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

Dr. Zsolt Saffer

Institute of Statistics and Mathematical Methods in Economics
Vienna University of Technology (TU Wien), Austria
Email: zsolt.saffer@tuwien.ac.at

May 30, 2021 to June 03, 2021 - AICT 2021 Valencia, Spain

TECHNISCHE UNIVERSITÄT WIEN
Vienna | Austria

IARIA

A short Resume of the Presenter

Dr. Zsolt Saffer is an assistant professor at the Institute of Statistics and Mathematical Methods in Economics of the Vienna University of Technology (TU Wien), Austria. He is lecturing in the areas of queueing theory and stochastic processes. He has more than 12 years R&D experience on machine learning and statistical methods. He holds a PhD in computer science from Budapest University of Technology and Economics (BUTE). His current research interests include queueing theory, performance evaluation of modern telecommunication networks, connected vehicles and optimization.

Outline

Cloud: from performance evaluation to cost optimization

Why do we need performance evaluation of Cloud ?

- Cloud provider needs control over the performance of Cloud in order to
  - get insights into the relationships among the used resources and the performance and
  - meet the performance requirements of the user (they want later probably also Service Level Agreements (SLAs) on Cloud performance).
- It provides the fundament for the Cloud cost optimization.

Why is Cloud cost optimization important ?

- Optimization enables the Cloud service provider the service provisioning at minimum cost.

Cloud cost optimization requires an energy efficient resource managent/allocation technique.

Research works on Cloud performance modeling

Performance evaluation

- Many research works on performance modeling of Clouds.
- An example is ”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.” proposing a multi-level interacting stochastic sub-models approach leading to numeric computational algorithms.
- An overview of existing research works on performance evaluation: ”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.”

Cloud depends on many factors $\Rightarrow$ Performance models

- are either too simplified to obtain meaningful relationships or
- lead to rather complex numeric solution $\Rightarrow$ no explicit relationships among the used resources and the performance.

Research works on Cloud optimization

Cloud optimization

- Summary of efficient resource management policies: ”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.”
- Efficient resource control mechanism based on hysteresis queue: ” 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.”
- Many other approaches leading to computational solution.

Cloud optimization is even more complex issue $\Rightarrow$ *The vast majority of works on cost optimization propose computational solutions.*

Goal & Approach

Goal: to provide ***an analytic model*** for Infrastructure-as-a-Service (***IaaS***) type ***Cloud service***, which is

- ***tractable*** and
- still ***suitable for practical use***.

Approach - to reach the goal

- IaaS Cloud model with Virtual Machines (VM) as parallel resources.
- IaaS Cloud service is modeled by an analytically tractable $M/M/m/K$ queue.
- A new resource control mechanism, called Shifted N-policy, is proposed, which is suitable to be used in practice due to providing energy efficiency.

$M/M/m$ queue is 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.

Shifted N-policy

Shifted N-policy - operation

- A predefined number of VMs (L) are always active.
- The remaining ones are
  - activated simultaneously when the number of requests reaches a threshold (N) and
  - deactivated when the number of requests falls below L.

Shifted N-policy - Properties

- A one threshold (N) based resource control mechanism.
- It has hysteresis-like characteristic upwards (in number of requests) $\Rightarrow$ suitable to be used for energy efficient resource control.
- It is simpler than the hysteresis queue $\Rightarrow$ it facilitates the developing of analytically tractable approximation.

The Cloud model

1. The IaaS Cloud model
   - The VMs are modeled as parallel resources - $M > 100$ VMs.
   - The Physical Machines (PMs) are grouped into two pools: active (running) and standby - either turned-on (but not ready) or turned-off machines.
2. The IaaS Cloud is modeled by an $M/M/m/K$ queue:
   - Poison arrival with rate $\lambda > 0$ and exponentially distributed service time with parameter $\mu > 0$.
   - Number of servers = $M \geq 1$ (=the total number of VMs).
   - The model has buffer with capacity for $K - M \geq 1$ VMs.
   - $\rho = \frac{\lambda}{M\mu}$ - utilization if $K \to \infty$.
3. Resource allocation is implemented by shifted N-policy
   - $0.1M \leq L \leq 0.5M$ VMs servers are always active,
   - simultaneous activation/deactivation of the remaining $M - L$ servers ($\uparrow$ at reaching the threshold $L + 1 \leq N \leq M$, $\downarrow$ falling below $L + 1$).

Cost model of cloud provider

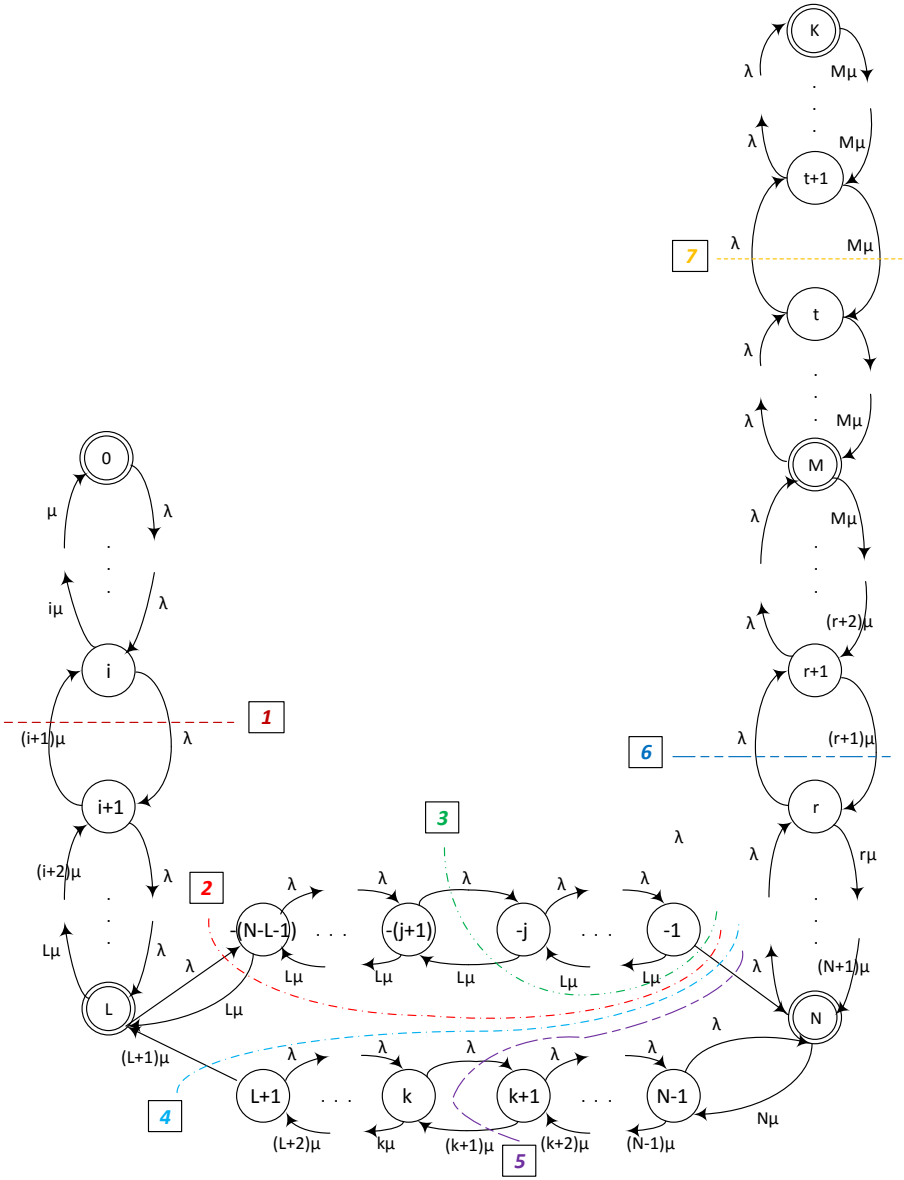
$$\begin{aligned} C_{cloud} &= E[\text{ number of active servers }]\, C_{on} \\ &+ E[\text{ number of standby servers }]\, C_{off} \\ &+ E[W]\, C_W + p_{loss}\, \lambda\, C_R, \\ &+ (\text{ activation rate of standby VMs })\ (M-L)\, C_A \\ &+ (\text{ deactivation rate of active VMs })\ (M-L)\, C_D, \end{aligned} \tag{1}$$

where

- 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,
- C_R - cost of loss of an arriving request,
- C_A - activation cost of a VM (changing from standby to active),
- C_D - deactivation cost of a VM (changing from active to standby).

$E[W]$ - expected waiting time of the request, p_{loss} - the probability of loss.

CTMC model of $M/M/m/K$ queue with shifted N-policy



- Notation: $n \geq 0$ - the number of requests in the system.
- $\Rightarrow$ The process $\{n(t), t \geq 0\}$ is a finite state Continuous-Time Markov chain (CTMC).
- The queue is always stable, since its CTMC has finite number of states.
- Notation of the states - contiguous range $[-(N-L-1), \ldots, -1, 0, \ldots, K]$ (notation of states, for which the $L < n < N$ - depends on the number of active servers).

Global balance equations

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

The selected set of states are marked on the state diagram for each equation

- by a separator line and
- an associated number in small square - to identify the case.

p_i - the stationary probability of state i, for $-(N-L-1) \leq i \leq K$.

Stationary distribution

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

$$
\begin{aligned}
P_L &= \frac{(\frac{\lambda}{\mu})^L}{L!} p_0, \\
p_{-1} &= \alpha\, p_L, \text{ where} \\
\alpha &= (\frac{\lambda}{L\mu})^{N-L-1} \frac{1 - \frac{\lambda}{L\mu}}{1 - (\frac{\lambda}{L\mu})^{N-L}}, \\
p_N &= \sum_{i=L}^{N-1} \frac{i!}{N!} (\frac{\lambda}{\mu})^{N-i} p_{-1} \\
&= \frac{(\frac{\lambda}{\mu})^N}{N!} s_{L,N}\, \alpha\, p_L, \text{ where} \\
s_{L,N} &= \sum_{i=L}^{N-1} \frac{i!}{(\frac{\lambda}{\mu})^i}, \\
p_M &= \frac{N!}{M!} (\frac{\lambda}{\mu})^{M-N} p_N. \qquad (2)
\end{aligned}
$$

Performance measures

$$p_{loss} = p_K = \left(\frac{\lambda}{M\mu}\right)^K \frac{M^M}{M!} \frac{N!}{\left(\frac{\lambda}{\mu}\right)^N} p_N, \tag{3}$$

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

$$E[W] = \sum_{k=-(N-L-1)}^{-1} (k+N-L)p_k + \sum_{k=M}^{K} (k-M)p_k = \tau p_L + \sigma p_M,$$

where

$$\tau = \frac{\frac{\lambda}{L\mu}}{\left(1-\left(\frac{\lambda}{L\mu}\right)^2} - (N-L)\frac{\left(\frac{\lambda}{L\mu}\right)^{N-L}}{1-\left(\frac{\lambda}{L\mu}\right)^{N-L}} \left(\frac{1}{1-\frac{\lambda}{L\mu}} + \frac{N-L-1}{2}\right), \tag{4}$$

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

p_{loss}, $p_{s1} = P\{\text{number of active VMs} = L\}$ and $E[W]$ influence the cloud cost.

Constructing the cost function

So far unknown terms arising in cost model of cloud provider can be expressed as

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

After performing several rearrangements we get the cost function, F_1 in terms of p_L and p_{s1} as

$$
\begin{aligned}
F_1 &= \left((\lambda(C_A + C_D)(M - L) + \eta\, s_{L,N})\alpha + C_W\tau\right) p_L\\
&- (C_{on} - C_{off})(M - L)p_{s1} + MC_{on}, \text{ where} \qquad (5)\\
\eta &= \left(C_R\lambda(\frac{\lambda}{M\mu})^K \frac{M^M}{M!} + C_W\sigma\frac{(\frac{\lambda}{\mu})^M}{M!} \right).
\end{aligned}
$$

Approximate equation for determining the local minimum

The minimization task $\min_N F_1(N)$ with positive integer N can be reduced to an approximate equation by performing the following steps:

1. utilizing the approximately N independent regions of p_0 ($M/L \gtrapprox 2$ and $\rho \gtrapprox 1.2\frac{L}{M}$),
2. applying approximations for α, τ and p_{s1} (when assuming $N - L \gg 1$) to cost function, F_1 (leading to F_{2app}) and
3. taking its difference with respect to N as well as setting $\Delta_N F_{2app} \approx 0$.

This results in the equation

$$\eta(1-\frac{L\mu}{\lambda})\frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}} = (C_{on}-C_{off})(M-L)+C_W\frac{\frac{L\mu}{\lambda}}{1-\frac{L\mu}{\lambda}}-C_W(N-L). \quad (6)$$

The last step: approximation of the discrete optimization task $\min_N F_1(N)$ with positive integer N through its continuous counterpart by assuming real N.

The structure of the equation

Simplifying the equation, by applying further approximations (assuming $K - M >> 1$), we get the equation with the following structure

$$\frac{(\frac{\lambda}{\mu})^M}{M!}\frac{(N-1)!}{(\frac{\lambda}{\mu})^{N-1}}u_0(\rho) = C_W\left(A(M-L) + \frac{1}{\rho\frac{M}{L}-1} - (N-L)\right), \quad (7)$$

where $u_0(\rho) = C_W\frac{\rho}{(1-\rho)^2}(1-\frac{1}{\rho\frac{M}{L}})$ and $A = \frac{C_{on}-C_{off}}{C_W}$.

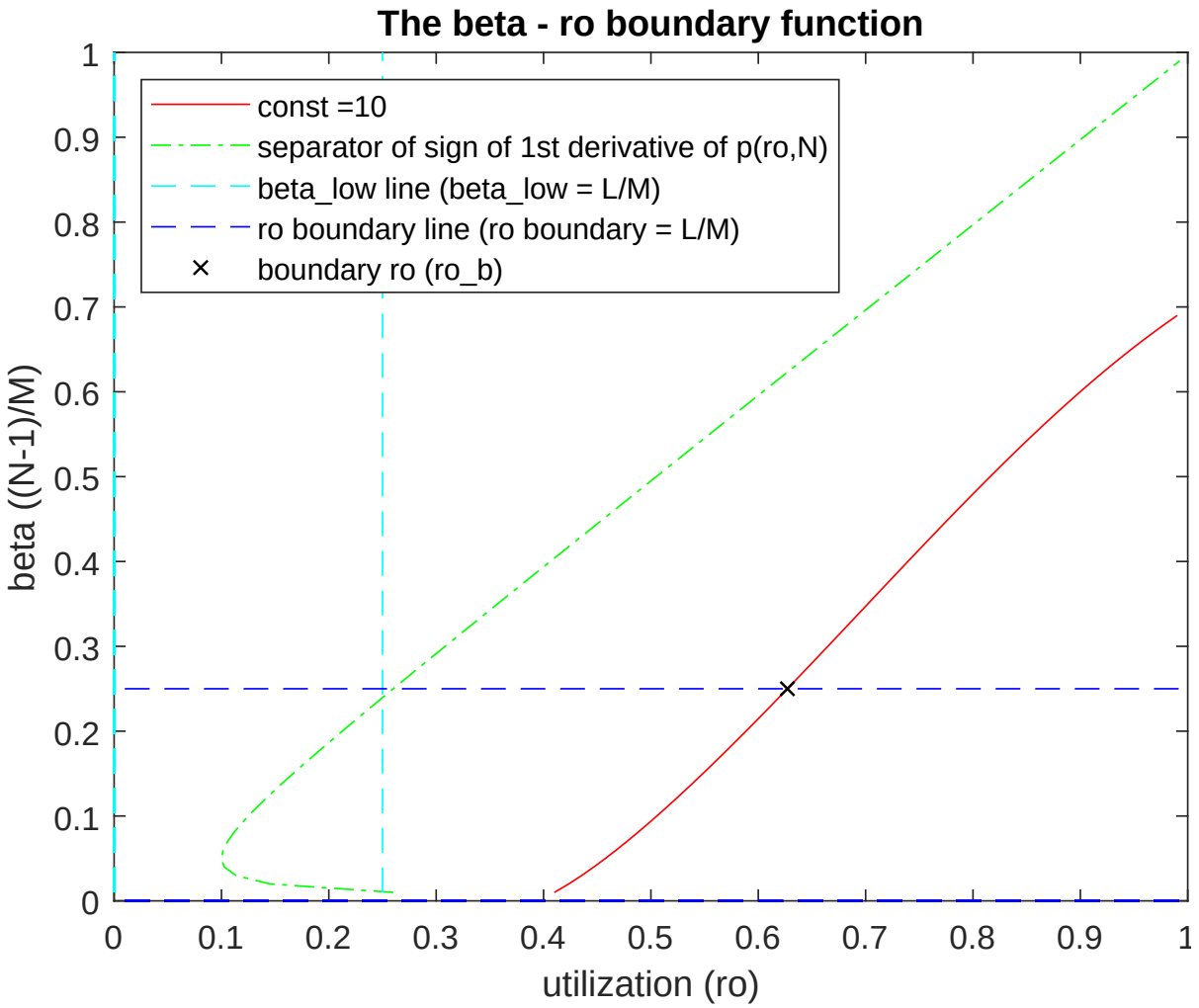
or in short form

$$p(\rho, N)u_0(\rho) = r(\rho, N). \quad (8)$$

Structural characteristics of the equation

- The function $p(\rho, N)$ with factorial terms constitutes the specific form of the equation and the optimization task.
- The right side of the equation, i.e. function $r(\rho, N)$ is linear with N.

The "low magnitude range"



.

Inside of "low magnitude range" with *const*

- = such values of $\beta = \frac{N-1}{M}$, for which $p(\rho, \beta) \leq e^{const}$,
- $\Rightarrow$ Illustration: the region above the red curve.

The idea of the solution

For the sake of better understanding - we consider a simplified form of the equation

$$p(\rho, N) = r(\rho, N). \tag{9}$$

Essence (The idea of the solution)

If the solution of $r(\rho, N) = 0$, N_s falls inside of the "low magnitude range" with $const = \ln(C_W) \Rightarrow$ 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)

$\Rightarrow N_s$ can be considered as approximate solution of the equation.

Approximate solution formula

Conditions

1. $100 \leq M$,
2. $0.1 \leq \beta_{low} \leq 0.5, \quad \beta_{low} = \frac{L}{M}$,
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. $A = \frac{C_{on} - C_{off}}{C_W} \geq \frac{45}{M-L}$.

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

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

$$ln(\rho_s) = \frac{2\ln(A) + \ln(M) + \ln(1 - \beta_{low}) - \ln \beta_{low}}{(1 - \beta_{low}) * M} - \frac{\beta_{low}}{1 - \beta_{low}} \ln(\beta_{low}) - 1 - \frac{1}{(1 - \beta_{low}) * 2 * M} \ln(\beta_{low}). \quad (10)$$

$\Rightarrow$ *The approximate optimal N does not depend on C_A, C_D and C_R, since they have no impact on N in the considered range of parameters.*

Numerical illustration

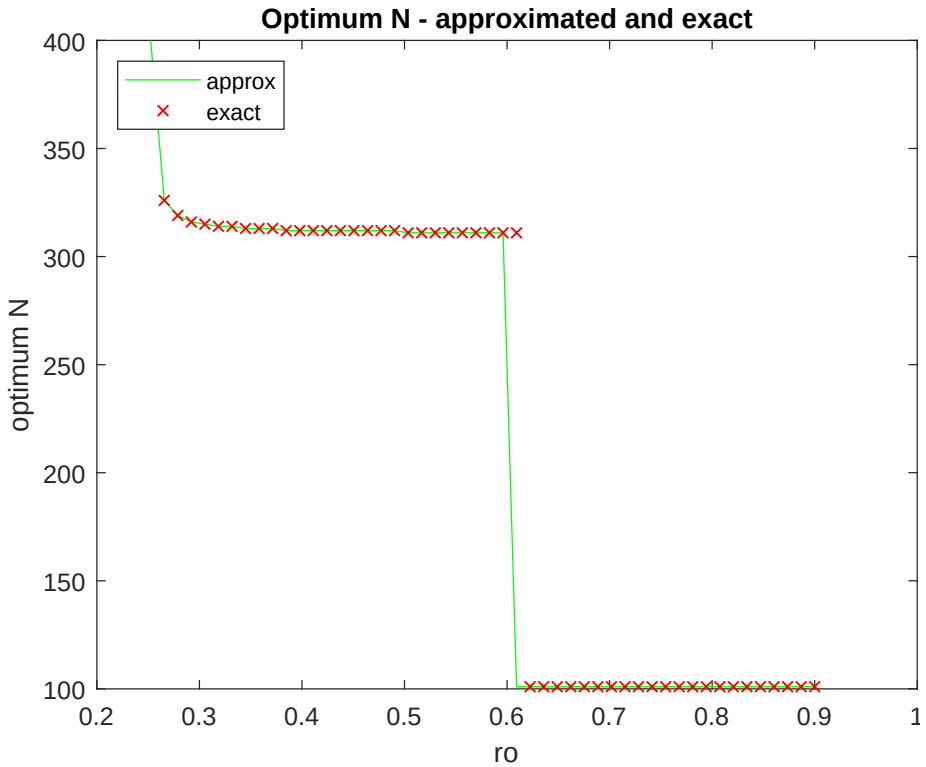


Figure: Exact and approximate optimal N (F_2) in dependency of ρ.

$M = 400$, $L = 100$, $K = 450$, $C_W = 50$, $\mu = 1$ and $\rho > 0.25 = \frac{L}{M}$

For both figures: $C_{on} = 50$, $C_{off} = 15$.

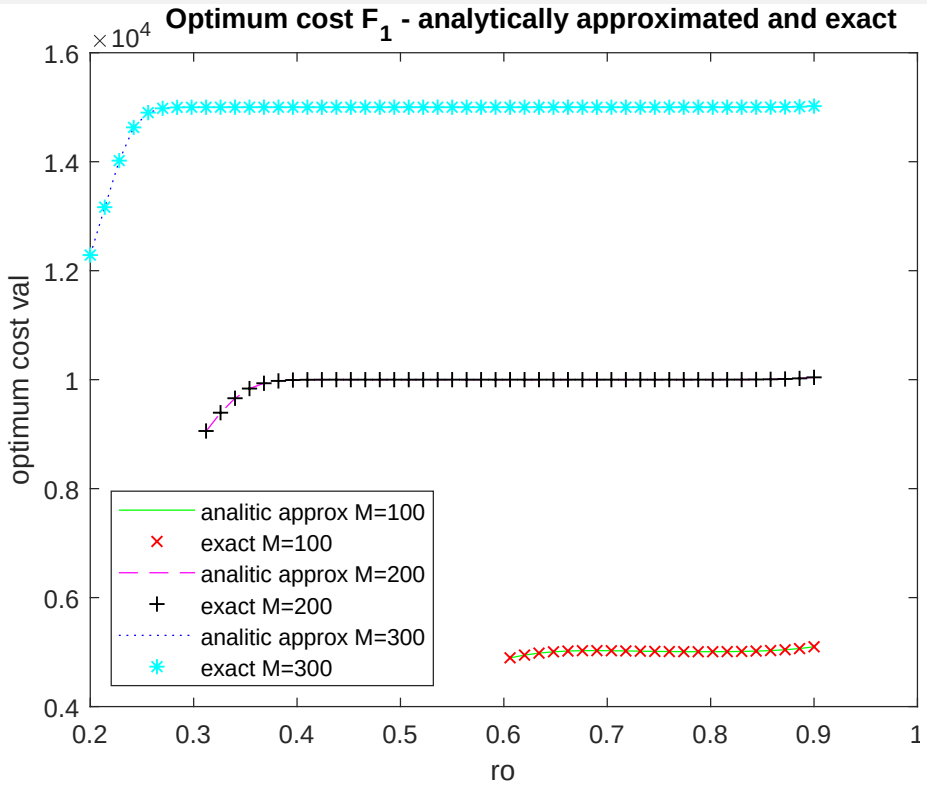


Figure: Exact and approximate optimal value (F_1) in dependency of ρ.

$L = 50$, $K = M + 100$, $C_W = 50$, $\mu = 1$ and $\rho > 0.25 = \frac{L}{M}$

Numerical validation

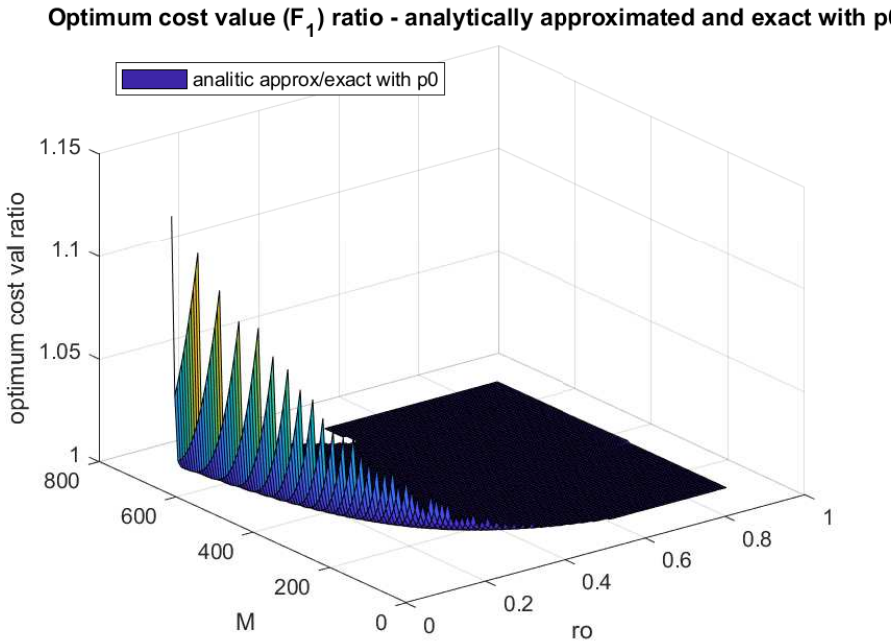


Figure: Ratio of the approximated and exact optimal value (F_1) for $100 \leq M \leq 700$ and $\frac{L}{M} < \rho$.

$L = 50$, $K = M + 100$, $C_W = 50$, $\mu = 1$

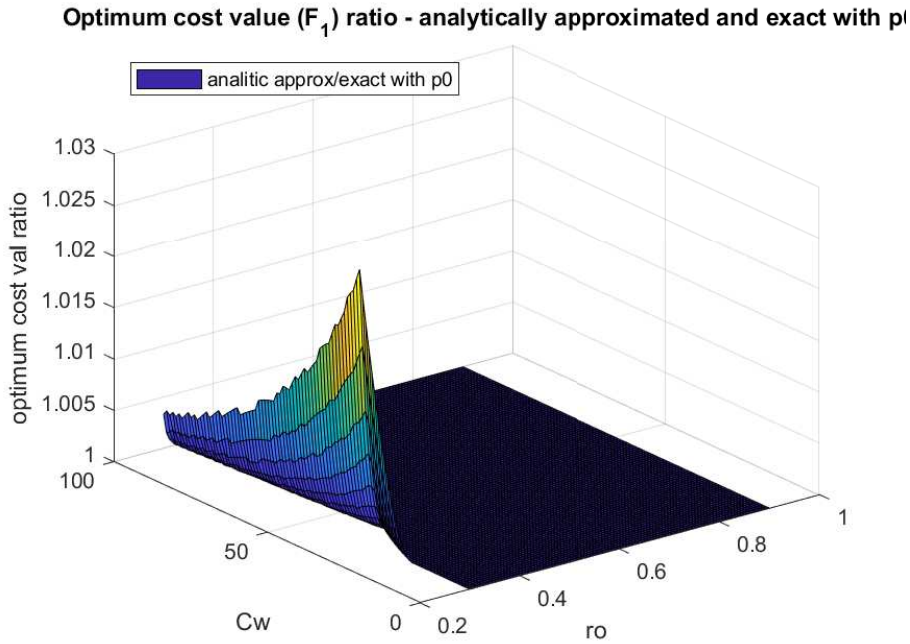


Figure: Ratio of the approximated and exact optimal value (F_1) for $0.1 \leq C_W \leq 100$ and $\frac{L}{M} = 0.25 < \rho$.

$L = 50$, $M = 200$, $K = 300$, $\mu = 1$

For both figures: $C_{on} = 50$, $C_{off} = 15$.

Contribution

1. Primary contribution:
   - the proposed shifted N-policy resource control mechanism and
   - the closed form approximate solution formula for the optimal value of the threshold N - under the most relevant range of parameters.

   $\Rightarrow$ It enables a very simple cloud resource management due to the approximate analytic formula for the optimal threshold.

2. Secondary contribution:
   The stationary analysis of the shifted N-policy $M/M/m/K$ model.

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.

Outlook

Future research topics

- Investigate an approximate solution also for the remaining parameter ranges not considered in this work (mainly $N < L$).
- Consider the joint optimization of parameters L and N (a more difficult issue).