_1$ is optimal because the demand rate increases after period 1 with higher probability. Because increasing the service rate after period 1 is relative expensive, it is better to choose a larger $\mu_1$ for those $a$.

For $a$ closer to one, the optimal $\mu_1$ is decreasing because also for the optimistic scenario (more demand in period 2), the demand in period 2 is not very high. The left kink in the graph is the point from where the service rate does not need to be reduced if the demand rate decreases after period 1.
In figure 16, we see that there is more cost for smaller values of \( a \). For smaller \( a \) it is likely that there is much demand in period 2, which gives more cost. However, for very small values of \( a \) (\( a < 0.005 \)), despite of the fact that there can be much demand, the probability of increasing demand becomes so small, that there is less cost for decreasing \( a \).

The differences between the two approaches are very small, as well for the optimal \( \mu_1 \) as for the cost. Hence, the difference between the two approaches is not only the result of asymmetry in the distribution for \( \lambda_2 \). Because the great demand in example 2 had a great probability, there was also much difference in costs.
2.9 Conclusions

We have investigated a queueing system with one queue and one server. In our model, we have two time periods: for the first time period the demand rate was assumed to be known, while we only had a distribution for the demand rate in the second period. We have optimized the service rates for both periods, such that we do not have too much expensive capacity, but enough capacity to serve all customers in a reasonable time.

For the second period, the optimal service rate can be given explicitly as a function of \( \mu_1 \) and \( \lambda_2 \). Using this, the optimal service rate for period 1 can be calculated numerically. We used Mathematica to do this.

We saw that for small systems, the optimal solution can be found within a few seconds. We analyzed the behavior of the optimal service rate and corresponding cost as function of different parameters and made a relation with Newsvendor problems. Also, we compared the optimal solution with a method where the capacity was optimized by using a weighted average parameter for the demand rate. In some cases the solutions were almost the same, but there are also examples for which using the average parameter for the demand rate only gives 20% more cost than our method. Therefore, at least for systems that are not too large, our method works very well.
3 Two customers, one server - First Come, First Served

In section 2, we investigated the optimal service rate for a queueing system with only one type of arrivals. Now, we extend this model to a system with two different customer types. How do the different customer types influence each other and what is the effect on the optimal service rates?

3.1 Model

Consider a queueing system in which two different types of customers arrive (or customers who ask for two different types of products). The customers are served by one server according to a First Come, First Served (FCFS) policy. The arrivals are according to a Poisson process for both customer types, but with different parameters. Again, we consider two time periods, for which the demand rates are known for the first period, but not for the second period. For period 2, only a (discrete) distribution for the demand rates is available. The goal is to minimize the total expected cost, by choosing optimal service rates for both periods. Here, the ratio of the service rates for the two customer types is fixed, (so for example, it is given that customers of type 1 are served twice as fast as type 2 customers, on average. This is clarified in section 3.2). In both period 1 and 2, costs are involved for the amount of capacity (service rate) and the number of customers in the queue. Also, costs are made for changing the service rates after period 1. The exact cost function is given in section 3.3.

3.2 Notation

In this section, we use the following notation:

- \( \lambda_i^{(j)} \) is the arrival rate of type \( j \) customers in period \( i \) for \( i = 1, 2 \) and \( j = 1, 2 \). The total arrival rate in period \( i \) is equal to \( \lambda_i = \lambda_i^{(1)} + \lambda_i^{(2)} \).

- \( \Lambda^{(j)}_2 \) is the random variable for the unknown demand rate \( \lambda_2^{(j)} \) in the second period for type \( j \) customers.

- The service rate for type 1 customers is \( \alpha_1 \mu_i \) and the service rate for type 2 customers is \( \alpha_2 \mu_i \) in period \( i \) (\( i = 1, 2 \)). Here, \( \alpha_1 \) and \( \alpha_2 \) are fixed and \( \mu_1 \) and \( \mu_2 \) need to be optimized. Note that the sum of \( \alpha_1 \) and \( \alpha_2 \) is not necessarily equal to 1.

- The occupation rate of the server is equal to \( \rho_i = \frac{\lambda_i}{\mu_i} \) in period \( i \) for \( i = 1, 2 \).

- \( c_i \) is the cost for waiting time (per time unit) for each product in period \( i \) for \( i = 1, 2 \).

- The capacity cost in period \( i \) is equal to \( k \mu_i \) for \( i = 1, 2 \).

- \( h_u \) and \( h_d \) are the costs for increasing and decreasing the service rate by one unit between period 1 and 2.

- \( L_i^Q \) is the mean number of customers in the queue in period \( i \) for \( i = 1, 2 \).

- \( W_i \) is the expected waiting time for a customer in period \( i \) for \( i = 1, 2 \).

The parameters \( k \) and \( c \) could be different in both periods, but to simplify notation, we choose them to have the same value for both periods in our model.
3.3 Cost function

Customers arrive according to Poisson processes and are served by a FCFS policy. Therefore, the service times have a hyperexponential distribution. Indeed, when considering the system at an arbitrary moment in period \(i\), with probability \(\lambda_i^{(1)}(1 + \lambda_i^{(2)})\) the next arrival is a type 1 customer with service rate \(\alpha_1 \mu_i\) and with probability \(\lambda_i^{(2)}(1 + \lambda_i^{(2)})\) it is a type 2 customer with service rate \(\alpha_2 \mu_i\).

Since we can consider the queue as an \(M/H/1\) queue, we can compute the waiting time in the same manner as in [4]. To keep this work self-contained, we also give the derivation here.

The waiting time for a general \(M/G/1\) queue [1] is equal to

\[
E(W) = \frac{\rho E(R)}{1 - \rho},
\]

where \(E(R)\), the mean residual service time, is equal to

\[
E(R) = \frac{E(B^2)}{2E(B)}.
\]

For a hyperexponential distribution of the service time \(B\), this gives

\[
E(R) = \frac{\sum_{i=1}^{n} p_i/\tilde{\mu}_i^2}{\sum_{i=1}^{n} p_i/\tilde{\mu}_i},
\]

where \(p_i\) is the probability that \(B\) is exponential distributed with parameter \(\tilde{\mu}_i\).

For our model, in period \(i(i = 1, 2)\) we have \(p_j = \frac{\lambda_i^{(j)}}{\lambda_i^{(1)} + \lambda_i^{(2)}}\) and \(\tilde{\mu}_j = \alpha_j \mu_i\). This leads to

\[
E(W_i) = \frac{\rho_i E(R)}{1 - \rho_i} = \frac{\rho_i}{1 - \rho_i} \left( \frac{p_1}{\tilde{\mu}_1} + \frac{p_2}{\tilde{\mu}_2} \right)
\]

\[
= \frac{\lambda_i^{(1)}}{\alpha_1 \mu_i} + \frac{\lambda_i^{(2)}}{\alpha_2 \mu_i} - \frac{\lambda_i^{(1)} \lambda_i^{(2)} \alpha_2 \mu_i}{\alpha_1 \mu_i \alpha_2 \mu_i (\lambda_i^{(1)} + \lambda_i^{(2)})} + \frac{\lambda_i^{(2)}}{\alpha_2 \mu_i (\lambda_i^{(1)} + \lambda_i^{(2)})}
\]

\[
= \frac{\lambda_i^{(1)}}{\alpha_1 \mu_i} + \frac{\lambda_i^{(2)}}{\alpha_2 \mu_i} - \frac{\lambda_i^{(1)}}{\alpha_1 \mu_i} \lambda_i^{(2)} - \frac{\lambda_i^{(2)}}{\alpha_2 \mu_i} \lambda_i^{(1)}
\]

\[
= \frac{\lambda_i^{(1)}}{\alpha_1 \mu_i} + \frac{\lambda_i^{(2)}}{\alpha_2 \mu_i}.
\]

We can now define the expected cost \(T\) per time unit with the following function:
\[ T(\mu_1, \mu_2, \lambda_2) = k\mu_1 + cE(L_1^2) + k\mu_2 + cE(L_2^2) + h_u(\mu_2 - \mu_1)^+ + h_d(\mu_1 - \mu_2)^+ \]  
\[ = k\mu_1 + c(\lambda_1^{(1)} + \lambda_1^{(2)})E(W_1) + k\mu_2 + c(\lambda_2^{(1)} + \lambda_2^{(2)})E(W_2) \]  
\[ + h_u(\mu_2 - \mu_1)^+ + h_d(\mu_1 - \mu_2)^+ \]  
\[ = k\mu_1 + c(\lambda_1^{(1)} + \lambda_1^{(2)}) \frac{\lambda_1^{(1)}}{\alpha_1\mu_1} + \frac{\lambda_1^{(2)}}{\alpha_2\mu_2} \]  
\[ + k\mu_2 + c(\lambda_2^{(1)} + \lambda_2^{(2)}) \frac{\lambda_2^{(1)}}{\alpha_1\mu_2} + \frac{\lambda_2^{(2)}}{\alpha_2\mu_2} \]  
\[ + h_u(\mu_2 - \mu_1)^+ + h_d(\mu_1 - \mu_2)^+. \]

### 3.4 Optimizing capacity in period 2

Our objective is to minimize the expected total cost as defined in section 3.3, by optimizing the service rates in both periods \((\mu_1 \text{ and } \mu_2)\).

Let us first consider the service rate in the second period. In our previous model, we formulated an explicit expression for \(\mu_2\) with the assumption that \(\lambda_2\) is known because \(\mu_2\) is determined at the beginning of period 2. In this model, expression for the waiting time is more difficult because of the hyperexponential distribution instead of the exponential distribution. Therefore, an exact expression is difficult to give and we solve the equation numerically.

Again, if the capacity for the first period is known, we need to consider the capacity cost for the second period, the waiting time for the second period and the cost for increasing/decreasing capacity between period 1 and 2. When increasing the capacity after period 1, the capacity cost for period 2 and cost for changing capacity increase, while the sojourn cost for period 2 decreases. Because the capacity cost and increasing cost increase proportionally to the amount of capacity, and because we have good beliefs that the waiting cost decreases in a convex way, it is optimal to increase the capacity until the capacity cost and increasing cost increase faster than the waiting cost decreases. In other words, until the derivative of \(k\mu_2 + c(\lambda_2^{(1)} + \lambda_2^{(2)}) \frac{\lambda_2^{(1)}}{\alpha_1\mu_2} + \frac{\lambda_2^{(2)}}{\alpha_2\mu_2} + h_u(\mu_2 - \mu_1)\) with respect to \(\mu_2\) is equal to zero. This gives

\[ k + h_u - c \left( \lambda_2^{(1)} + \lambda_2^{(2)} \right) \frac{1}{\left(1 - \frac{\lambda_1^{(1)}}{\alpha_1\mu_2} - \frac{\lambda_1^{(2)}}{\alpha_2\mu_2} \right) \lambda_2^{(2)}}. \]  
\[ \left( 1 - \frac{\lambda_2^{(1)}}{\alpha_1\mu_2} - \frac{\lambda_2^{(2)}}{\alpha_2\mu_2} \right) \left( \frac{-2\alpha_1\alpha_2\mu_2(\alpha_1^2\lambda_1^{(1)} + \alpha_1^2\lambda_2^{(2)})}{\alpha_1^2\alpha_2^2\mu_2^2} \right) + \right. 
\[ \left. \left( \frac{\lambda_2^{(1)}}{\alpha_1\mu_2} + \frac{\lambda_2^{(2)}}{\alpha_2\mu_2} \right) \left( \frac{(\alpha_2\lambda_2^{(1)} + \alpha_1\lambda_2^{(2)})\alpha_1\alpha_2}{\alpha_1^2\alpha_2^2\mu_2^2} \right) \right) = 0. \]
Because this equation is difficult to solve for $\mu_2$, we let Mathematica do it numerically. In this way, we find a certain value. Now, just as in section 2.6, if $\mu_1$ is smaller than this value, then set $\mu_2$ equal to this value. By increasing the service rate from $\mu_1$ to this optimal $\mu_2$, increasing costs (for capacity and increasing service rate) are less the savings for decreasing waiting times. It is not profitable to increase the service rate further because this will give not enough savings for the increased extra costs. If $\mu_1$ is already greater than the solution of the equation, it is not profitable to increase the service rate at all.

Remark: to find the appropriate solution of (30) in Mathematica, we add the constraint that $\rho$ must be smaller than one.

For the consideration of decreasing the service rate after period 1, we have a similar equation. It is profitable to decrease the service rate as long as cost for capacity decreases faster as cost for waiting time and cost for decreasing the service rate increase. In other words, decrease the service rate until the derivative of $k\mu_2 + c(\lambda_2^{(1)} + \lambda_2^{(2)})\frac{\alpha_2^{(1)}\mu_2 + \alpha_2^{(2)}\mu_2}{\alpha_1\mu_2 - \alpha_2\mu_2} + h_d(\mu_1 - \mu_2)$ with respect to $\mu_2$ is equal to zero. This gives

\[
\frac{k - h_d - c}{\left(1 - \frac{\lambda_2^{(1)}}{\alpha_1\mu_2} - \frac{\lambda_2^{(2)}}{\alpha_2\mu_2}\right)}\left(1 - \frac{\lambda_2^{(1)}}{\alpha_1\mu_2} - \frac{\lambda_2^{(2)}}{\alpha_2\mu_2}\right)^2 - 2\frac{\lambda_2^{(1)}}{\alpha_1\mu_2} + \frac{\lambda_2^{(2)}}{\alpha_2\mu_2} + h_d(\mu_1 - \mu_2) \right) + h_d(\mu_1 - \mu_2) = 0,
\]

provided that $h_d < k$. If $h_d > k$, it is never profitable to decrease the service rate because the savings on capacity cost are less than the cost for decreasing the service rate, and the increasing cost for waiting time gives even more total cost.

After solving this equation by using Mathematica, we have the value to which it is profitable to decrease the service rate after period 1. If $\mu_1$ is already smaller than this value, it is not profitable to decrease the service rate at the beginning of the second period.

If $\mu_1$ is greater than the solution of (30), but smaller than the solution of (33), it is optimal to set $\mu_2$ equal to $\mu_1$. The service rate is already close enough to the optimum value to keep it the same.

### 3.5 Optimizing capacity in period 1

Since we do not have an explicit expression for the optimal service rate in period 2, we need to include the method for finding the optimal $\mu_2$ in our program. We get the following function that needs to be optimized for $\mu_1$: 
\[
T(\mu_1, \lambda_2^{(1)}, \lambda_2^{(2)}) = \mu_1 + c(\lambda_1^{(1)} + \lambda_1^{(2)}) \frac{\lambda_1^{(1)}}{\alpha_1^{(1)}} + \frac{\lambda_1^{(2)}}{\alpha_2^{(2)}} - \frac{\lambda_1^{(1)}}{\alpha_1^{(1)}} - \frac{\lambda_1^{(2)}}{\alpha_2^{(2)}}
\]
\[
+ k\mu_2 + c(\lambda_1^{(1)} + \lambda_1^{(2)}) \frac{\lambda_1^{(1)}}{\alpha_1^{(1)}} + \frac{\lambda_1^{(2)}}{\alpha_2^{(2)}} - \frac{\lambda_1^{(1)}}{\alpha_1^{(1)}} - \frac{\lambda_1^{(2)}}{\alpha_2^{(2)}}
\]
\[
+ h_u(\mu_2^* - \mu_1)^+ + h_d(\mu_1 - \mu_2^*)^+,
\]

with $\mu_2^*$ the optimal capacity as a function of $\mu_1$, $\lambda_2^{(1)}$, and $\lambda_2^{(2)}$ as described in section 3.4.

It appears that despite of this method for $\mu_2^*$ instead of an explicit expression, at least small problems (two scenarios for each demand rate in period 2) can be solved within a few seconds by Mathematica. This gives hope that also larger instances can be solved within reasonable time. The program can calculate the expected total cost for each possible scenario. Combined with the probability of each scenario, the general expected total cost can be computed and minimized for $\mu_1$. The program is given in appendix B.

3.6 Analysis

Since we have a script to evaluate instances of our capacity assignment problem with two customer types, we can now analyze different instances and compare the results.

3.6.1 Basic example

For the analysis of different instances, we start with a basic example. By changing parameters in this basic example, we can evaluate the behavior of the optimal solution and corresponding cost. For our basic example, we choose the following parameters:

- $k = 2$
- $c = 1$
- $\lambda_1^{(1)} = 100$
- $\lambda_1^{(2)} = 200$
- $\lambda_2^{(1)} = \begin{cases} 90 & \text{with probability } 0.6 \\ 110 & \text{with probability } 0.4 \end{cases}$
- $\lambda_2^{(2)} = \begin{cases} 220 & \text{with probability } 0.8 \\ 180 & \text{with probability } 0.2 \end{cases}$
- $\alpha_1 = 2/3$
- $\alpha_2 = 1/3$
- $h_u = 5$
• $h_d = 1$

The optimal solution for this instance is $\mu_1 = 811.3$ with corresponding expected cost 3364, but it is more interesting to see what happens with this solution if we change one or more parameters.

### 3.6.2 Different ratios for the service rates

The ratio of the service rates for the two customer types is fixed, and given by $\alpha_1$ and $\alpha_2$. We investigate what happens if we change this ratio. To do so, we let $\alpha_1$ run from 0 to 1 and set $\alpha_2$ equal to $1 - \alpha_1$ (in the original problem, the sum of $\alpha_1$ and $\alpha_2$ is not necessarily equal to 1).

Figures 17 and 18 show the optimal service rates and corresponding costs for the different ratios.

![Figure 17: Optimal $\mu_1$ for different service ratios](image)
The capacity of the server needs to be divided over the two customer types. If $\alpha_1$ is small (close to 0), the optimal $\mu_1$ tends to infinity because only a small part of the capacity is used for type 1 customers. The same holds for $\alpha_1$ close to 1. Then, $\alpha_2$ is close to zero and many capacity is needed to serve all the type two customers.

For $\alpha_1$ not close to 0 or 1, the optimal capacity is much lower. Least capacity is needed for an $\alpha_1$ value smaller than 0.5 ($\alpha_1$ equal to 0.4), because the demand rate for type 2 customers is greater than the demand rate for type 1 customers. Therefore, less capacity is needed as a greater part can be used for type 2 customers.

In the graph for the corresponding (optimal) total cost, we see the same shape: if the capacity is divided well over the customer types, then the expected total cost is lower.
3.6.3 Demand rate for queue 1 in period 1

In our model, the ratio of the service rates for the two queues is assumed to be fixed. To reduce costs, this ratio should match the demand rates of the queues. For example, if there are much more type 1 customers than type 2 customers, it would be convenient if the server is faster in serving customer 1 types. In figure 19, we see what the optimal service rate is for increasing $\lambda_1^{(1)}$.

![Figure 19: Optimal $\mu_1$ for different demand rates](image)

For $\lambda_1^{(1)}$ smaller than 110, the optimal value for $\mu_1$ is 810. This is the capacity needed to serve the type 2 customers fast enough. When the value of $\lambda_1^{(1)}$ exceeds 110, the optimal $\mu_1$ is increasing in $\lambda_1^{(1)}$, because more capacity is needed to serve all customers of type 1 fast enough.
3.6.4 Cost for increasing capacity

When we increase $h_u$ in the model with two queues and one server, the optimal server capacity is increasing as shown in figure 20. This is the same behavior as in the model with only one queue, so for an analysis we refer to section 2.7.2.

![Figure 20: Optimal $\mu_1$ for different $h_u$](image)

3.7 Conclusions

We extended the model by adding a queue, where orders arrive with their own demand and service rate. There is still one server, which orders the both queues according to a First Come, First Served policy.

For the model with two queues, expressions and functions become much more difficult. Therefore, in contrast to the model in the previous section, $\mu_2$ can not explicitly be expressed as a function of $\mu_1$ and the demand rates for the second period anymore. However, also when calculating both $\mu_1$ and $\mu_2$ numerically, Mathematica can find the optimal solution within a few seconds for small systems. This makes our models also appropriate for systems with different product types.

We also analyzed the behavior of the optimal service rate and corresponding cost, and saw for example that the ratio of the service rates for both queues (which we assumed to be fixed) can have a major impact on the optimal solution.
4 Multiple customers, one server - polling model

Until now, we made the assumption that the customers are served according to a FCFS policy. Here, the server could change customer type without having any switchover time or switchover cost. However, in practice it usual takes some time and/or cost to adjust the server for serving other customer/product types. Therefore, it is better not to work according to a FCFS policy, but to model the system as a polling model. In a polling model, there is one server which serves multiple queues with each its own distributions for arrivals and services.

In a polling model, the server can visit the queues according to different policy. One policy is the exhaustive policy, at which the server serves customers in one queue until this queue is empty. After that, the server will start serving the next queue. Another policy is the gated policy. In this policy the server serves all customers which are in the queue at the moment the server arrives.

In our model, we use the $k$-limited policy: the server serves $k_i$ customers when visiting queue $i$ or serves until the queue becomes empty, whichever occurs first. Here, $k_i$ is called the service limit. Advantages of this policy is that cycle times are limited, such that it never takes too long before certain customer types are served again, and that priorities can be assigned to certain queues by choosing appropriate service limits. A disadvantage is that relatively not many explicit expressions are known for this polling model, so we must use some approximations.

Remark: strictly speaking, the model in the previous section is actually also a polling model. However, in mathematics, the exhaustive, gated and $k$-limited policy are more associated with polling models.

4.1 Model

Consider a polling model with $N$ queues. We assume that the server visit the queues in cyclic order $1, 2, ..., N$. In each queue, customers are served in exponential distributed and independent service times. As we have a $k$-limited polling model, the server leaves queue $i$ after $k_i$ customers are served or when the queue becomes empty, whichever occurs first.

Because a $k$-limited polling model is a relative difficult model, we start with considering only one time period (instead of the two time periods in the previous models) with given demand rates. After the model with one time period, we investigate a model with two time periods and unknown demand rates.

As in the previous model, in which customers were served FCFS, the ratio of the service time is fixed.

Our goal is again to minimize total expected cost by optimizing the speed of the server. As in the previous model, in which customers were served FCFS, the ratio of the service times are fixed.

The total cost in this model consists of capacity cost and cost for waiting time of the customers.
4.2 Notation

In this section, we use the following notation:

- \( \lambda_i \) is the arrival rate of the customers in queue \( i \) (type \( i \) customers) for \( i = 1, 2, ..., N \).
- \( \alpha_i \mu \) is the service rate for queue \( i \) for \( i = 1, 2, ..., N \) (not that \( \sum_{i=1}^{N} \alpha_i \) is not necessarily equal to 1).
- \( \rho_i = \frac{\lambda_i}{\alpha_i \mu} \) is the fraction of the time at which the server is serving type \( i \) customers for \( i = 1, 2, ..., N \).
- The occupation rate of the server is equal to \( \rho = \sum_{i=1}^{N} \rho_i \).
- \( c \) is the cost for waiting time (per time unit) for each product.
- \( k \mu \) is the capacity cost.
- The service limit of queue \( i \) is equal to \( k_i \) for \( i = 1, 2, ..., N \). This is the maximum number of customers that is served at one visit of the server. The server leaves queue \( i \) if \( k_i \) customers are served or if the queue becomes empty, whichever occurs first.
- The switchover durations are independent random variables with mean \( s_i \). This is the time between ending service at queue \( i \) and starting service at queue \( (i \mod N) + 1 \). For the computations in this section, the distribution of the switchover durations do not matter.
- The total switchover time during one cycle of the server has mean \( s = \sum_{i=1}^{N} s_i \).

4.3 Cost function

The expected total cost consists of cost for the amount of capacity and cost for the expected waiting time:

\[
T(\mu) = k\mu + \sum_{i=1}^{N} c_i \lambda_i E(W_i). \tag{38}
\]

For the \( k \)-limited polling model, no exact expression for the expected waiting time is known. Therefore, we use an approximation for the waiting times in our model. This approximation is given in section 4.3.1.

4.3.1 Expected waiting time

In [5], the problem of finding the optimal service limits in a cyclic polling system with the \( k \)-limited service discipline is studied. Here, it is found that the Fuhrmann and Wang approximation [7] is ‘very effective in finding the optimal service limits’ and does not need too much computation time. In numerical experiments, Borst et al. ‘have observed that the waiting cost according to the..."
Fuhrmann and Wang approximation sometimes differs dramatically from the “true” waiting cost obtained by the psa (power series algorithm), but that still the optimal service limits according to the Fuhrmann and Wang approximation agree with the “true” optimal service limits obtained from the psa’. The word ‘true’ is in quotation marks because the power series algorithm is also an approximation. The psa is a very time-consuming, but also a very accurate algorithm and therefore appropriate for judging other approximations.

Unless the approximation sometimes differs from the true waiting time, Fuhrmann and Wang state in [7] that their approximations are ‘largely heuristic but show very good accuracy in cases where the system parameters are not extremely asymmetric and the switchover times not largely relative to the service times’. Also, because the approximation is very appropriate for optimizing the service limits, it is a good approximation in some way. Together with the fact that the Fuhrmann and Wang approximation is easy to use, this makes it convenient to use in our model.

The Fuhrmann and Wang approximation for the waiting time in the $k$-limited polling model is given by

$$E(W_i) \approx \frac{(1 - \rho_i)(1 - \rho) + \frac{\rho_i}{k_i^2}(2 - \rho)}{1 - \rho - \frac{\lambda_i}{k_i}} \cdot \frac{D + \frac{s}{1 - \rho} \sum_{j=1}^{N} \rho_j^2}{\sum_{j=1}^{N} \left[ \rho_j(1 - \rho_j) + \frac{\rho_j^2 (2 - \rho)}{k_j (1 - \rho_j)} \right]}$$

with

$$D = \rho \sum_{i=1}^{N} \lambda_i \rho_i^{(2)} + \rho \frac{s^{(2)}}{2s} + \frac{s}{2(1 - \rho)} \left[ \rho^2 - \sum_{i=1}^{N} \rho_i^2 \right].$$

In (40), $\rho_i^{(2)}$ is the second moment of the service rate of queue $i$, which is in our case equal to $2/(\alpha_i \mu_i)^2$. Approximation (39) also holds for queues where the service rates do not have exponential distributions.

In [5], also another approximation is given. This approximation yields an explicit expression for the optimal service limits. However, this approximation is a result for constrained waiting cost optimization, while (39) is for unconstrained waiting cost optimization. For constrained waiting cost optimization, there is the constraint that the weighted sum of the service limits may not exceed a certain value:

$$\sum_{i=1}^{N} \gamma_i k_i \leq K,$$

where $k_i$ are the service limits and $\gamma_i$ are arbitrary parameters. However, in our model there is no constraint on the service limits, which make this approximation less appropriate. The approximation only gives good results for a small $K$, because the available capacity for the $k_i$ is distributed over the queues. When $K$ is increasing, the values for $k_i$ are kept in approximately the same ratios. Subsequently, for large $K$, all $k_i$ tend to infinity. This gives an exhaustive polling system in which the advantages of a $k$-limited system are lost.
Also, it is shown in [5] that the approximation (39) of $E(W_i)$ is decreasing in $k_i$ and increasing in $k_j, j \neq i$. This supports the use of the approximation in trying to obtain the optimal service limit values. After some tests, it seems that there is only a global minimum and no other local minima, which makes it easier to find the optimal service rates (at least we will not get stuck in a local minimum).

All together, the unconstrained approximation (39) seems the most useful approximation for the waiting time in our model.

4.4 Optimizing capacity

Because a polling model is more difficult than the models in sections 2 and 3, we start with optimizing the capacity in a polling model in which the demand rates are given. The Mathematica program is given in appendix C. First, let us consider two examples.

4.4.1 Two examples

For the first example, we have the following parameters:

- $N = 2$
- $k = 1$
- $c = 1$
- $\lambda_1 = 0.75$
- $\lambda_2 = 0.75$
- $\alpha_1 = 0.6$
- $\alpha_2 = 0.4$
- $s_1 = 10$
- $s_2 = 20$

After running the program, we find that the optimal service limits are $k_1 = k_2 = \infty$ and that the optimal capacity is $\mu = 9.28$. The corresponding expected total cost is 60.35. Note that both service limits are $\infty$, so we have an exhaustive polling model in the optimal case.

For the second example, we have the following parameters:

- $N = 2$
- $k = 1$
- $c = 1$
- $\lambda_1 = 0.75$
• $\lambda_2 = 0.75$
• $\alpha_1 = 1.11$
• $\alpha_2 = 10$
• $s_1 = 0.1$
• $s_2 = 0.1$

In this example, the optimal service limits are $k_1 = 2$ and $k_2 = \infty$. The optimal capacity is $\mu = 1.71$ and the corresponding expected total cost is 2.69.

The main reason that not both service limits are equal to $\infty$, is the difference in service speed of the two queues. The server can serve customers of type 2 much faster than type 1 customers. Therefore, the server should serve at queue 2 for a longer time (if customers are present) and not stay too long at queue 1. Switchover times are disadvantageous, but if customers can be served fast after a switchover, it may be worth it.

In [5], it is shown that in a $k$-limited polling model, always at least one of the service limits is equal to $\infty$.

### 4.4.2 Switchover times

Because the main difference between the polling model and the FCFS-model are the switchover times, we are interested in their influence on the optimal capacity and corresponding cost. In the approximation for the expected waiting times, the values of $s_i$ are not taken into account separately, so we only need to vary the total switchover time per cycle $s$.

![Figure 21: Optimal $\mu$ for different switchover times](image)
In figure 21, the optimal $\mu$ is given as a function of $s$. Other parameters are the same as in example 1. When $s$ increases, it takes more time for the server to move to another queue. This increases the waiting times for the customers, which can be caught by increasing the service rate. For smaller $s$, $\mu$ is increasing faster, because then the influence of increasing $s$ is relatively bigger.

In figure 22, we see that the optimal cost is increasing in $s$ in a linear way.

### 4.4.3 Service time ratios

We assumed that the ratio of the service rates for the two queues are fixed. In this subsection, we investigate the influence of this ratio. Therefore, we again take the parameters of example 1, but now we vary the service rates of the two queues. We let $\alpha_2 = 1 - \alpha_1$, so the service rate for queue 1 is equal to $\mu_1 = \alpha_1 \mu$, and the service rate for queue 2 is equal to $\mu_2 = \alpha_2 \mu = (1 - \alpha_1) \mu$. In figures 23 and 24, we let $\alpha_1$ run between 0 and 1 and show the corresponding optimal $\mu$ and total cost. Note that $\alpha_1$ may not be equal to 0 or 1, because then there would be one queue that cannot be served, which makes the system instable.

The graphs are symmetric and the minimum is reached for $\alpha_1 = 0.5$. For small $\alpha_1$, the service rate for queue 1 is low so a high $\mu$ is needed for that queue, while for large $\alpha_1$, the service rate for queue 2 is low so a high $\mu$ is needed for that queue. For $\alpha_1 = 0.5$, the service rate is the same for both queues and a smaller $\mu$ is sufficient and optimal. Note that this is due to the symmetry of the instance: the demand rates are the same for both queues and the approximation for the waiting times only takes the total switchover time per cycle into account (not dependant on the specific queue). If the demand rates are not the same for both queues, the optimal $\alpha_1$ is in general not exactly equal to 0.5 (the queue with more demand will be assigned a higher service rate). However,
also in that case the capacity will tend to infinity for $\alpha_1$ close to 0 or 1.
4.5 Uncertain demand

For the polling model, it is also possible to include uncertain demand in the system. By considering all possible scenarios with corresponding probabilities, the expected total cost can be minimized in the server capacity en service limits for both periods. The Mathematica program to compute the optimal service speed and service limits for the first time period can be found in appendix C. When computing this optimum, the costs for the second time period are also taken into account.

Because the functions and expressions for the polling system are more complex than in the previous models, it takes a few minutes (instead of a few seconds for the previous models) to compute the optimal solution for a model with \( N = 2 \) and two scenarios for the demand rate for each queue. For more extensive models, it must be considered if the large computation time is worth the better solution. If the optimal solution must be implemented fast, it may be better to use a simpler approximation for the waiting times in a polling model (despite the slightly worse solution).

4.5.1 Impact of the uncertainty

To study the impact of uncertainty on the decisions for the \( k \)-limited polling model, we considering an example with the following parameters:

- \( N = 2 \)
- \( k = 1 \)
- \( c = 2 \)
- Period 1: \( \lambda_1 = \lambda_2 = 0.75 \)
- Period 2: \( \lambda_1 = \begin{cases} 0.75 + a & \text{with probability 0.5} \\ 0.75 - a & \text{with probability 0.5} \end{cases} \) and \( \lambda_2 = \begin{cases} 1 & \text{with probability 0.4} \\ 0.9 & \text{with probability 0.6} \end{cases} \)
- \( \alpha_1 = 0.6 \)
- \( \alpha_2 = 0.4 \)
- \( s_1 = 20 \)
- \( s_2 = 10 \)
- \( h_u = 5 \)
- \( h_d = 1 \)

In this example, \( a \) can take any value smaller than 0.75. When \( a \) increases, more uncertainty is incorporated in the model. Figures 25 and 26 show the optimal value of \( \mu_1 \) for different values of \( a \). For larger \( a \), there is more uncertainty in the model, which leads to more capacity to cover a possible higher demand.

Note that only the uncertainty of queue 1 is increasing in \( a \), for queue 2 the distribution is fixed. For small values of \( a \), the uncertainty of queue 2 is dominant, such that the impact of the value of \( a \) is limited. For larger \( a \), the demand rate of queue 1 can increase more in period 2, such that the optimal value of \( \mu_1 \) is increasing faster in \( a \).
Figure 25: Optimal $\mu_1$ for different demand distributions

Figure 26: Optimal cost for different demand distributions
4.6 Conclusions

In this section, we changed our model into a more realistic model by adding switchover times. If one server has to serve multiple customer types, it needs in general a certain time period (and possibly also some cost) to prepare the server for serving another customer type. This leads to a polling model with switchover times. For our polling model, we have chosen for a $k$-limited policy, because this suits best with the application in the chemical industry where our problem is based on. Advantages of the $k$-limited policy are the opportunity to assign priorities to queues and to limit the total cycle time. On the other hand, relative few expressions and functions are known explicitly for the $k$-limited polling model. Therefore, we have to use an approximation for the expected waiting times.

The approximation given by Fuhrmann and Wang seems to fit best in our model. However, because the expression of this approximation is quite large and complex, it takes some time for Mathematica to compute the optimum. Next to the difficult expressions, Mathematica must not only optimize the server capacity, but also the service limits for the queues. Because the larger computation times for this model, one should decide whether a faster or a better solution is more important.
5 General conclusions and recommendations

In this thesis, we have created and investigated a number of queueing models to gain insight in the robustness of supply chains.

A supply chain can be made robust at multiple components and in many different ways. In particular, we investigated the robustness for uncertain demand in the long term and have looked at the impact of design choices on cost, and throughout time. We started with a simple model with one queue and one server. In the next sections, we extended the model with multiple customer types and later also with switchover times for the server. For all models, we assumed two time periods: the demand rates in period 1 were assumed to be known, but for period 2 only (discrete) distributions were given.

For the model in section 2, with one queue and one server, the optimal service rate for time period 2 is given explicitly as function of $\mu_1$ and $\lambda_2$. Using this function, the optimal service rate for the first period is optimized numerically by using Mathematica. In section 2.8, the value of our model became visible. When comparing our method to the optimal solution where only the average demand rate was taken into account, the cost difference was up to more than 20%. This shows that it can be very profitable to make supply chains robust in the right way.

We extended the model by adding a queue in section 3. Because expressions for the waiting time became more complex, both $\mu_1$ and $\mu_2$ were calculated numerically in this section. When examining some examples, we found that for both models in sections 2 and 3, the optimal solution can be calculated fast. This offers opportunities to use those models in practice.

In section 4, we introduced the $k$-limited polling model with switchover times. The main advantage of this model is that it is more realistic than the other ones, because in general, it takes some time for production systems to switch to another product. Disadvantages are the more difficult expressions and approximations for the waiting times. Because of the difficult expressions, more computation time is needed. When less time is available, one can consider using a polling model with another policy. For example, an exhaustive polling model is in many production applications also much more realistic than a model without switchover times, and more and simpler expressions have been developed for this model.

For now, we can conclude that for using the $k$-limited polling model, one should decide if the larger computation time outweighs the better solution.

The main purpose of our models was to gain insight in the design of robust supply chains. The models in this thesis are not for direct use in practical problems, because they are oversimplifying the operations of real production systems. However, when designing a supply chain, parts of the models can certainly be used to create the insight where potential issues and their solutions may be found, in order to create more robustness in the supply chain. When designing a supply chain, one has to investigate for the specific situation which parts of the models can be used to make the supply chain more robust.
The models in this thesis were a first step to create insight in robust supply chains. For more understanding and to be applicable in more situations, they may be extended in several ways. In our models, we assumed exponential arrival and service rate. It would be interesting to see what happens if other distributions are used. Also, switchover costs could be added to the polling model, other service policies could be used or systems with multiple servers could be examined. This would open the possibility to study larger networks with more customers, products and production facilities.
References


A Mathematica program - one customer, one server

Clear["Global`*"]

(*Optimal value for $\mu_2$ as a function of $\mu_1$ and $\lambda_2$*)

$M[\mu_1_, \lambda_2_] = \text{If}[\mu_1 < \lambda_2 + \frac{\sqrt{c + \lambda_2}}{k + hu},$

$\frac{\sqrt{c + \lambda_2}}{k + hu}, \text{If}[hd < k + \lambda_1 > \lambda_2 + \frac{\sqrt{c + \lambda_2}}{k - hd}, \frac{\sqrt{c + \lambda_2}}{k - hd}]]];$

(*Input*)
k = 2; (*capacity costs*)
c = 10; (*sojourn costs*)
hu = 5; (*cost for increasing capacity*)
hd = 1; (*cost for decreasing capacity*)
$\lambda_1 = 10; \text{(*demand rate for period 1*)}$
$\lambda_2 = \{5, 10, 15\}; \text{(*possible demand rates for period 2*)}$
p = \{0.7, 0.15, 0.15\}; (*probabilities corresponding to demand rates for period 2*)

(*Compute optimal server capacity:* )

Minimize\[
\{k \cdot m1 + c \cdot \frac{1}{m1 - \lambda_1} + p[[1]] \cdot \left(k \cdot M[m1, \lambda_2[[1]]] + c \cdot \frac{\lambda_2[[1]]}{M[m1, \lambda_2[[1]]] - \lambda_2[[1]]} + \right.$

$\text{Max}[0, m1 - M[m1, \lambda_2[[1]]]] + hu \cdot \text{Max}[0, M[m1, \lambda_2[[1]]] - m1]\}$,

$p[[2]] \cdot \left(k \cdot M[m1, \lambda_2[[2]]] + c \cdot \frac{\lambda_2[[2]]}{M[m1, \lambda_2[[2]]] - \lambda_2[[2]]} + \right.$

$\text{Max}[0, m1 - M[m1, \lambda_2[[2]]]] + hu \cdot \text{Max}[0, M[m1, \lambda_2[[2]]] - m1]\}$,

$p[[3]] \cdot \left(k \cdot M[m1, \lambda_2[[3]]] + c \cdot \frac{\lambda_2[[3]]}{M[m1, \lambda_2[[3]]] - \lambda_2[[3]]} + hd \cdot \text{Max}[0, m1 - M[m1, \lambda_2[[3]]]] + hu \cdot \text{Max}[0, M[m1, \lambda_2[[3]]] - m1]\}, m1 > \lambda_1, \{m1\}]$\]

{93.542, \{m1 \[Rightarrow] 17.1492\}}
B Mathematica program - two customers, one server - First Come, First Served

Clear["Global`*"]
k = 2; (*capacity cost*)
hd = 1; (*cost for decreasing capacity*)
c = 1; (*waiting cost*)
111 = 100; (*demand rate customer 1, period 1*)
121 = 200; (*demand rate customer 2, period 1*)
a11 = 2/3; (*factor capacity product 1*)
a21 = 1/3; (*factor capacity product 2*)
111 = 100; (*demand rate customer 1, period 1*)
121 = 200; (*demand rate customer 2, period 1*)
121 = 90; (*demand rate customer 1, period 2, scenario 1*)
p121 = 0.6; (*corresponding probability for 11121*)
1122 = 110;
p122 = 0.4;
1221 = 220;
p221 = 0.8;
p122 = 180;
p222 = 0.2;

(*Optimal capacity for period 2 as a function of the capacity in period 1 and demand rates for period 2*)
M[ml_, 11_, 12_] := If[ml < NSolve[0 < 11 / (a11*l2 + m2) + 12 / (a21*m2) < 1, d]...

(*NSolve*)
(*Printed by Wolfram Mathematica Student Edition*)

46
\begin{aligned}
\text{FindMinimum}\{ & k \cdot \text{hd} = c \cdot (11 + 12) + \\
& \left(1 - \frac{11}{a1 + m2} \cdot \frac{12}{a2 + m2}\right) \cdot \\
& \left(\frac{11}{a1 + m2} \cdot \frac{12}{a2 + m2}\right) \cdot \left(11 \cdot a2 + 12 \cdot a1 + 2 \cdot a1^2 \cdot a2 + a1^2 \cdot a2^2 \cdot m2\right) + \\
& \left(\frac{11}{a1^2 + m2^2} \cdot \frac{12}{a2^2 + m2^2}\right) \cdot \left(11 \cdot a2 + 12 \cdot a1 + 2 \cdot a1^2 \cdot a2 + a1^2 \cdot a2^2 \cdot m2\right) + \\
& m2\begin{bmatrix} (1) & (1) & (1) \end{bmatrix}; (\text{functie voor m_2})\}
\end{aligned}
C Mathematica program - multiple customers, one server - polling model

Clear["Global`*"]

k = 2; (*capacity costs*)
l11 = 0.75; (*demand rate for period 1, queue 1*)
l12 = 0.75; (*demand rate for period 1, queue 2*)
l21 = 0.9; (*demand rate for period 2, queue 1, scenario 1*)
l22 = 0.3; (*corresponding probability to l211*)
p211 = 0.8; (*corresponding probability to l221*)
p222 = 0.4; (*corresponding probability to l222*)
p211 = 0.9; (*corresponding probability to l221*)
p222 = 0.6; (*corresponding probability to l222*)
d = 0.6; (*capacity factor queue 1*)
s = 0.4; (*capacity factor queue 2*)
m11 = s1; (*service rate period 1, queue 1*)
m12 = s2; (*service rate period 1, queue 2*)
m21 = 0.75; (*service rate period 2, queue 1*)
m22 = 0.7; (*service rate period 2, queue 2*)

Clear[r1, r2, l1, l2, m1, m2]

DO[r1, r2, l1, l2, m1, m2] :=

W1[r1, r2, l1, l2, m1, m2, k1, k2] :=

W2[r1, r2, l1, l2, m1, m2, k1, k2] :=

(*Expected total cost for period 2,
for given service rate for period 1 and demand rates for period 2,
and for optimal service rate choice in period 2:*)

MMimize[If[l1 / (a1 + m2) + l2 / (a2 + m2) + Max[l1 * s1 / k1, l2 * s2 / k2] + 1 ||

k1 <= 0 || k2 <= 0 || m2 <= 0, $MaxMachineNumber,

k + m2 + l1, l2 / (a1 + m2), l2 / (a2 + m2), l1 / (a1 + m2), l2 / (a2 + m2), l1, l2,

a1 + m2, a2 + m2, k1, k2)] + W2[l1 / (a1 + m2) + l2 / (a2 + m2), l1 / (a1 + m2), l2 / (a2 + m2), l1, l2,

a1 + m2, a2 + m2, k1, k2] + W1[l1 / (a1 + m2) + l2 / (a2 + m2), l1 / (a1 + m2), l2 / (a2 + m2), l1, l2,

a1 + m2, a2 + m2, k1, k2] ;
\[
\frac{l_2}{a_2 m_2}, l_1, l_2, a_1 m_2, a_2 m_2, k_1, k_2 + h_d \cdot \text{Max}[0, m_1 - m_2] + h_u \cdot \text{Max}[0, m_2 - m_1], \{m_2, k_1, k_2\}, \text{MaxIterations} \to 1000][[1]]
\]

Clear[{k_1, k_2, m_1}];
NMinimize[
If[l_11/(a_1 m_1) + l_12/(a_2 m_2) + Max[l_11 s/k_1, l_12 s/k_2] \geq 1 \land \text{Max}[0, m_1] \leq 0 || k_1 \leq 0 || k_2 \leq 0 || m_1 \leq 0, \text{MachineNumber}, k_1 + l_11/(a_1 m_1), l_12/(a_2 m_2), l_11, l_12, a_1 + m_1, a_2 + m_2, k_1, k_2 + l_11 + l_12/(a_1 + m_1)] + l_11/(a_1 + m_1), l_12/(a_2 + m_2), l_11/(a_1 + m_1), l_12/(a_2 + m_2), l_11, l_12, a_1 + m_1, a_2 + m_2, k_1, k_2 + p_211 \cdot p_221 \cdot M_2[m_1, 1211, 1221] + p_212 \cdot p_222 \cdot M_2[m_1, 1212, 1222], \{m_1, k_1, k_2\}, \text{MaxIterations} \to 100]

\{227.6, \{m_1 \to 28.0685, k_1 \to 95.2151, k_2 \to 187.833\}\}