Control of queues with MAP servers: experimental results

Levente Bodrog
Politecnico di Milano
bodrog@elet.polimi.it

Marco Gribaudo
Politecnico di Milano
gribaudo@elet.polimi.it

Gábor Horváth
Technical Univ. of Budapest
ghorvath@hit.bme.hu

András Mészáros
Technical Univ. of Budapest
meszarosa@hit.bme.hu

Miklós Telek
Technical Univ. of Budapest
telek@hit.bme.hu

ABSTRACT

The paper considers simple queueing systems with multiple MAP servers, where the incoming customers can be freely assigned with service unit in case of more than one available free servers. In case of a well defined service policy the analysis of such queueing systems is a standard matrix analytic procedure, but the optimal control of those queues is rather complex. We do not directly optimize the service policy, but consider the number of all service policies in case of small models and evaluate their main performance parameter, which is the mean system time (waiting time + service time) in this work. As results we report some surprising optimal policies for small $M/MAP(k)/n$ queues.

Keywords: M/MAP/n queue, control of queue, numerical optimization.

1. INTRODUCTION

Queueing systems with multiple servers allow a degree of freedom to assign incoming customers with one of the idle servers when there is more than one. When the service units have completely memoryless behaviour, then the assignment of the incoming customer with an idle server does not affect the queueing behaviour, but when the servers are not memoryless, the overall queueing performance is subject to optimal server selection.

We focus on the analysis of $M/MAP(k)/n$ queues, where customers arrive to a queuing system according to a Poisson process with rate λ, the queueing system is composed by n server units and an infinite buffer. The service units are identical and their service times are characterized by a Markov arrival process of order k (MAP(k)) [1]. In case of n service units, a customer arriving to the queue with $m < n-1$ customers in the system finds $n-m > 1$ free servers and the system assigns the customer with one of the idle servers freely. In this work we assume that the system knows the phase of the free service units and makes its choice based on that.

Generalization of this model to more complex queues, e.g., to the $MAP/MAP(k)/n$ queue, is straightforward and preserves the strange optimal behaviour as the one obtained for $M/MAP(2)/2$ queues.

2. THE MATRIX ANALYTIC MODEL OF THE M/MAP(K)/N QUEUE

The number of customers in a M/MAP(k)/n queue and the phase of the service processes can be characterized by a continuous time Markov chain and due to the fact that customers arrive and depart one by one this Markov chain has a quasi birth death (QBD) structure [1]. The backward, the local and the forward matrices (denoted by $\mathbf{B}$, $\mathbf{L}$ and $\mathbf{F}$ respectively) of this QBD are level independent in case of more than n customers in the system. Consequently, the QBD can be solved using the matrix geometric stationary behaviour of the level independent part [1]. In the next subsection we specialize this solution for the M/MAP(2)/2 queue.

2.1 The M/MAP(2)/2 queue

The M/MAP(2)/2 model translates to a quasi birth-death (QBD) process, which is the multi-phase extension of the M/M/1 model. The state space is partitioned into so-called levels according to the number of customers in the system, which implies the block tri-diagonal form of the infinitesimal generator

$$\mathbf{Q} = \begin{pmatrix} \mathbf{L}_0 & \mathbf{F}_0 & \mathbf{0} & \cdots & & \\ \mathbf{B}_1 & \mathbf{L}_1 & \mathbf{F}_1 & \mathbf{0} & \cdots & \\ 0 & \mathbf{B}_2 & \mathbf{L} & \mathbf{F} & \mathbf{0} & \\ \vdots & \mathbf{0} & \mathbf{B} & \mathbf{L} & \mathbf{F} & \ddots \\ & & \ddots & \ddots & \ddots & \ddots \end{pmatrix}, \qquad (1)$$

where

$$\mathbf{L}_0 = -\lambda \mathbf{I} \otimes \mathbf{I}, \mathbf{F}_0 = \lambda \left((\mathbf{I} \otimes \mathbf{I}) \mathbf{P}, \quad (\mathbf{I} \otimes \mathbf{I}) (\mathbf{I} - \mathbf{P}) \right),$$

$$\mathbf{B}_1 = \begin{pmatrix} \mathbf{I} \otimes \mathbf{S}_1 \\ \mathbf{S}_1 \otimes \mathbf{I} \end{pmatrix}, \mathbf{L}_1 = \begin{pmatrix} -\lambda \mathbf{I} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{S}_0 & 0 \\ 0 & -\lambda \mathbf{I} \otimes \mathbf{I} + \mathbf{S}_0 \otimes \mathbf{I} \end{pmatrix},$$

$$\mathbf{F}_1 = \lambda \begin{pmatrix} \mathbf{I} \otimes \mathbf{I} \\ \mathbf{I} \otimes \mathbf{I} \end{pmatrix}, \mathbf{B}_2 = \begin{pmatrix} \mathbf{I} \otimes \mathbf{S}_1 \\ \mathbf{S}_1 \otimes \mathbf{I} \end{pmatrix}, \mathbf{L} = -\lambda \mathbf{S}_0 \oplus \mathbf{S}_0,$$

$$\mathbf{F} = \mathbf{A}_1 \otimes \mathbf{I} \otimes \mathbf{I}, \mathbf{B} = \mathbf{I} \otimes \mathbf{S}_1 + \mathbf{S}_1 \otimes \mathbf{I} \, .$$

and

$$\mathbf{P} = \mathrm{diag}\,(1/2, p, 1-p, 1/2)\,. \tag{2}$$

According to matrix $\mathbf{P}$, at a customer arrival to the empty system the customer is directed to the server in phase 1 with probability p and to the server in phase 2 with probability $1-p$ if the servers are in different phases and if idle servers are in the same phase the service units are chosen evenly.

The steady state solution of the system is partitioned according to the levels as

$$\boldsymbol{\pi} = \begin{pmatrix}\boldsymbol{\pi}_0, & \boldsymbol{\pi}_1, & \boldsymbol{\pi}_2, & \ldots\end{pmatrix}. \tag{3}$$

Due to the level independent behaviour of (1) for $i \geq 2$ we have

$$\boldsymbol{\pi}_i = \boldsymbol{\pi}_2 \mathbf{R}^{i-2}, \quad i \geq 2, \tag{4a}$$

Where $\mathbf{R}$ is the minimal non-negative solution of the quadratic matrix equation [1]

$$\mathbf{0} = \mathbf{F} + \mathbf{R}\mathbf{L} + \mathbf{B}\mathbf{R}^2.$$

Based on (1) and matrix $\mathbf{R}$, the irregular part of the steady state distribution is the solution of the linear system

$$\begin{pmatrix}\boldsymbol{\pi}_0 & \boldsymbol{\pi}_1 & \boldsymbol{\pi}_2\end{pmatrix} \begin{pmatrix}\mathbf{L}_0 & \mathbf{F}_0 & 0 \\ \mathbf{B}_1 & \mathbf{L}_1 & \mathbf{F}_1 \\ 0 & \mathbf{B}_2 & \mathbf{L}+\mathbf{R}\mathbf{B}\end{pmatrix} = 0, \tag{4b}$$

with normalization condition

$$1 \quad = \quad \boldsymbol{\pi}_0 \mathbb{1} \quad + \quad \boldsymbol{\pi}_1 \mathbb{1} \quad + \quad \boldsymbol{\pi}_2 \left(\mathbf{I}-\mathbf{R}\right)^{-1} \mathbb{1} \quad . \tag{4c}$$

Using the steady state distribution (4), the mean number of customers in the system can be expressed as

$$\begin{aligned}\mathrm{E}\,(N) &= \sum_{i=0}^{\infty} i\boldsymbol{\pi}_i \mathbb{1} = \boldsymbol{\pi}_1 \mathbb{1} + \sum_{i=2}^{\infty} i\boldsymbol{\pi}_2 \mathbf{R}^{i-2} \mathbb{1} \\ &= \boldsymbol{\pi}_1 \mathbb{1} + 2\boldsymbol{\pi}_2 \left(\mathbf{I}-\mathbf{R}\right)^{-1} + \boldsymbol{\pi}_2 \mathbf{R} \left(\mathbf{I}-\mathbf{R}\right)^{-2} \mathbb{1},\end{aligned}$$

and the mean system time as

$$T = \frac{\mathrm{E}\,(N)}{\lambda}. \tag{5}$$

3. CASE STUDIES

In this section we examine some small M/MAP(k)/n queues and obtain numerical results for their optimal control.

3.1 M/MAP(2)/2 systems

The the M/MAP(2)/2 queue is the simplest meaningful M/MAP(k)/n queue, in which there are two servers with the same order 2 MAP service units characterized by

$$\mathbf{D}_0 = \begin{pmatrix}-1/10 & 1/20 \\ 0 & -100\end{pmatrix}, \mathbf{D}_1 = \begin{pmatrix}1/20 & 0 \\ 5 & 95\end{pmatrix}, \tag{6}$$

and the customers arrive according to a Poisson process with intensity $\lambda = 1.5$. In this queue there is one simple question to be answered: If both servers are idle, one of them is in phase 1 and the other one is in phase 2, which server has to process the next arriving customer to have a minimal average system time? In other words what is the optimal value of p in (2)? The first intuitive answer to choose the server which can serve the customer faster. This means that we compare the mean service time starting from phase 1 and phase 2, i.e., $(1,0)(-\mathbf{D_0})^{-1}\mathbb{1}$ and $(0,1)(-\mathbf{D_0})^{-1}\mathbb{1}$, and if the first expression is smaller, we choose the server in phase 1 ($p = 1$), otherwise the one in phase 2 ($p = 0$). This greedy decision can be motivated by the fact that we would like to serve the customer as fast as possible to have an idle queue as soon as possible. The numerical results, however, show that the opposite choice is the optimal, as can be seen in Figure 1(a), i.e., it is better to choose the server which serves the customer slower.

This counter-intuitive result can be interpreted in the following way. If we use the faster server for the first customer, the probability of finishing the service before a new arrival is high, as the mean service time of the faster state is significantly smaller than the mean inter-arrival time of a new customer. Upon service there is a chance that the server moves to the slower state, leaving the system with two servers in the phase with higher service time. In this state there is a higher chance that more than 2 consecutive customers arrive before the first customer can be served, which leads to a higher average system time. In other words assigning the customer with the faster server leads to a more deteriorated state after service completion. While assigning the customer with the server in the slower phase, there is a chance that the server will move to the faster state upon service, thus the state of the system improves. One can think of this effect as the repair of the server at the cost of a slower service.

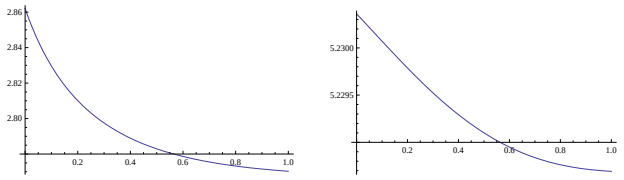
The presented behaviour is quite typical. Our investigations show that choosing the server with higher service time (with probability 1) is optimal regardless of other characteristics of the servers and the intensity of arrivals. For example the MAP characterized by $\mathbf{D}_0$ and $\mathbf{D}_1$ has a positive lag-1 correlation. Replacing $\mathbf{D}_1$ with $\mathbf{D}_1' = \begin{pmatrix}1/20 & 0 \\ 95 & 5\end{pmatrix}$ results in a MAP with negative lag-1 correlation, whose associated system time is depicted in Figure 1(b).

3.2 M/MAP(3)/2 queue

In our second example we also have two servers, but the service time is a MAP(3) with

$$\mathbf{D}_0 = \begin{pmatrix}-1 & 0 & 0 \\ 0 & -r_1 & 0 \\ 0 & 0 & -r_2\end{pmatrix}, \mathbf{D}_1 = \begin{pmatrix}0 & 1 & 0 \\ 0 & 0 & r_1 \\ r_2 & 0 & 0\end{pmatrix} \tag{7}$$

In this case, just like in the previous example, we can make decision only if both servers are idle and their phases are different. However, while for MAP(2) this meant the determination of a single priority relation, here we have three relations (priority between phase 1 and 2, 1 and 3, 2 and 3). Our numerical experiments suggests that it is sufficient to consider only strict (the server in the phase with higher priority serves the new customer with a probability of 1) and transitive priority relations (if phase 1 has priority over phase 2 and phase 2 has priority over phase 3, then 1 has priority over phase 3). Applying this assumption we have 6 possible service policies and by evaluating all of them we can determine the best service policy for any M/MAP(3)/2 queue. We denote the priority of the server in phase i by



(a) Positive correlated service times

(b) Negative correlated service times

Figure 1: Mean system time of M/MAP(2)/2 queue as a function of p

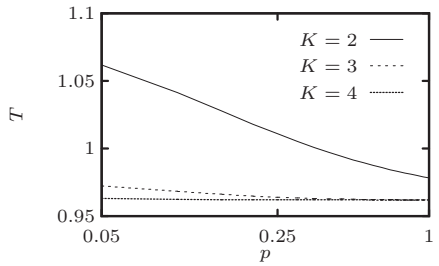


Figure 2: The system time vs. p for several K values

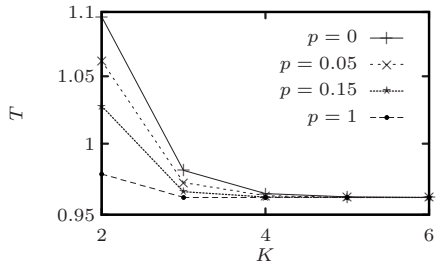


Figure 3: The system time versus the number of cores, for several p parameters

pr_i, where $pr_i > pr_j$ if we choose the server in phase i over the one in phase j.

First, let $r_1 = 100$, $r_2 = 10$ and $\lambda = 1.5$. In this case the results are in accordance with the observations obtained for M/MAP(2)/2 queue. The best priority scenario is $pr_1 > pr_3 > pr_2$. The reasoning is the same as before. By using the server in the slowest phase, we guarantee that its next service time will be faster (probabilistically), i.e. we improve the state of the system. While choosing the server in phase 2 and phase 3 both worsens the state of the system, the deterioration is greater when the server transitions from phase 3 to phase 1. If $r_1 = 10$, $r_2 = 100$, and $\lambda = 1.5$, however, the optimal priority scenario remains the same ($pr_1 > pr_3 > pr_2$) although the previous reasoning would suggest that $pr_1 > pr_2 > pr_3$ is better. In this case we have to lower r_1 to 2.2 to get $pr_1 > pr_2 > pr_3$ for optimum. If we change the arrival rate the switching point between the optimal scenarios also changes. For $\lambda = 1$ the switching point is around $r_1 = 2.4$. These results imply that the optimal priority scenario is determined by conflicting effects and the intuitive understanding of the optimal decision is more complicated.

3.3 M/MAP(2)/n with more than 2 servers

In this section we investigate the cases with higher number of MAP(2) servers. The numerical analysis is based on the Markov chains built on the analogy of (1). In case of more than 2 servers, at a customer arrival we need to choose between the idle servers if more than one is available. We identify an idle server by its phase that is held during the idle period and we number the phases such that the server in phase 1 is the "slower". The number of all possible strict and transitive priority cases increases exponentially with the number of servers. To simplify the investigation we apply a uniform probabilistic server selection shame. If upon a customer arrival there is at least one idle server in phase 1 and one in phase 2 we choose the one in phase 1 (the slower) with probability p. We compare the uniform probabilistic server selection shames for different p values.

We evaluated the cases when the arrival process is a Poisson process with parameter $\lambda = 1/10$ and the service time is characterized by the MAP(2) in (6). The results are depicted in Figure 2 and 3. Figure 2 indicates that for all evaluated number of servers the best uniform probabilistic policy is to choose the slowest server, which is the case at $p = 1$. According to our investigations the correlation of the MAP(2) service does not play a role in the optimal probabilistic decision and the same conclusion holds. It can be seen in Figure 3 that, as number of servers tends to infinity, the system time tends to a constant limiting value independent of parameter p. According to the expectations the limiting constant value is the mean service time of the MAP(2), $T_\infty = 100/101$.

The evaluations of uniform probabilistic server selection shames suggest that the conclusion obtained for the M/MAP(2)/2 case might extend to M/MAP(2)/n queues with more than 2 servers. In general, it can be interpreted as follows. At any levels of system saturation (any number of customers in the queue), it worth to choose the slower server, because it results in a better system state for higher levels of system saturation.

4. REFERENCES

[1] G. Latouche and V. Ramaswami. *Introduction to Matrix Analytic Methods in Stochastic Modeling.* ASA-SIAM Series on Statistics and Applied Probability. Society for Industrial and Applied Mathematics, 1999.