EVALUATION OF EXPECTED OPERATING COSTS PER TIME UNIT FOR A DISCRETE-TIME QUEUING SYSTEM WITH MULTIPLE LEVELS OF EFFICIENCY

F. A. Zhydok¹, A. Yu. Pilipenko²

¹Kyiv Academic University, Kyiv, Ukraine ²Institute of Mathematics of NAS of Ukraine, Kyiv, Ukraine *f.zhydok@kau.edu.ua, pilipenko.ay@gmail.com*

Due to the abundance of computing resource providers, the task of evaluating the operating costs of some task processing system with dynamic resource allocation is quite relevant. In this problem the processing system considered has several levels of processing efficiencies. If the system on some level of efficiency is overflowed by tasks to process, a higher level of efficiency is employed. If the number of tasks reduces significantly, a lower level of efficiency is employed. The higher the level of efficiency becomes, then the tasks are processed faster and the the price for system usage becomes higher. Also, within the framework of this problem, the system operating costs include transition costs between different levels.

The simplified representation of such a processing system can be done using the discrete-time Markov chain with certain state space E and transition matrix P. Assume that such a processing system has R levels of processing efficiencies and $\forall r = 1, ..., R : E_r = \{(r, k), ..., (r, n)\} \subset E$, where E_r is a set of states in the processing level r. From the levels' first states (2, k), ..., (R, k) happens the transition to respective states in the lower level, which are denoted as (1, m), ..., (R - 1, m). From the levels' last states (1, n)..., (R - 1, n) happens the transition to respective states (1, n)..., (R - 1, n) happens the transition to respective states (1, n)..., (R - 1, n) happens the transition to respective states (1, n)..., (R - 1, n) happens the transition to respective states (1, n)..., (R - 1, n) happens the transition to respective states (1, n)..., (R - 1, n) happens the transition to respective states (1, n)..., (R - 1, n) happens the transition to respective states (1, n)..., (R - 1, n) happens the transition to respective states in the upper level, which are denoted as (2, l), ..., (R, l). Denote level efficiencies as $q_r \in (0, 1) \forall r = 1, ..., R$, which are, respectively, probabilities to jump from state (r, j) to (r, j - 1) for j > k and to jump from (r, k) to (r - 1, m) for r = 2, ..., R (for (1, k)) it is the probability to go back to (1, k)). Respective intensity of incoming tasks is denoted as $p_r = 1 - q_r$ corresponds to the probability to jump from (r, j) to (r, j + 1) for j < n and to jump from (r, n) to (r + 1, l) for r = 1, ..., R - 1 (for (R, n) it is the probability to go back to (R, n)). Importantly, it was assumed that $\forall r = 2, ..., R - 1 : l < m$. Markov chain, which corresponds to some arbitrary processing system, is further denoted as Q_n . An example for the processing system with R = 3 is on figure 1.



Figure 1: Example of Markov chain Q_n with R = 3

Also, let x_r be the price for one unit of time using the efficiency of level r, where r = 1, ..., R. And let the price for transition from level r to r + 1 be denoted as y_r , where r = 1, ..., R - 1. The price for respective transitions from r + 1 to r is the same as from r to r + 1. Now the problem can be formulated.

http://www.imath.kiev.ua/~young/youngconf2025

Problem 1. Consider Markov chain Q_n which corresponds to the specific task processing system with R number of levels and corresponding level usage prices $x_1, ..., x_R$ and level transition prices $y_1, ..., y_{R-1}$. Then, find $\lim_{n\to\infty} \frac{M_n}{n}$, where M_n is a cumulative cost for the system usage for n steps.

The most obvious solution to this problem is to find the stationary distribution from the transition matrix P and then use the corollary from the ergodic theorem [1, p. 121] on the function $f: E \times E \to \mathbb{R}$, which would map states to the corresponding level prices and transition prices.

Theorem 1. Let $\{X_n\}_{n\geq 0}$ be an irreducible positive recurrent Markov chain with the stationary distribution π , and let $f: E \times E \to \mathbb{R}$ be such that:

$$\sum_{i \in E} \sum_{j \in E} |f(i,j)| * \pi(i) * p_{ij} < \infty$$

Then for any initial distribution μ , P_{μ} -a.s.:

$$\lim_{N \to \infty} \frac{1}{N} \sum_{k=1}^{N} f(X_k, X_{k+1}) = \sum_{i \in E} \sum_{j \in E} f(i, j) * \pi(i) * p_{ij}$$
(1)

With plausible transition matrix P, the best computational complexity of this solution is $\mathcal{O}(zn^2)$, where n is the number of states in E and z is the number of iterations of transition matrix multiplication. For large systems, such a solution is inefficient.

Instead, a more optimal approach was developed. Looking closely at Q_n , it can be noticed that segments of Markov chains between states $E = \{(1, m), (2, l), (2, m), \dots, (R-1, l), (R-1,$ (1, m), (R, l) closely resemble birth-death Markov chains. Using this fact and recurrent equations, respective transition probabilities between each neighbouring pair of states in E and expected time to reach any of the neighbouring states in \tilde{E} are obtained. After that, the nested Markov chain on E is obtained, which is also a birth-death Markov chain and is denoted as G_n . Then, the function $q: \hat{E} \times \hat{E} \to \mathbb{R}$ can be introduced, which represents the cost of transitions inside the nested Markov chain, which is equivalent to the cost of all the steps that occurred in Q_n , during specific transitions in G_n . Even though the number of steps are random, the specific variation of the corollary from the ergodic theorem can be used, and q(i, j) can be set as a conditional expectation of the number of steps in Q_n from i to j, without getting to the other neighbouring state. After defining g(i, j), theorem 1 is applied to the g(i, j) on G_n and equation (1) denoted as L_q would describe the limit of the sum of costs of Q_n divided by the number of steps in G_n . As the limit needs to be normalised, analogous to the g(i, j), the function $t: \hat{E} \times \hat{E} \to \mathbb{R}$ is introduced and represents the number of steps in Q_n during one step in G_n . After using the same argumentation on t(i, j) as on g(i, j), the limit L_t is obtained. Ratio $\frac{L_g}{L_t}$ represents the expected cost of operating Q_n per unit of time. The time complexity of this method is $\mathcal{O}(R)$, which is better than the complexity of the straightforward approach.

- 1. Bremaud P. Discrete Probability Models and Methods: Probability on Graphs and Trees, Markov Chains and Random Fields, Entropy and Coding. Cham: Springer, 2017, XIV+559 p.
- 2. Chung K.L. Markov Chains: with Stationary Transition Probabilities. Berlin, Heidelberg: Springer-Verlag, 1967, XI+301 p.