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

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

***Abstract*—In this paper, we present the performance analysis and cost optimization of an Infrastructure-as-a-Service (IaaS) cloud model with a capacity control policy. The Virtual Machines (VM) are modeled 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 VMs. We define a cost model, that the cloud provider encounters. It 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 VMs 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 VMs. We call it as shifted N-policy. We provide the stationary analysis of the model. We derive closed form results for the distribution of the number of requests and 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 reduce the problem to a nonlinear equation with a specific structure including factorial terms. We provide the approximate solution of the optimization task. The major result of the work is the closed form approximate solution formula, which gives the optimal threshold under the most relevant range of parameters. The formula gives insight into the dependency of the optimum on the model and cost parameters. We provide also illustrating examples for the most important approximations and validate the approximate solution formula by numeric optimization.**

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

I. Introduction

Cloud computing [1] [2] is a distributed computing paradigm gaining more importance in the last decade. This is driven by rapidly growing demand for computational resources needed by applications in many areas, like e.g., business, science or web-applications. In this work, we deal with Infrastructure-as-a-Service (IaaS) type Cloud service, in which computing resources are delivered to customers. One of the key attribute of Cloud services is the virtualization which enables to decouple the computing resources from the physical hardware and deliver them to customers as Virtual Machines (VM).

Performance evaluation of Cloud services plays a central role for Cloud service providers to get insights into the relationships among the used resources and the performance in order to meet the performance requirements of the user. The users want guaranteed performance and probably will also require Service Level Agreements (SLAs) on Cloud performance in a later, mature phase of business models for Cloud service. However, Cloud depends on many factors, which makes its performance evaluation to a complex issue. 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 modeling of Clouds. In [3], a multi-level interacting stochastic sub-models approach is proposed, which provides 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 [4] and the references herein.

Cloud cost optimization enables the Cloud service provider the service provisioning at minimum cost. It requires an energy efficient resource management technique. Such resource management and allocation policies for Clouds are summarized in [5] [6]. One efficient resource control mechanism for Clouds is the threshold based activation and deactivation of VMs, which can be modeled by hysteresis queue. Such resource control is proposed in [7], in which computational algorithms are provided for computing the optimal thresholds. Another numerical approaches to cloud cost optimization are presented in [8] and [9]. Optimization of Clouds is even more complex issue than its performance evaluation. Hence it is not surprising, that 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 an approximate explicit formula for determining the only threshold of the control mechanism. The formula holds in most relevant range of parameters. The newly introduced resource control 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 (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 [10], 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.

Copyright (c) IARIA, 2021. ISBN: 978-1-61208-860-0

We present closed form results for the stationary distribution of the number of requests and for several performance measures in the shifted N-policy $M/M/m/K$ model. The cost model leads to a discrete optimization task, which can be approximated by a nonlinear continuous optimization task. It turns out that the objective function is not convex everywhere on its definition range. After applying several approximations, including Stirling's formula, we reduce the problem to a nonlinear equation with a specific structure including factorial terms. We provide the approximate solution of the optimization task for a bounded range of parameters. The major contribution of the work is the proposed shifted N-policy resource control and the closed form approximate solution formula for the optimal value of the threshold N under the most relevant range of parameters. The secondary contribution of the work is the stationary analysis of the shifted N-policy $M/M/m/K$ model. The advantage of using the proposed shifted N-policy is that it makes the cloud resource management very simple due to the approximate analytic formula 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, like 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”, like e.g., computing-as-a-service, analytics or Business Intelligence(BI)-as-a-service.

We also provide illustrating examples for the most important approximations and validate the approximate solution formula 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 approximate minimization is discussed in Section V. In Section VI, we give illustrative examples for the approximations and provide the numeric validation of the approximate solution formula. The work is concluded in Section VII.

II. Cloud model description

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 $K - M \geq 1$ VMs. When all active VMs are busy upon arrival of a new request then the request 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 modeling 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 $m = M \geq 1$ servers and buffer capacity for $K - M \geq 1$ requests.

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 upwards (in number of requests), which makes it suitable to be used for 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, since it can be modeled by a finite state Continuous-Time Markov chain (CTMC). The utilization of the system, denoted by ρ is given by

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

C. 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 (=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

$$\begin{aligned} \mathcal{C}_{cloud} = & E[\text{ number of active servers }]\ C_{on} \\ & + E[\text{ number of standby servers }]\ C_{off} \\ & + E[W]\ C_W + p_{loss}\ \lambda\ C_R, \\ & + (\text{ activation rate of standby VMs })\ (M-L)\ C_A \\ & + (\text{ deactivation rate of active VMs })\ (M-L)\ C_D. \end{aligned} \tag{2}$$

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 formula for computing the threshold N (via parameter A see in subsection V-D).

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.

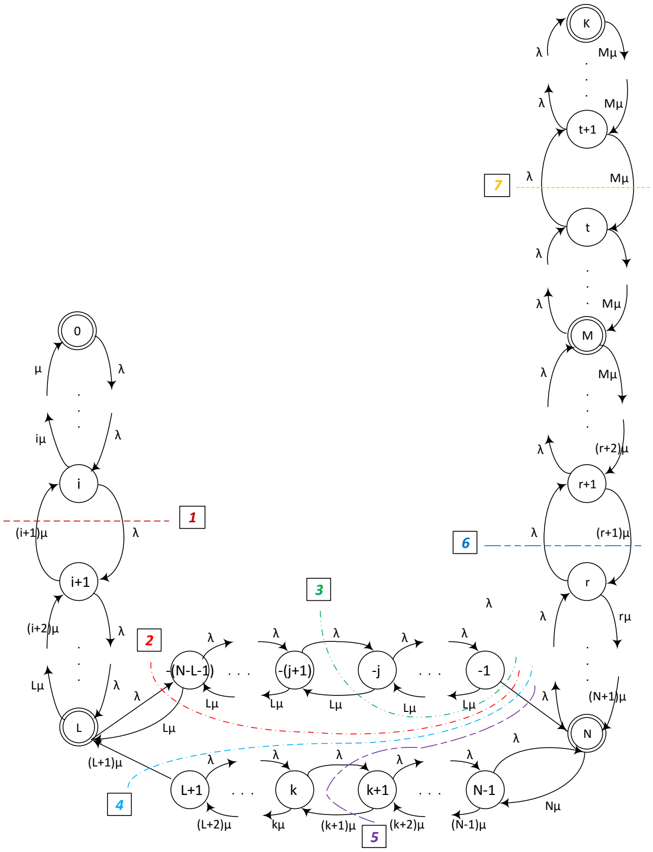

Figure 1. State diagram.

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 are used. On this way the states can be described as a contiguous range $[-(N-L-1), \ldots, K]$.

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 we get the stationary distribution of the number of requests as

$$
\begin{aligned}
p_k &= \frac{(\frac{\lambda}{\mu})^k}{k!} p_0, \text{ for } k = 0, \ldots, L, \\
p_k &= (\frac{\lambda}{L\mu})^{N-L} \frac{(\frac{\lambda}{L\mu})^k - 1}{1 - (\frac{\lambda}{L\mu})^{N-L}} p_L, \\
&\quad \text{for } k = -(N-L-1), \ldots, -1, \\
p_k &= \sum_{i=L}^{k-1} \frac{i!}{k!} (\frac{\lambda}{\mu})^{k-i} p_{-1}, \text{ for } k = L+1, \ldots, N, \\
p_k &= \frac{N!}{k!} (\frac{\lambda}{\mu})^{k-N} p_N, \text{ for } k = N+1, \ldots, M, \\
p_k &= (\frac{\lambda}{M\mu})^{k-M} p_M, \text{ for } k = M+1, \ldots, K. \qquad (3)
\end{aligned}
$$

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

$$
\begin{aligned}
P_L &= \frac{(\frac{\lambda}{\mu})^L}{L!} p_0, \\
p_{-1} &= \alpha \; p_L, \text{ where } \alpha = (\frac{\lambda}{L\mu})^{N-L-1} \frac{1 - \frac{\lambda}{L\mu}}{1 - (\frac{\lambda}{L\mu})^{N-L}}, \\
p_N &= \sum_{i=L}^{N-1} \frac{i!}{N!} (\frac{\lambda}{\mu})^{N-i} p_{-1} = \frac{(\frac{\lambda}{\mu})^N}{N!} s_{L,N} \; \alpha \; p_L, \\
&\quad \text{where } s_{L,N} = \sum_{i=L}^{N-1} \frac{i!}{(\frac{\lambda}{\mu})^i}, \\
p_M &= \frac{N!}{M!} (\frac{\lambda}{\mu})^{M-N} p_N, \qquad (4)
\end{aligned}
$$

3) Performance measures: The performance measures p_{loss}, $p_{s1} = P\{$ the number of active VMs = L $\}$ and $E[W]$ influence the cloud cost. They are given by

ISBN: 978-1-61208-860-0

$$p_{loss} = p_K = (\frac{\lambda}{M\mu})^{K-M} p_M = (\frac{\lambda}{M\mu})^K \frac{M^M}{M!} \frac{N!}{(\frac{\lambda}{\mu})^N} p_N \quad (5)$$

$$\begin{aligned} p_{s1} &= \sum_{k=0}^{L} p_k + \sum_{k=-(N-L-1)}^{-1} p_k \qquad (6) \\ &= \sum_{k=0}^{L} \frac{(\frac{\lambda}{\mu})^k}{k!} p_0 + \sum_{k=-(N-L-1)}^{-1} (\frac{\lambda}{L\mu})^{N-L} \frac{(\frac{\lambda}{L\mu})^k - 1}{1-(\frac{\lambda}{L\mu})^{N-L}} p_L \\ &= \sum_{k=0}^{L} \frac{(\frac{\lambda}{\mu})^k}{k!} p_0 + \sum_{k=1}^{N-L-1} \frac{(\frac{\lambda}{L\mu})^k - (\frac{\lambda}{L\mu})^{N-L}}{1-(\frac{\lambda}{L\mu})^{N-L}} p_L \\ &= \sum_{k=0}^{L} \frac{(\frac{\lambda}{\mu})^k}{k!} p_0 + \frac{\frac{\frac{\lambda}{L\mu} - (\frac{\lambda}{L\mu})^{N-L}}{1-\frac{\lambda}{L\mu}} - (N-L-1)(\frac{\lambda}{L\mu})^{N-L}}{1-(\frac{\lambda}{L\mu})^{N-L}} p_L . \end{aligned}$$

$$\begin{aligned} E[W] = &\sum_{k=-(N-L-1)}^{-1} (k+N-L)p_k + \sum_{k=M}^{K} (k-M)p_k \\ &= \sum_{k=1}^{N-L-1} k \; p_{-(N-L)+k} + \sum_{k=M}^{K} (k-M)p_k \\ &= \tau p_L + \sigma p_M, \qquad (7) \end{aligned}$$

where

$$\begin{aligned} \tau &= \frac{\frac{\lambda}{L\mu}}{(1-(\frac{\lambda}{L\mu})^2} \qquad (8) \\ &- (N-L) \; \frac{(\frac{\lambda}{L\mu})^{N-L}}{1-(\frac{\lambda}{L\mu})^{N-L}} \left(\frac{1}{1-\frac{\lambda}{L\mu}} + \frac{N-L-1}{2} \right), \\ \sigma &= \frac{\lambda}{M\mu} \frac{1-(\frac{\lambda}{M\mu})^{K-M+1}}{(1-\frac{\lambda}{M\mu})^2} - (K-M+1) \frac{(\frac{\lambda}{M\mu})^{K-M+1}}{1-\frac{\lambda}{M\mu}} . \end{aligned}$$

IV. Cost function

A. Constructing the cost function

The cost function, to be optimized, can be constructed by applying the cost model (2) to the shifted N-policy queue. The so far unknown terms arising in (2) can be expressed with the help of parameters, stationary probabilities and performance measures of the shifted N-policy queue as follows.

$$\begin{aligned} &E[\text{ number of active servers }] = L + (1-p_{s1})(M-L), \quad (9) \\ &E[\text{ number of standby servers }] = p_{s1}(M-L), \\ &(\text{ activation rate of standby VMs }) = p_{-1}\lambda, \\ &(\text{ deactivation rate of active VMs }) = p_{L+1}(L+1)\mu. \end{aligned}$$

Substituting the expressions (9) into (2) we get the cost function, F_1 as

$$\begin{aligned} F_1 = &p_{-1}\lambda \; (M-L) \; C_A + p_{L+1}(L+1)\mu \; (M-L) \; C_D \\ &+ (L + (1-p_{s1})(M-L)) \; C_{on} + p_{s1}(M-L) \; C_{off} \\ &+ E[W] \; C_W + p_{loss} \; \lambda \; C_R . \qquad (10) \end{aligned}$$

After performing several rearrangements on (10) and using the balance equation $(L+1)\mu p_{L+1} = \lambda p_{-1}$ as well as (4), (5) and (7) we get the cost function in terms of p_L and p_{s1} as

$$\begin{aligned} F_1 &= ((\lambda(C_A + C_D)(M-L) + \eta \; s_{L,N})\alpha + C_W \tau) \, p_L \\ &\quad - (C_{on} - C_{off})(M-L)p_{s1} + MC_{on}, \text{ where} \qquad (11) \\ \eta &= \left(C_R \lambda (\frac{\lambda}{M\mu})^K \frac{M^M}{M!} + C_W \sigma \frac{(\frac{\lambda}{\mu})^M}{M!} \right) . \end{aligned}$$

B. Approximating the cost function

The optimization of (11) with respect to N 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 p_{s1}. Therefore we establish approximation for (11), which on the other hand restricts the parameter range, for which it holds.

1) Approximations for α, τ and p_{s1}: 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 $1-(\frac{\lambda}{L\mu})^{N-L}$ and $(N-L-1)$ can be approximated by $-(\frac{\lambda}{L\mu})^{N-L}$ and $(N-L)$, respectively. Utilizing it in the expression of α, τ and p_{s1} ((4), (8) and (6)) gives the approximation α^*, τ^* and p^*_{s1}, respectively as

$$\begin{aligned} \alpha^* &\approx 1 - \frac{L\mu}{\lambda}, \\ \tau^* &\approx \frac{\frac{L\mu}{\lambda}}{1-\frac{L\mu}{\lambda}} \left(\frac{1}{1-\frac{L\mu}{\lambda}} - (N-L) \right) + \frac{(N-L)(N-L)}{2}, \\ p^*_{s1} &\approx (N-L)p_L, \qquad (12) \end{aligned}$$

where at evaluating p^*_{s1} we also used 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 $L \gg 1$.

2) Utilizing the approximately N independent regions of p_0: Unfortunately p_0, which is involved in almost every term of (11) via the expression of p_L, depends on N. Now we identify parameter regions, in which p_0 is approximately independent of N. This leads to further restriction on the parameter range. By defining the probability sums

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

p_0 can be given by $p_0 = \frac{1}{p_{sw}}$ with $p_{sw} = p_{s1w} + p_{s2w} + p_{s3w} + p_{s4w}$. It can be seen by taking the difference of p_{s2w}

and p_{s3w} with respect to N that the sum $p_{s2w}+p_{s3w}$ is approximately independent of N and equals to $\sum_{i=L+1}^{M}\frac{(\frac{\lambda}{\mu})^k}{k!}\ \alpha^*$. Furthermore it can be seen that the magnitude of $p_{s2w}+p_{s3w}$ increases rapidly with ρ and for $M/L \gtrapprox 2$ with $\rho \gtrapprox 1.2\frac{L}{M}$ it is much higher than the one of p_{s1w}, which depends on N approximately linearly due to $p_{s1w} \approx (N-L)\frac{p_L}{p_0}$. Moreover the term p_{s4w} is independent of N in this parameter range. We omit the details here due to the limitation on the size of the paper. Summarizing all the above, if $M/L \gtrapprox 2$ and $\rho \gtrapprox 1.2\frac{L}{M}$ then p_{sw} and therefore also p_0 is approximately independent of N. For this case the minimizing task reduces to find the minimum of the function F_2, which can be obtained from (11) by omitting the N independent term MC_{on} and dividing it by p_L. This results in

$$\begin{aligned} F_2 = {} & ((\lambda(C_A+C_D)(M-L)+\eta\ s_{L,N})\alpha + C_W\tau) \\ & - (C_{on}-C_{off})(M-L)\frac{p_{s1}}{p_L}. \end{aligned} \tag{13}$$

3) Applying the approximations for α, τ and p_{s1}: The minimizing task can be further reduced to find the minimum of the objective function F_{2app}, which can be obtained by applying the approximations (12) in (13). This leads to

$$\begin{aligned} F_{2app} = {} & (\lambda(C_A+C_D)(M-L)+\eta\ s_{L,N})\,(1-\frac{L\mu}{\lambda}) \\ & + C_W\frac{\frac{L\mu}{\lambda}}{1-\frac{L\mu}{\lambda}}\left(\frac{1}{1-\frac{L\mu}{\lambda}}-(N-L)\right) \\ & + C_W\frac{(N-L)(N-L)}{2} \\ & - (C_{on}-C_{off})(M-L)(N-L). \end{aligned} \tag{14}$$

C. Approximate equation for determining the local minimum

We obtain an approximate equation for determining the local minimum of (13) by taking its difference with respect to N and setting $\Delta_N F_{2app} \approx 0$. Using $\Delta_N s_{L,N} = \frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}}$ and $\Delta(N-L)(N-L) \approx 2(N-L)$ this leads to the equation

$$\begin{aligned} \eta(1-\frac{L\mu}{\lambda})\frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}} = {} & (C_{on}-C_{off})(M-L) \\ & + C_W\frac{\frac{L\mu}{\lambda}}{1-\frac{L\mu}{\lambda}} - C_W(N-L). \end{aligned} \tag{15}$$

V. Approximate minimization of the cost function

In order to get closer to the solution of equation (15) first we investigate its structure.

A. Structure of the equation

To identify the structure of equation (15), 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

$$\begin{aligned} \sigma = {} & \frac{\lambda}{M\mu}\frac{1-(\frac{\lambda}{M\mu})^{K-M+1}}{(1-\frac{\lambda}{M\mu})^2} - (K-M+1)\frac{(\frac{\lambda}{M\mu})^{K-M+1}}{1-\frac{\lambda}{M\mu})} \\ \approx {} & \frac{\rho}{(1-\rho)^2}. \end{aligned} \tag{16}$$

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

$$\begin{aligned} \eta = {} & \left(C_R\lambda(\frac{\lambda}{M\mu})^{K-M}\frac{(\frac{\lambda}{\mu})^M}{M!} + C_W\sigma\frac{(\frac{\lambda}{\mu})^M}{M!}\right) \\ \approx {} & C_W\frac{\rho}{(1-\rho)^2}\frac{(\frac{\lambda}{\mu})^M}{M!}. \end{aligned} \tag{17}$$

Using (17) in the equation (15) and further rearrangement gives the simplified form of the equation as

$$\frac{(\frac{\lambda}{\mu})^M}{M!}\frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}}u_0(\rho) = r(\rho,N), \text{ where} \tag{18}$$

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

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

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

The term $\frac{(\frac{\lambda}{\mu})^M}{M!}\frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}}$ on the left hand side (lhs) of (18) 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

$$p(\rho,N) = \frac{(\frac{\lambda}{\mu})^M}{M!}\frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}}, \tag{19}$$

the equation (18) can be given in a short form as

$$p(\rho,N)u_0(\rho) = r(\rho,N). \tag{20}$$

B. 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 (19) gives an approximation for $p(\rho,N)$ as

$$\begin{aligned} p(\rho,N) = {} & \frac{(\frac{\lambda}{\mu})^M}{M!}\frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}} = (\frac{\lambda}{\mu M})^{(M-N+1)}\frac{M^M}{M!}\frac{(N-1)!}{M^{N-1}} \\ \approx {} & \rho^{(M-N+1)}e^{(M-N+1)}\sqrt{\frac{N-1}{M}}(\frac{N-1}{M})^{N-1}. \end{aligned} \tag{21}$$

It can be seen from (21) that the dependency of $p(\rho, N)$ on ρ is exponential. 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 (21) we get

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

By introducing the notation

$$\beta = \frac{N-1}{M}. \tag{22}$$

this can be rewritten as

$$\begin{aligned} &\ln(p(\rho, \beta)) = \\ &M\left((1-\beta)(\ln(\rho)+1) + (\beta + \frac{1}{2*M})\ln(\beta)\right). \end{aligned} \tag{23}$$

Taking its first derivative with respect to β gives

$$\frac{d\ln(p(\rho,\beta))}{d\beta} = M\left(\ln(\frac{\beta}{\rho}) + \frac{1}{2*M*\beta}\right) \approx M\ln(\frac{\beta}{\rho}), \tag{24}$$

since in the typical model parameter ranges $M >> 100$ and thus the term $\frac{1}{2*M*\beta}$ can be neglected. The first derivative of $(p(\rho, N)$ with respect to N comes by using $\frac{d(p(\rho,N))}{dN} = \frac{d(e^{\ln(p(\rho,N))})}{dN} = p(\rho,N)\frac{d\ln(p(\rho,\beta))}{d\beta}\frac{d\beta}{dN} = p(\rho,N)\frac{1}{M} * \frac{d\ln(p(\rho,\beta))}{d\beta}$, which yields

$$\frac{d(p(\rho,N))}{dN} \approx p(\rho,N)\ln(\frac{\beta}{\rho}). \tag{25}$$

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

$$p(\rho, N) \text{ is } \left\{ \begin{array}{l} \text{monotone decreasing, if } \beta < \rho \\ \text{monotone increasing, if } \beta \geq \rho \end{array} \right\}. \tag{26}$$

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

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

$$M\left((1-\beta)\;(\ln(\rho)+1) + (\beta + \frac{1}{2*M})\;ln(\beta)\right) = const. \tag{27}$$

Observe that this equation implicitly defines a boundary function $\beta(\rho)$, 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 $p(\rho, N)$ 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 (27) we get the expression of $\ln(\rho)$ along the boundary function as

$$\begin{aligned} ln(\rho) = \frac{const}{(1-\beta)*M} - \frac{\beta}{1-\beta}\ln(\beta) - 1 \\ - \frac{1}{(1-\beta)*2*M}\ln(\beta). \end{aligned} \tag{28}$$

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

$$\zeta = \frac{1}{(1-\beta)*M}. \tag{29}$$

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 (25)) along the boundary function can be obtained as

$$\begin{aligned} \ln(\frac{\beta}{\rho}) = \ln(\beta) - \ln(\rho) = \ln(\beta) + \frac{\beta}{1-\beta}\ln(\beta) + 1 \\ - \left(\frac{const}{(1-\beta)*M} - \frac{1}{(1-\beta)*2*M}\ln(\beta)\right) \\ \leq \frac{1}{1-\beta}\ln(\beta) + 1 \leq -\frac{1}{2}(1-\beta) < 0. \end{aligned} \tag{30}$$

where we used the inequality $\ln(\beta) \leq -(1-\beta) - \frac{1}{2}(1-\beta)^2$ and that the term in the brackets is non-negative. Hence 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 relevant region of the $\beta - \rho$ plane is restricted by $\beta > \beta_{low} = \frac{L}{M}$ and $\rho \geq \beta_{low}$ due to the limitations $N > L \Leftrightarrow \frac{N}{M} > \frac{L}{M}$ and $\frac{\lambda}{\mu} > L \Leftrightarrow \rho > \frac{L}{M}$, respectively. The cross point of the horizontal $\beta = \beta_{low}$ and the boundary curve is called boundary ρ and denoted by ρ_b. All these are shown on the illustrating example Figure 2.

C. Constructing the approximate minimization

1) Solution regimes: For the sake of better understanding the idea of the solution, first we consider a modified form of the equation (20) as

$$p(\rho, N) = r(\rho, N). \tag{31}$$

The idea of the approximate solution is based on the concept of "low magnitude range". When setting the r.h.s of (31) to 0 and the solution of $r(\rho, N) = 0$, let us say N_s, falls inside of the "low magnitude range" with $const = \ln(C_W)$, then it ensures that the value of $r(\rho, N)$ reaches the value of $p(\rho, N) \leq e^{const} = C_W$ by decreasing N not more than 1, since $\frac{d(r(\rho,N))}{dN} = -C_W$ and both the value of $p(\rho, N)$ and its first derivative are $<< C_W$ in a large portion of the "low magnitude range" (up to close to its boundary). Therefore, N_s can be considered as approximate solution of (31).

More precise specification of the inside area of the needed boundary requires both $p(\rho, N) < C_W$ and $\frac{d(p(\rho,N))}{dN} \approx$

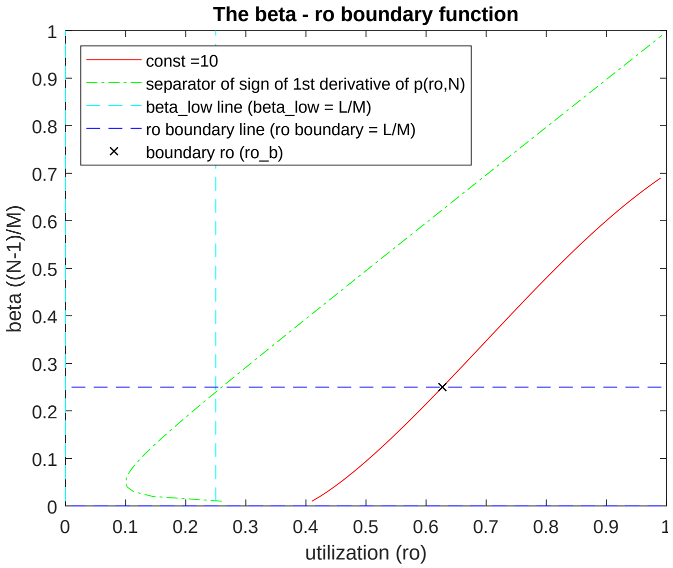

Figure 2. The $\beta - \rho$ boundary function, const = 10, M=200, L=50.

$p(\rho, N)\ln(\frac{\beta}{\rho}) > -C_W$ to be hold. However, the second condition leads to curve on $\beta - \rho$ plane very close to the boundary curve of the ”low magnitude range” with $const = \ln(C_W)$. This is because

$$M\left((1-\beta)\ (\ln(\rho)+1) + (\beta + \frac{1}{2*M})\ ln(\beta)\right) + \ln(-\ln(\frac{\beta}{\rho})) = \ln(C_W) \tag{32}$$

leads to a change in $\ln(\rho)$ in absolute value as $\frac{\ln(-\ln(\frac{\beta}{\rho}))}{(1-\beta)*M}$, which is very small under the most relevant range of parameters. For example it is ≤ 0.04 in absolute value for $M \geq 100$ and $-2.3 \leq \ln(\frac{\beta}{\rho}) \leq -0.35$ due to $0.1 \leq \frac{\beta}{\rho}$ for $\beta_{low} \geq 0.1$ as well as using $\beta \leq 0.7$, which can be shown from the properties of this second $\beta - \rho$ curve. Therefore, the second curve can be neglected from the specification of the required inside area and hence it is enough to specify the needed boundary by $p(\rho, N) = const$ for any $const$.

We denote the boundary ρ under the specific condition $const = \ln(C_W)$ by ρ_0. Approximately at $N = L$, the first derivative of $p(\rho_0, N)$ equals to $-C_W$. At this point $r(\rho_0, L) > p(\rho_0, L)$. By decreasing N, from that point the first derivative of $p(\rho_0, N)$ is in absolute value greater than that one of $r(\rho_0, N)$, and hence an other cross point of the functions $p(\rho_0, N)$ and $r(\rho_0, N)$ must arise, let us say at $N = N_1$. This is a maximum point of the cost function, since (in N) below this point the sign of $p(\rho, N) - r(\rho, N)$ changes from negative to positive. Further decreasing N it reaches the point $N = N_2$, where the value of the cost function is less then at N_s. The situation is illustrated on Figure 3.

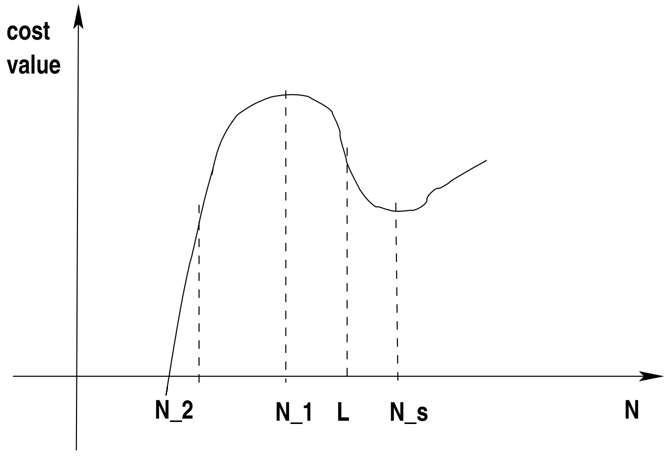

Figure 3. Example cost function.

The above discussed decrease in any range of N, in which $p(\rho_0, N)$ is monotone decreasing with respect to N, causes an increase in the value of $p(\rho_0, N)$, which equivalently can be also considered as a change in $const$ of (27) while keeping N unchanged. This change in $const$ corresponds to a shift of the boundary curve to right. If $\rho > \rho_0$ then the point in β corresponding to N_1 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 at β_{low} (corresponding to N=L+1). If $\rho > \rho_0$, it can also happen that β_s falls outside of the ”low magnitude range” (= under the boundary curve). In this case $|\frac{d(p(\rho,N))}{dN}| > C_W$ and there is no cross point at all, the cost function is monotone increasing with respect to N and hence the global minimum at 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 $\rho < \rho_0$ and $\beta_{low} \leq \beta_s < 1$. Above ρ_0 there is a gap in ρ until a specific point, ρ_s, at which N_2 reaches the β_{low} line and hence the global minimum of the cost function is still at N_s (for $\beta_{low} \leq \beta_s < 1$). 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 is the change in $const$ causing a shift of the boundary ρ from ρ_0 to ρ_s.

2) The magnitude of $\Delta const$*:* The solution of $r(\rho, N) = 0$, N_s can be given from (18)) as

$$N_s = A(M - L) + \frac{1}{\rho\frac{M}{L} - 1} + L. \tag{33}$$

We use the notation

$$\Delta N = N_s - L = A(M - L) + \frac{1}{\rho\frac{M}{L} - 1}. \tag{34}$$

The magnitude of $\Delta const$ is about $2\ln(\Delta N)$. The first $\ln(\Delta N)$ stands for the increase $p(\rho_0, L) \rightarrow p(\rho_0, N_1)$, i.e., from C_w up to $(N_s - N_1)C_W \approx (N_s - L)C_W =$

ΔNC_w (= the value of $r(\rho_0, N)$ at β_{low}), on ln level which is $\ln(\frac{\Delta NC_w}{C_w})$. The second one stands for increase $p(\rho_0, N_1) \rightarrow p(\rho_0, N_2)$, on ln level. During $N_1 \rightarrow N_2$ the cost function F_{2app} decreases so much as its increases during $N_s \rightarrow N_1$, which is approximately $(N_s - N_1) \times |$ maximum value of $\frac{d\ F_{2app}}{dN}$ in $[N_s, N_1]| = (N_s - N_1)|p(\rho_0, L) - r(\rho_0, L)| \approx (N_s - N_1)|C_w - (N_s - L)C_w| \approx (N_s - N_1)\Delta NC_w$. On the other hand the change of the cost function F_{2app} during $N_1 \rightarrow N_2$ is in the magnitude of $p(\rho_0, N_2) - p(\rho_0, N_1)$ (again due to the exponential character of function $p(\rho_0, N)$, but we omit the details here due to the limitation on the size of the paper). Putting all these together $\ln \frac{p(\rho_0,N_2)}{p(\rho_0,N_1)} = \ln(\frac{(N_s-N_1)\Delta NC_w}{(N_s-N_1)C_w} + 1) \approx \ln(\Delta N)$. Note that $(N_s - N_1) \times |$ maximum value of $\frac{d\ F_{2app}}{dN}$ in $[N_s, N_1]|$ overestimates the increase of the cost function F_{2app} during $N_s \rightarrow N_1$ and hence $2\ln(\Delta N)$ also overestimates $\Delta const$.

In order to estimate $2\ln(\Delta N)$, we impose a condition on A, which ensures that the term $A(M-L)$ dominates over $\frac{1}{\rho\frac{M}{L}-1}$. For this purpose an upper bound is set on $\frac{1}{\rho\frac{M}{L}-1}$, which can be obtained by setting a lower bound for ρ as $\beta_{low}\xi < \rho < 1$. With $\xi = 1.2$ this gives $\frac{1}{\rho\frac{M}{L}-1} \leq 5$. We set $A(M-L)/($A (M-L)$+\frac{1}{\rho\frac{M}{L}-1}) \geq 0.9$, which causes a difference of $2\ln(0.9) = -0.2$ in the value of $\Delta const$ corresponding to difference of $\frac{-0.2}{(1-0.5)100} = 0.004$ on $\ln(\rho)$ level when assuming $M \geq 100$ and $\beta_{low} < 0.5$. With this setting we get $A(M-L) \geq 45$ which implies the condition on A as

$$A \geq \frac{45}{M-L}, \tag{35}$$

under which $A(M-L) + \frac{1}{\rho\frac{M}{L}-1} \approx A(M-L)$.

Now we can estimate $2\ln(\Delta N)$ as

$$\begin{aligned} 2\ln(\Delta N) &\approx \ln(A(M-L)) \\ &= 2\ln(A) + \ln(M) + \ln(1-\beta). \end{aligned} \tag{36}$$

3) Relation for ρ_s: So far we discussed the way of solution without considering the term $u_0(\rho)$ on the lhs of equation (20). Now taking into account also the term $u_0(\rho)$, the relation for the boundary curve crossing the β_{low} line at ρ_s can be given by

$$\begin{aligned} &M\left((1-\beta_{low})\ (\ln(\rho_s)+1) + (\beta_{low} + \frac{1}{2*M})\ ln(\beta_{low})\right) \\ &+ \ln(u_0(\rho_s)) = \ln(C_W) + 2\ln(\Delta N). \end{aligned} \tag{37}$$

By substituting the expression of $u_0(\rho)$ from (18) 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

$$\begin{aligned} &M\left((1-\beta_{low})\ (\ln(\rho_s)+1) + (\beta_{low} + \frac{1}{2*M})\ ln(\beta_{low})\right) \\ &+ \ln(C_W) + \ln(\rho_s) + \ln(\frac{1}{(1-\rho_s)^2}) + \ln(\beta_{low}) - \ln(\rho_s) \\ &+ \ln(\frac{\rho_s}{\beta_{low}} - 1) = \ln(C_W) + 2\ln(\Delta N). \end{aligned}$$

Rearranging yields

$$\begin{aligned} &M\left((1-\beta_{low})\ (\ln(\rho_s)+1) + (\beta_{low} + \frac{1}{2*M})\ ln(\beta_{low})\right) \\ &= 2\ln(\Delta N) - \ln(\beta_{low}) - \ln(\frac{1}{(1-\rho_s)^2}) - \ln(\frac{\rho_s}{\beta_{low}} - 1). \end{aligned} \tag{38}$$

We approximate the term $\frac{1}{(1-\rho_s)^2}(\frac{\rho_s}{\beta_{low}} - 1)$ by setting 1, which gives an uncertainty of ≈ 7 on right hand side (rhs) of (38) ($1 \leq \frac{1}{(1-\rho_s)^2} \leq 100$ for $\rho_s \leq 0.9$ and $0.28 \leq (\frac{\rho_s}{\beta_{low}} - 1) \leq 9$ for $\beta_{low} \geq 0.1$ and $\frac{\rho_s}{\beta_{low}} \geq 1.28$ following from (30) with $\beta_{low} \leq 0.5$ and thus $ln(9 \times 100) < 7$) corresponding to difference of $\frac{7}{(1-0.5)200} \approx 0.07$ on $\ln(\rho_s)$ level when assuming $M \geq 200$ and again $\beta_{low} < 0.5$. The relation $\rho_s \leq 0.9$ can be justified by the approximate solution of (38) for ρ_s by assuming that its rhs ≤ 14 and setting $\beta = \max(\beta_{low}) = 0.5$, since the solution in ρ_s is monotone increasing with respect to β_{low}. Using the above approximation and (36) we get the final form of the relation for ρ_s as

$$\begin{aligned} &M\left((1-\beta_{low})\ (\ln(\rho_s)+1) + (\beta_{low} + \frac{1}{2*M})\ ln(\beta_{low})\right) \\ &= 2\ln(A) + \ln(M) + \ln(1-\beta_{low}) - \ln\beta_{low}, \end{aligned} \tag{39}$$

D. Approximate solution formula

Now putting all together 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) $\rho \geq \beta_{low}\xi$ with $\xi = 1.2$,
4) $N - L >> 1$, practically $N > L + 10$,
5) $K - M >> 1$, practically $K > M + 10$,
6) $A \geq \frac{45}{M-L}$

Solution formula

If **Conditions** 1-6 hold, then

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

where

$$\begin{aligned} ln(\rho_s) &= \frac{2\ln(A) + \ln(M) + \ln(1-\beta_{low}) - \ln\beta_{low}}{(1-\beta_{low})*M} \\ &- \frac{\beta_{low}}{1-\beta_{low}}\ln(\beta_{low}) - 1 \\ &- \frac{1}{(1-\beta_{low})*2*M}\ln(\beta_{low}). \end{aligned} \tag{40}$$

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 $K - M$.

VI. Numerical comparisons

In this Section, we illustrate the approximations and validate the approximate solution formula by numeric optimization. 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_{off}, C_{on} have impact to the solution formula only via the parameter A, which was varied through C_W. The parameters C_a, C_d and C_R have no impact on the approximate solution formula in the considered range of (other) parameters. We applied $100 < M < 1000$ for all experiments.

A. Illustrating the approximations

1) N independent region of p_0: Figure 4 shows the dependency of p_0 for the parameter setting $M = 300$, $L = 100$, $K = 350$ and $\rho = 0.6$. It can be seen on the figure that p_0 is independent of N for $N \gtrapprox 120$, which corresponds to $N - L \approx 20 \gg 1$.

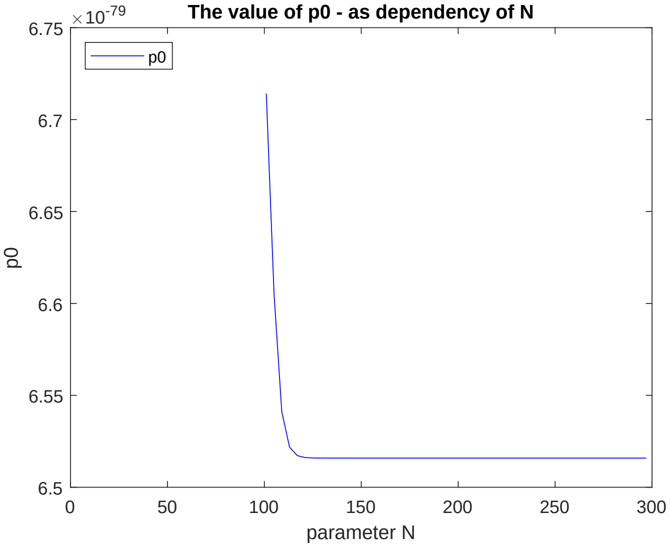

Figure 4. Probability p_0 in dependency of threshold N.

2) Approximation of F_2 by F_{2app}: Figure 5 illustrates the approximation of the cost function F_2 (without taking into account p_0) by F_{2app} in dependency of threshold N for the parameter setting $M = 300$, $L = 100$, $K = 350$, $C_W = 50$, $\mu = 1$ 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.

B. Illustration of the approximate solution formula

The comparison of the exact and approximate optimal N of F_2 can be seen in Figure 6 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 7 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. The small bias between the exact and approximated ρ_s in Figure 6 can be explained by the uncertainity introduced by setting $\frac{1}{(1-\rho_s)^2}(\frac{\rho_s}{\beta_{low}} - 1)$ to 1.

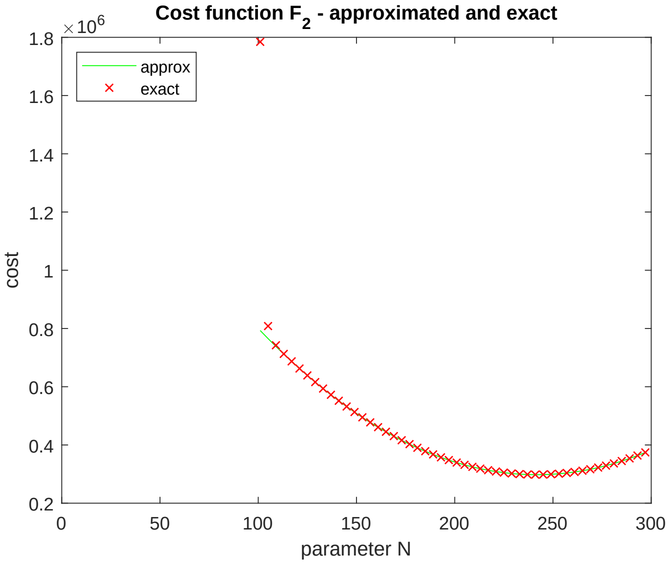

Figure 5. Exact and approximate values of the cost function F_2 in dependency of threshold N.

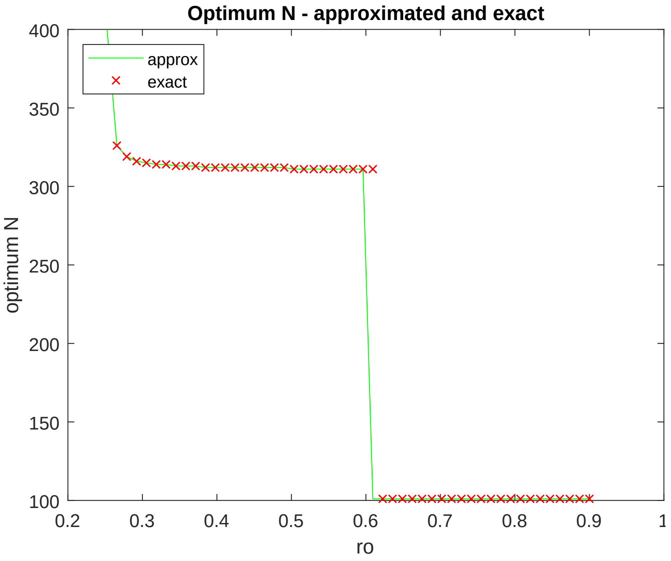

Figure 6. Exact and approximate optimal N (F_2) in dependency of ρ.

C. Validation of the approximate formula

We validated the approximate solution formula by numeric optimization in the considered range of parameters. Figure 8 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 $L = 50$, $K = M + 100$, $C_W = 50$, $\mu = 1$.

Similarly Figure 9 shows the ratio of the approximated and the exact optimal value of F_1 for the range of parameters

ISBN: 978-1-61208-860-0

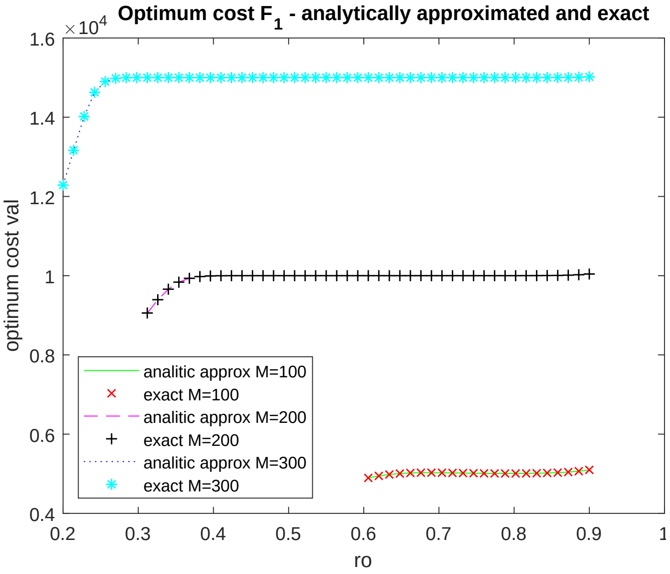

Figure 7. Exact and approximate optimal value (F_1) in dependency of ρ for different values of M.

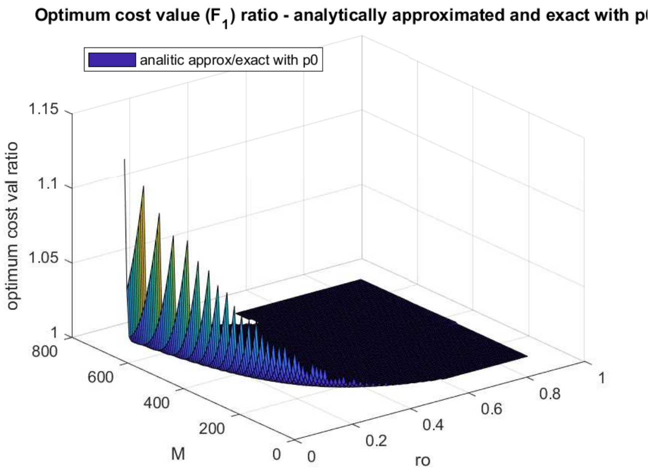

Figure 8. Ratio of the approximated and exact optimal value (F_1) for $100 \leq M \leq 700$ and $\frac{L}{M} < \rho$.

$0.1 \leq C_W \leq 100$ 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.

VII. Conclusion

In this paper, we proposed shifted N-policy for a simple, but energy efficient control of number of active VMs in the IaaS cloud. Besides of the stationary analysis of the underlying queueing model, we provided an approximate formula for computing the optimal threshold N, which minimizes the cloud provider's cost, in the most relevant parameter range. The validation of the approximate solution formula by means of numeric optimization shows a good match in the considered parameter range. The closed form approximate solution formula enables a simple management of the cloud and gives an insight into the dependency of the optimal threshold N on the model and cost parameters.

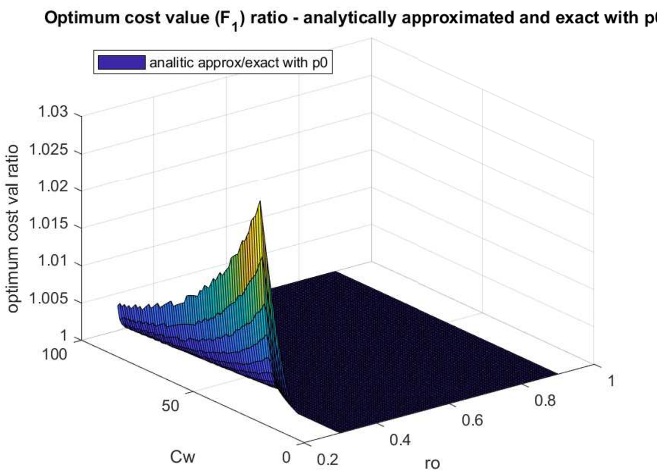

Figure 9. Ratio of the approximated and exact optimal value (F_1) for $0.1 \leq C_W \leq 100$ and $\frac{L}{M} = 0.25 < \rho$.

A future research work is to investigate an approximate solution also for the remaining parameter ranges not considered in this work. Another, more difficult future research topic is the joint optimization of parameters L and N.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

[8] B. Wan, J. Dang, Z. Li, H. Gong, F. Zhang and S. Oh, "Modeling Analysis and Cost-PerformanceRatio Optimization of Virtual Machine Scheduling in CloudComputing," IEEE Transactions on Parallel and DistributedSystems, vol. 31, no. 7, pp. 1518–1532, 2020.

[9] 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.

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