Cloud Cost Optimization Based on Shifted N-policy M/M/M/K Queue Model

Zsolt Saffer
Institute of Statistics and Mathematical Methods in Economics
Vienna University of Technology
Vienna, Austria
Email: zsolt.saffer@tuwien.ac.at

***Abstract*—Cloud cost optimization is an important issue having also impact on the economy of the Cloud service, since it enables the Cloud service provider the service provisioning at minimum cost. This paper provides the performance analysis and cost optimization of an Infrastructure-as-a-Service Cloud model with a capacity control policy. The Virtual Machines are modelled as parallel resources, which can be either in active or in standby state. The capacity of the cloud is controlled by changing the number of active Virtual Machines. The cost model, which the cloud provider encounters, takes into account both energy consumption and performance measures. The major objective of the work is to provide a tractable analytic model, which is suitable for practical use. For this purpose, we model the Cloud services by an $M/M/M/K$ queue. We propose a simple control policy, in which a predefined portion of Virtual Machines are always active. The remaining ones are activated simultaneously when the number of requests reaches a threshold and deactivated when the number of requests falls below the predefined portion of active Virtual Machines. We call this policy as shifted N-policy. We provide the stationary analysis of the model and derive closed form results for the distribution of the number of requests as well as for several performance measures. The cost model leads to a discrete optimization task, which we approximate by a nonlinear continuous optimization task. After applying numerous approximations, we provide, in form of solution formulas, approximate solutions of the optimization task in the two most relevant traffic ranges. We give illustrating examples for the most important approximations and properties of the model as well as validate the approximate solution formulas by numeric optimization. The major results of the work are the closed form approximate solution formulas, which give the optimal threshold under the most relevant ranges of parameters and provide insight into the dependency of the optimum on the model and cost parameters.**

***Keywords*—optimization; cloud model; queueing model; N-policy.**

I. Introduction

Cloud cost optimization is an important issue having also impact on the economy of the Cloud service, since it enables the Cloud service provider the service provisioning at minimum cost. This paper is an extension of our previous work [1], in which we provided a performance evaluation and optimization of an Infrastructure-as-a-Service (IaaS) Cloud model with a proposed simple threshold based resource control. The major objective of this research is to establish a tractable analytic model, which is also suitable for practical use.

The growing demand for computational resources lead to a concept of Cloud computing. With a wide spreaded use of computer networks, new applications emerged in the '90-s in many areas, e.g., business, science or web-applications. They created a growing demand for computational resources, which does not necessarily locate locally. This lead to a new distributed computing paradigm called Cloud computing [2] [3] [4] [5]. In this work, we focus on IaaS type Cloud service, in which computing resources are delivered to customers. Besides of the physically distributed character of Cloud services, another key attribute of clouds is the virtualization, which enables to decouple the computing resources from the physical hardware and deliver them to customers as Virtual Machines (VM).

The users want guaranteed performance and the Cloud service provider want to ensure it and supervise the operation. These require proper performance modelling and evaluation enabling to get insights into the relationships among the used resources and the performance.

However, the performance evaluation of Cloud services is a complex issue, since it depends on many factors. Analytic models are either too simplified to obtain meaningful relationships or lead to rather complex numeric solution, which does not provide an explicit relationships among the used resources and the performance. There are many research works on performance modelling of clouds. An advanced work on performance analysis is [6], which provides a numeric solution. An outstanding work is [7], in which a multi-level interacting stochastic sub-models approach is proposed, leading to a numeric method to compute the performance measures. For an overview on research works on performance evaluation of clouds the reader is referred to the survey [8] and the references herein.

Cloud cost optimization requires a resource management technique. The work [9] provides an approach to predict resource usage in Cloud computing, which can be seen as a simplest way to enable the managing of the system. Resource scheduling techniques are proposed in [10], [11] and [12]. Energy-aware resource allocation mechanism for management of clouds is proposed in [13]. Energy efficient resource management and allocation policies for clouds are summarized in [14] [15]. One recent efficient resource control mechanism for clouds is the threshold based activation and deactivation of VMs, proposed e.g., in [16]. This mechanism can be modelled by hysteresis queue and in [16] computational algorithms are provided for computing the optimal thresholds. For another numerical approaches to cloud cost optimization we refer to [17] and [18]. As expected, optimization of clouds is even more complex issue than its performance evaluation. Therefore, the vast majority of works on Cloud cost optimization

proposes a computational solution.

In this paper, we present a performance evaluation and optimization of an IaaS Cloud model with a proposed simple threshold based resource control, but in contrast to the vast majority of relevant works, we provide approximate explicit formulas for determining the only threshold of the control mechanism. The formulas hold in most relevant range of parameters. The resource control mechanism, introduced in [1], is called as shifted N-policy. According to this policy, a predefined portion of VMs are always active. The remaining ones are activated simultaneously when the number of requests reaches a threshold N (like in N-policy) and deactivated when the number of requests falls below the predefined portion of active VMs. This explains the name of the policy. The cloud is modeled by multi-server $M/M/M/K$ queue. Note that, as pointed out in [19], the $M/M/m$ queue can be an acceptable approximation of the $GI/GI/m$ queue until the coefficient of variations of both the interarrival and the service times are not far from 1.

In [1], we presented closed form results for the stationary distribution of the number of requests and several performance measures in the shifted N-policy $M/M/M/K$ model. The cost model lead to a discrete optimization task, which we approximated by a nonlinear continuous optimization task. We provided a closed form approximate solution formula for the high traffic range, in which the utilization is higher than a model parameter dependent threshold.

In this work, we recall the results of [1] and extend the solution of the optimization task to the low traffic range, i.e., in which the utilization is lower than the above model parameter dependent threshold. Additionally, we generalize the former approximate solution formula for the high traffic range by relaxing its restrictions by omitting the condition on cost parameters. Moreover, we also provide the details of the stationary analyis and the derivations both in the former and new optimization parts.

The major contributions of this research are the closed form approximate solution formulas for the optimal value of the threshold N under the most relevant ranges of parameters. The secondary contribution of this research is the proposal of the shifted N-policy control mechanism and its stationary analysis in the context of $M/M/M/K$ model. The advantage of using the proposed shifted N-policy control is that it makes the cloud resource management very simple due to the approximate analytic formulas for the optimal threshold, i.e., no need for computational algorithm. On the other hand, it leads to somewhat higher optimal cost than other more complex computational solutions, e.g., the hysteresis policy with multi-thresholds. The proposed optimization can be used for example for the use case "Enabling add-on services on top of the infrastructure", e.g., computing-as-a-service, analytics or Business Intelligence(BI)-as-a-service.

We also provide illustrating examples for the approximations and the most important properties of the model as well as validate the approximate solution formulas by numeric optimization in the relevant range of parameters.

The rest of this paper is organized as follows. Section II is devoted to the description of the model. The stationary analysis of the queueing model is given in Section III. In Section IV, we construct the cost function to be optimized. The dependency of the probability of the empty system on threshold N, as fundamental building block of the optimization, is investigated in Section V. The approximate minimization for high traffic range is presented in Section VI. This is followed by establishing the approximate minimization for low traffic range in Section VII. In Section VIII, we illustrate the approximate solution formulas as well as provide their numeric validation. The work is concluded in Section IX.

II. Cloud model description

In this section, we give the description of the IaaS Cloud model and the shifted N-policy queueing model.

A. IaaS Cloud model

The IaaS Cloud delivers low-level computational resources to the users. The Physical Machines (PMs) are grouped into two pools: active (running) and standby machines. The PMs in standby can represent either turned-on (but not ready) or turned-off machines. The computational resources are provided to users in the form of VMs. Total number of available VMs is $M > 100$, from which $0.1M \leq L \leq 0.5M$ VMs are always active. The resource control is realized by threshold based activation and deactivation of the remaining $M - L$ VMs. The model has buffer with capacity for $K - M \gg 1$ users. When all active VMs are busy upon arrival of a new request then it is directed into the buffer, where it waits until getting an access to a VM becoming free. When the buffer is full upon arrival of a new request, then the request is lost.

B. Shifted N-policy queueing model

The queueing system modelling the IaaS Cloud is an $M/M/M/K$ queue with shifted N-policy. In the queueing context the VMs are called as servers. The request arrive according to Poisson process with rate $\lambda > 0$ and the service times are exponentially distributed with parameter $\mu > 0$. The arrival process and the service process are assumed to be mutually independent. The system has $M \geq 1$ servers and buffer capacity for $K - M \gg 1$ requests. When the servers and the buffer are full upon arrival of a new request, then the request is lost.

The control of the VMs is realized by the newly proposed shifted N-policy. According to this policy $L < M$ servers are always active. When the queueing system is empty then the remaining $M - L$ servers are in standby. They will be activated simultaneously when the number of requests in the system reaches the threshold $L + 1 \leq N \leq M$. After having all the M servers active, $M - L$ servers will be deactivated simultaneously, when the number of requests in the system reaches again L. This policy has hysteresis-like characteristic (in number of requests), which makes it suitable to use as energy efficient resource control. However, it is much simpler

than the hysteresis queue, which could facilitate the developing of analytically tractable approximation.

The queue is always stable due to the finiteness of the underlying Continuous-Time Markov chain (CTMC) model (see in Section III). The (approximate) utilization of the system, denoted by ρ is given by

$$\rho = \frac{\lambda}{M\mu}. \tag{1}$$

Although this expression would be exact only in case of $K - M = \infty$, otherwise the blocking probability should be also taken into account to get an exact expression, under the model assumption $K - M \gg 1$ the expression (1) can be considered as a good approximation of the utilization.

III. Analysis of the Queuing Model

Let $n \geq 0$ be the number of requests in the system. The process $\{n(t), t \geq 0\}$ is a finite state CTMC.

A. State diagram

The state diagram of the $M/M/M/K$ queue with shifted N-policy can be seen in Figure 1.

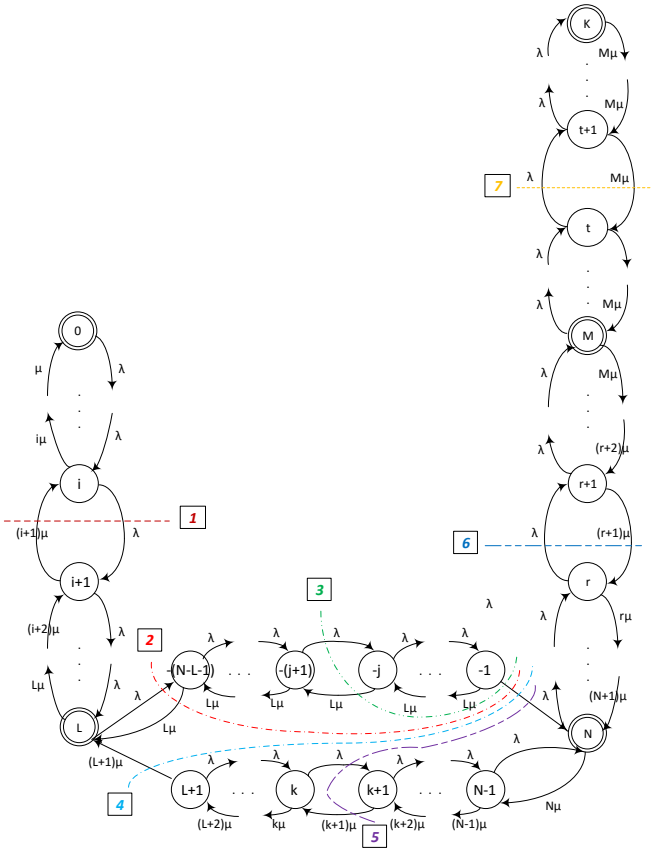

Figure 1. State diagram.

Basically the states are denoted according to the number of requests in the system. However, the notation of the states, in which the $L < n < N$, depends on the number of active servers. If there are L active servers then the states are denoted by the number $-(N-n)$. Otherwise (i.e., there are M active servers) the default numbering, n is used. On this way the states can be described as a contiguous range $[-(N-L-1), \ldots, -1, 0, 1, \ldots, K]$.

B. Stationary analysis

We perform the stationary analysis rather by utilizing the principle of global balance equations instead of applying the standard way by means of equilibrium equations. This results in shorter derivations for the stationary distribution of the number of requests in the system. We define the stationary probability, p_i as the probability that the system is in state i, for $-(N-L-1) \leq i \leq K$.

1) Global balance equations: We marked the selected set of states used for the balance equations on the state diagram. Each case is marked by a separator line and an associated number in small square, which is used to identify the case.

1) $(i+1)\mu p_{i+1} = \lambda p_i$, $i = 0, \ldots, L-1$,
2) $L\mu p_{-(N-L-1)} + \lambda p_{-1} = \lambda p_L$,
3) $L\mu p_j + \lambda p_{-1} = \lambda p_{j-1}$, $j = -(N-L-2), \ldots, -1$,
4) $(L+1)\mu p_{L+1} = \lambda p_{-1}$,
5) $(k+1)\mu p_{k+1} = \lambda p_k + \lambda p_{-1}$, $k = L+1, \ldots, N-1$,
6) $(r+1)\mu p_{r+1} = \lambda p_r$, $r = N, \ldots, M-1$,
7) $M\mu p_{t+1} = \lambda p_t$, $t = M, \ldots, K-1$.

2) Stationary distribution of the number of requests: By solving the balance equations by applying standard techniques (see in Appendix (I)), we get the stationary distribution of the number of requests as

$$\begin{aligned}
p_k &= \frac{(\frac{\lambda}{\mu})^k}{k!} p_0, \text{ for } k = 1, \ldots, L, \\
p_k &= \left(\frac{\lambda}{L\mu}\right)^{N-L} \frac{(\frac{\lambda}{L\mu})^k - 1}{1 - (\frac{\lambda}{L\mu})^{N-L}} p_L, \\
&\quad \text{for } k = -(N-L-1), \ldots, -1, \\
p_k &= \sum_{i=L}^{k-1} \frac{i!}{k!} \left(\frac{\lambda}{\mu}\right)^{k-i} p_{-1}, \text{ for } k = L+1, \ldots, N, \\
p_k &= \frac{N!}{k!} \left(\frac{\lambda}{\mu}\right)^{k-N} p_N, \text{ for } k = N+1, \ldots, M, \\
p_k &= \left(\frac{\lambda}{M\mu}\right)^{k-M} p_M, \text{ for } k = M+1, \ldots, K
\end{aligned} \tag{2}$$

and p_0 can be determined from the normalization condition $\sum_{k=-(N-L-1)}^{K} p_k = 1$.

The probabilities p_L, p_{-1}, p_N and p_M are probabilities of events representing some boundary in the operation of the considered queueing model. They are given by

$$
\begin{aligned}
P_L &= \frac{(\frac{\lambda}{\mu})^L}{L!} p_0, \\
p_{-1} &= \alpha\ p_L, \text{ where } \alpha = \left(\frac{\lambda}{L\mu}\right)^{N-L-1} \frac{1-\frac{\lambda}{L\mu}}{1-(\frac{\lambda}{L\mu})^{N-L}}, \\
p_N &= \sum_{i=L}^{N-1} \frac{i!}{N!} \left(\frac{\lambda}{\mu}\right)^{N-i} p_{-1} = \frac{(\frac{\lambda}{\mu})^N}{N!} s_{L,N}\ \alpha\ p_L, \\
&\text{where } s_{L,N} = \sum_{i=L}^{N-1} \frac{i!}{(\frac{\lambda}{\mu})^i}, \\
p_M &= \frac{N!}{M!} \left(\frac{\lambda}{\mu}\right)^{M-N} p_N. \qquad (3)
\end{aligned}
$$

IV. Cost function

In this section, we establish the cost model of the cloud provider and apply it to the shifted N-policy queue. This results in the cost function.

A. Cost model

The cloud provider encounters different type of costs with different weights. These are taken into account by the help of cost parameters, which are defined by

- C_{on} - cost of an active VM/time unit,
- C_{off}- cost of a standby VM/time unit,
- C_W - cost of waiting of a request/time unit (= holding a request in the buffer/time unit),
- C_R - cost of loss of an arriving request,
- C_A - activation cost of a VM (changing from standby to active state),
- C_D - deactivation cost of a VM (changing from active to standby state).

Using these parameters the cloud cost can be specified by the following function

$$
\begin{aligned}
\mathcal{C}_{cloud} = {} & E[\text{ number of active servers }]\ C_{on} \qquad (4) \\
& + E[\text{ number of standby servers }]\ C_{off} \\
& + E[W]\ C_W + p_{loss}\ \lambda\ C_R, \\
& + (\text{ activation rate of standby VMs })\ (M-L)\ C_A \\
& + (\text{ deactivation rate of active VMs })\ (M-L)\ C_D,
\end{aligned}
$$

where $E[\,]$ stands for the expected value of a random variable, W is the waiting time of the requests in the buffer and p_{loss} is the probability of loss.

Note that the operation of N-policy implies that one of the major trade-off of the model is the relation $C_{on} - C_{off}$ versus C_W, which in fact appears also in the approximate solution formulas for computing the threshold N (via parameter A see in subsections VI-F and VII-D, where parameters b and c depend on A).

B. Constructing the cost function

The cost function, to be optimized, can be constructed by applying the cost model (4) to the shifted N-policy queue.

1) Adapting the cost model to the shifted N-policy queue: The so far unknown terms arising in (4) can be expressed with the help of parameters, stationary probabilities and performance measures of the shifted N-policy queue as follows.

$$
\begin{aligned}
&E[\text{ number of active servers }] = L + (1-p_{s1})(M-L), \qquad (5) \\
&E[\text{ number of standby servers }] = p_{s1}(M-L), \\
&(\text{ activation rate of standby VMs }) = \lambda\ p_{-1}, \\
&(\text{ deactivation rate of active VMs }) = (L+1)\mu\ p_{L+1},
\end{aligned}
$$

where $p_{s1} = P\{\text{ the number of active VMs = L }\}$.

Substituting the expressions (5) into (4) we get the cost function, F_1 as

$$
\begin{aligned}
F_1 = {} & \Big(L + (1-p_{s1})(M-L)\Big)\ C_{on} + p_{s1}(M-L)\ C_{off} \\
& + E[W]\ C_W + p_{loss}\ \lambda\ C_R \\
& + \lambda\ p_{-1}\ (M-L)\ C_A + (L+1)\mu\ p_{L+1}(M-L)\ C_D. \qquad (6)
\end{aligned}
$$

2) Performance measures: The performance measures p_{s1}, p_{loss} and $E[W]$ influence the cloud cost. They are given by

$$
\begin{aligned}
p_{s1} &= \sum_{k=0}^{L} p_k + \sum_{k=-(N-L-1)}^{-1} p_k \qquad (7) \\
&= \sum_{k=0}^{L} \frac{(\frac{\lambda}{\mu})^k}{k!} p_0 + \sum_{k=-(N-L-1)}^{-1} \left(\frac{\lambda}{L\mu}\right)^{N-L} \frac{(\frac{\lambda}{L\mu})^k - 1}{1-(\frac{\lambda}{L\mu})^{N-L}} p_L \\
&= \sum_{k=0}^{L} \frac{(\frac{\lambda}{\mu})^k}{k!} p_0 + \sum_{k=1}^{N-L-1} \frac{(\frac{\lambda}{L\mu})^k - (\frac{\lambda}{L\mu})^{N-L}}{1-(\frac{\lambda}{L\mu})^{N-L}} p_L \\
&= \sum_{k=0}^{L} \frac{(\frac{\lambda}{\mu})^k}{k!} p_0 + \frac{\frac{\frac{\lambda}{L\mu} - (\frac{\lambda}{L\mu})^{N-L}}{1-\frac{\lambda}{L\mu}} - (N-L-1)(\frac{\lambda}{L\mu})^{N-L}}{1-(\frac{\lambda}{L\mu})^{N-L}} p_L.
\end{aligned}
$$

$$
p_{loss} = p_K = \left(\frac{\lambda}{M\mu}\right)^{K-M} p_M = \left(\frac{\lambda}{M\mu}\right)^{K} \frac{M^M}{M!} \frac{N!}{(\frac{\lambda}{\mu})^N} p_N. \qquad (8)
$$

$$
\begin{aligned}
E[W] &= \sum_{k=-(N-L-1)}^{-1} (k+N-L) p_k + \sum_{k=M+1}^{K} (k-M) p_k \\
&= \sum_{k=1}^{N-L-1} k\ p_{-(N-L)+k} + \sum_{k=M+1}^{K} (k-M) p_k \\
&= \tau p_L + \sigma p_M, \qquad (9)
\end{aligned}
$$

where

$$\tau = \frac{\frac{\lambda}{L\mu}}{(1-\frac{\lambda}{L\mu})^2} - (N-L)\ \frac{(\frac{\lambda}{L\mu})^{N-L}}{1-(\frac{\lambda}{L\mu})^{N-L}}\left(\frac{1}{1-\frac{\lambda}{L\mu}} + \frac{N-L-1}{2}\right), \quad (10)$$

$$\sigma = \frac{\lambda}{M\mu}\frac{1-(\frac{\lambda}{M\mu})^{K-M+1}}{(1-\frac{\lambda}{M\mu})^2} - (K-M+1)\frac{(\frac{\lambda}{M\mu})^{K-M+1}}{1-\frac{\lambda}{M\mu}}.$$

The derivation of the above expression of $E[W]$ can be found in Appendix (II).

3) Final form of the cost function F_1: Applying the balance equation $(L+1)\mu p_{L+1} = \lambda p_{-1}$, the expression of p_{-1} from (3), (9) and (8) in (6) as well as performing rearrangements yields

$$\begin{aligned}
F_1 &= \lambda\ \alpha p_L(M-L)\ C_A + \lambda\alpha p_L\ (M-L)\ C_D \\
&\quad + (\tau p_L + \sigma p_M)\ C_W + \left[\left(\frac{\lambda}{M\mu}\right)^K \frac{M^M}{M!}\frac{N!}{(\frac{\lambda}{\mu})^N}p_N\right]\ \lambda\ C_R \\
&\quad - (C_{on}-C_{off})(M-L)p_{s1} + MC_{on} \\
&= \Big(\lambda(C_A+C_D)(M-L)\alpha + C_W\tau\Big)p_L \\
&\quad + C_W\sigma p_M + C_R\ \lambda\ \left[\left(\frac{\lambda}{M\mu}\right)^K \frac{M^M}{M!}\frac{N!}{(\frac{\lambda}{\mu})^N}\right]p_N \\
&\quad - (C_{on}-C_{off})(M-L)p_{s1} + MC_{on}. \qquad (11)
\end{aligned}$$

The last but one line of (11) can be rearranged by using the expressions of p_M and p_N from (3) leading to

$$\begin{aligned}
&C_W\sigma p_M + C_R\ \lambda\ \left[\left(\frac{\lambda}{M\mu}\right)^K \frac{M^M}{M!}\frac{N!}{(\frac{\lambda}{\mu})^N}\right]p_N \\
&= \left[C_W\sigma\frac{(\frac{\lambda}{\mu})^M}{M!} + C_R\lambda\left(\frac{\lambda}{M\mu}\right)^K\frac{M^M}{M!}\right]\frac{N!}{(\frac{\lambda}{\mu})^N}p_N \\
&= \left[C_R\lambda\left(\frac{\lambda}{M\mu}\right)^K\frac{M^M}{M!} + C_W\sigma\frac{(\frac{\lambda}{\mu})^M}{M!}\right]s_{L,N}\alpha\ p_L. \qquad (12)
\end{aligned}$$

Substituting (12) back into (11) we get the final form of the cost function in terms of p_L and p_{s1} as

$$\begin{aligned}
F_1 &= \left[\Big(\lambda(C_A+C_D)(M-L) + \eta\ s_{L,N}\Big)\alpha + C_W\tau\right]p_L \\
&\quad - (C_{on}-C_{off})(M-L)p_{s1} + MC_{on},\ \text{where} \qquad (13) \\
\eta &= \left[C_R\lambda\left(\frac{\lambda}{M\mu}\right)^K\frac{M^M}{M!} + C_W\sigma\frac{(\frac{\lambda}{\mu})^M}{M!}\right].
\end{aligned}$$

V. Characterizing p_0 as a dependency of N

Unfortunately p_0, which is involved in almost every term of (13) via the expression of p_L, depends on N. Now we characterize p_0 as a dependency of N in order to identify parameter regions, in which p_0 is approximately independent of N. This leads to further restriction on the parameter range. We define the probability coefficients

$$\begin{aligned}
p_{s1w} &= \frac{1}{p_0}p_{s1} \\
p_{s2w} &= \frac{1}{p_0}\sum_{k=L+1}^{N} p_k = \sum_{k=L+1}^{N}\frac{(\frac{\lambda}{\mu})^k}{k!}\sum_{i=L}^{k-1}\frac{i!}{(\frac{\lambda}{\mu})^i}\alpha\frac{p_L}{p_0}, \\
p_{s3w} &= \frac{1}{p_0}\sum_{k=N+1}^{M} p_k = \frac{N!}{(\frac{\lambda}{\mu})^N}\sum_{k=N+1}^{M}\frac{(\frac{\lambda}{\mu})^k}{k!}\frac{p_N}{p_0} \\
&= \frac{N!}{(\frac{\lambda}{\mu})^N}\sum_{k=N+1}^{M}\frac{(\frac{\lambda}{\mu})^k}{k!}\frac{(\frac{\lambda}{\mu})^N}{N!}s_{L,N}\ \alpha\ \frac{p_L}{p_0} \\
&= \sum_{k=N+1}^{M}\frac{(\frac{\lambda}{\mu})^k}{k!}\sum_{i=L}^{N-1}\frac{i!}{(\frac{\lambda}{\mu})^i}\ \alpha\ \frac{p_L}{p_0}, \\
p_{s4w} &= \frac{1}{p_0}\sum_{k=M+1}^{K} p_k = \sum_{k=M+1}^{K}\rho^{k-M}\frac{p_M}{p_0} \\
&= \frac{\rho-\rho^{K-M+1}}{1-\rho}\frac{N!}{M!}\left(\frac{\lambda}{\mu}\right)^{M-N}\frac{p_N}{p_0}, \qquad (14)
\end{aligned}$$

which are the terms of the coefficient of p_0 in the normalization equation, i.e., $p_{sw}p_0 = 1$, where $p_{sw} = p_{s1w} + p_{s2w} + p_{s3w} + p_{s4w}$ and it is referred as psum. Moreover, the values p_{s1w}, p_{s2w}, p_{s3w} and p_{s4w} are referred as psum1, psum2, psum3 and psum4, respectively, or simple psum parts.

We assume $K-M \gg 1$. Under this assumption $\rho^{K-M} \ll 1$ holds for the whole traffic range $\rho < 1$ and thus the term ρ^{K-M} can be neglected comparing to 1. Using it, the approximation of p_{s4w}, denoted by p^*_{s4w}, for the whole traffic range $\rho < 1$ can be given by

$$\begin{aligned}
p_{s4w^*} &= \frac{\rho}{1-\rho}\frac{N!}{M!}\left(\frac{\lambda}{\mu}\right)^{M-N}\frac{p_N}{p_0} \\
&= \frac{\rho}{1-\rho}\frac{N!}{M!}\left(\frac{\lambda}{\mu}\right)^{M-N}\frac{(\frac{\lambda}{\mu})^N}{N!}s_{L,N}\ \alpha\ \frac{p_L}{p_0} \\
&= \frac{\rho}{1-\rho}\frac{(\frac{\lambda}{\mu})^M}{M!}\sum_{i=L}^{N-1}\frac{i!}{(\frac{\lambda}{\mu})^i}\ \alpha\ \frac{p_L}{p_0}. \qquad (15)
\end{aligned}$$

The evaluation of the N dependency of the psum parts requires further approximations. However, these approximations themself depend on the traffic range. Therefore, in the following we split the investigation of the N dependency of p_0 according to the traffic ranges determining the appropriate approximations.

A. Traffic range $\frac{\lambda}{L\mu} > 1$

We assume $N-L \gg 1$. Under this assumption $(\frac{\lambda}{L\mu})^{N-L} \gg 1$ holds for the traffic range $\frac{\lambda}{L\mu} > 1$ and thus the term $(\frac{\lambda}{L\mu})^{N-L}$ dominates over 1. We make use of this in the approximations.

1) Approximation for p_{s1} and α: Using $(\frac{\lambda}{L\mu})^{N-L} \gg 1$ the second term of p_{s1} in (7) can be approximated as

$$\begin{aligned}
&\frac{\frac{\lambda}{L\mu}-(\frac{\lambda}{L\mu})^{N-L}}{1-\frac{\lambda}{L\mu}} - (N-L-1)(\frac{\lambda}{L\mu})^{N-L}}{1-(\frac{\lambda}{L\mu})^{N-L}} p_L \\
&\approx \frac{\frac{(\frac{\lambda}{L\mu})^{N-L}}{\frac{\lambda}{L\mu}-1} - (N-L-1)(\frac{\lambda}{L\mu})^{N-L}}{-(\frac{\lambda}{L\mu})^{N-L}} p_L \\
&= \left((N-L-1) - \frac{\frac{L\mu}{\lambda}}{1-\frac{L\mu}{\lambda}}\right) p_L.
\end{aligned}$$

This together with the upper limit $\sum_{k=0}^{L} \frac{(\frac{\lambda}{\mu})^k}{k!} \leq \frac{1}{1-\frac{L\mu}{\lambda}} \frac{(\frac{\lambda}{\mu})^L}{L!}$ for $\frac{\lambda}{\mu} > L$ (see in Appendix III-A applied to $q = \frac{\lambda}{\mu}$) gives the estimation for p_{s1} as

$$\begin{aligned}
p_{s1} &\approx \frac{1}{1-\frac{L\mu}{\lambda}} \frac{(\frac{\lambda}{\mu})^L}{L!} p_0 + \left((N-L-1) - \frac{\frac{L\mu}{\lambda}}{1-\frac{L\mu}{\lambda}}\right) p_L \\
&= \left((N-L-1) + \frac{1}{1-\frac{L\mu}{\lambda}} - \frac{\frac{L\mu}{\lambda}}{1-\frac{L\mu}{\lambda}}\right) p_L \\
&= (N-L) p_L.
\end{aligned}$$

Note that $\frac{1}{1-\frac{L\mu}{\lambda}}$, for $\frac{\lambda}{L\mu} > \xi$ with $\xi = 1.2$, is small compared to $(N-L)$ due to $N-L \gg 1$ and hence the effect of the overestimation by the upper limit can be neglected. Similarly, by using $(\frac{\lambda}{L\mu})^{N-L} \gg 1$, α can be approximated in this traffic range as

$$\alpha \approx \left(\frac{\lambda}{L\mu}\right)^{N-L-1} \frac{\frac{\lambda}{L\mu}-1}{(\frac{\lambda}{L\mu})^{N-L}} = \left(1-\frac{L\mu}{\lambda}\right).$$

By using the above approximations of p_{s1} and α, the approximations of the terms p_{s1w}, p_{s2w} and p_{s3w}, denoted by p^*_{s1w}, p^*_{s2w} and p^*_{s3w}, respectively, can be given as

$$\begin{aligned}
p^*_{s1w} &= (N-L)\frac{p_L}{p_0} \\
p^*_{s2w} &= \sum_{k=L+1}^{N} \frac{(\frac{\lambda}{\mu})^k}{k!} \sum_{i=L}^{k-1} \frac{i!}{(\frac{\lambda}{\mu})^i} \left(1-\frac{L\mu}{\lambda}\right) \frac{p_L}{p_0}, \\
p^*_{s3w} &= \sum_{k=N+1}^{M} \frac{(\frac{\lambda}{\mu})^k}{k!} \sum_{i=L}^{N-1} \frac{i!}{(\frac{\lambda}{\mu})^i} \left(1-\frac{L\mu}{\lambda}\right) \frac{p_L}{p_0}.
\end{aligned}$$

2) Behavior of $p_{s2w} + p_{s3w}$ in dependency of N: Hereinafter it will be shown that the sum $p^*_{s2w} + p^*_{s3w}$ is approximately independent of N when ρ is above some threshold, which depends on $\frac{M}{L}$.

Taking the difference of p^*_{s2w} and p^*_{s3w} with respect to N gives

$$\begin{aligned}
\Delta_N p^*_{s2w} &= p^*_{s2w}(N) - p^*_{s2w}(N-1) \\
&= \frac{(\frac{\lambda}{\mu})^N}{N!} \sum_{i=L}^{N-1} \frac{i!}{(\frac{\lambda}{\mu})^i} \left(1-\frac{L\mu}{\lambda}\right) \frac{p_L}{p_0}. \\
\Delta_N p^*_{s3w} &= p^*_{s3w}(N) - p^*_{s3w}(N-1) \\
&= \sum_{k=N+1}^{M} \frac{(\frac{\lambda}{\mu})^k}{k!} \sum_{i=L}^{N-1} \frac{i!}{(\frac{\lambda}{\mu})^i} \left(1-\frac{L\mu}{\lambda}\right) \frac{p_L}{p_0} \\
&\quad - \sum_{k=N}^{M} \frac{(\frac{\lambda}{\mu})^k}{k!} \sum_{i=L}^{N-2} \frac{i!}{(\frac{\lambda}{\mu})^i} \left(1-\frac{L\mu}{\lambda}\right) \frac{p_L}{p_0} \\
&= \sum_{k=N+1}^{M} \frac{(\frac{\lambda}{\mu})^k}{k!} \sum_{i=L}^{N-1} \frac{i!}{(\frac{\lambda}{\mu})^i} \left(1-\frac{L\mu}{\lambda}\right) \frac{p_L}{p_0} \\
&\quad - \sum_{k=N}^{M} \frac{(\frac{\lambda}{\mu})^k}{k!} \sum_{i=L}^{N-1} \frac{i!}{(\frac{\lambda}{\mu})^i} \left(1-\frac{L\mu}{\lambda}\right) \frac{p_L}{p_0} \\
&\quad + \sum_{k=N}^{M} \frac{(\frac{\lambda}{\mu})^k}{k!} \frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}} \left(1-\frac{L\mu}{\lambda}\right) \frac{p_L}{p_0} \\
&= -\frac{(\frac{\lambda}{\mu})^N}{N!} \sum_{i=L}^{N-1} \frac{i!}{(\frac{\lambda}{\mu})^i} \left(1-\frac{L\mu}{\lambda}\right) \frac{p_L}{p_0} \\
&\quad + \sum_{k=N}^{M} \frac{(\frac{\lambda}{\mu})^k}{k!} \frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}} \left(1-\frac{L\mu}{\lambda}\right) \frac{p_L}{p_0}.
\end{aligned}$$

Thus for $\Delta_N(p^*_{s2w} + p^*_{s3w})$ we get

$$\Delta_N(p^*_{s2w} + p^*_{s3w}) = \sum_{k=N}^{M} \frac{(\frac{\lambda}{\mu})^k}{k!} \frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}} \left(1-\frac{L\mu}{\lambda}\right) \frac{p_L}{p_0}.$$

The sum $p^*_{s2w} + p^*_{s3w}$ at $N = L+1$, i.e., at the smallest possible value of N can be expressed as

$$\begin{aligned}
&p^*_{s2w}(L+1) + p^*_{s3w}(L+1) = \frac{(\frac{\lambda}{\mu})^{L+1}}{(L+1)!} \frac{L!}{(\frac{\lambda}{\mu})^L} \left(1-\frac{L\mu}{\lambda}\right) \frac{p_L}{p_0} \\
&+ \sum_{k=L+2}^{M} \frac{(\frac{\lambda}{\mu})^k}{k!} \frac{L!}{(\frac{\lambda}{\mu})^L} \left(1-\frac{L\mu}{\lambda}\right) \frac{p_L}{p_0} \\
&= \sum_{k=L+1}^{M} \frac{(\frac{\lambda}{\mu})^k}{k!} \frac{L!}{(\frac{\lambda}{\mu})^L} \left(1-\frac{L\mu}{\lambda}\right) \frac{p_L}{p_0} \\
&= \sum_{k=L+1}^{M} \frac{(\frac{\lambda}{\mu})^k}{k!} \left(1-\frac{L\mu}{\lambda}\right),
\end{aligned}$$

where we used $\frac{p_L}{p_0} = \frac{(\frac{\lambda}{\mu})^L}{L!}$.

Now we compare $\Delta_N(p^*_{s2w} + p^*_{s3w})$ to $p^*_{s2w}(L+1) + p^*_{s3w}(L+1)$. The expression of $\Delta_N(p^*_{s2w} + p^*_{s3w})$ and $p^*_{s2w}(L+1) + p^*_{s3w}(L+1)$ can be rewritten by applying the formula for $\sum_{k=A+1}^{B} \frac{q^k}{k!} \frac{A!}{q^A}$ (see in Appendix III-B) with $q = \frac{\lambda}{\mu}$ to $A = N-1$, $B = M$ and $A = L$, $B = M$, respectively as

$$\Delta_N(p^*_{s2w}+p^*_{s3w}) = \sum_{k=N}^{M} \frac{(\frac{\lambda}{\mu})^k}{k!} \frac{N-1!}{(\frac{\lambda}{\mu})^{N-1}} \frac{p_L}{p_0}\left(1-\frac{L\mu}{\lambda}\right)$$
$$= f_1^* \frac{1-(f_1^*)^{M-N+1}}{1-f_1^*}\left(1-\frac{L\mu}{\lambda}\right)\frac{p_L}{p_0},$$
where $f_1^* = \frac{\frac{\lambda}{\mu}}{f_0^*}$ and $N \le f_0^* \le M$,

$$p^*_{s2w}(L+1)+p^*_{s3w}(L+1) = \sum_{k=L+1}^{M} \frac{(\frac{\lambda}{\mu})^k}{k!}\left(1-\frac{L\mu}{\lambda}\right)$$
$$= \frac{(\frac{\lambda}{\mu})^L}{L!}\sum_{k=L+1}^{M}\frac{(\frac{\lambda}{\mu})^k}{k!}\frac{L!}{(\frac{\lambda}{\mu})^L}\left(1-\frac{L\mu}{\lambda}\right)$$
$$= f_1\frac{1-(f_1)^{M-L}}{1-f_1}\left(1-\frac{L\mu}{\lambda}\right)\frac{p_L}{p_0},$$
where $f_1 = \frac{\frac{\lambda}{\mu}}{f_0}$ and $L+1 \le f_0 \le M$,

as well as $f_1^* < f_1$ due to $L < N-1$ for $N \ge L+2$, where $\Delta_N(p^*_{s2w}+p^*_{s3w})$ can be interpreted. We investigate the case when $f_1 > 1$ holds. In this case $p^*_{s2w}(L+1)+p^*_{s3w}(L+1) \approx \frac{f_1^{M-L+1}}{f_1-1}(1-\frac{L\mu}{\lambda})\frac{p_L}{p_0}$ due to $M-L \gg 1$. If $f_1^* < 1$ then

$$\frac{p^*_{s2w}(L+1)+p^*_{s3w}(L+1)}{\Delta_N(p^*_{s2w}+p^*_{s3w})} \approx \frac{f_1^{M-L+1}}{f_1-1}\frac{1-f_1^*}{f_1^*} \gg 1,$$

since the dominating term $f_1^{M-L+1} \gg 1$. If $f_1^* > 1$ then the quotient $\frac{p^*_{s2w}(L+1)+p^*_{s3w}(L+1)}{\Delta_N(p^*_{s2w}+p^*_{s3w})}$ can be approximated as

$$\frac{p^*_{s2w}(L+1)+p^*_{s3w}(L+1)}{\Delta_N(p^*_{s2w}+p^*_{s3w})} \approx \frac{f_1^{M-L+1}}{(f_1^*)^{M-N+2}}\frac{f_1^*-1}{f_1-1}$$
$$= \left(\frac{f_1}{f_1^*}\right)^{M-N+2} f_1^{N-L-1}\frac{f_1^*-1}{f_1-1} \gg 1,$$

since for the major term of the expression $\left(\frac{f_1}{f_1^*}\right)^{M-N+2} f_1^{N-L-1} \gg 1$ holds, because either $M-N+2 \gg 1$ (recall that $f_1^* < f_1$) or $N-L-1 \gg 1$ depending on the value of N. Thus for the traffic range, for which $f_1 > 1$ holds, $\Delta_N(p^*_{s2w}+p^*_{s3w})$ can be neglected comparing to $p^*_{s2w}(L+1)+p^*_{s3w}(L+1)$. Applying it recursively to $p^*_{s2w}(N)+p^*_{s3w}(N)$ for $N = L+1, \dots M-1$ we get

$$p^*_{s2w}(N)+p^*_{s3w}(N) \approx \sum_{k=L+1}^{M}\frac{(\frac{\lambda}{\mu})^k}{k!}\left(1-\frac{L\mu}{\lambda}\right)$$
$$\text{for } N = L+1, \dots, M, \text{ for } f_1 > 1.$$

This means that the sum $p_{s2w}+p_{s3w}$ approximately independent of N in the traffic range, for which $f_1 > 1$ holds.

However, in traffic range, for which $f_1 < 1$ holds, $p^*_{s2w}(L+1)+p^*_{s3w}(L+1) \approx \frac{f_1}{1-f_1}(1-\frac{L\mu}{\lambda})\frac{p_L}{p_0}$ due to $M-L \gg 1$ and hence the quotient $\frac{p^*_{s2w}(L+1)+p^*_{s3w}(L+1)}{\Delta_N(p^*_{s2w}+p^*_{s3w})}$ can be approximated as

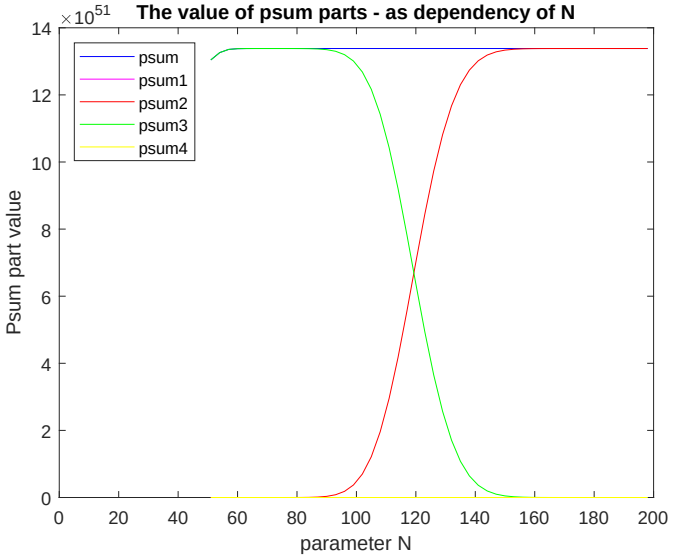

Figure 2. The magnitudes of $p_{s1w}, p_{s2w}, p_{s3w}, p_{s4w}$ and their sum, $M = 200, L = 50, \rho = 0.6$.

$$\frac{p^*_{s2w}(L+1)+p^*_{s3w}(L+1)}{\Delta_N(p^*_{s2w}+p^*_{s3w})} \approx \frac{f_1}{f_1^*-(f_1^*)^{M-N+2}}\frac{1-f_1^*}{1-f_1}.$$

This formula shows that, depending on the values of $f_1^* < f_1$ and $f_1 < 1$, the above quotient can fall in the magnitude of 1, and therefore $\Delta_N(p^*_{s2w}+p^*_{s3w})$ can not be neglected compared to $p^*_{s2w}(L+1)+p^*_{s3w}(L+1)$. This implies that the value $p_{s2w}+p_{s3w}$ depends on N in the traffic range, for which $f_1 < 1$ holds.

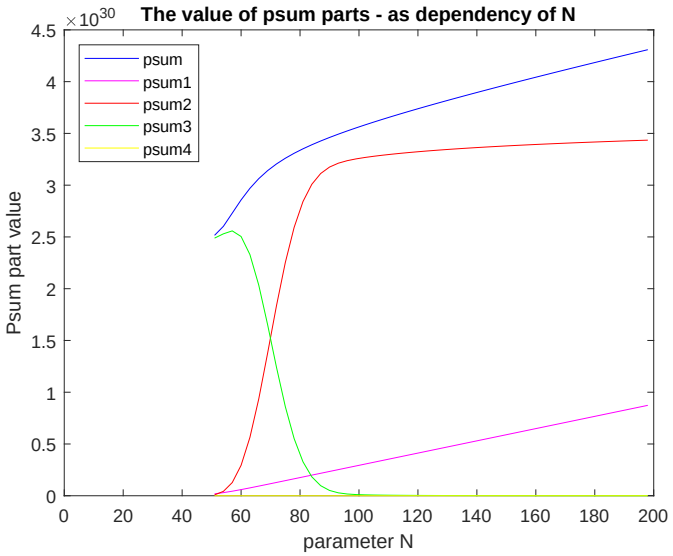

Figure 3. The magnitudes of $p_{s1w}, p_{s2w}, p_{s3w}, p_{s4w}$ and their sum, $M = 200, L = 50, \rho = 0.35$.

For the completeness, the traffic range boundary, for which f_1 is close to 1, still have to be estimated. An upper limit can be obtained for $\rho M = \frac{\lambda}{\mu}$ by applying the upper limit

for q at f_1 close to 1 (see in Appendix III-C) to $q = \frac{\lambda}{\mu}$ and $A = L$ leading to $\frac{\lambda}{\mu} < Le^{\frac{5.5}{L}}$. In order to ensure the validity of the above approximations the value of f_1 must be somewhat above 1, thus we set $f_1 = 1.1$. Taking into account that, by neglecting the dependency of f_0 on $\frac{\lambda}{\mu}$, f_1 is approximately proportional to $\frac{\lambda}{\mu}$ we get an upper limit at the traffic range boundary, denoted by ρ_c, as

$$\rho_c = \frac{\lambda}{M\mu} < \frac{L}{M} 1.1 e^{\frac{5.5}{L}} \lessapprox \frac{L}{M} 1.2 \text{ for } L \geq 60.$$

Although the above factor $1.1e^{\frac{5.5}{L}} \approx 1.2$ increases for $L < 60$, the quotient $\frac{M}{L}$ also increases and thus it causes a counter-effect due to higher overestimation of $\frac{\lambda}{\mu}$ by the upper limit, which together justify the keeping of the approximate factor 1.2.

3) The sum $p_{s2w}+p_{s3w}$ versus p_{s1w}: If $f_1 > 1$ then $p^*_{s2w} + p^*_{s3w} \approx \frac{f_1^{M-L+1}}{f_1-1}(1 - \frac{L\mu}{\lambda})\frac{p_L}{p_0}$ due to $M - L \gg 1$. In this case the major term $f_1^{M-L+1} \gg (N-L)$ and hence $p^*_{s2w}+p^*_{s3w} \gg (N-L)\frac{p_L}{p_0} = p^*_{s1w}$, which implies that p^*_{s1w} can be neglected comparing to $p^*_{s2w} + p^*_{s3w}$.

On the other hand, if $f_1 < 1$ then $p^*_{s2w} + p^*_{s3w} \approx \frac{f_1}{1-f_1}(1-\frac{L\mu}{\lambda})\frac{p_L}{p_0}$ and the coefficient $\frac{f_1}{1-f_1}(1-\frac{L\mu}{\lambda})$ becomes less than $N - L$. It follows that p^*_{s1w} dominates over $p^*_{s2w}+p^*_{s3w}$ and hence $p^*_{s1w} + p^*_{s2w} + p^*_{s3w}$ depends linearly on N in the corresponding traffic range.

4) Effect of p_{s4w}: Observe that the expression of p_{s4w^*} in (15) without the first multiplication factor $\frac{\rho}{1-\rho}$ is the same as the value of the expression of p_{s3w} in (14) for $k = M$. The item $\frac{(\frac{\lambda}{\mu})^k}{k!}$ in the sum $\sum_{k=L+1}^{M} \frac{(\frac{\lambda}{\mu})^k}{k!}$ takes its maximum at $k \approx \frac{\lambda}{\mu}$, since $\frac{\frac{\lambda}{\mu}}{k} > 1$ for $k < \frac{\lambda}{\mu}$ and $\frac{\frac{\lambda}{\mu}}{k} < 1$ for $k > \frac{\lambda}{\mu}$. It follows that the major part of p^*_{s3w} is determined by the items $(\frac{\lambda}{\mu} - \Delta) \leq k \leq \frac{\lambda}{\mu} + \Delta)$ for some Δ. The last item with $k = M$ can contribute to the sum practically only if $\frac{\lambda}{\mu} > M - \Delta$. For such a case $\rho = \frac{\lambda}{M\mu}$ is close to 1 for which also the multiplication factor in (15), $\frac{\rho}{1-\rho}$, has a value $\gg 1$.

It follows that p_{s4w} can emerge to the magnitude of the value of p_{s3w}, and thus that of the sum $p_{s2w} + p_{s3w}$, only for ρ close to 1, otherwise it is negligible compared to that sum.

On the other hand, we show that p_{s4w} is independent of N for values of ρ close to 1. Applying the formula for $\sum_{k=A+1}^{B} \frac{k!}{q^k}\frac{q^A}{A!}$ (see in Appendix III-D) to $A = L - 1$, $B = N - 1$, the expression of p^*_{s4w}, (15), can be rearranged as

$$\begin{aligned} p_{s4w^*} &= \frac{\rho}{1-\rho}\frac{(\frac{\lambda}{\mu})^M}{M!}\sum_{i=L}^{N-1}\frac{i!}{(\frac{\lambda}{\mu})^i}\ \alpha\ \frac{p_L}{p_0} \\ &= \frac{\rho}{1-\rho}\frac{(\frac{\lambda}{\mu})^M}{M!}\sum_{i=L}^{N-1}\frac{i!}{(\frac{\lambda}{\mu})^i}\frac{(\frac{\lambda}{\mu})^L}{L!}\ \alpha \\ &= \frac{\rho}{1-\rho}\ g_1\frac{1-(g_1)^{N-L}}{1-g_1}\ \frac{\frac{\lambda}{\mu}}{L}\ \frac{(\frac{\lambda}{\mu})^M}{M!}\ \alpha, \\ &\text{where } g_1 = \frac{g_0}{\frac{\lambda}{\mu}} \text{ and } L \leq g_0 \leq N-1. \end{aligned}$$

It can be seen from the above formula that, for values of ρ close to 1, for which $\rho > \frac{N}{M} > \frac{g_0}{M} \Leftrightarrow \frac{\lambda}{\mu} > g_0$ holds, p_{s4w^*} can be approximated as $\frac{\rho}{1-\rho}\ \frac{g_1}{1-g_1}\ \frac{\frac{\lambda}{\mu}}{L}\ \frac{(\frac{\lambda}{\mu})^M}{M!}\ (1 - \frac{L\mu}{\lambda})$, which is independent of N.

5) Illustrating the relations of the magnitudes of $p_{s1w}, p_{s2w}, p_{s3w}, p_{s4w}$ and their sum: The magnitudes of $p_{s1w}, p_{s2w}, p_{s3w}, p_{s4w}$ and their sum compared to each other are illustrated on Figures 2, 3 and 4. The figures were created by varying ρ under the parameter setting $M = 200, L = 50, K = 300, \mu = 1$. It can be seen on Figure 2 that the sum $p_{s3w} + p_{s4w}$ is approximately independent of N and dominates over the other terms as expected due to $\rho = 0.6 > 1.2\frac{L}{M} = 0.3$. These relations start to change as ρ gets closer to $1.2\frac{L}{M}$, where the sum $p_{s3w} + p_{s4w}$ starts to depend on N and p_{s1w} starts to emerge comparing to $p_{s3w} + p_{s4w}$ as well as starts to affect the overall sum p_{sw} to become linear with N.

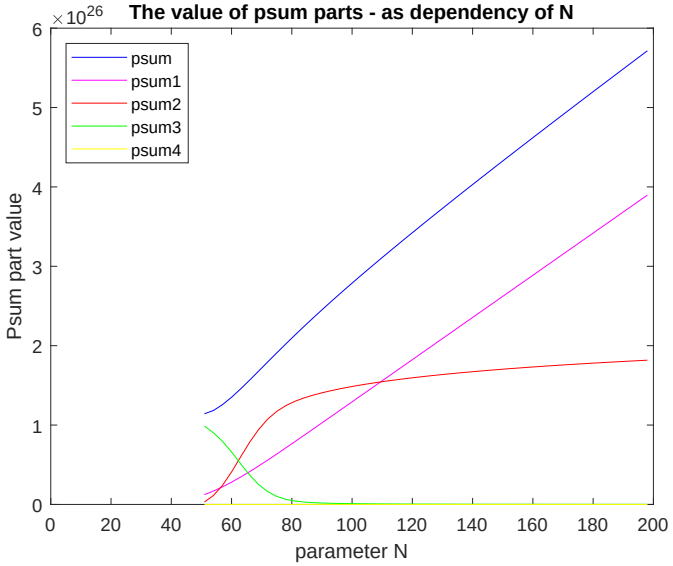

Figure 4. The magnitudes of $p_{s1w}, p_{s2w}, p_{s3w}, p_{s4w}$ and their sum, $M = 200, L = 50, \rho = 0.3$.

This can be seen on Figure 3 with $\rho = 0.35$ and even more on Figure 4 with $\rho = 0.3$, where the dependency of the overall sum p_{sw} on N becomes clearly linear.

B. Traffic range $\frac{\lambda}{L\mu} < 1$

We assume again $N - L \gg 1$. Under this assumption $(\frac{\lambda}{L\mu})^{N-L} \ll 1$ holds for the traffic range $\frac{\lambda}{L\mu} < 1$ and thus the term $(\frac{\lambda}{L\mu})^{N-L}$ can be neglected compared to 1.

1) Approximation for p_{s1} and α: Using $(\frac{\lambda}{L\mu})^{N-L} \ll 1$ the expression of p_{s1} in (7) can be approximated as

$$\begin{aligned} p_{s1} \approx &\sum_{k=0}^{L}\frac{(\frac{\lambda}{\mu})^k}{k!}p_0 \\ &+ \left(\frac{\frac{\lambda}{L\mu}}{1-\frac{\lambda}{L\mu}} - (N-L-1)\left(\frac{\lambda}{L\mu}\right)^{N-L}\right)p_L. \quad (16) \end{aligned}$$

Using the notations $h=\frac{\lambda}{L\mu}$ and $z=N-L$ the derivative of the major part of the last term in (16) with respect to z is given by $(zh^z)'=h^z(z\ln(h)+1)$. Thus, a condition for $z\ln(h)+1<0$ is $z>\frac{-1}{\ln(h)}$, since $\ln(h)<0$. The inequalities $\ln(h)<-0.22$ and $\frac{-1}{\ln(h)}<4.5$ hold for $h<0.8$, and hence $z>10$ is sufficient for $(zh^z)'<0$. It follows that for $N-L\gg 1$, e.g., for $N-L>10$, the nonnegative term $(N-L)(\frac{\lambda}{L\mu})^{N-L}$ is monotone decreasing with increasing $N-L$. Then it is enough to compute the values of the middle and the last terms of (16) for several values of $\frac{\lambda}{L\mu}$ with $N-L=10$ to see the magnitude of the last term comparing to the middle one. It turns out that except for high $\frac{\lambda}{L\mu}$ (above 0.8) and low $(N-L)$ ($\lessapprox 40$) the last term of (16) can be neglected compared to the middle one. Taking it into account, we get another approximation for p_{s1} as

$$p_{s1}\approx\sum_{k=0}^{L}\frac{(\frac{\lambda}{\mu})^k}{k!}p_0+\left(\frac{\frac{\lambda}{L\mu}}{1-\frac{\lambda}{L\mu}}\right)p_L. \tag{17}$$

Using $(\frac{\lambda}{L\mu})^{N-L}\ll 1$ in the expression of α in (3) we get the approximation of α in this traffic range as

$$\alpha\approx\left(1-\frac{\lambda}{L\mu}\right)\left(\frac{\lambda}{L\mu}\right)^{N-L-1}. \tag{18}$$

2) The properties of p_{s1w}: Using (17), the sum p_{s1w} can be approximated as

$$p_{s1w}\approx\sum_{k=0}^{L}\frac{(\frac{\lambda}{\mu})^k}{k!}+\left(\frac{\frac{\lambda}{L\mu}}{1-\frac{\lambda}{L\mu}}\right)\frac{p_L}{p_0}. \tag{19}$$

The formula (19) also shows that p_{s1w} is approximately independent of N in the traffic range of $\frac{\lambda}{L\mu}<1$, except for the rather small sub-range mentioned at the derivation of (17).

In the following we establish a lower limit for the first term of p_{s1w} for $\frac{\lambda}{\mu}=tL$ and $1\le L_0\le L$.

$$\begin{aligned}
&\sum_{k=0}^{L}\frac{(\frac{\lambda}{\mu})^k}{k!}=\sum_{k=0}^{L}\frac{(\frac{\lambda}{\mu})^{k-L}}{k!/L!}\frac{(\frac{\lambda}{\mu})^L}{L!}=\sum_{k=0}^{L}\frac{L!/k!\ p_L}{(\frac{\lambda}{\mu})^{L-k}p_0}\\
&=\left(1+\frac{L}{\frac{\lambda}{\mu}}+\frac{L(L-1)}{(\frac{\lambda}{\mu})^2}+\ldots+\frac{L!}{(\frac{\lambda}{\mu})^L}\right)\frac{p_L}{p_0}\\
&=\left(1+\frac{1}{t}+\frac{(1-\frac{1}{L})}{t^2}+\ldots+\frac{(1-\frac{1}{L})\ldots(1-\frac{L-1}{L})}{t^L}\right)\frac{p_L}{p_0}\\
&\ge\left(1+\frac{1}{t}+\frac{(1-\frac{1}{L})}{t^2}+\ldots+\frac{(1-\frac{1}{L})\ldots(1-\frac{L_0-1}{L})}{t^{L_0}}\right)\frac{p_L}{p_0}\\
&\ge\left(1+\frac{1}{t}+\frac{(1-\frac{1}{L_0})}{t^2}+\ldots+\frac{(1-\frac{1}{L_0})\ldots(1-\frac{L_0-1}{L_0})}{t^{L_0}}\right)\frac{p_L}{p_0}.
\end{aligned}$$

This limit can be computed for a given values of t and L_0. For a realistic range it yields

$$\sum_{k=0}^{L}\frac{(\frac{\lambda}{\mu})^k}{k!}>8.22\ \frac{p_L}{p_0},\quad \frac{\lambda}{\mu}\le 0.8L,\quad L\ge 10. \tag{20}$$

Therefore, the magnitude of the multiplication factor of $\frac{p_L}{p_0}$ in the first term of p_{s1w} falls in the magnitude of ten and above.

3) The sum $p_{s2w}+p_{s3w}$ compared to p_{s1w}: The approximate expression p^*_{s2w} and p^*_{s3w} can be obtained by applying (18) in the expression of p_{s2w} and p_{s3w} in (14), respectively.

We investigate the sum $p^*_{s2w}+p^*_{s3w}$ at $N=L+1$, i.e., at the smallest possible value of N.

$$\begin{aligned}
p^*_{s2w}(L+1)+p^*_{s3w}(L+1)&=\frac{(\frac{\lambda}{\mu})^{L+1}}{(L+1)!}\frac{L!}{(\frac{\lambda}{\mu})^L}\left(1-\frac{\lambda}{L\mu}\right)\frac{p_L}{p_0}\\
&+\sum_{k=L+2}^{M}\frac{(\frac{\lambda}{\mu})^k}{k!}\frac{L!}{(\frac{\lambda}{\mu})^L}\ \left(1-\frac{\lambda}{L\mu}\right)\ \frac{p_L}{p_0}\\
&=\sum_{k=L+1}^{M}\frac{(\frac{\lambda}{\mu})^k}{k!}\frac{L!}{(\frac{\lambda}{\mu})^L}\left(1-\frac{\lambda}{L\mu}\right)\frac{p_L}{p_0}.
\end{aligned}$$

Applying the formula for $\sum_{k=A+1}^{B}\frac{q^k}{k!}\frac{A!}{q^A}$ (see in Appendix III-B) with $q=\frac{\lambda}{\mu}$ to $A=L$, $B=M$ we get an upper limit for the sum $p^*_{s2w}(L+1)+p^*_{s3w}(L+1)$ as

$$\begin{aligned}
&p^*_{s2w}(L+1)+p^*_{s3w}(L+1)=\frac{\lambda}{f_0\mu}\frac{1-(\frac{\lambda}{f_0\mu})^{M-L}}{1-\frac{\lambda}{f_0\mu}}\\
&*\left(1-\frac{\lambda}{L\mu}\right)\frac{p_L}{p_0}<\frac{\frac{\lambda}{L\mu}-(\frac{\lambda}{L\mu})^{M-L+1}}{1-\frac{\lambda}{L\mu}}\left(1-\frac{\lambda}{L\mu}\right)\frac{p_L}{p_0}\\
&=\left[\frac{\lambda}{L\mu}-\left(\frac{\lambda}{L\mu}\right)^{M-L+1}\right]\frac{p_L}{p_0}<\frac{\lambda}{L\mu}\frac{p_L}{p_0}<\frac{p_L}{p_0},
\end{aligned} \tag{21}$$

where we utilized that $L<f_0$ and $M-L\gg 1$.

Now we investigate the effect of increasing N by 1 on $p_{s2w}(N)+p_{s3w}(N)$.

$$\begin{aligned}
&p_{s2w}(N+1)+p_{s3w}(N+1)=\frac{\alpha(N+1,L)}{\alpha(N,L)}p_{s2w}(N)\\
&+\frac{(\frac{\lambda}{\mu})^{N+1}}{(N+1)!}\sum_{i=L}^{N}\frac{i!}{(\frac{\lambda}{\mu})^i}\alpha(N+1,L)\frac{p_L}{p_0}+\frac{\alpha(N+1,L)}{\alpha(N,L)}p_{s3w}(N)\\
&+\sum_{k=N+1}^{M}\frac{(\frac{\lambda}{\mu})^k}{k!}\frac{N!}{(\frac{\lambda}{\mu})^N}\alpha(N+1,L)\frac{p_L}{p_0}\\
&-\frac{(\frac{\lambda}{\mu})^{N+1}}{(N+1)!}\sum_{i=L}^{N}\frac{i!}{(\frac{\lambda}{\mu})^i}\alpha(N+1,L)\frac{p_L}{p_0}\\
&=\frac{\lambda}{L\mu}\Big(p_{s2w}(N)+p_{s3w}(N)\Big)\\
&+\sum_{k=N+1}^{M}\frac{(\frac{\lambda}{\mu})^k}{k!}\frac{N!}{(\frac{\lambda}{\mu})^N}\alpha(N+1,L)\frac{p_L}{p_0}.
\end{aligned}$$

Applying again the formula for $\sum_{k=A+1}^{B}\frac{q^k}{k!}\frac{A!}{A^k}$ (see in Appendix III-B) with $q=\frac{\lambda}{\mu}$ to $A=N$, $B=M$ and assuming $p^*_{s2w}(N)+p^*_{s3w}(N)<\frac{p_L}{p_0}$ we get an upper limit for $p^*_{s2w}(N+1)+p^*_{s3w}(N+1)$ as

$$p^*_{s2w}(N+1)+p^*_{s3w}(N+1) < \frac{\lambda}{L\mu}\left(p^*_{s2w}(N)+p^*_{s3w}(N)\right)$$
$$+\frac{\lambda}{N\mu}\frac{1-(\frac{\lambda}{N\mu})^{M-N}}{1-\frac{\lambda}{N\mu}}\left(1-\frac{\lambda}{L\mu}\right)\left(\frac{\lambda}{L\mu}\right)^{N-L}\frac{p_L}{p_0}$$
$$<\left(p^*_{s2w}(N)+p^*_{s3w}(N)\right)+\frac{\frac{\lambda}{N\mu}}{1-\frac{\lambda}{N\mu}}\left(1-\frac{\lambda}{L\mu}\right)\left(\frac{\lambda}{L\mu}\right)^{N-L}$$
$$\times\frac{p_L}{p_0} < \frac{\lambda}{L\mu}\frac{p_L}{p_0}+\frac{\frac{\lambda}{N\mu}}{1-\frac{\lambda}{N\mu}}\left(1-\frac{\lambda}{N\mu}\right)\left(\frac{\lambda}{L\mu}\right)^{N-L}\frac{p_L}{p_0}$$
$$=\left[\frac{\lambda}{L\mu}+\frac{\lambda}{N\mu}\left(\frac{\lambda}{L\mu}\right)^{N-L}\right]\frac{p_L}{p_0} < \frac{\lambda}{L\mu}\left[1+\left(\frac{\lambda}{L\mu}\right)^{N-L}\right]\frac{p_L}{p_0},$$

where we used $0<\frac{\lambda}{N\mu}<\frac{\lambda}{L\mu}<1$ and $M-N\geq 0$.

The factor $\frac{\lambda}{L\mu}\left(1+(\frac{\lambda}{L\mu})^{N-L}\right)$ is monotone increasing with $\frac{\lambda}{\mu}$ and less than 1 for the range of $\frac{\lambda}{\mu}<0.8L$ and $N-L\gg 1$, e.g., $N-L>10$. Thus, starting with (21) and applying mathematical induction we get

$$p_{s2w}(N)+p_{s3w}(N) < \frac{p_L}{p_0} \quad \text{for } N=L+1,\ldots,M. \quad (22)$$

Comparing (22) to (20) shows that $p_{s2w}+p_{s3w}$ is negligible comparing to p_{s1w} in the major part of the range $\frac{\lambda}{L\mu}<1$.

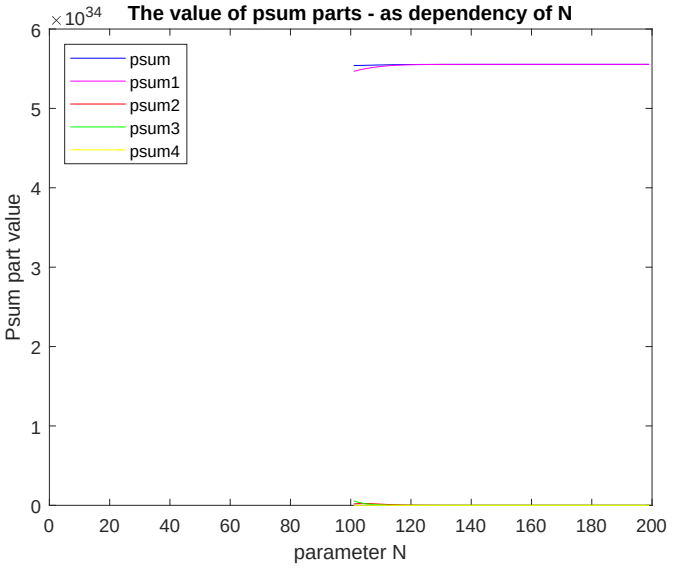

Figure 5. The magnitudes of $p_{s1w}, p_{s2w}, p_{s3w}, p_{s4w}$ and their sum, $M=200, L=100, \rho=0.4$.

4) The sum p_{s4w} compared to p_{s1w}: We assume $\frac{M}{L}\geq 2$. This implies $\rho\leq 0,5$ in the traffic range of $\frac{\lambda}{L\mu}<1$. Hence, the relation $\frac{\rho}{1-\rho}\leq 1$ holds. Using it in (15) leads to an upper limit for p_{s4w} as

$$p_{s4w^*}=\frac{\rho}{1-\rho}\frac{N!}{M!}\left(\frac{\lambda}{\mu}\right)^{M-N}\frac{p_N}{p_0}\leq\frac{N!}{M!}\left(\frac{\lambda}{\mu}\right)^{M-N}\frac{p_N}{p_0}.$$

Observe that it equals to the value of the expression of p_{s3w} in (14) for $k=M$. It follows that $p_{s4w^*}<p_{s3w}$ and therefore p_{s4w} can be neglected compared to p_{s1w} due to $p_{s2w}+p_{s3w}\ll p_{s1w}$.

5) Illustrating the relations of the magnitudes of $p_{s1w}, p_{s2w}, p_{s3w}, p_{s4w}$ and their sum: The magnitudes of $p_{s1w}, p_{s2w}, p_{s3w}, p_{s4w}$ and their sum compared to each other are illustrated on Figures 5, 6 and 7. The figures were created by varying ρ under the parameter setting $M=200, L=100, K=250, \mu=1$. It can be seen on Figure 5 that the probability coefficient p_{s1w} is approximately independent of N and dominates over the other terms as expected due to $\frac{\lambda}{L\mu}\leq 0.8 \Leftrightarrow \rho\leq 0.8\frac{L}{M}=0.4$. This independence start to change as ρ becomes higher than $0.8\frac{L}{M}$, where the probability coefficient p_{s1w} and thus also the overall sum p_{sw} start to depend on N for low $N-L$.

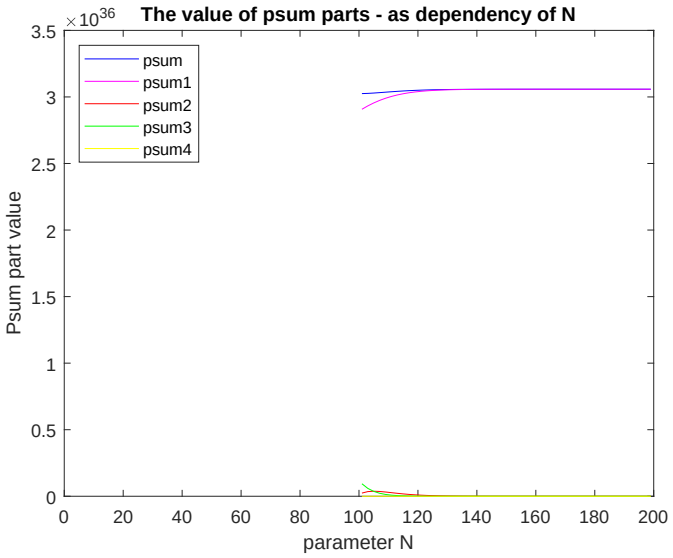

Figure 6. The magnitudes of $p_{s1w}, p_{s2w}, p_{s3w}, p_{s4w}$ and their sum, $M=200, L=100, \rho=0.42$.

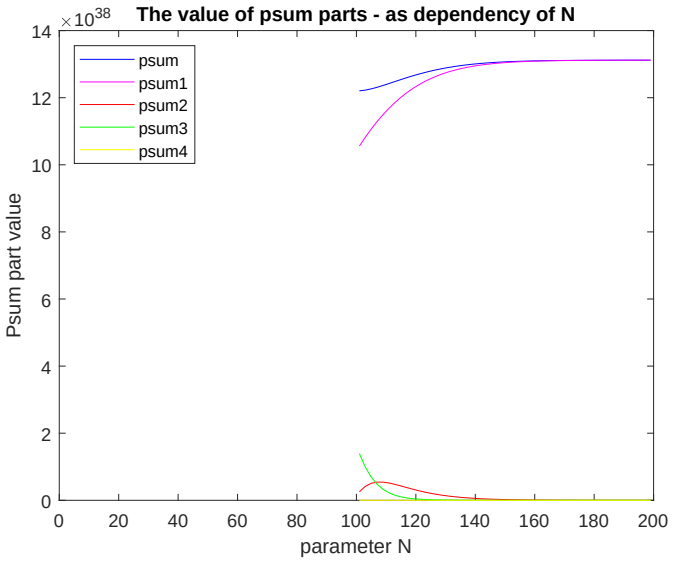

Figure 7. The magnitudes of $p_{s1w}, p_{s2w}, p_{s3w}, p_{s4w}$ and their sum, $M=200, L=100, \rho=0.45$.

This can be seen on Figure 6 with $\rho = 0.42$ and even more on Figure 7 with $\rho = 0.45$, where also the starting increase of the magnitude of sum $p_{s2w} + p_{s3w}$ can be observed.

C. Approximately N independent regions of p_0

Figure 8 illustrates the dependency of p_0 in the traffic range of $\frac{\lambda}{L\mu} > 1$ for the parameter setting $M = 400$, $L = 100$, $K = 450, \mu = 1$ and $\rho = 0.6$. It can be seen on the figure that p_0 is independent of N for $N \gtrapprox 110$, which corresponds to $N - L \approx 10 \gg 1$.

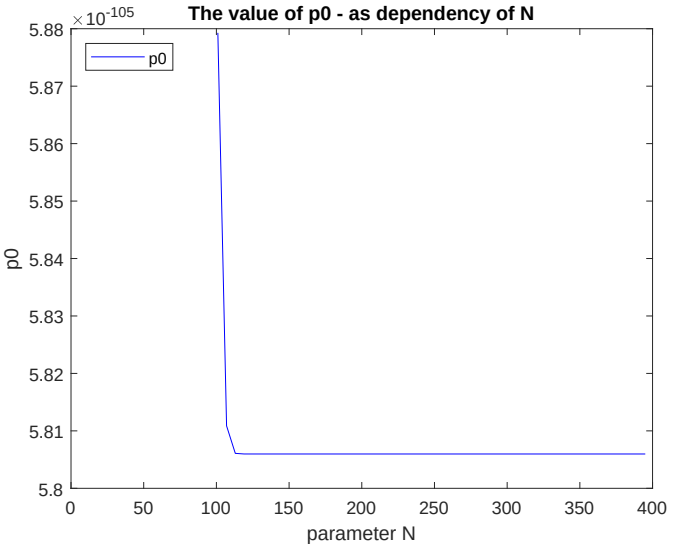

Figure 8. Probability p_0 in dependency of threshold N, high traffic range.

Similarly, the dependency of p_0 in the traffic range of $\frac{\lambda}{L\mu} < 1$ is illustrated on Figure 9 for the parameter setting $M = 200$, $L = 100$, $K = 250, \mu = 1$ and $\rho = 0.4$.

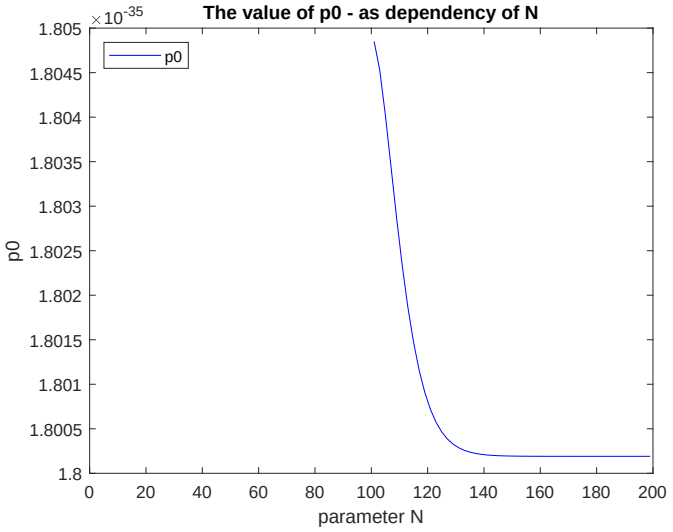

Figure 9. Probability p_0 in dependency of threshold N, low traffic range.

For the first sight it seems that there is a dependency on N in the range of $100 \leq N \lessapprox 130$. However, the value of p_0 changes in relative less then 0.3%. Therefore, the figure shows that p_0 is independent of N practically in the whole range of $N > 100$.

VI. Approximate minimization - in traffic range $\frac{\lambda}{L\mu} > 1$

In the N independent regions of p_0 also p_L is independent of N. Hence, the function to be minimized, (13), can be reduced by omitting the N independent term MC_{on} and dividing it by p_L. This results in a function F_2, to be minimized, as

$$F_2 = \left[\left(\lambda(C_A + C_D)(M - L) + \eta\ s_{L,N}\right)\alpha + C_W\tau\right] - (C_{on} - C_{off})(M - L)\frac{p_{s1}}{p_L}. \quad (23)$$

The optimization of (23) with respect to N still seems not to be tractable on analytic way due to the complex dependency of several of its terms on N, like $s_{L,N}$ or τ. Therefore, we simplify the optimization task by applying approximations, like the ones for α and p_{s1}. On the other hand, these approximations will restrict the parameter range, for which they hold.

A. Approximating the function to be minimized

Besides of the approximations for α and p_{s1}, we need approximation also for τ.

1) Approximation for τ: When $N - L \gg 1$ then $(\frac{\lambda}{L\mu})^{N-L} \gg 1$ holds for the traffic range $\frac{\lambda}{L\mu} > 1$ and thus the term $1 - (\frac{\lambda}{L\mu})^{N-L}$ can be approximated by $-(\frac{\lambda}{L\mu})^{N-L}$. Utilizing it in the expression of τ in (10) gives the approximation for τ as

$$\tau^* \approx \frac{\frac{\lambda}{L\mu}}{(1 - \frac{\lambda}{L\mu})^2} - (N - L)\ \frac{1}{\frac{\lambda}{L\mu} - 1} + \frac{(N - L)(N - L - 1)}{2} = \frac{1}{\frac{\lambda}{L\mu} - 1}\left(\frac{\frac{\lambda}{L\mu}}{\frac{\lambda}{L\mu} - 1} - (N - L)\right) + \frac{(N - L)(N - L - 1)}{2}.$$

2) Applying the approximations for p_0, α, τ and p_{s1}: The minimizing task can be significantly reduced by applying the approximations of α, τ and p_{s1} in (23). This leads to the approximate objective function F_{2app} as

$$F_{2app} = \left(\lambda(C_A + C_D)(M - L) + \eta\ s_{L,N}\right)\left(1 - \frac{L\mu}{\lambda}\right) + C_W\frac{1}{\frac{\lambda}{L\mu} - 1}\left(\frac{\frac{\lambda}{L\mu}}{\frac{\lambda}{L\mu} - 1} - (N - L)\right) + C_W\frac{(N - L)(N - L - 1)}{2} - (C_{on} - C_{off})(M - L)(N - L). \quad (24)$$

Figure 10 illustrates the approximation of the cost function F_2 by F_{2app} in dependency of threshold N for the parameter

setting $M = 200, L = 50, K = 300, \mu = 1, C_W = 50, C_{off} = 15, C_{on} = 50, C_a = 30, C_d = 20, C_R = 20$, and $\rho = 0.6$. The figure shows a very good match. The mismatch on the left side of the curve is caused by violating the condition $N - L >> 1$ as N becomes close to L.

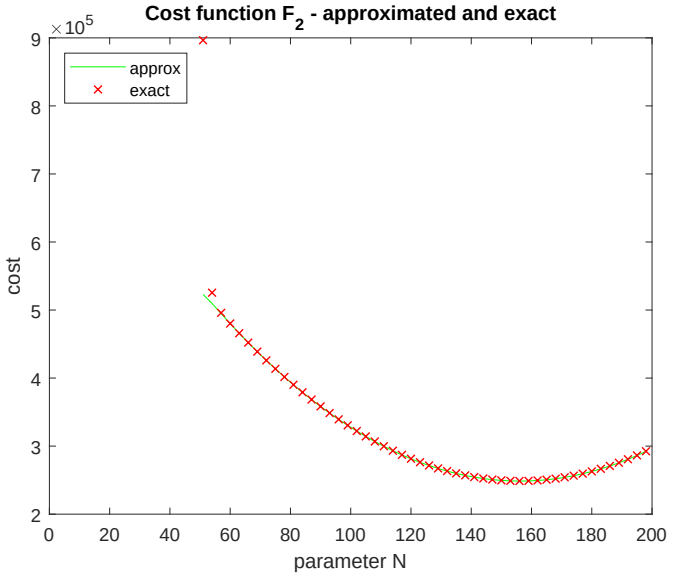

Figure 10. Exact and approximate values of the cost function F_2 in dependency of threshold N.

B. Approximate equation for determining the local minimum

We obtain an approximate equation for determining the local minimum of (24) by taking its difference with respect to N and setting $\Delta_N F_{2app} \approx 0$. Evaluating $\Delta\Big((N-L)(N-L)\Big)$ gives

$$\Delta\Big((N-L)(N-L)\Big) = (N-L)^2 - (N-1-L)^2 = 2(N-L-1)+1 = 2(N-L)-1.$$

Using $\Delta_N s_{L,N} = \frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}}$ and the above expression for $\Delta(N-L)(N-L)$ leads to the equation

$$\eta\left(1-\frac{L\mu}{\lambda}\right)\frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}} = (C_{on}-C_{off})(M-L) + C_W\frac{1}{\frac{\lambda}{L\mu}-1} - C_W(N-L-1). \quad (25)$$

In order to get closer to the solution of equation (25) first we investigate its structure.

C. Structure of the equation

To identify the structure of equation (25), we simplify its form by applying further approximations. The relation $K - M - 1 >> 1$ holds usually under practical settings. Hence, the term $(\frac{\lambda}{M\mu})^{K-M+1}$ can be neglected due to $\rho = \frac{\lambda}{M\mu} < 1$, which gives an approximation for σ as

$$\sigma = \frac{\lambda}{M\mu}\frac{1-(\frac{\lambda}{M\mu})^{K-M+1}}{(1-\frac{\lambda}{M\mu})^2} - (K-M+1)\frac{(\frac{\lambda}{M\mu})^{K-M+1}}{1-\frac{\lambda}{M\mu}} \approx \frac{\lambda}{M\mu}\frac{1}{(1-\frac{\lambda}{M\mu})^2} = \frac{\rho}{(1-\rho)^2}.$$

Applying rearrangement on the expression of η and applying again the negligibility of the term $(\frac{\lambda}{M\mu})^{K-M}$ in it, leads to an approximation for η as

$$\eta = \left[C_R\lambda\left(\frac{\lambda}{M\mu}\right)^{K-M}\frac{(\frac{\lambda}{\mu})^M}{M!} + C_W\sigma\frac{(\frac{\lambda}{\mu})^M}{M!}\right] \approx C_W\frac{\rho}{(1-\rho)^2}\frac{(\frac{\lambda}{\mu})^M}{M!}. \quad (26)$$

Using (26) in the equation (25) and further rearrangement gives the simplified form of the equation as

$$\frac{(\frac{\lambda}{\mu})^M}{M!}\frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}}u_0(\rho) = r(\rho,N), \text{ where} \quad (27)$$

$$u_0(\rho) = C_W\frac{\rho}{(1-\rho)^2}\left(1-\frac{1}{\rho\frac{M}{L}}\right) \quad \text{and}$$

$$r(\rho,N) = C_W\left(A(M-L) + \frac{1}{\rho\frac{M}{L}-1} - (N-L-1)\right)$$

$$\text{with } A = \frac{C_{on}-C_{off}}{C_W}.$$

The term $\frac{(\frac{\lambda}{\mu})^M}{M!}\frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}}$ on the left hand side (lhs) of (27) constitutes the structure of the equation. Its magnitude varies in a huge range for larger M and N depending on the value of the parameters. Therefore, we also use its natural logarithm in the course of the analysis. By introducing the notation

$$p(\rho,N) = \frac{(\frac{\lambda}{\mu})^M}{M!}\frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}},$$

the equation (27) can be given in a short form as

$$p(\rho,N)u_0(\rho) = r(\rho,N). \quad (28)$$

D. Properties of function $p(\rho, N)$

The approximate global solution of the considered minimization task requires the knowledge of several properties of function $p(\rho, N)$.

1) Dependency on ρ: Applying the Stirling formula $n! \approx \sqrt{2\pi}n^{(n+1/2)}e^{-n}$ to both M and $N-1$ in the expression of $p(\rho, N)$ gives an approximation as

$$p(\rho,N) = \frac{(\frac{\lambda}{\mu})^M}{M!}\frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}} = \left(\frac{\lambda}{\mu M}\right)^{(M-N+1)}\frac{M^M}{M!}\frac{(N-1)!}{M^{N-1}} \approx \rho^{(M-N+1)}e^{(M-N+1)}\sqrt{\frac{N-1}{M}}\left(\frac{N-1}{M}\right)^{N-1}. \quad (29)$$

It can be seen from (29) that the dependency of $p(\rho, N)$ on ρ follows power law. This leads to rapid changes under the typical model parameter settings, e.g., increasing ρ by 2.5% at $M - N + 1 = 95$ leads to 10 times multiplication due to $1.025^{95} \approx 10$.

2) Dependency of $p(\rho, N)$ on N: Taking the natural logarithm of (29) we get

$$\ln\left[p(\rho, N)\right] = (M - N + 1)\left(\ln(\rho) + 1\right) + \left((N-1) + \frac{1}{2}\right)\ln\left(\frac{N-1}{M}\right).$$

By introducing the notation

$$\beta = \frac{N-1}{M} \tag{30}$$

this can be rewritten as

$$\ln\left[p(\rho, \beta] = M\left[(1-\beta)\left(\ln(\rho) + 1\right) + \left(\beta + \frac{1}{2\ M}\right)\ln(\beta)\right].$$

Taking its first derivative with respect to β gives

$$\frac{\partial \ln\left[p(\rho, \beta)\right]}{\partial \beta} = M\left[\ln\left(\frac{\beta}{\rho}\right) + \frac{1}{2\ M\ \beta}\right] \approx M \ln\left(\frac{\beta}{\rho}\right),$$

since in the typical model parameter ranges $\beta \geq \frac{L}{M} \geq 0.1$ and $M > 100$ and thus, except in the small sub-range $\beta \approx \rho$, the term $\frac{1}{2*M*\beta}$ can be neglected. The first derivative of $p(\rho, N)$ with respect to N comes by using $\frac{\partial\ p(\rho,N)}{\partial N} = \frac{\partial\left(e^{\ln[p(\rho,N)]}\right)}{\partial N} = p(\rho, N)\frac{\partial \ln[p(\rho,\beta)]}{\partial \beta}\frac{d\beta}{dN} = p(\rho, N)\frac{1}{M} * \frac{\partial \ln[p(\rho,\beta)]}{\partial \beta}$, which yields

$$\frac{\partial\ p(\rho, N)}{\partial N} \approx p(\rho, N) \ln\left(\frac{\beta}{\rho}\right). \tag{31}$$

The sign of $\ln(\frac{\beta}{\rho})$ divides the $\beta - \rho$ plane into two disjunct sub-areas regarding the characteristic of $p(\rho, N)$ with respect to N as

$$p(\rho, N) \text{ is } \left\{ \begin{array}{l} \text{monotone decreasing, if } \beta < \rho \\ \text{monotone increasing, if } \beta \geq \rho \end{array} \right\}. \tag{32}$$

Moreover the dependency of $p(\rho, N)$ on N is faster than exponential, since $|\ln(\frac{\beta}{\rho})|$ is increasing with decreasing N (increasing N) in the range $\beta < \rho$ ($\beta \geq \rho$).

3) The "low magnitude range": We investigate the case when $p(\rho, N) = e^{const}$ holds, where $const$ is a given real constant. With the notation of β this equation can be given by

$$M\left[(1-\beta)\left(\ln(\rho) + 1\right) + \left(\beta + \frac{1}{2\ M}\right)\ ln(\beta)\right] = const. \tag{33}$$

Observe that this equation implicitly defines a boundary function $\beta(\rho)$ (or equivalently $\rho(\beta)$), which separates the "low magnitude range" $p(\rho, N) \leq e^{const}$ from the complementer range, in which $p(\rho, N) > e^{const}$. In the range $p(\rho, N) \leq e^{const}$ the magnitude of $\ln(p(\rho, N))$ is less than $const$, which explains the name "low magnitude range". We say that a $\beta - \rho$ point is inside and outside of the "low magnitude range" if $p(\rho, \beta) \leq e^{const}$ holds and does not hold for that point, respectively. By rearranging (33) we get the expression of $\ln(\rho)$ along the boundary function as

$$ln(\rho) = \frac{const}{(1-\beta)\ M} - \frac{\beta}{1-\beta}\ln(\beta) - 1 - \frac{1}{(1-\beta)\ 2\ M}\ln(\beta).$$

Therefore, the sensitivity of $ln(\rho)$ with respect to the $const$, ζ is given by

$$\zeta = \frac{1}{(1-\beta)\ M}. \tag{34}$$

An upper limit for the factor $\ln(\frac{\beta}{\rho})$ determining the relation between $p(\rho, N)$ and its first derivative with respect to N (see (31)) along the boundary function can be obtained as

$$\begin{aligned} \ln\left(\frac{\beta}{\rho}\right) &= \ln(\beta) - \ln(\rho) = \ln(\beta) + \frac{\beta}{1-\beta}\ln(\beta) + 1 \\ &\quad - \left(\frac{const}{(1-\beta)\ M} - \frac{1}{(1-\beta)\ 2\ M}\ln(\beta)\right) \\ &\leq \frac{1}{1-\beta}\ln(\beta) + 1 \leq -\frac{1}{2}(1-\beta) < 0. \end{aligned} \tag{35}$$

where we used the non-negativity of the term in the brackets and the inequality $\ln(\beta) \leq -(1-\beta) - \frac{1}{2}(1-\beta)^2$. Hence,

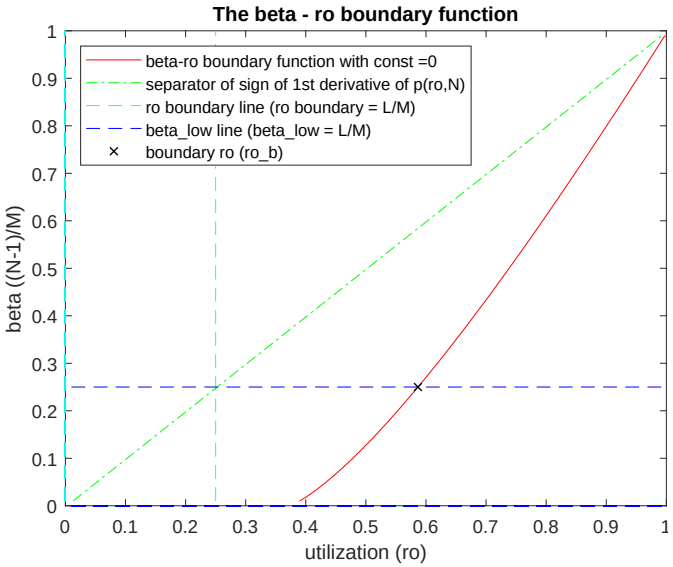

Figure 11. The $\beta - \rho$ boundary function, const = 0, M=200, L=50.

the boundary curve lies under the line separating the $\beta - \rho$ plane into parts with monotone decreasing and increasing $p(\rho, N)$ with respect to N. The permitted region of the $\beta - \rho$ plane is restricted by $\beta \geq \beta_{low} = \frac{L}{M}$ and $\rho > \beta_{low}$ due to the limitations $N - 1 \geq L \Leftrightarrow \frac{N-1}{M} \geq \frac{L}{M}$ and $\frac{\lambda}{\mu} > L \Leftrightarrow \rho > \frac{L}{M}$, respectively. The ρ at cross point of the horizontal line $\beta = \beta_{low}$ and the boundary curve is called boundary ρ and denoted by ρ_b. All these are shown on the illustrating example Figure 11. Since $\frac{\partial\ p(\rho,N)}{\partial N} < 0$ on the

boundary curve, the "low magnitude range" is located above the red marked boundary curve. Note that for $const \geq 0$, the whole range above the boundary curve belongs to the "low magnitude range", since $p(\rho, \beta)$ is monotone increasing with β above $\beta = \rho$ and $\ln(p(\rho, 1)) = 0 \leq const$.

4) Monotonicity of $\rho(\beta)$: The first derivative of the boundary function $\rho(\beta)$ with respect to β can be determined from (33) by applying the implicit function's derivative rule. This leads to

$$\frac{d\rho(\beta)}{d\beta} = -\frac{\frac{\partial \ln[p(\rho,\beta)]}{\partial \beta}}{\frac{\partial \ln[p(\rho,\beta)]}{\partial \rho}} = -\frac{M\left[\ln(\frac{\beta}{\rho}) + \frac{1}{2\ M\ \beta}\right]}{M(1-\beta)\frac{1}{\rho}}$$
$$= \frac{\ln(\frac{\rho}{\beta}) - \frac{1}{2M\beta}}{1-\beta}\rho > 0, \quad (36)$$

since $\ln(\frac{\rho}{\beta}) > \frac{1}{2}(1-\beta)$ due to (35) and $(1-\beta) > \frac{1}{M\beta}$ for $0.5 - \sqrt{0.5 - \frac{1}{M}} < 0.1 \leq \beta < 0.5 + \sqrt{0.5 - \frac{1}{M}} \approx 0.99$. The relation (36) implies that ρ is monotone increasing with respect to β on the boundary curve up to $\beta \approx 0.99$.

E. Constructing the approximate minimization

1) Solution regimes: For the sake of better understanding of the idea of the solution, first we consider a modified form of the equation (28) as

$$p(\rho, N) = r(\rho, N). \quad (37)$$

The idea of the approximate solution is based on the concept of "low magnitude range". Let N_s stand for the solution of $r(\rho, N) = 0$. Since $\frac{d\ r(\rho,N)}{dN} = -C_W$, the value of $r(\rho, N)$ changes from 0 up to C_W while decreasing N from N_s to $N_s - 1$. On the other hand, the value of $p(\rho, N)$ falls between 0 and C_W everywhere in the "low magnitude range" with $const = \ln(C_W)$. Hence, if $N_s - 1$ falls inside of the "low magnitude range" then it follows that $r(\rho, N)$ must cross $p(\rho, N)$ somewhere between N_s and $N_s - 1$ due to the continuity of $r(\rho, N)$. Therefore, N_s can be considered as approximate solution of (37).

We denote the value of N on the boundary curve by N_b, which depends on ρ. Due to (35), $|\ \ln(\frac{\beta}{\rho})\ | \lessapprox 0.5$ and hence, along the boundary curve, the first derivative of $p(\rho, N)$ is in the magnitude of $-C_W$ and the dependency of $p(\rho, N)$ on N is faster than exponential. Let us investigate a case, for which $N_s - 1$ falls inside the "low magnitude range", i.e., it locates above the boundary curve for some value of ρ. By decreasing N, at $N = N_b$, we have $r(\rho, N_b) > C_W = p(\rho, N_b)$. By further decreasing N, the first derivative of $p(\rho, N)$ becomes in absolute value greater than that one of $r(\rho, N)$, and hence, afterwards an other cross point of the functions $p(\rho, N)$ and $r(\rho, N)$ must exist, let us say at $N = N_1$. This is a maximum point of the cost function, since at this point (in N), with decreasing N, the sign of $p(\rho, N) - r(\rho, N)$ changes from negative to positive. After further decreasing N it reaches the point, $N = N_2$, where the value of the cost function falls under N_s. The situation is illustrated on Figure 12.

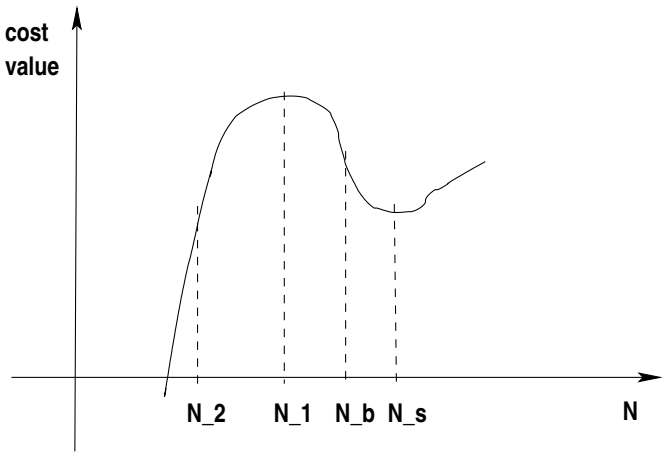

Figure 12. Example cost function.

The above discussed decrease from N_b to N_2 in any range of N, in which $p(\rho, N)$ is monotone decreasing with respect to N, causes an increase in the value of $p(\rho, N)$, which equivalently can be also considered as an increase in $const$ of (33). This change in $const$ corresponds to a shift of the boundary curve to right (see (34)). Hence, the points with N_2 in a dependency of ρ lie on another boundary curve with an increased $const$, where $\Delta const$ (= the increase in $const$) is determined by the change $p(\rho, N_b) \rightarrow p(\rho, N_2)$.

If $\rho > \rho_b$ then the point in β corresponding to N_1 (locating under the boundary curve) can fall over the β_{low} line. Until N_2 falls still below the β_{low} line, the value of the cost function at β_{low} is still higher than at β_s (corresponding to N_s), and therefore the global minimum of the cost function is still at N_s. However, if N_2 also falls above β_{low} line then the global minimum of the cost function is just above β_{low} (corresponding to $N = L + 1$). If $\rho > \rho_b$, it can also happen that β_s falls outside of the "low magnitude range" (= under the boundary curve and above the β_{low} line). In this case $|\frac{d\ p(\rho,N)}{dN}|$ is either still $\leq C_W$ or it is also possible, with β_s far under the boundary curve, that it is $> C_W$. In the later case there is no cross point at all, as well as the cost function is monotone increasing with respect to N and hence the global minimum is just above β_{low}. Note that in the range $N > N_s$ there can not be any cross point of the functions $p(\rho, N)$ and $r(\rho, N)$, since $p(\rho, N) > 0$ and $r(\rho, N) < 0$ in that range.

It follows from the above argumentation that the global minimum of the cost function is approximately at N_s in the range of $\rho < \rho_b$ and $\beta_{low} \leq \beta_s < 1$, since in this case the boundary curve locates under the horizontal β_{low} line and hence β_s always falls in the "low magnitude range". Above ρ_b there is a gap in ρ until a specific point, ρ_s, at which N_2 reaches the β_{low} line and hence in this gap β_s can fall also below the boundary curve. However, the derivative of $p(\rho, N)$ with respect to N is in absolute still less then C_W while N_s not much less than N_b (see (35)) and therefore we assume that the global minimum of the cost function still close to N_s. Finally above ρ_s the position of the global minimum of the

cost function changes to $N = L + 1$.

The position of ρ_s depends on $\Delta const$, which, as change in $const$, causes a shift of the boundary ρ, from ρ_b to ρ_s.

2) The magnitude of $\Delta const$: The change in $const$ is a sum of the changes due to $p(\rho, N_b) \to p(\rho, N_1)$ and $p(\rho, N_1) \to p(\rho, N_2)$ on ln level.

The first part of the sum counts for the increase $p(\rho, N_b) \to p(\rho, N_1)$ on ln level. During $N_b \to N_1$ the value of $p(\rho, N)$ increases from C_w up to $(N_s - N_1)C_W$. Hence, the first part of change in $const$ is given by

$$\ln\left(\frac{p(\rho, N_1)}{p(\rho, N_b)}\right) = \ln\left(\frac{(N_s - N_1)C_w}{C_w}\right) = \ln(N_s - N_1). \quad (38)$$

The second part of the sum stands for increase $p(\rho_0, N_1) \to p(\rho, N_2)$, on ln level. During $N_1 \to N_2$ the cost function F_{2app} decreases so much as its increases during $N_s \to N_1$. The major term of ΔF_{2app} during the transition $N_s \to N_1$ is $r(\rho_0, N)$, which is linear with N, and its value changes from 0 to $(N_s - N_1)C_W$. Therefore, the increase of the cost function F_{2app} during $N_s \to N_1$ is overestimated as

$$F_{2app}(N_1) - F_{2app}(N_s) \lessapprox \frac{(N_s - N_1)^2 C_W}{2}. \quad (39)$$

In the range $N \leq N_1$, the change of F_{2app} is dominated by $p(\rho, N)$ due to its over-exponential character against the linear character of $r(\rho, N)$. Therefore, we model the change of F_{2app} during the transition $N_2 \to N_1$ by $p(\rho, N) \gtrapprox e^{\chi N}$ with χ equal to the derivative factor at $N = N_1$, i.e., $\chi = \ln\left(\frac{\beta_{N_1}}{\rho}\right) < 0$, where β_{N_1} is the value of β at $N = N_1$. This yields

$$\begin{aligned} &F_{2app}(N_1) - F_{2app}(N_2) \gtrapprox \int_{N=N_2}^{N_1} e^{\chi N} \\ &= \frac{1}{\chi}\Big(p(\rho, N_1) - p(\rho, N_2)\Big) = \frac{1}{|\chi|}\Big(p(\rho, N_2) - p(\rho, N_1)\Big) \\ &= \ln\left(\frac{\rho}{\beta_{N_1}}\right)\Big(p(\rho, N_2) - p(\rho, N_1)\Big). \end{aligned} \quad (40)$$

Combining (39) and (40) we have

$$p(\rho, N_2) - p(\rho, N_1) \lessapprox \frac{1}{\ln(\frac{\rho}{\beta_{N_1}})} \frac{(N_s - N_1)^2 C_W}{2},$$

and therefore

$$\begin{aligned} &\ln\left(\frac{p(\rho, N_2)}{p(\rho, N_1)}\right) = \ln\left(\frac{p(\rho, N_2) - p(\rho, N_1)}{p(\rho, N_1)} + 1\right) \\ &\lessapprox \ln\left[\frac{(N_s - N_1)^2 C_W}{2\ln\left(\frac{\rho}{\beta_{N_1}}\right)(N_s - N_1)C_W} + 1\right] \\ &\approx \ln\left[\frac{(N_s - N_1)}{2\ln\left(\frac{\rho}{\beta_{N_1}}\right)}\right] \\ &= \ln(N_s - N_1) - \ln\left[2\ln\left(\frac{\rho}{\beta_{N_1}}\right)\right]. \end{aligned} \quad (41)$$

Putting (38) and (41) together we get an estimation for $\Delta const$ as

$$\Delta const \approx 2\ln(N_s - N_1) - \ln\left[2\ln\left(\frac{\rho}{\beta_{N_1}}\right)\right]. \quad (42)$$

Due to the over-exponential character of $p(\rho, N)$ the value of N_1 is close to N_2 in general. We will use the estimation of $\Delta const$ to determine the value of ρ_s at which the boundary curve (with $const$ including also $\Delta const$ as additive term) crosses the β_{low} line. Thus around that specific value of ρ_s we have $N_2 \approx L$. Therefore, N_1 can be approximated by L in (42), which results in the estimate for the magnitude of $\Delta const$ as

$$\Delta const \approx 2\ln(N_s - L) - \ln\left[2\ln\left(\frac{\rho}{\beta_{low}}\right)\right]. \quad (43)$$

Based on (35) (and that $const << M$ holds in the final relation for ρ_s (51)), an approximation can be given for $\ln(\frac{\rho_s}{\beta_L})$ as

$$\ln\left(\frac{\rho_s}{\beta_{low}}\right) \gtrapprox \frac{1}{2}(1 - \beta_{low}). \quad (44)$$

Applying (44) to (43) we get the final form of the estimation for the magnitude of $\Delta const$ as

$$\begin{aligned} \Delta const \approx 2\ln(\Delta N) - \ln(1 - \beta_{low}), \\ \text{where } \Delta N = N_s - L. \end{aligned} \quad (45)$$

3) Estimation of $\Delta const$: The value of N_s, the solution of $r(\rho, N_s) = 0$, can be given from (27) as

$$N_s = A(M - L) + \frac{1}{\rho\frac{M}{L} - 1} + L + 1. \quad (46)$$

Hence, ΔN can be expressed as

$$\Delta N = N_s - L = A(M - L) + \frac{1}{\rho\frac{M}{L} - 1} + 1,$$

which can be rearranged as

$$\Delta N = AM\left[(1 - \beta_{low}) + \frac{1}{A\ M}\left(\frac{1}{\rho\frac{M}{L} - 1} + 1\right)\right].$$

The term $\frac{1}{\rho\frac{M}{L}-1} + 1$ can be limited as $1 < \frac{1}{\frac{\rho}{\beta_{low}}-1} + 1 \leq 6$ assuming $\rho \geq \beta_{low}\xi$ with $\xi = 1.2$. This term counts in $\ln(\Delta N)$ only if $\frac{1}{A\ M}\left(\frac{1}{\rho\frac{M}{L}-1} + 1\right)$ dominates over $(1-\beta_{low})$, in which case it causes an uncertainity around $\ln(6) \approx 1.79$ on $\Delta const$ level. In order to minimize this uncertainity, we approximate this term by $\sqrt{6} \approx 2.45$ causing an uncertainity not higher than $\ln(\sqrt{6}) \approx \frac{1.79}{2} \approx 0.9$ on $\Delta const$ level. Based on this, $\ln(\Delta N)$ can be estimated as

$$\ln(\Delta N) = \ln(A) + \ln(M) + \ln\left(1 - \beta_{low} + \frac{2.45}{AM}\right). \quad (47)$$

Now using (47) in (45) we can estimate $\Delta const$ as

$$\Delta const \approx 2\left[\ln(A)+\ln(M)+\ln\left(1-\beta_{low}+\frac{2.45}{AM}\right)\right] - \ln\left(1-\beta_{low}\right). \quad (48)$$

4) Relation for ρ_s: So far we discussed the way of solution without considering the term $u_0(\rho)$ on the lhs of equation (28). Now we take into account also the term $u_0(\rho)$. In this case $\ln[p(\rho,1)]\ln[u_0(\rho)]$ can be greater than $\ln(C_W)$ for β close to 1, but N_s can be considered as approximate solution also in this case, since the derivatives of $r(\rho,N)$ and $p(\rho,N)u_0(\rho)$ have different signs in that range. By including the term $u_0(\rho)$, the relation for the boundary curve crossing the β_{low} line at ρ_s can be given by

$$M\left[(1-\beta_{low})\left(\ln(\rho_s)+1\right)+\left(\beta_{low}+\frac{1}{2\ M}\right)\ ln(\beta_{low})\right] + \ln[u_0(\rho_s)] = \ln(C_W)+\Delta const.$$

By substituting the expression of $u_0(\rho)$ from (27) and using $(1-\frac{1}{\rho_s\frac{M}{L}}) = (1-\frac{\beta_{low}}{\rho_s}) = \frac{\beta_{low}}{\rho_s}(\frac{\rho_s}{\beta_{low}}-1)$ we get

$$M\left[(1-\beta_{low})\left(\ln(\rho_s)+1\right)+\left(\beta_{low}+\frac{1}{2\ M}\right)\ ln(\beta_{low})\right] + \ln(C_W)+\ln(\rho_s)+\ln\left(\frac{1}{(1-\rho_s)^2}\right)+\ln(\beta_{low})-\ln(\rho_s) + \ln\left(\frac{\rho_s}{\beta_{low}}-1\right) = \ln(C_W)+\Delta const.$$

Rearranging yields

$$M\left[(1-\beta_{low})\left(\ln(\rho_s)+1\right)+\left(\beta_{low}+\frac{1}{2\ M}\right)\ ln(\beta_{low})\right] = -\ln(\beta_{low})-\ln\left(\frac{1}{(1-\rho_s)^2}\right)-\ln\left(\frac{\rho_s}{\beta_{low}}-1\right)+\Delta const. \quad (49)$$

By using (44) the term $\ln(\frac{\rho_s}{\beta_{low}}-1)$ can be rearranged as

$$\ln\left(\frac{\rho_s}{\beta_{low}}-1\right)\approx\ln\left(e^{\frac{1}{2}(1-\beta_{low})}-1\right)\approx\ln\left(\frac{1}{2}(1-\beta_{low})\right), \quad (50)$$

where we utilized that $\frac{1}{2}(1-\beta_{low}) \le 0.45$ for $\beta_{low} \ge 0.1$. The term $-\ln(\frac{1}{(1-\rho_s)^2})$ gives an uncertainty of ≈ -4.6 on rhs of (49) ($1 \le \frac{1}{(1-\rho_s)^2} \le 100$ for $\rho_s \le 0.9$ and thus $ln(100) = 4.6$) corresponding to difference of $\frac{4.6}{(1-0.5)100} \lessapprox 0.09$ on $\ln(\rho)$ level (according to the first term on the rhs of (34)) when assuming $M \ge 100$ and again $\beta_{low} \le 0.5$. The relation $\rho_s \le 0.9$ can be justified by the approximate solution of (49) for ρ_s by assuming that its rhs $\le 0.1M$. Utilizing that ρ_s is monotone increasing with respect to β_{low} (see (36)), and setting $\max(\beta_{low}) = 0.5$, according to (34) we get $\ln(\rho_s) \lessapprox \frac{0.1}{0.5} - \ln(0.5) - 1 \lessapprox -0.1 \Leftrightarrow \rho_s \lessapprox 0.9$. In order to minimze the above uncertainty of the term $-\ln(\frac{1}{(1-\rho_s)^2})$, we approximate it by -2.3. Using this approximation and (50) as well as (48), we get the final form of the relation for ρ_s as

$$M\left[(1-\beta_{low})\left(\ln(\rho_s)+1\right)+\left(\beta_{low}+\frac{1}{2\ M}\right)\ ln(\beta_{low})\right] = 2\left[\ln(A)+\ln(M)+\ln\left(1-\beta_{low}+\frac{2.45}{AM}\right)\right] - \ln\left(\frac{1}{2}\right)-2\ln\left(1-\beta_{low}\right)-\ln(\beta_{low})-2.3. \quad (51)$$

F. Approximate solution formula

Now putting all together we get the approximate solution formula.

Conditions

1) $100 \le M$,
2) $0.1 \le \beta_{low} \le 0.5$ with $\beta_{low} = \frac{L}{M}$,
3) $\rho \ge \beta_{low}\xi$ with $\xi = 1.2$,
4) $N - L >> 1$, practically $N > L + 10$,
5) $K - M >> 1$, practically $K > M + 10$.

Solution formula

If **Conditions** 1-5 hold, then

$$N_{opt} = \left\{\begin{array}{ll}\min_{(\lfloor A(M-L)+\frac{1}{\rho\frac{M}{L}-1}+L+1\rfloor, M)} & \text{if } \rho \le \rho_s,\\ L+1 & \text{if } \rho_s < \rho < 1,\end{array}\right\}$$

where

$$ln(\rho_s) = \frac{const}{(1-\beta_{low})*M} - \frac{\beta_{low}}{1-\beta_{low}}\ln(\beta_{low}) - 1 - \frac{1}{(1-\beta_{low})*2*M}\ln(\beta_{low}) \text{ and}$$

$$const = 2\left[\ln(A)+\ln(M)+\ln\left(1-\beta_{low}+\frac{2.45}{AM}\right)\right] - \ln\left(\frac{1}{2}\right)-2\ln\left(1-\beta_{low}\right)-\ln(\beta_{low})-2.3. \quad (52)$$

Note that the condition $N - L \gg 1$ refers to N_s and hence $N_{opt} = L + 1 < L + 10$ in the solution formula does not violate this condition.

Observe that the approximate optimal N does not depend on C_A, C_D and C_R. This is because they have no impact on N in the considered range of parameters. The cost parameters C_A, C_D influence N only via p_0 and hence they effect the optimal N in the range, in which p_0 depends on N. The cost parameter C_R has impact on the optimal N via η and hence it is effective only for small values of $K - M$ (see in (26)).

VII. Approximate minimization - in traffic range $\frac{\lambda}{L\mu} < 1$

For the N independent regions of p_0 the function to be minimized, (13), can be reduced to the minimization of (23) like we did it for the traffic range $\frac{\lambda}{L\mu} > 1$. In order to further simplify the optimization task, we establish an approximate equation for determining the local minimum for (23). Of course this approximation restricts the parameter range, for which it holds.

A. Necessary approximations

In this subsection, we establish the approximations, which are needed to establish the approximate equation for determining the local minimum. Again, we make use that under the assumption $N-L \gg 1$ the relation $(\frac{\lambda}{L\mu})^{N-L} \ll 1$ holds for the traffic range $\frac{\lambda}{L\mu} < 1$ and thus the term $(\frac{\lambda}{L\mu})^{N-L}$ can be neglected compared to 1.

1) Approximation for τ: Using $(\frac{\lambda}{L\mu})^{N-L} \ll 1$ the expression of τ in (10) can be approximated as

$$\tau = \frac{\frac{\lambda}{L\mu}}{(1-\frac{\lambda}{L\mu})^2} - \left(\frac{\lambda}{L\mu}\right)^{N-L} (N-L) \left(\frac{1}{1-\frac{\lambda}{L\mu}} + \frac{N-L-1}{2}\right). \tag{53}$$

2) Approximate $\Delta_N\alpha$, $\Delta_N \frac{p_{s1}}{p_L}$ and $\Delta_N\tau$: Taking Δ_N on (18) we get

$$\begin{aligned}\Delta_N\alpha &\approx \left(1-\frac{\lambda}{L\mu}\right)\left(\frac{\lambda}{L\mu}\right)^{N-L-1} - \left(1-\frac{\lambda}{L\mu}\right)\left(\frac{\lambda}{L\mu}\right)^{N-L-2} \\ &= \left(1-\frac{\lambda}{L\mu}\right)\left(\frac{\lambda}{L\mu}-1\right)\left(\frac{\lambda}{L\mu}\right)^{N-L-2} \\ &= -\left(1-\frac{\lambda}{L\mu}\right)^2\left(\frac{\lambda}{L\mu}\right)^{N-L-2}.\end{aligned} \tag{54}$$

It can be seen from (16) that the first two terms of $\frac{p_{s1}}{p_L}$ are independent of N. Thus, using the expression based on the second approximation form (17) for taking Δ_N would neglect the major term of $\Delta_N \frac{p_{s1}}{p_L}$, which would lead to incorrect approximation. Therefore, Δ_N must be taken on the expression based on the first approximation of p_{s1} in (16), which yields

$$\begin{aligned}&\Delta_N \frac{p_{s1}}{p_L} \\ &\approx (N-L-2)\left(\frac{\lambda}{L\mu}\right)^{N-L-1} - (N-L-1)\left(\frac{\lambda}{L\mu}\right)^{N-L} \\ &= \left(\frac{\lambda}{L\mu}\right)^{N-L-1}\left((N-L-2) - \frac{\lambda}{L\mu}(N-L-1)\right) \\ &= \left(\frac{\lambda}{L\mu}\right)^{N-L-1}\left((N-L-1)(1-\frac{\lambda}{L\mu}) - 1\right).\end{aligned} \tag{55}$$

Taking Δ_N on (53) leads to

$$\begin{aligned}\Delta_N\tau =& -\left(\frac{\lambda}{L\mu}\right)^{N-L}(N-L)\left[\frac{1}{1-\frac{\lambda}{L\mu}} + \frac{N-L-1}{2}\right] \\ &+\left(\frac{\lambda}{L\mu}\right)^{N-L-1}(N-L-1)\left[\frac{1}{1-\frac{\lambda}{L\mu}} + \frac{N-L-2}{2}\right] \\ =&\left(\frac{\lambda}{L\mu}\right)^{N-L-1}(N-L-1) \\ &\times\left[\left(\frac{1}{1-\frac{\lambda}{L\mu}} + \frac{N-L-1}{2}\right)\left(1-\frac{\lambda}{L\mu}\right) - \frac{1}{2}\right] \\ &-\left(\frac{\lambda}{L\mu}\right)^{N-L}\left(\frac{1}{1-\frac{\lambda}{L\mu}} + \frac{N-L-1}{2}\right) \\ =&\left(\frac{\lambda}{L\mu}\right)^{N-L-1}\left[(N-L-1)\left(1-\frac{\lambda}{L\mu}\right) - \frac{\lambda}{L\mu}\right] \\ &\times\left[\frac{N-L-1}{2} + \frac{1}{1-\frac{\lambda}{L\mu}}\right] - \left(\frac{\lambda}{L\mu}\right)^{N-L-1}\frac{N-L-1}{2}.\end{aligned} \tag{56}$$

3) Limits on $s_{L,N}$: Applying a lower limit on every term of the sum in the expression of $s_{L,N}$ we get a lower limit on it as

$$\begin{aligned}s_{L,N} =& \sum_{i=L}^{N-1}\frac{i!}{(\frac{\lambda}{\mu})^i} \\ =&\frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}}\left(1 + \frac{\frac{\lambda}{\mu}}{N-1} + \ldots + \frac{(\frac{\lambda}{\mu})^{N-1-L}}{(N-1)\ldots(L+1)}\right) \\ \geq&\frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}}\sum_{i=0}^{N-1-L}\left(\frac{\lambda}{(N-1)\mu}\right)^i \\ =&\frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}}\frac{1-\left(\frac{\lambda}{(N-1)\mu}\right)^{N-L}}{1-\frac{\lambda}{(N-1)\mu}} \approx \frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}}\frac{1}{1-\frac{\lambda}{(N-1)\mu}},\end{aligned} \tag{57}$$

where in the last step we utilized that $\left(\frac{\lambda}{(N-1)\mu}\right)^{N-L} \ll 1$ for $N-L \gg 1$ due to $\frac{\lambda}{(N-1)\mu} \leq \frac{\lambda}{L\mu} < 1$.

Similarly, applying an upper limit on every term of the sum in the expression of $s_{L,N}$, we get a upper limit on it as

$$\begin{aligned}s_{L,N} =& \sum_{i=L}^{N-1}\frac{i!}{(\frac{\lambda}{\mu})^i} \\ =&\frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}}\left(1 + \frac{\frac{\lambda}{\mu}}{N-1} + \ldots + \frac{(\frac{\lambda}{\mu})^{N-1-L}}{(N-1)\ldots(L+1)}\right) \\ <&\frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}}\sum_{i=0}^{N-1-L}\left(\frac{\lambda}{L\mu}\right)^i = \frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}}\frac{1-\left(\frac{\lambda}{L\mu}\right)^{N-L}}{1-\frac{\lambda}{L\mu}}.\end{aligned} \tag{58}$$

B. Establishing an approximation for $\Delta_N F_2$

1) Basic form of $\Delta_N F_2$: Taking ΔN on the function (23) to be minimized and using $\Delta_n a(n)b(n) = a(n)\Delta_n b(n) + \Delta_n a(n)b(n-1)$ we get

$$\Delta_N F_2 = \lambda(C_A + C_D)(M-L)\Delta_N\alpha + \eta\Big(\Delta_N s_{L,N}\alpha(N-1,L) + s_{L,N}\Delta_N\alpha(N,L)\Big) + C_W\Delta_N\tau - (C_{on}-C_{off})(M-L)\Delta_N\frac{p_{s1}}{p_L}.$$

2) Applying the approximations: Using $\Delta_N s_{L,N} = \frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}}$ and applying the approximations for α, $\Delta_N\alpha$, $\Delta_N\frac{p_{s1}}{p_L}$ and $\Delta_N\tau$, i.e., (18), (54), (55), (56) yields

$$\begin{aligned}\Delta_N F_2 \approx{}& \eta\Bigg[\frac{(N-1)!}{\left(\frac{\lambda}{\mu}\right)^{N-1}}\left(1-\frac{\lambda}{L\mu}\right)\left(\frac{\lambda}{L\mu}\right)^{N-L-2} - s_{L,N}\left(1-\frac{\lambda}{L\mu}\right)^2\left(\frac{\lambda}{L\mu}\right)^{N-L-2}\Bigg] \\ &- (C_{on}-C_{off})(M-L) \\ &\times\left(\frac{\lambda}{L\mu}\right)^{N-L-1}\left[(N-L-1)\left(1-\frac{\lambda}{L\mu}\right)-1\right] \\ &+ C_W\left(\frac{\lambda}{L\mu}\right)^{N-L-1}\left[(N-L-1)\left(1-\frac{\lambda}{L\mu}\right)-\frac{\lambda}{L\mu}\right] \\ &\times\left(\frac{N-L-1}{2}+\frac{1}{1-\frac{\lambda}{L\mu}}\right) - C_W\left(\frac{\lambda}{L\mu}\right)^{N-L-1}\frac{N-L-1}{2} \\ &- \lambda(C_A+C_D)(M-L)\left(1-\frac{\lambda}{L\mu}\right)^2\left(\frac{\lambda}{L\mu}\right)^{N-L-2}.\end{aligned} \tag{59}$$

The exact and approximated values of the function $\Delta_N F_2$ by F_{2app} are shown on Figure 13 in dependency of threshold N for the parameter setting $M = 200, L = 100, K = 250, \mu = 1, C_W = 10, C_{off} = 0.01, C_{on} = 0.02, C_a = 0.03, C_d = 0.02, C_R = 200$, and $\rho = 0.6$.

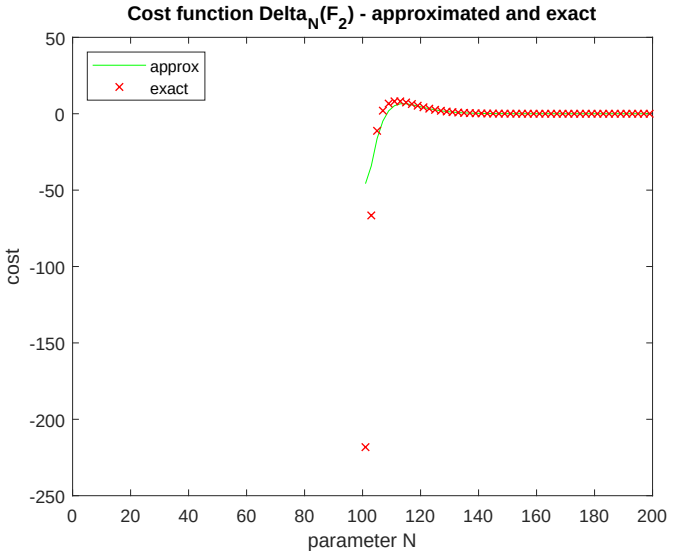

Figure 13. Exact and approximate values of the function $\Delta_N F_2$ in dependency of threshold N.

The figure shows a very good match. The mismatch on the left side of the curve is caused by violating the condition $N - L >> 1$, which was utilized by the approximations, as N becomes close to L.

Due to $K - M \gg 1$ the approximation for η in (26) holds also in traffic range $\frac{\lambda}{L\mu} < 1$. Using it in (59) results in

$$\begin{aligned}\Delta_N F_2 \approx{}& C_W\frac{\rho}{(1-\rho)^2}\frac{(\frac{\lambda}{\mu})^M}{M!}\left[\frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}} - s_{L,N}\left(1-\frac{\lambda}{L\mu}\right)\right] \\ &\times\left(1-\frac{\lambda}{L\mu}\right)\left(\frac{\lambda}{L\mu}\right)^{N-L-2} - C_W\left(\frac{\lambda}{L\mu}\right)^{N-L-1} \\ &\times\Bigg\{\frac{C_{on}-C_{off}}{C_W}(M-L)\left[(N-L-1)\left(1-\frac{\lambda}{L\mu}\right)-1\right] \\ &-\left[(N-L-1)\left(1-\frac{\lambda}{L\mu}\right)-\frac{\lambda}{L\mu}\right]\left(\frac{N-L-1}{2}+\frac{1}{1-\frac{\lambda}{L\mu}}\right) \\ &+\frac{N-L-1}{2}+\frac{\lambda(C_A+C_D)}{C_W}(M-L)\frac{L\mu}{\lambda}\left(1-\frac{\lambda}{L\mu}\right)^2\Bigg\}.\end{aligned} \tag{60}$$

3) Limits on the term $\left[\frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}} - s_{L,N}\left(1-\frac{\lambda}{L\mu}\right)\right]$*:* The term $\left[\frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}} - s_{L,N}\left(1-\frac{\lambda}{L\mu}\right)\right]$ on the lhs of (59) is positive. This can be shown by the help of the upper limit on $s_{L,N}$, (58) as

$$\begin{aligned}&\frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}} - s_{L,N}\left(1-\frac{\lambda}{L\mu}\right) \\ &\geq \frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}} - \frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}}\frac{1-(\frac{\lambda}{L\mu})^{N-L}}{1-\frac{\lambda}{L\mu}}\left(1-\frac{\lambda}{L\mu}\right) \\ &= \frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}}\left\{1-\left[1-\left(\frac{\lambda}{L\mu}\right)^{N-L}\right]\right\} \\ &= \frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}}\left(\frac{\lambda}{L\mu}\right)^{N-L} > 0.\end{aligned}$$

On the other hand, by using the lower limit on $s_{L,N}$, (57) we get an upper limit on $\left[\frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}} - s_{L,N}\left(1-\frac{\lambda}{L\mu}\right)\right]$ as

$$\begin{aligned}&\frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}} - s_{L,N}\left(1-\frac{\lambda}{L\mu}\right) \\ &\lessapprox \frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}} - \frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}}\frac{1}{1-\frac{\lambda}{(N-1)\mu}}\left(1-\frac{\lambda}{L\mu}\right) \\ &= \frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}}\left[1-\left(1-\frac{\lambda}{L\mu}\right)\frac{1}{1-\frac{\lambda}{(N-1)\mu}}\right] \\ &\leq \frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}}\left[1-\left(1-\frac{\lambda}{L\mu}\right)\right] = \frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}}\frac{\lambda}{L\mu},\end{aligned} \tag{61}$$

where in the last but one step we utilized $\frac{1}{1-\frac{\lambda}{(N-1)\mu}} > 1$ due to $\frac{\lambda}{(N-1)\mu} \leq \frac{\lambda}{L\mu} < 1$.

4) Final form of the approximation for $\Delta_N F_2$: We introduce the notation

$$B = \frac{\lambda(C_A + C_D)}{C_W}.$$

Replacing the term $\left[\frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}} - s_{L,N}\left(1-\frac{\lambda}{L\mu}\right)\right]$ by its upper limit (61) in (60), using the notations A (introduced in (27)) and B and performing rearrangements leads to an upper limit on $\Delta_N F_2$ as

$$\begin{aligned}\Delta_N F_2 &\lessapprox \left(\frac{\lambda}{L\mu}\right)^{N-L-1} \frac{(\frac{\lambda}{\mu})^M}{M!}\frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}} \\ &\times C_W \frac{\rho}{(1-\rho)^2}\left(1-\frac{\lambda}{L\mu}\right) - \left(\frac{\lambda}{L\mu}\right)^{N-L-1} C_W \\ &\times \left\{A(M-L)\left[(N-L-1)\left(1-\frac{\lambda}{L\mu}\right)-1\right]\right. \\ &-\left[(N-L-1)\left(1-\frac{\lambda}{L\mu}\right)-\frac{\lambda}{L\mu}\right]\left(\frac{N-L-1}{2}+\frac{1}{1-\frac{\lambda}{L\mu}}\right) \\ &\left.+\frac{N-L-1}{2}+B(M-L)\frac{L\mu}{\lambda}\left(1-\frac{\lambda}{L\mu}\right)^2\right\}.\end{aligned}$$

We define x as

$$x = N - L - 1.$$

Using this notation and performing further rearrangements we get the final form of the approximation for $\Delta_N F_2$ as

$$\begin{aligned}\Delta_N F_2 &< \left(\frac{\lambda}{L\mu}\right)^{N-L-1} C_W \\ &\times\left(\frac{(\frac{\lambda}{\mu})^M}{M!}\frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}}\, u_1(\rho) + q(\rho,N)\right),\end{aligned}$$

where

$$u_1(\rho) = \frac{\rho}{(1-\rho)^2}\left(1-\frac{\rho}{\frac{L}{M}}\right) \quad \text{and}$$

$$q(\rho,N) = ax^2 - bx + c, \text{ and } x = N-L-1, \tag{62}$$

$$a = \frac{1}{2}\left(1-\frac{\lambda}{L\mu}\right),\ b = \left[\left(A(M-L)-\frac{1}{2}\right)\left(1-\frac{\lambda}{L\mu}\right)\right],$$

$$c = \left\{\left[A - B\frac{L\mu}{\lambda}\left(1-\frac{\lambda}{L\mu}\right)^2\right](M-L) - \frac{\frac{\lambda}{L\mu}}{1-\frac{\lambda}{L\mu}}\right\}.$$

C. Constructing the approximate minimization

As we will see, the construction of the approximate minimization is based on the existence of range of N, in which the cost function is almost constant, and the properties of the quadratic equation $q(\rho, N) = 0$.

1) Existence of almost constant cost ranges: The term $(\frac{\lambda}{L\mu})^{N-L-1}$ decreases rapidly with increasing N in the range of $\frac{\lambda}{L\mu} < 1$. Hence, after enough large values of N, at the latest when $N-L-1$ reaches several tens, it supresses the rhs of (62). Thus in that range $\Delta F_2 \approx 0$ and therefore the cost function is almost constant. The range of N, in which the cost function is almost constant will be called as "almost constant cost range".

In general the value of the factorial terms in (62) can vary in a huge magnitude range. However $p(\rho, N) = \frac{(\frac{\lambda}{\mu})^M}{M!}\frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}}$ is monotone increasing in the range of $\beta > \rho$ (see (32), which is the case due to $\beta \geq \beta_{low} \Leftrightarrow N-1 \geq L$ and $1 > \frac{\lambda}{L\mu} \Leftrightarrow \beta_{low} > \rho$, and thus $p(\rho, N) \leq p(\rho, M) = 1$. Also the term $u_1(\rho)$ is upper limited by $u_1(\rho) < \frac{\rho}{(1-\rho)^2} < \frac{0.5}{(1-0.5)^2} = 2$ for $\rho < \beta_{low} \leq 0.5$. It follows that the term $\frac{(\frac{\lambda}{\mu})^M}{M!}\frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}}u_1(\rho)$ can not grow any large in the allowed traffic range and hence it will be also suppressed by term $(\frac{\lambda}{L\mu})^{N-L-1}$ when $N-L-1$ reaches several tens depending on the value of ρ.

Figures 14 and 15 illustrate the existence of "almost constant cost range" for two typical form cost function curves.

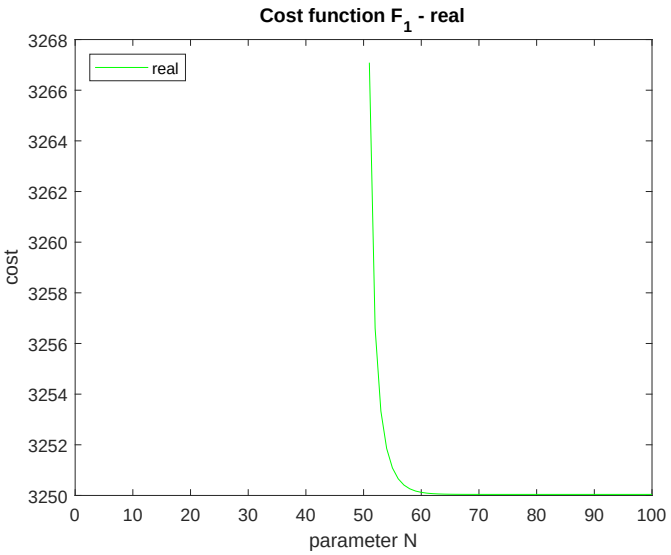

Figure 14. Typical cost function curve - Type 1.

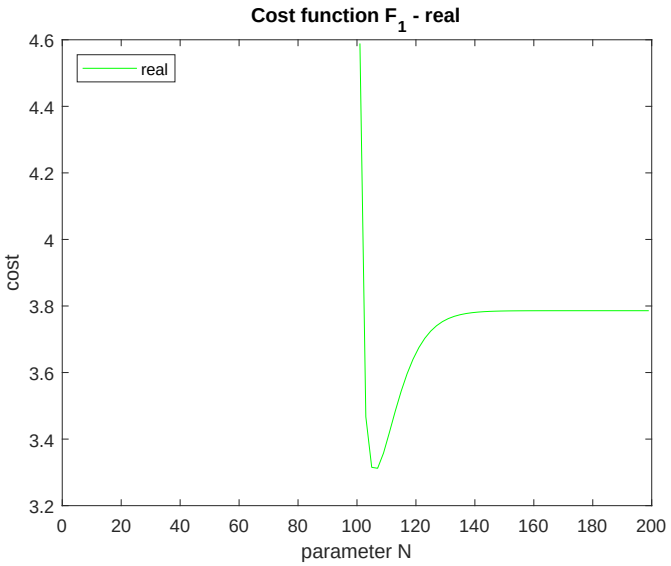

Figure 15. Typical cost function curve - Type 2.

Figure 14 and 15 were created by using the parameter settings $M = 100, L = 50, K = 150, \mu = 1, C_W = 50, C_{off} = 15, C_{on} = 50, C_a = 30, C_d = 20, C_R = 200, \rho = 0.3$ and $M = 200, L = 100, K = 250, \mu = 1, C_W = 10, C_{off} = 0.01, C_{on} = 0.02, C_a = 0.03, C_d = 0.02, C_R = 200, \rho = 0.4$, respectively.

In the range of $N-L-1 \gg 1$, e.g., $N-L-1 > 10$, the term $\frac{(\frac{\lambda}{\mu})^M}{M!}\frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}}u_1(\rho)$ falls in a very low magnitude range and thus it can be neglected comparing to $q(\rho, N)$. Hence, in that range the maximum value of ΔF_2 is determined by $q(\rho, N)$. Due to the parabolic character of $q(\rho, N)$ the maximum of its absolute value in the range of $x = 0, \dots M - L - 1$, denoted by q_{ul}, is upper limited by the maximum of its absolute value at $x = 0$, at $x = M - L - 1$ and at its local extremum, which is at $x = \frac{b}{2a}$ due to $\frac{\partial q(\rho,N)}{\partial x} = 0 \Leftrightarrow 2ax - b = 0$. Therefore, q_{ul} can be given by

$$q_{ul} = \max\left[\mid c \mid, \mid a(M-L-1)^2 - b(M-L-1) + c \mid, \mid c - \frac{b^2}{4a} \mid\right]. \quad (63)$$

Let N_{uccl} stand for an upper limit on the lower boundary of the "almost constant cost range". This upper limit is determined by the equation

$$q_{ul}\left(\frac{\lambda}{L\mu}\right)^{N_{uccl}-L-1} = \epsilon,$$

where ϵ is the required precision, e.g., $\epsilon = 0.01$. Solving the above equation results in N_{uccl} as

$$N_{uccl} = \lfloor \frac{\ln(\frac{\epsilon}{q_{ul}})}{\ln(\frac{\lambda}{L\mu})} + L + 1 \rfloor \text{ with } \epsilon < q_{ul}.$$

2) The concept of the approximate solution: The existence of almost constant cost ranges implies that either there is a cost minimum at N_{opt} below the lower boundary of the "almost constant cost range" or the cost at any N in the whole "almost constant cost range" can be considered as minimal.

The situation is determined by the properties of $\Delta_N F_2$ in the range of N below the lower boundary of the "almost constant cost range". In the majority of that range the term $\frac{(\frac{\lambda}{\mu})^M}{M!}\frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}}u_1(\rho)$ can be neglected comparing to $q(\rho, N)$ and therefore the properties (sign and position of roots) of the parabola $q(\rho, N)$ determines the situation as follows:

Let x_l and x_h stand for the real roots of the quadratic equation $q(\rho, N) = 0$ with $x_l < x_h$, if they exists. Similarly, let N_l and N_h stand for the corresponding values in N, i.e., $N_l = x_l + L + 1$ and $N_h = x_h + L + 1$.The coefficient of x^2 in the above quadratic equation is positive, since a is positive due to $\frac{\lambda}{L\mu} < 1$. Therefore, in the range $x < x_l$ and $x > x_h$ the value of $q(\rho, N)$ is positive. Similarly, in the range $x_l < x < x_h$ the value of $q(\rho, N)$ is negative.

It follows from the above argumentation that the approximate solution can be constructed as follows.

1) The quadratic equation $q(\rho, N)$ has two real roots and $x_h \geq 0 \Leftrightarrow N_h \geq L + 1$.
   - If $N_h < N_{uccl}$ then the cost function F_2 has local minimum at N_h,
   - otherwise any N in the "almost constant cost range" is a local minimum.

   Moreover if $x_l < 0$ or $x_l \geq 0$, but it lies close enough to 0 then the above local minimum places are also global. Otherwise it is also possible that the value of F_2 at $x = 0$ is less then at the above local minimum places.
2) In any other case (two real roots and $x_h < 0$, one real root or no real root) $q(\rho, N) > 0$ and thus F_2 is monotone increasing in the range of $x > 0 \Leftrightarrow N \geq L + 1$, and hence it has global minimum at $N = L + 1$,

3) Condition for discriminant of the quadratic equation to be positive: The coefficients of the quadratic equation $q(\rho, N) = (ax^2 - bx + c) = 0$ can be rearranged as

$$\begin{aligned}
a &= \frac{1}{2}\left(1 - \frac{\lambda}{L\mu}\right), \quad (64)\\
b &= \left[\left(A(M-L) - \frac{1}{2}\right)\left(1 - \frac{\lambda}{L\mu}\right)\right] = \left[2A(M-L) - 1\right]a,\\
c &= \left\{\left[A - B\frac{L\mu}{\lambda}\left(1 - \frac{\lambda}{L\mu}\right)^2\right](M-L) - \frac{\frac{\lambda}{L\mu}}{1 - \frac{\lambda}{L\mu}}\right\}\\
&= \left(A(M-L) - \frac{\frac{\lambda}{L\mu}}{1 - \frac{\lambda}{L\mu}}\right) - 4B(M-L)\frac{\beta_{low}}{\rho}a^2\\
&= \left(A(M-L) + 1 - \frac{1}{2a}\right) - 4B(M-L)\frac{\beta_{low}}{\rho}a^2.
\end{aligned}$$

We rearrange the discriminant of the equation $D = b^2 - 4ac$ as

$$\begin{aligned}
D &= \left[\left(A(M-L) - \frac{1}{2}\right)\left(1 - \frac{\lambda}{L\mu}\right)\right]^2\\
&+ 16B(M-L)\frac{\beta_{low}}{\rho}a^3\\
&- 2\left(1 - \frac{\lambda}{L\mu}\right)\left[\left((A(M-L) - \frac{1}{2}\right) + \left(\frac{1}{2} - \frac{\frac{\lambda}{L\mu}}{1 - \frac{\lambda}{L\mu}}\right)\right]\\
&= \left[\left(A(M-L) - \frac{1}{2}\right)\left(1 - \frac{\lambda}{L\mu}\right) - 1\right]^2 - 1\\
&- 2\left(1 - \frac{\lambda}{L\mu}\right)\left(\frac{1}{2} - \frac{\frac{\lambda}{L\mu}}{1 - \frac{\lambda}{L\mu}}\right) + 16B(M-L)\frac{\beta_{low}}{\rho}a^3\\
&= \left[\left(A(M-L) - \frac{1}{2}\right)\left(1 - \frac{\lambda}{L\mu}\right) - 1\right]^2\\
&- \left[3\left(1 - \frac{\lambda}{L\mu}\right) - 1\right] + 16B(M-L)\frac{\beta_{low}}{\rho}a^3\\
&= (b-1)^2 - \left(-16B(M-L)\frac{\beta_{low}}{\rho}a^3 + 6a - 1\right).
\end{aligned}$$

We introduce the notation

$$E = \left(-16B(M-L)\frac{\beta_{low}}{\rho}a^3 + 6a - 1\right). \quad (65)$$

With this notation we have

$$D = (b-1)^2 - E. \tag{66}$$

If $E < 0$ then $D > 0$.
Otherwise, i.e., for $E \geq 0$, ensuring $D > 0$ gives

$$(b-1)^2 > E.$$

If $(b-1) \geq 0$ then it leads to

$$\begin{aligned} b-1 &> \sqrt{E} \;\Leftrightarrow \\ \Big(2A(M-L)-1\Big)a &> 1+\sqrt{E} \;\Leftrightarrow \\ A(M-L) &> \frac{1}{2} + \frac{1+\sqrt{E}}{2a}. \end{aligned}$$

In the other case of $(b-1) < 0$ we get

$$\begin{aligned} 1-b &> \sqrt{E} \;\Leftrightarrow \\ \Big(2A(M-L)-1\Big)a &< 1-\sqrt{E} \;\Leftrightarrow \\ A(M-L) &< \frac{1}{2} + \frac{1-\sqrt{E}}{2a}. \end{aligned}$$

Summarizing the necessary and sufficient condition for $D > 0$ can be given by

$$\left\{ \begin{array}{ll} - & \text{if } E < 0, \\ \Big[A(M-L) > \frac{1}{2} + \frac{1+\sqrt{E}}{2a} \text{ or} & \\ \quad A(M-L) < \frac{1}{2} + \frac{1-\sqrt{E}}{2a} \quad \Big] & \text{if } E \geq 0 \end{array} \right\}, \tag{67}$$

where E is defined in (65).

4) Conditions for nonnegative upper root: The roots of the quadratic equation $q(\rho, N) = (ax^2 - bx + c) = 0$ are given by

$$x_{1,2} = \frac{b \pm \sqrt{b^2 - 4ac}}{2a}.$$

Assuming $D > 0$ the quadratic equation has two roots. The upper root x_h can be nonnegative in two cases.

Case 1. The condition $b < 0$ holds.
In this case a second condition $D \geq b^2$ is also required to have nonnegative upper root. It also ensures $D \geq 0$.
Due to $(1 - \frac{\lambda}{L\mu}) > 0$ the condition $b < 0$ yields

$$A(M-L) < \frac{1}{2}.$$

The second condition is equivalent to $4ac \leq 0$, and due to $a > 0$ leads to $c \leq 0$, which results in the condition

$$\begin{aligned} \left(A(M-L) + 1 - \frac{1}{2a}\right) &\leq 4B(M-L)\frac{\beta_{low}}{\rho}a^2 \;\Leftrightarrow \\ B(M-L) &\geq \frac{\rho}{\beta_{low}} \frac{1}{4a^2} \left(A(M-L) + 1 - \frac{1}{2a}\right). \end{aligned}$$

Hence, the condition for ensuring both $b < 0$ and $D \geq b^2$ can be summarized as

$$\left\{ \begin{array}{l} A(M-L) < \frac{1}{2} \text{ and} \\ B(M-L) \geq \frac{\rho}{\beta_{low}} \frac{1}{4a^2} \left(A(M-L) + 1 - \frac{1}{2a}\right) \end{array} \right\}. \tag{68}$$

Note that in this case the lower root is always negative and hence the local minimum at N_h or in the "almost constant cost range" is also a global one.

Case 2. The condition $b \geq 0$ holds.
In this case $x_h = b + \sqrt{D} > 0$ holds always due to $D > 0$. Thus no additional condition is necessary. The condition $b \geq 0$ leads to

$$A(M-L) \geq \frac{1}{2}.$$

The conditions both for $b \geq 0$ and $D > 0$ for this case can be summarized as

$$\left\{ \begin{array}{ll} A(M-L) \geq \frac{1}{2} & \text{if } E < 0, \\ \Big[A(M-L) > \frac{1}{2} + \frac{1+\sqrt{E}}{2a} \quad \text{or} & \\ \Big(A(M-L) \geq \frac{1}{2} \text{ and } A(M-L) < \frac{1}{2} + \frac{1-\sqrt{E}}{2a} \Big) \Big] & \\ & \text{if } E \geq 0 \end{array} \right\}, \tag{69}$$

where E is defined in (65).

5) The effect of the lower root on the global minimum: Herein we investigate the magnitude of the lower root for the case $b \geq 0$ and $D > 0$. First, we provide a lower limit for D by applying $D > 0$ in (66). If $E < 0$ then $D = (b-1)^2 - E > (b-1)^2$. Otherwise $E \geq 0$ and we can establish a lower limit

$$\begin{aligned} D &= (b-1)^2 - E = (b-1)^2 - (\sqrt{E})^2 \\ &= (b-1-\sqrt{E})(b-1+\sqrt{E}) \\ &\geq \Big(\min(|\,(b-1-\sqrt{E})\,|, |\,(b-1+\sqrt{E})\,|)\Big)^2. \end{aligned}$$

Hence, the lower limit for $D > 0$ can be summarized as

$$D \geq \left\{ \begin{array}{ll} (b-1)^2, & \text{if } E < 0, \\ (b-1-\sqrt{E})^2, & \text{if } b \geq 1 \text{ and } E \geq 0, \\ (b-1+\sqrt{E})^2, & \text{if } 0 < b < 1 \text{ and } E \geq 0 \end{array} \right\}.$$

Based on it, we can give an upper limit for x_l for the case $b \geq 0$ and $D > 0$ as

$$\begin{aligned} x_l &= \frac{b - \sqrt{D}}{2a} \\ &\leq \left\{ \begin{array}{ll} \frac{b-|b-1|}{2a} = \frac{1}{2a}, & \text{if } E < 0 \text{ and } b \geq 1, \\ \frac{b-|b-1|}{2a} = \frac{2b-1}{2a}, & \text{if } E < 0 \text{ and } 0 < b < 1, \\ \frac{b-(b-1-\sqrt{E})}{2a} = \frac{1+\sqrt{E}}{2a}, & \text{if } E \geq 0 \text{ and } b \geq 1, \\ \frac{b-(1-b-\sqrt{E})}{2a} = \frac{2b-1+\sqrt{E}}{2a}, & \text{if } E \geq 0 \text{ and } 0 < b < 1 \end{array} \right\}. \end{aligned}$$

Assuming $\frac{\lambda}{L\mu} \leq \psi$ with $\psi = 0.8$, we have $\frac{1}{2a} < \frac{1}{(1-0.8)} = 5$. Moreover $E \leq 6a - 1 \leq 2$ and thus $\sqrt{E} < \sqrt{2} < 1.5$. Thus

for the first and third cases we have $x_l \leq \frac{1}{2a} \leq \frac{1+\sqrt{E}}{2a} < 5 * 2.5 = 12.5$. In the second and fourth cases $x_l \leq \frac{2b-1+\sqrt{E}}{2a} \leq \frac{1+\sqrt{E}}{2a} < 12.5$, since $b < 1$. Therefore, x_l is low in all the four cases comparing to the range of N, and thus the chance of having lower cost at $N = L+1$ than the local minimum at N_h or in the "almost constant cost range" can be neglected. It follows that the local minimum can be considered as a global one.

D. Approximate solution formula

Now taking into account the necessary limitations and arguments we get the approximate solution formula.

Conditions

1) $100 \leq M$,
2) $0.1 \leq \beta_{low} \leq 0.5$ with $\beta_{low} = \frac{L}{M}$,
3) $\rho \leq \beta_{low}\psi$ with $\psi = 0.8$,
4) $N - L >> 1$, practically $N > L + 10$,
5) $K - M >> 1$, practically $K > M + 10$,

Solution formula

If **Conditions** 1-5 hold, then

- If

either

$$\left[A(M-L) < \frac{1}{2} \text{ and } B(M-L) \geq \frac{\rho}{\beta_{low}} \frac{1}{4a^2}\left(A(M-L) + 1 - \frac{1}{2a}\right)\right]$$

or

$$\left\{\begin{array}{ll} A(M-L) \geq \frac{1}{2} & \text{if } E < 0, \\ \left[A(M-L) > \frac{1}{2} + \frac{1+\sqrt{E}}{2a} \text{ or } \left(A(M-L) \geq \frac{1}{2} \text{ and } A(M-L) < \frac{1}{2} + \frac{1-\sqrt{E}}{2a}\right)\right] & \text{if } E \geq 0 \end{array}\right\}$$

holds then

$$N_{opt} = \left\{\begin{array}{ll} N_h & \text{if } N_h < N_{uccl} \\ \text{any } N \in [N_{uccl}, M] & \text{if } N_{uccl} \leq N_h \end{array}\right\},$$

where

$$N_{uccl} = \lfloor \frac{\ln(\frac{\epsilon}{q_{ul}})}{\ln(\frac{\lambda}{L\mu})} + L + 1 \rfloor \text{ with } \epsilon < q_{ul},$$

$$N_h = \lfloor \frac{b + \sqrt{b^2 - 4ac}}{2a} + L + 1 \rfloor, \quad (70)$$

and q_{ul}, a, b, c and E is given in (63), (64) and (65), respectively.

- Otherwise

$$N_{opt} = L + 1.$$

The approximate optimal N does not depend on C_R also in this traffic range. The cost parameter C_R has impact on the optimal N via η and hence it is effective only for small values of $K - M$.

VIII. Numerical comparisons

In this section, we illustrate the approximations and validate the approximate solution formula by numeric optimization.

A. High traffic range - $\frac{\lambda}{L\mu} > 1$

The setting $C_{on} = 50$, $C_{off} = 15$ $C_a = 30$, $C_d = 20$ and $C_R = 20$ was used for all experiments. The parameters C_a, C_d and C_R have no impact on the approximate solution formula in the considered range of parameters.

B. Illustration of the approximate solution formula

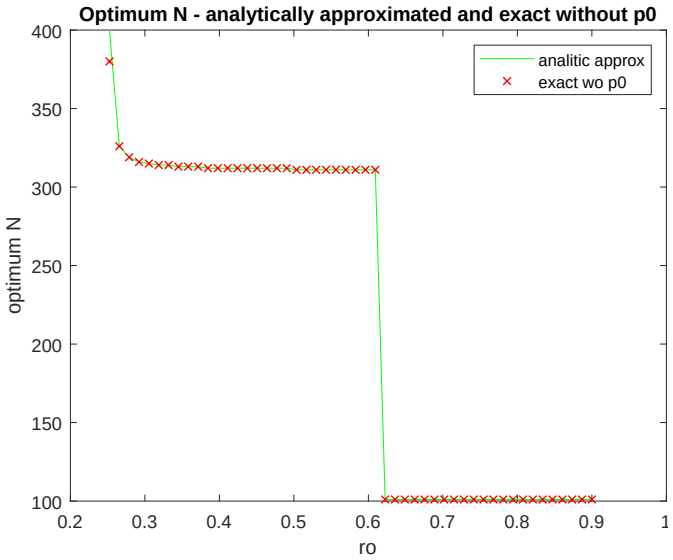

Figure 16. Exact and approximate optimal N (F_2) in dependency of ρ.

The comparison of the exact and approximate optimal N of F_2 can be seen in Figure 16 in dependency of ρ for the parameter setting $M = 400$, $L = 100$, $K = 450$, $C_W = 50$, $\mu = 1$ and $\rho > 0.25 = \frac{L}{M}$.

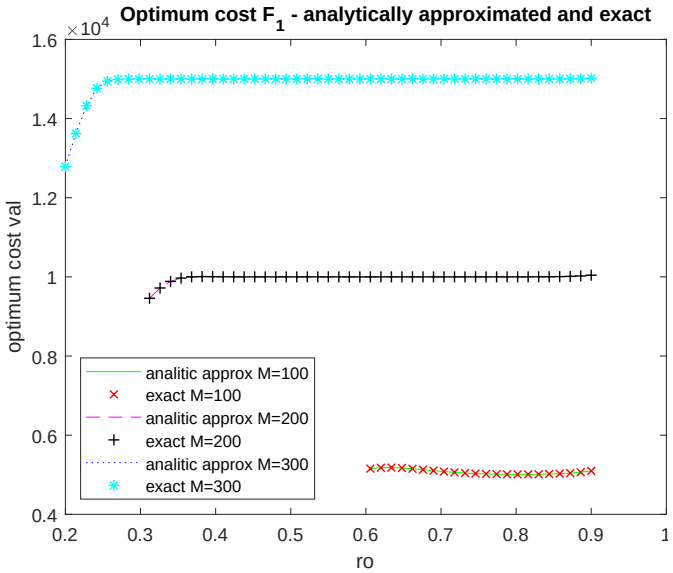

Figure 17. Exact and approximate optimal value (F_1) in dependency of ρ for different values of M.

Figure 17 shows the exact and approximate optimal value of F_1 in dependency of ρ for different values of M with the parameter setting $L = 50$, $K = M + 100$, $C_W = 50$, $\mu = 1$ and $\rho > 0.25 = \frac{L}{M}$.

Both figures show a very good match.

C. Validation of the approximate formula

We validated the approximate solution formula by numeric optimization in the considered range of parameters. Figure 18 shows the ratio of the approximated and the exact optimal value of F_1 for the range of parameters $100 \leq M \leq 700$ and $\rho > \frac{L}{M}$ with the parameter setting $L = 50$, $K = M + 100$, $C_W = 50$, $\mu = 1$.

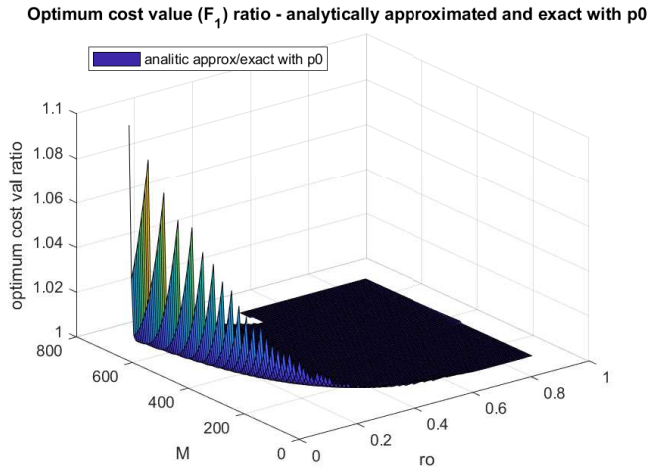

Figure 18. Ratio of the approximated and exact optimal value (F_1) for $100 \leq M \leq 700$ and $\frac{L}{M} < \rho$.

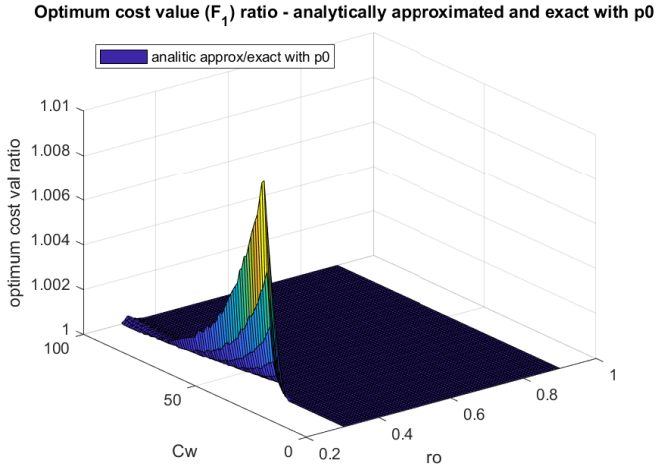

Figure 19. Ratio of the approximated and exact optimal value (F_1) for $0.1 \leq C_W \leq 100$ and $\frac{L}{M} = 0.25 < \rho$.

Similarly, Figure 19 shows the ratio of the approximated and the exact optimal value of F_1 for the range of parameters $0.1 \leq C_W \leq 100$ and $\rho > 0.25 = \frac{L}{M}$ with the parameter setting $L = 50$, $M = 200$, $K = 300$, $\mu = 1$.

Both figures show a very good match until approaching the ρ boundary $\frac{L}{M}$, where the condition 3, does not hold any more.

D. Low traffic range - $\frac{\lambda}{L\mu} < 1$

We used two basic parameter settings and their variations for all experiments. One of them is $M = 100, L = 50, K = M + 50, \mu = 1, C_W = 50, C_{off} = 15, C_{on} = 50, C_a = 30, C_d = 20$, which represents a typical cost function curve of Type 1. The other one is $M = 200, L = 100, K = M + 50, \mu = 1, C_W = 10, C_{off} = 0.01, C_{on} = 0.02, C_a = 0.03, C_d = 0.02$ and it belongs to cost function curves of Type 2. Although the parameter C_R was set to 200 in both cases, it has no impact on the approximate solution formula in the considered range of parameters due to condition 5.

E. Illustration of the approximate solution formula

First, we illustrate the approximate solution formula for typical cost function curve of Type 2 (see Figure 15) under the parameter setting $M = 200, L = 100, K = M + 50, \mu = 1, C_W = 10, C_{off} = 0.01, C_{on} = 0.02, C_a = 0.03, C_d = 0.02, C_R = 200$, while ρ is varied in the low traffic range. Figure 20 shows the approximate optimal N and the upper limit on the lower boundary of the "almost constant cost range", N_{uccl}, both as a function of ρ.

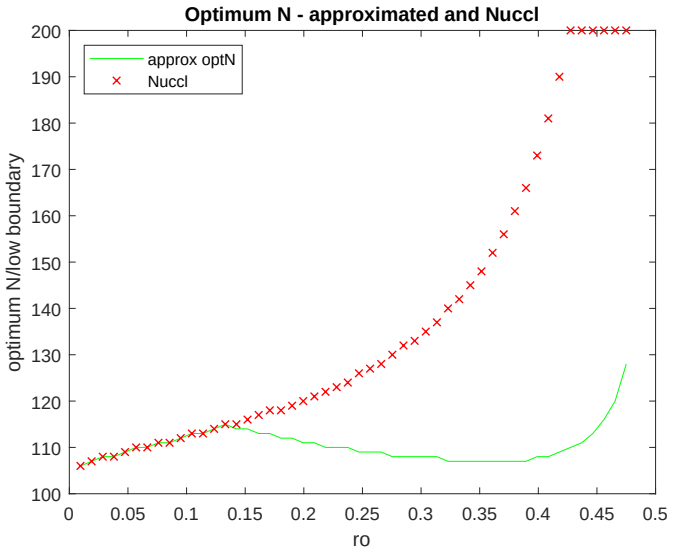

Figure 20. Approximate optimal N and N_{uccl} in dependency of ρ - Type 2 cost function.

It can be seen on the figure that higher the ρ, higher the upper limit on the lower boundary of the "almost constant cost range". Additionally it can be also observed on the figure that there exists a global optimum point above some ρ, for cost function curves of Type 2, which is also expected from the form of the cost curve.

Figure 21 shows the exact and approximate optimal N in dependency of ρ, while keeping the above parameter setting for the cost function curve of Type 2.

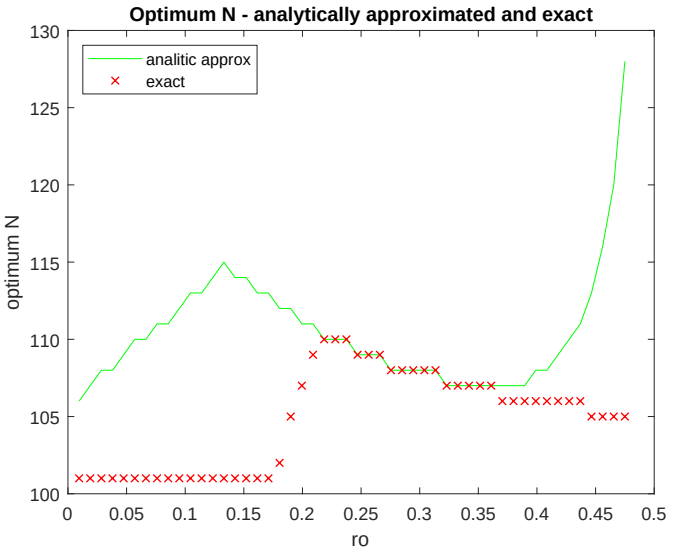

Figure 21. Exact and approximate optimal N in dependency of ρ - Type 2 cost function.

If the approximate optimal N equals to N_{uccl} then the whole "almost constant cost range" is a range of optimal N-s, for which the value of the cost function is approximately the same, i.e., it changes only in a negligible magnitude. Note that this range can fall also below N_{uccl}, since N_{uccl} is only an upper limit on the lower boundary of the "almost constant cost range". This explains the mismatch between the exact optimal N and the approximate one, which is set simple to N_{uccl}, in the traffic range $\lessapprox 0.15$. In fact the lower boundary of the "almost constant cost range" can be below its upper limit, N_{uccl}, also for the values of ρ somewhat above 0.15, which explains that the mismatch continues up to ≈ 0.21.

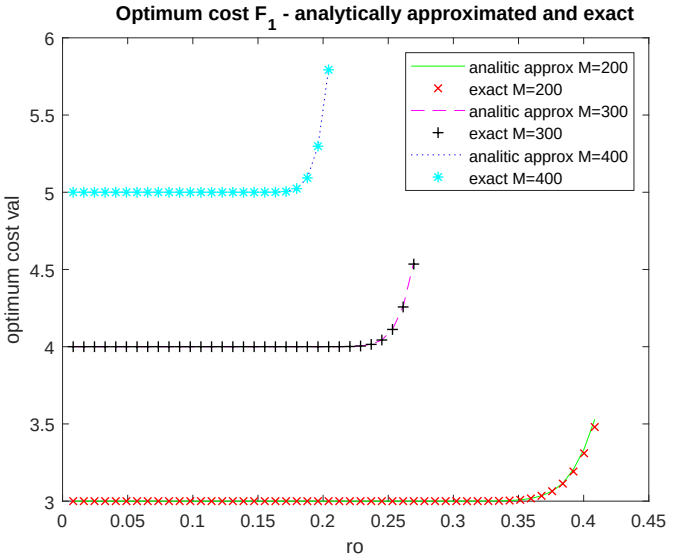

Figure 22. Exact and approximate optimal value (F_1) in dependency of ρ for different values of M.

However this mismatch does not have any significance, since the value of the cost function is approximately the same in the whole range of optimal N-s. It follows that the approximate optimal N for the traffic range $\rho \lessapprox 0.21$ is irrelevant and the exact and approximate optimal value of F_1 show a good match, as it can be seen on the lowest curve (with $M = 200$) in Figure 22. In the next traffic range $\rho > 0.21$ the exact and approximate optimal N show a good match until approaching $\rho = 0.4$. Above that point the condition 3. does not hold any more, which causes a mismatch between the exact and approximate values not only in optimal N but also in optimal value. Therefore, we focus on the traffic range $\rho \leq \frac{L}{M}\psi$ with $\psi = 0.8$.

Figure 22 compares the exact and approximate optimal value of the cost function in dependency of ρ for different values of M with the above parameter setting and $\rho \leq \frac{L}{M}\psi$. The figure shows a good match for all three values of M.

Next we illustrate the approximate solution formula for typical cost function curves of Type 1 (see Figure 14) under the parameter setting $M = 100, L = 50, K = M + 50, \mu = 1, C_W = 50, C_{off} = 15, C_{on} = 50, C_a = 30, C_d = 20, C_R = 200$, while ρ is varied in the low traffic range.

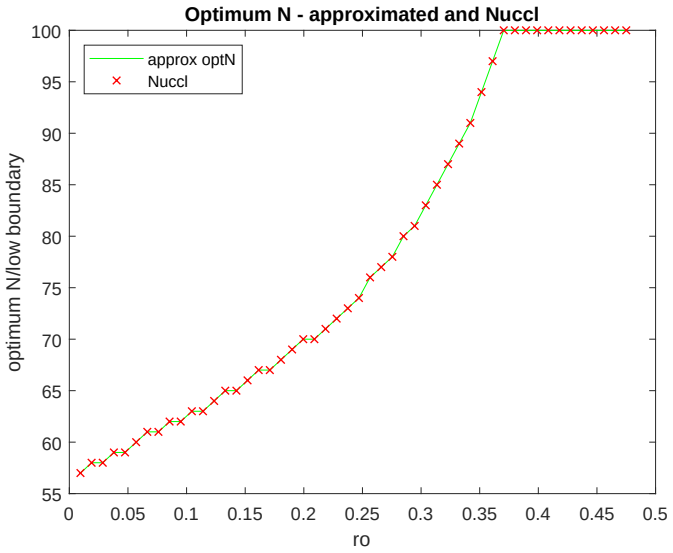

Figure 23. Approximate optimal N and N_{uccl} in dependency of ρ - Type 1 cost function.

Figure 23 plots the approximate optimal N and N_{uccl} as a dependency of ρ. The figure shows that there exists a range of optimal N-s for cost function curves of Type 1 with any value of ρ, which is expected again from the form of the cost curve. In this case the approximate optimal N is irrelevant since any value in the range of optimal N-s can be considered as optimal N. Therefore, we focus on the approximate optimal value.

The exact and approximate optimal value of the cost function in dependency of ρ are plotted on Figure 24 for different values of C_w, with the above parameter setting and $\rho \leq \frac{L}{M}\psi$. The figure shows again a good match.

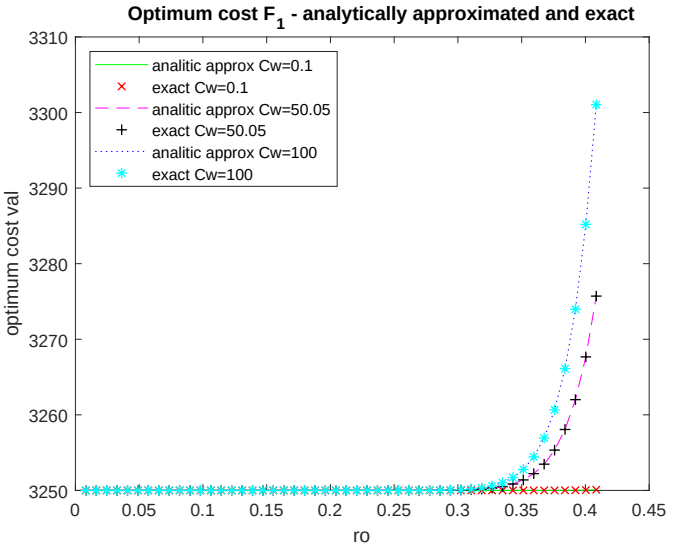

Figure 24. Exact and approximate optimal value (F_1) in dependency of ρ for different values of C_W.

F. Validation of the approximate formula

We validated the approximate solution formula for the traffic range $\frac{\lambda}{L\mu} < 1$ again by numeric optimization in the considered range of parameters.

First, we validate it by the help of the parameter set for cost function curve of Type 2, $M = 200, L = 100, K = M + 50, \mu = 1, C_W = 10, C_{off} = 0.01, C_{on} = 0.02, C_a = 0.03, C_d = 0.02, C_R = 200$.

Figure 25 and 26 show the ratio of the approximated and the exact optimal N and the ratio of the approximated and the exact optimal value of F_1, respectively, for the range of parameters $200 \leq M \leq 800$ and $\rho \leq \frac{L}{M}\psi$.

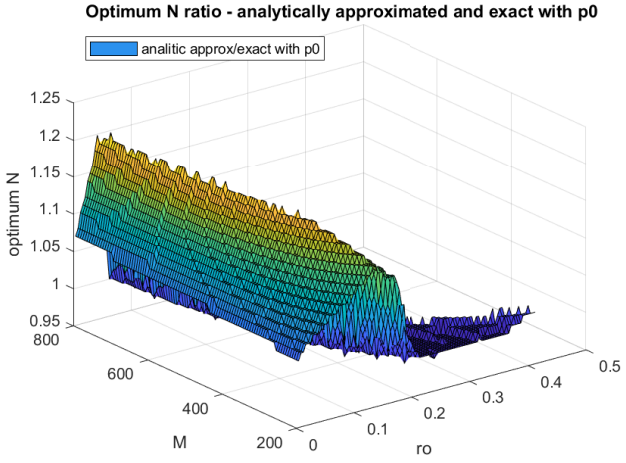

Figure 25. Ratio of the approximated and exact optimal N for $200 \leq M \leq 800$ and $\rho \leq \frac{L}{M}\psi$.

It can be seen on Figure 25 that the approximated optimal N deviates from its exact value by a factor $0.95 - 1.2$ for low values of ρ. This is due to the existence of range of optimal N making the approximate optimal N irrelevant as explained at Figure 21. For higher values of ρ (up to $\rho \leq \frac{L}{M}\psi$) the exact and approximated optimal N show a good match.

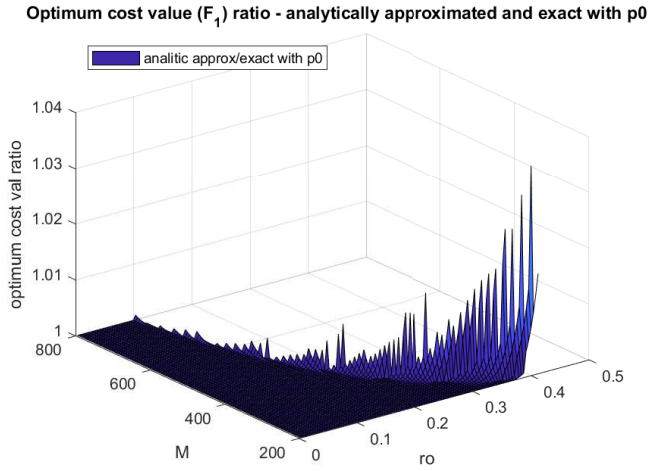

Figure 26. Ratio of the approximated and exact optimal value (F_1) for $200 \leq M \leq 800$ and $\rho \leq \frac{L}{M}\psi$.

The approximated and exact optimal value of F_1 show a very good match, as it can be seen on Figure 26.

Now we validate the approximate solution formula also by the help of the parameter setting for cost function curve of Type 1, $M = 100, L = 50, K = M + 50, \mu = 1, C_W = 50, C_{off} = 15, C_{on} = 50, C_a = 30, C_d = 20, C_R = 200$. In this case the approximate optimal N is irrelevant for the whole range of ρ due to the existence of a range of optimal N-s and therefore we validate only the optimal value of the cost function.

Figure 27 shows the ratio of the approximated and the exact optimal value of F_1 for the range of parameters $100 \leq M \leq 700$ and $\rho \leq \frac{L}{M}\psi$.

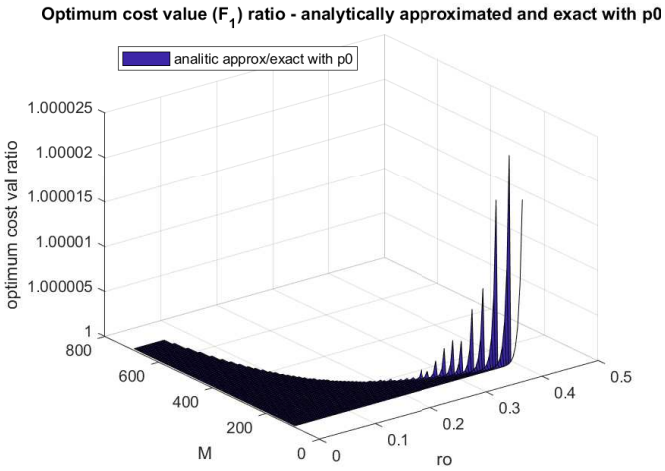

Figure 27. Ratio of the approximated and exact optimal value (F_1) for $100 \leq M \leq 700$ and $\rho < \frac{\frac{L}{M}}{\xi}$.

The approximated and exact optimal value of the cost function F_1 show a very good match. A small mismatch (like

also in Figure 26) can be observed in the parameter area of approaching the traffic boundary $\frac{L}{M}\psi$, above which the condition 3, does not hold any more.

Keeping the parameter setting for the cost function curve of Type 1, Figure 28 shows the ratio of the approximated and the exact optimal value of the cost function F_1 for the range of parameters $0.1 \leq C_W \leq 100$ and $\rho \leq \frac{L}{M}\psi$.

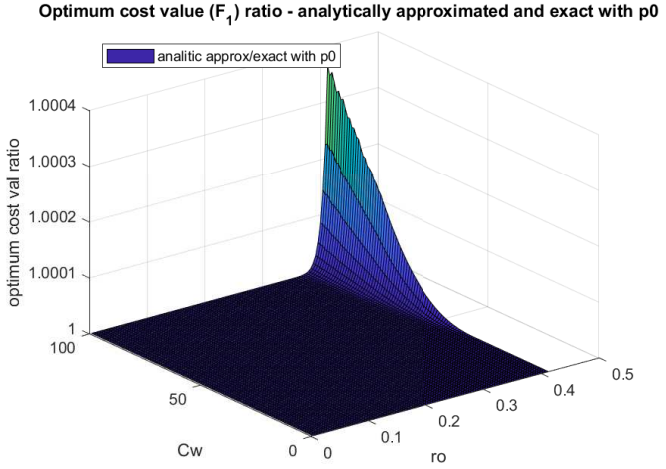

Figure 28. Ratio of the approximated and exact optimal value (F_1) for $0.1 \leq C_W \leq 100$ and $\rho \leq \frac{L}{M}\psi$.

The figure shows a very good match of the approximated and exact optimal values of cost function F_1.

IX. Conclusion and Future Work

In this paper, we continued our previous work [1]. We extended the solution to the low traffic range and generalized the approximate solution formula for the high traffic range by omitting the condition on the cost parameters. Moreover, we also provided the details of the stationary analysis and the derivations both in the former and new optimization parts. The first contribution of this research is the proposal of the shifted N-policy for a simple, but energy efficient control of number of active VMs in the IaaS Cloud. A secondary contribution is the stationary analysis of the underlying queueing model. However, the major contributions are the approximate formulas for computing the optimal threshold N, which minimizes the cloud provider's cost, in the most relevant parameter ranges. The validation of the approximate solution formulas by means of numeric optimization show good match. The closed form approximate solution formulas enable a simple management of the cloud and give an insight into the dependency of the optimal threshold N on the model and cost parameters.

A future research work is to investigate the validity of the solution formulas outside of the parameter ranges defined by the conditions of the solution formulas. A second potential research topic is to establish an approximate solution also for the traffic range around $\frac{\lambda}{\mu} = L \Leftrightarrow \rho = \frac{L}{M}$. Further future research topics are the optimization of L besides fixed N and the rather more difficult joint optimization of parameters L and N.

Appendix I
Derivation of Stationary Distribution of the Number of Requests

1) p_k for $k = 1, \ldots, L$: From the balance equation (1) we have

$$p_{i+1} = \frac{\lambda}{(i+1)\mu} p_i, \ i = 0, \ldots L-1 \ \Leftrightarrow$$
$$p_i \ = \frac{\lambda}{i\mu} p_{i-1}, \ i = 1, \ldots L.$$

Solving it recursively for $i = 1, \ldots L$ gives

$$p_k = \frac{(\frac{\lambda}{\mu})^k}{k!} p_0, \text{ for } k = 1, \ldots, L.$$

2) p_k for $k = -(N-L-1), \ldots, -1$: The balance equation (3) can be rearranged as

$$p_j = \frac{\lambda}{L\mu}(p_{j-1} - p_{-1}) \text{ for } j = -(N-L-2), \ldots, -1.$$

Solving it recursively for $j = -(N-L-2), \ldots, -1$ we get

$$p_{-(N-L)+k} = \left(\frac{\lambda}{L\mu}\right)^{k-1} p_{-(N-L-1)} - \sum_{j=1}^{k-1} \left(\frac{\lambda}{L\mu}\right)^j p_{-1}$$
$$= \left(\frac{\lambda}{L\mu}\right)^{k-1} p_{-(N-L-1)} - \frac{\lambda}{L\mu} \frac{1 - \left(\frac{\lambda}{L\mu}\right)^{k-1}}{1 - \frac{\lambda}{L\mu}} p_{-1}$$
$$\text{for } k = 2, \ldots, (N-L-1).$$

Setting $k = N-L-1$ gives

$$p_{-1} = \left(\frac{\lambda}{L\mu}\right)^{N-L-2} p_{-(N-L-1)} - \frac{\frac{\lambda}{L\mu} - \left(\frac{\lambda}{L\mu}\right)^{N-L-1}}{1 - \frac{\lambda}{L\mu}} p_{-1},$$

from which p_{-1} can be expressed in terms of $p_{-(N-L-1)}$ as

$$p_{-1} = \frac{1 - \frac{\lambda}{L\mu}}{1 - \left(\frac{\lambda}{L\mu}\right)^{N-L-1}} \left(\frac{\lambda}{L\mu}\right)^{N-L-2} p_{-(N-L-1)},$$
$$= \frac{1}{\sum_{j=0}^{N-L-2} \left(\frac{\lambda}{L\mu}\right)^j} \left(\frac{\lambda}{L\mu}\right)^{N-L-2} p_{-(N-L-1)}.$$

Applying the expression of $p_{-(N-L-1)}$ from the balance equation (2) in the above relation and rearrangement leads to

$$p_{-1} \sum_{j=0}^{N-L-2} \left(\frac{\lambda}{L\mu}\right)^j = \left(\frac{\lambda}{L\mu}\right)^{N-L-2} \frac{\lambda}{L\mu}(p_L - p_{-1}) \ \Leftrightarrow$$
$$p_{-1} \sum_{j=0}^{N-L-1} \left(\frac{\lambda}{L\mu}\right)^j = \left(\frac{\lambda}{L\mu}\right)^{N-L-1} p_L,$$

which results in the expression of p_{-1} as

$$p_{-1} = \left(\frac{\lambda}{L\mu}\right)^{N-L-1} \frac{1-\frac{\lambda}{L\mu}}{1-\left(\frac{\lambda}{L\mu}\right)^{N-L}}\ p_L$$
$$= \left(\frac{\lambda}{L\mu}\right)^{N-L-1} \frac{1}{\sum_{j=0}^{N-L-1}\left(\frac{\lambda}{L\mu}\right)^{j}}\ p_L.$$

Substituting it back into the expression of $p_{-(N-L-1)}$ from the balance equation (2) leads to the expression of $p_{-(N-L-1)}$ as

$$p_{-(N-L-1)} = \frac{\lambda}{L\mu}\left(1 - \frac{\left(\frac{\lambda}{L\mu}\right)^{N-L-1}}{\sum_{j=0}^{N-L-1}\left(\frac{\lambda}{L\mu}\right)^{j}}\right)\ p_L$$
$$= \frac{\lambda}{L\mu}\frac{\sum_{j=0}^{N-L-2}\left(\frac{\lambda}{L\mu}\right)^{j}}{\sum_{j=0}^{N-L-1}\left(\frac{\lambda}{L\mu}\right)^{j}}\ p_L$$
$$= \frac{\lambda}{L\mu}\frac{1-\left(\frac{\lambda}{L\mu}\right)^{N-L-1}}{1-\left(\frac{\lambda}{L\mu}\right)^{N-L}}\ p_L.$$

Now substituting the expressions of p_{-1} and $p_{-(N-L-1)}$ back into the expression of $p_{-(N-L)+k}$ and rearranging it leads to

$$p_{-(N-L)+k} = \left(\frac{\lambda}{L\mu}\right)^{k-1}\frac{\lambda}{L\mu}\frac{1-\left(\frac{\lambda}{L\mu}\right)^{N-L-1}}{1-\left(\frac{\lambda}{L\mu}\right)^{N-L}}\ p_L$$
$$-\frac{\lambda}{L\mu}\frac{1-\left(\frac{\lambda}{L\mu}\right)^{k-1}}{1-\frac{\lambda}{L\mu}}\left(\frac{\lambda}{L\mu}\right)^{N-L-1}\frac{1-\frac{\lambda}{L\mu}}{1-\left(\frac{\lambda}{L\mu}\right)^{N-L}}\ p_L$$
$$=\frac{\left(\frac{\lambda}{L\mu}\right)^{k}\left[1-\left(\frac{\lambda}{L\mu}\right)^{N-L-1}\right]-\left(\frac{\lambda}{L\mu}\right)^{N-L}\left[1-\left(\frac{\lambda}{L\mu}\right)^{k-1}\right]}{1-\left(\frac{\lambda}{L\mu}\right)^{N-L}}\ p_L$$
$$=\frac{\left(\frac{\lambda}{L\mu}\right)^{k}-\left(\frac{\lambda}{L\mu}\right)^{N-L}}{1-\left(\frac{\lambda}{L\mu}\right)^{N-L}}\ p_L \text{ for } k=1,\ldots,(N-L-1),$$

where we utilized that this formula holds also for $k = 1$. Applying the reindexing $-(N-L)+k \Rightarrow k$, we get

$$p_k = \left(\frac{\lambda}{L\mu}\right)^{N-L}\frac{\left(\frac{\lambda}{L\mu}\right)^{k}-1}{1-\left(\frac{\lambda}{L\mu}\right)^{N-L}}\ p_L$$
$$\text{for } k=-(N-L-1),\ldots,-1.$$

3) p_k for $k = L+1,\ldots,N$: The balance equation (5) can be rearranged as

$$p_{k+1}=\frac{\lambda}{(k+1)\mu}p_k+\frac{\lambda}{(k+1)\mu}p_{-1}, k=L+1,\ldots,N-1 \Leftrightarrow$$
$$p_k \quad =\frac{\lambda}{k\mu}p_{k-1}+\frac{\lambda}{k\mu}p_{-1},\ k=L+2,\ldots N.$$

Solving it recursively for $k = L+2,\ldots N$ we get

$$p_k = \prod_{i=L+2}^{k}\frac{\lambda}{i\mu}p_{L+1}+\sum_{i=L+2}^{k}\prod_{\ell=i}^{k}\frac{\lambda}{\ell\mu}p_{-1}$$
$$=\frac{(L+1)!}{k!}\left(\frac{\lambda}{\mu}\right)^{k-(L+1)}p_{L+1}+\sum_{i=L+2}^{k}\frac{(i-1)!}{k!}\left(\frac{\lambda}{\mu}\right)^{k-i+1}p_{-1},$$
$$k=L+2,\ldots N.$$

Using the expression of p_{L+1} from balance equation (4) and rearrangement gives

$$p_k = \frac{(L+1)!}{k!}\left(\frac{\lambda}{\mu}\right)^{k-(L+1)}\frac{\lambda}{(L+1)\mu}p_{-1}$$
$$+\sum_{i=L+2}^{k}\frac{(i-1)!}{k!}\left(\frac{\lambda}{\mu}\right)^{k-i+1}p_{-1}$$
$$=\left[\frac{L!}{k!}\left(\frac{\lambda}{\mu}\right)^{k-L}+\sum_{i=L+2}^{k}\frac{(i-1)!}{k!}\left(\frac{\lambda}{\mu}\right)^{k-i+1}\right]p_{-1}$$
$$=\sum_{i=L+1}^{k}\frac{(i-1)!}{k!}\left(\frac{\lambda}{\mu}\right)^{k-(i-1)}p_{-1}$$
$$=\sum_{i=L}^{k-1}\frac{i!}{k!}\left(\frac{\lambda}{\mu}\right)^{k-i}p_{-1},\ k=L+1,\ldots N,$$

where we utilized that this formula holds also for $k = L+1$ due to balance equation (4).

4) p_k for $k = N+1,\ldots,M$: From the balance equation (6) we have

$$p_{r+1} = \frac{\lambda}{(r+1)\mu}p_r,\ r=N,\ldots M-1 \ \Leftrightarrow$$
$$p_r \quad = \frac{\lambda}{r\mu}p_{r-1},\ r=N+1,\ldots M.$$

Solving it recursively for $r = N+1,\ldots M$ gives

$$p_k = \frac{N!}{k!}\left(\frac{\lambda}{\mu}\right)^{k-N}p_N,\ \text{for } k=N+1,\ldots,M.$$

5) p_k for $k = M+1,\ldots,K$: From the balance equation (7) we have

$$p_{t+1} = \frac{\lambda}{M\mu}p_t,\ t=M,\ldots K-1 \ \Leftrightarrow$$
$$p_t \quad = \frac{\lambda}{M\mu}p_{t-1},\ t=M+1,\ldots K.$$

Solving it recursively for $t = M+1,\ldots K$ gives

$$p_k = \left(\frac{\lambda}{M\mu}\right)^{k-M}p_M,\ \text{for } k=M+1,\ldots,K.$$

APPENDIX II
DERIVATION OF $E[W]$

The expected waiting time of the requests is defined by

$$E[W] = \sum_{k=1}^{N-L-1} k\ p_{-(N-L)+k} + \sum_{k=M+1}^{K} (k-M)p_k.$$

The first term of $E[W]$ can be rearranged as

$$\sum_{k=1}^{N-L-1} k\ p_{-(N-L)+k} = \sum_{k=1}^{N-L-1} k\ \frac{(\frac{\lambda}{L\mu})^k - (\frac{\lambda}{L\mu})^{N-L}}{1-(\frac{\lambda}{L\mu})^{N-L}} p_L$$

$$= \frac{1}{1-(\frac{\lambda}{L\mu})^{N-L}} \left[\frac{\lambda}{L\mu} \frac{1-(\frac{\lambda}{L\mu})^{N-L}}{(1-\frac{\lambda}{L\mu})^2} - (N-L)\frac{(\frac{\lambda}{L\mu})^{N-L}}{1-\frac{\lambda}{L\mu}} \right.$$

$$\left. - \frac{(N-L)(N-L-1)}{2} \left(\frac{\lambda}{L\mu}\right)^{N-L} \right] p_L = \left\{ \frac{\frac{\lambda}{L\mu}}{(1-\frac{\lambda}{L\mu})^2} \right.$$

$$\left. - (N-L)\ \frac{(\frac{\lambda}{L\mu})^{N-L}}{1-(\frac{\lambda}{L\mu})^{N-L}} \left[\frac{1}{1-\frac{\lambda}{L\mu}} + \frac{N-L-1}{2} \right] \right\} p_L,$$

where we used the formula

$$\sum_{k=1}^{n} k\ q^k = q\frac{1-q^{n+1}}{(1-q)^2} - (n+1)\frac{q^{n+1}}{1-q}.$$

The second term of $E[W]$ can be rearranged as

$$\sum_{k=M+1}^{K} (k-M)p_k = \sum_{k-M=1}^{K-M} (k-M) \left(\frac{\lambda}{M\mu}\right)^{k-M} p_M$$

$$= \sum_{i=0}^{K-M} ((i+1)-1) \left(\frac{\lambda}{M\mu}\right)^{i} p_M$$

$$= \left[\frac{d\left(\sum_{i=0}^{K-M} (\frac{\lambda}{M\mu})^{i+1}\right)}{d(\frac{\lambda}{M\mu})} - \sum_{i=0}^{K-M} \left(\frac{\lambda}{M\mu}\right)^{i} \right] p_M$$

$$= \left[\frac{d\left(\frac{\frac{\lambda}{M\mu} - (\frac{\lambda}{M\mu})^{KM+2}}{1-\frac{\lambda}{M\mu}}\right)}{d(\frac{\lambda}{M\mu})} - \frac{1-(\frac{\lambda}{M\mu})^{K-M+1}}{1-\frac{\lambda}{M\mu}} \right] p_M$$

$$= \left[\frac{\left(1-(K-M+2)(\frac{\lambda}{M\mu})^{K-M+1}\right)(1-\frac{\lambda}{M\mu})}{(1-\frac{\lambda}{M\mu})^2} \right.$$

$$+ \frac{\lambda}{M\mu} \frac{\left(1-(\frac{\lambda}{M\mu})^{K-M+1}\right)}{(1-\frac{\lambda}{M\mu})^2}$$

$$\left. - \frac{1-\frac{\lambda}{M\mu} - \left(1-\frac{\lambda}{M\mu}\right)(\frac{\lambda}{M\mu})^{K-M+1}}{(1-\frac{\lambda}{M\mu})^2} \right] p_M$$

$$= \left[\frac{\lambda}{M\mu} \frac{\left(1-(\frac{\lambda}{M\mu})^{K-M+1}\right)}{(1-\frac{\lambda}{M\mu})^2} - (K-M+1)\frac{(\frac{\lambda}{M\mu})^{K-M+1}}{1-\frac{\lambda}{M\mu}} \right] p_M.$$

Putting all these together as well as using τ and σ defined in (10), we get the final expression of $E[W]$ as

$$E[W] = \tau p_L + \sigma p_M.$$

APPENDIX III
AUXILIARY RELATIONS

A. Upper limit for $\sum_{k=0}^{L} \frac{q^k}{k!}$

Statement:

$$\sum_{k=0}^{L} \frac{q^k}{k!} \le \frac{1}{1-\frac{L}{q}} \frac{q^L}{L!} \text{ for } L \in \mathbb{N}, 0 < q \text{ and } L < q.$$

The statement can be shown as

$$\sum_{k=0}^{L} \frac{q^k}{k!} = \frac{q^L}{L!} \sum_{k=0}^{L} \frac{q^k}{k!} \frac{L!}{q^L}$$

$$= \frac{q^L}{L!} \left(1 + \frac{L}{q} + \frac{L(L-1)}{q^2} + \ldots + \frac{L\ldots 1}{q^L}\right)$$

$$\le \frac{q^L}{L!} \left(1 + \frac{L}{q} + \frac{L}{q^2}\frac{L}{} + \ldots + \frac{L\ldots L}{q^L}\right)$$

$$= \frac{q^L}{L!} \sum_{k=0}^{L} (\frac{L}{q})^k = \frac{q^L}{L!} \frac{1-(\frac{L}{q})^{L+1}}{1-\frac{L}{q}} \le \frac{1}{1-\frac{L}{q}} \frac{q^L}{L!},$$

where in the last step we utilized $0 \le \frac{L}{q} < 1$ and $0 < L+1$.

B. Formula for $\sum_{k=A+1}^{B} \frac{q^k}{k!} \frac{A!}{q^A}$

Statement: The following formula holds for $\sum_{k=A+1}^{B} \frac{q^k}{k!} \frac{A!}{q^A}$ for $A, B \in \mathbb{N}$, $A < B$, $0 < q$ as well as $q \neq A+1$ and $q \neq B$

$$\sum_{k=A+1}^{B} \frac{q^k}{k!} \frac{A!}{q^A} = f_1 \frac{1-f_1^{B-A}}{1-f_1},$$

where $f_1 = \frac{q}{f_0}$ and $A+1 \le f_0 \le B$.

The sum in the statement can be rearranged as

$$\sum_{k=A+1}^{B} \frac{q^k}{k!} \frac{A!}{q^A}$$

$$= \left(\frac{q}{A+1} + \frac{q^2}{(A+1)(A+2)} + \ldots + \frac{q^{B-A}}{(A+1)\ldots B}\right).$$

If all terms in the denominators are replaced by $A+1$ then we get an upper limit as

$$\sum_{k=A+1}^{B} \frac{q^k}{k!} \frac{A!}{q^A}$$

$$\le \left(\frac{q}{A+1} + \frac{q^2}{(A+1)(A+1)} + \ldots + \frac{q^{B-A}}{(A+1)\ldots(A+1)}\right)$$

$$= \sum_{i=1}^{B-A} \left(\frac{q}{A+1}\right)^i = \frac{q}{A+1} \frac{1-(\frac{q}{A+1})^{B-A}}{1-\frac{q}{A+1}}.$$

Similarly, replacing all terms in the denominator by B, we get a lower limit as

$$\sum_{k=A+1}^{B} \frac{q^k}{k!} \frac{A!}{q^A} \ge \left(\frac{q}{B} + \frac{q^2}{B\ B} + \ldots + \frac{q^{B-A}}{B\ldots B}\right)$$

$$= \sum_{i=1}^{B-A} \left(\frac{q}{B}\right)^i = \frac{q}{B} \frac{1-(\frac{q}{B})^{B-A}}{1-\frac{q}{B}}.$$

It follows from the development of the function $x\frac{1-(x)^{B-A}}{1-x}$ in the range $\frac{q}{B} \leq x \leq \frac{q}{A+1}$ that there exists an $f_1 = \frac{q}{f_0}$ with $A+1 \leq f_0 \leq B$, for which

$$\sum_{k=A+1}^{B} \frac{q^k}{k!}\frac{A!}{q^A} = f_1\frac{1-f_1^{B-A}}{1-f_1}$$

holds.

Note that keeping B (or A) constant a higher A (or B) implies higher f_0 and thus lower f_1.

C. Upper limit for q at f_1, in the formula for $\sum_{k=A+1}^{B} \frac{q^k}{k!}\frac{A!}{q^A}$, close to 1

Statement: An upper limit for q, for which f_1 in the formula in III-B has a value close to 1, can be given as

$$q < Ae^{\frac{5.5}{A}}.$$

where $A \in \mathbb{N}$, $0 < A$, $0 < q$ as well as $q \neq A+1$.

The sum $\sum_{k=A+1}^{B} \frac{q^k}{k!}\frac{A!}{q^A}$ can be rearranged as

$$\sum_{k=A+1}^{B} \frac{q^k}{k!}\frac{A!}{q^A} = \frac{q}{A(1+\frac{1}{A})} + \frac{q^2}{A^2(1+\frac{1}{A})(1+\frac{2}{A})} + \ldots + \frac{q^{B-A}}{A^{B-A}(1+\frac{1}{A})\ldots(1+\frac{B-A}{A})}.$$

Using $1 + x \lessapprox e^x$ and $\sum_{k=1}^{n} k = \frac{n(n+1)}{2}$ in the above expression we get

$$\begin{aligned}
&\sum_{k=A+1}^{B} \frac{q^k}{k!}\frac{A!}{q^A} \\
&\lessapprox \frac{q}{A\ e^{\frac{1}{A}}} + \frac{q^2}{A^2\ e^{\frac{1}{A}}e^{\frac{2}{A}}} + \ldots + \frac{q^{B-A}}{A^{B-A}\ e^{\frac{1}{A}}\ldots e^{\frac{B-A}{A}}} \\
&= \frac{q}{A\ e^{\frac{1}{A}}} + \frac{q^2}{A^2\ e^{\frac{2\times 3}{2A}}} + \ldots + \frac{q^{B-A}}{A^{B-A}\ e^{\frac{(B-A)(B-A+1)}{2A}}} \\
&= \frac{q/A}{e^{\frac{1}{A}}} + \frac{(q/A)^2}{(e^{\frac{1}{A}}e^{\frac{1}{2A}})^2} + \ldots + \frac{(q/A)^{B-A}}{(e^{\frac{1}{A}}e^{\frac{B-A-1}{2A}})^{(B-A)}}.
\end{aligned}$$

The k-th term of the sum has the form $\left(\frac{q/A}{e^{\frac{1}{A}}}\frac{1}{e^{\frac{k-1}{2A}}}\right)^k$ and thus it decreases with incresing k due to $e^{\frac{1}{2A}} > 1$. Let us determine the value of k, as a function of q and A, at which the k-th term equals to 1.

$$\frac{q/A}{e^{\frac{1}{A}}}\frac{1}{e^{\frac{k-1}{2A}}} = 1 \ \Leftrightarrow\ \frac{q/A}{e^{\frac{1}{A}}} = e^{\frac{k-1}{2A}} \ \Leftrightarrow\ \frac{k-1}{2A} = \ln(q/A) - \frac{1}{A} \ \Leftrightarrow\ k = 2A\ln(q/A) - 1.$$

Around $f_1 \approx 1$ the magnitude of the sum $\sum_{k=A+1}^{B} \frac{q^k}{k!}\frac{A!}{q^A}$ changes from $\ll 1$ to $\gg 1$ with increasing f_1 (and q). This can be taken into account by setting $k \gg 1$, since in this case the first k terms are ≥ 1 implying $\sum_{k=A+1}^{B} \frac{q^k}{k!}\frac{A!}{q^A} > k \gg 1$. In fact the k-th term will be even higher than 1 due to the the overestimations of the terms with the form $1 + x$ by e^x. Therefore, setting a value lower than $k \gg 1$ is enough to have $\sum_{k=A+1}^{B} \frac{q^k}{k!}\frac{A!}{q^A} > k \gg 1$ and thus $2A\ln(q/A) - 1$ is a lower limit for k. We count for $k \gg 1$ by setting $k = 10$. This leads to the upper limit of q as

$$10 > 2A\ln(q/A) - 1 \ \Leftrightarrow\ \ln(q/A) < \frac{5.5}{A} \ \Leftrightarrow\ q < Ae^{\frac{5.5}{A}}.$$

Note that with increasing $\frac{B-A}{A} = \frac{B}{A} - 1$ increases also the overestimation of q by the above limit.

D. Formula for $\sum_{k=A+1}^{B} \frac{k!}{q^k}\frac{q^A}{A!}$

Statement: The following formula holds for $\sum_{k=A+1}^{B} \frac{k!}{q^k}\frac{q^A}{A!}$ for $A, B \in \mathbb{N}$, $A < B$ and $0 < q$ as well as $q \neq A+1$ and $q \neq B$

$$\sum_{k=A+1}^{B} \frac{k!}{q^k}\frac{q^A}{A!} = g_1\frac{1-g_1^{B-A}}{1-g_1},$$

where $g_1 = \frac{g_0}{q}$ and $A+1 \leq g_0 \leq B$.

The sum in the statement can be rearranged as

$$\sum_{k=A+1}^{B} \frac{k!}{q^k}\frac{q^A}{A!} = \left(\frac{A+1}{q} + \frac{(A+1)(A+2)}{q^2} + \ldots + \frac{(A+1)\ldots B}{q^{B-A}}\right).$$

If all terms in the nominators are replaced by $A+1$ then we get a lower limit as

$$\begin{aligned}
&\sum_{k=A+1}^{B} \frac{k!}{q^k}\frac{q^A}{A!} \\
&\geq \left(\frac{A+1}{q} + \frac{(A+1)(A+1)}{q^2} + \ldots + \frac{(A+1)\ldots(A+1)}{q^{B-A}}\right) \\
&= \sum_{i=1}^{B-A}\left(\frac{A+1}{q}\right)^i = \frac{A+1}{q}\frac{1-(\frac{A+1}{q})^{B-A}}{1-\frac{A+1}{q}}.
\end{aligned}$$

Similarly, replacing all terms in the nominator by B, we get an upper limit as

$$\begin{aligned}
\sum_{k=A+1}^{B} \frac{k!}{q^k}\frac{q^A}{A!} &\leq \left(\frac{B}{q} + \frac{B\ B}{q^2} + \ldots + \frac{B\ldots B}{q^{B-A}}\right) \\
&= \sum_{i=1}^{B-A}\left(\frac{B}{q}\right)^i = \frac{B}{q}\frac{1-(\frac{B}{q})^{B-A}}{1-\frac{B}{q}}.
\end{aligned}$$

It follows from the development of the function $x\frac{1-(x)^{B-A}}{1-x}$ in the range $\frac{A+1}{q} \leq x \leq \frac{B}{q}$ that there exists an $g_1 = \frac{g_0}{q}$ with $A+1 \leq g_0 \leq B$, for which

$$\sum_{k=A+1}^{B} \frac{k!}{q^k}\frac{q^A}{A!} = g_1\frac{1-g_1^{B-A}}{1-g_1}$$

holds.

Note that keeping B (or A) constant a higher A (or B) implies higher g_0 and thus higher g_1.

REFERENCES

[1] Z. Saffer, "Optimization of Cloud Model Based on Shifted N-policy $M/M/m/K$ Queue," in E. Borcoci, editor, IARIA, AICT2021: 17th Advanced International Conference on Telecommunications, Valencia, Spain, pp. 11-20, 2021.

[2] F. Durao, J. Fernando, S. Carvalho, A. Fonseka and V. C. Garcia, "A systematic review on cloud computing," J. Supercomput., vol. 68, no. 3, pp. 1321–1346, 2014.

[3] C. Chapman, W. Emmerich, F. G. Mrquez, S. Clayman and A. Galis, "Software architecture definition for on-demand cloud provisioning," Cluster Comput., vol. 15, no. 2, pp. 79–100, 2011.

[4] N. A. Sultan, "Reaching for the cloud: How SMEs can manage," Int. J. Inf. Manage., vol. 31, pp. 272-278, 2011.

[5] W. Christof et al., "Cloud computing - A classification business models and research directions," J. Bus. Inf. Syst. Eng., vol. 1, no. 5, pp. 391-399, 2009.

[6] H. Khazaei, J. Misic and v.B. Misic, "Performance analysis of Cloud computing centers using M/G/m/m+r queuing systems," IEEE Trans. Parallel Distrib. Syst., vol. 23, no. 5, pp. 936-943, 2012.

[7] R. Ghosh, F. Longo, V.K. Naik and K.S. Trivedi, "Modeling and performance analysis of large scale IaaS clouds Future Generation Computer Systems," Future Generation Computer Systems, vol. 29, pp. 1216–1234, 2013.

[8] Q. Duan, " Cloud service performance evaluation: status, challenges, and opportunities a survey from the system modeling perspective," Digital Communications and Networks, vol. 3, no. 2, pp. 101–111, 2017.

[9] G. Kaur, A. Bala and I. Chana, "An intelligent regressive ensemble approach for predicting resource usage in cloud computing," Journal of Parallel and Distributed Computing, vol. 123, pp. 1 12, 2019.

[10] R. Yang, X. Ouyang, Y. Chen, P. Townend and J. Xu, "Intelligent resource scheduling at scale: A machine learning perspective," in2018 IEEE Symposium on Service-Oriented System Engineering (SOSE),3, pp. 132141, 2018.

[11] M. Ghobaei-Arani, S. Jabbehdari and M. A. Pourmina, "An autonomic resource provisioning approach for service-based cloud applications: A hybrid approach," Future Generation Computer Systems, vol. 78, pp. 191 210, 2018.

[12] J. N. Witanto, H. Lim and M. Atiquzzaman, "Adaptive selection of dynamic vm consolidation algorithm using neural network for cloud resource management," Future Generation Computer Systems, vol. 87, pp. 35 42, 2018.

[13] A. Beloglazov, J. Abawajy and R. Buyya, "Energy-aware resource allocation heuristics for efficient management of data centers for Cloud computing," Future Generation Comput. Syst., Vol. 28, no. 5, pp. 755-68, 2012.

[14] F. Nzanywayingoma and Y. Yang, "Efficient resource management techniques in cloud computing environment: a review and discussion," International Journal of Computers and Applications, vol. 41, no. 3, pp. 165–182, 2019.

[15] T. Ma, Y. Chu, L. Zhao and O. Ankhbayar, "Resource Allocation and Scheduling in Cloud Computing: Policy and Algorithm," IETE Technical Review, vol. 31, no. 1, pp. 4–16, 2014.

[16] T. Tournaire, H. Castel-Taleb, E. Hyon and T. Hoche, " Generating optimal thresholds in a hysteresis queue: application to a cloud model," MASCOTS 2019: 27th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems, Rennes, France, Oct 2019, pp.283–294.

[17] B. Wan, J. Dang, Z. Li, H. Gong, F. Zhang and S. Oh, "Modeling Analysis and Cost-Performance Ratio Optimization of Virtual Machine Scheduling in Cloud Computing," IEEE Transactions on Parallel and Distributed Systems, vol. 31, no. 7, pp. 1518–1532, 2020.

[18] Y. Mansouri, A. N. Toosi and R. Buyya, "Cost Optimization for Dynamic Replication and Migration of Data in Cloud Data Centers," in IEEE Transactions on Cloud Computing, vol. 7, no. 3, pp. 705-718, 2019.

[19] W. Whitt, "Approximations for the GI/G/m queue," Production Oper. Management, vol. 2, pp. 114–161, 1993.