

The $c\mu/\theta$ rule

(Invited Paper)

Rami Atar
Department of Electrical
Engineering
Technion — IIT
Haifa 32000, Israel
atar@ee.technion.ac.il

Chanit Giat
Department of Electrical
Engineering
Technion — IIT
Haifa 32000, Israel
gchanit@tx.technion.ac.il

Nahum Shimkin
Department of Electrical
Engineering
Technion — IIT
Haifa 32000, Israel
shimkin@ee.technion.ac.il

ABSTRACT

We consider a multi-class queueing system with customer abandonment. For class i , the holding cost per unit time, the service rate and the abandonment rate are denoted by c_i , μ_i and θ_i , respectively. Our results show that under a many-server fluid scaling and heavy traffic conditions, a routing policy that assigns non-preemptive priority to classes according to their index $c_i\mu_i/\theta_i$, is asymptotically optimal for minimizing the overall long run average holding cost.

Keywords

Multi-class multi server queues, many servers asymptotic regime, asymptotically optimal control, fluid limits

1. INTRODUCTION

The usefulness of the well known $c\mu$ rule of routing control stems from its simplicity and its robustness. This routing policy and its generalizations have been proved to be optimal (in a precise [8] or an asymptotic sense [5], [6], [7]) for delay and queue-length costs, in a variety of settings. Although these settings are quite general, they do not include ones where customers may abandon while waiting to be served. Abandonment phenomena has been widely discussed in the recent literature, as it is a significant modeling aspect in applications, and particularly in call centers (for recent developments on these applications and related models see [1] and [4]). In this paper, we introduce a routing rule for models which include abandonment, to which we refer as the $c\mu/\theta$ rule. Like the $c\mu$ rule, the $c\mu/\theta$ rule is simple on one hand, and performs well on the other hand. In particular, our results show that it asymptotically minimizes the long run average holding cost for many-server models with abandonment, in fluid regime. Our aim here is only to report and discuss this result; the proof will appear in [2].

The model considered here consists of a fixed number of customer classes and a server pool with homogeneous

servers. Arrivals occur according to Poisson processes, services are exponential and so are the abandonment times. The arrival, service and abandonment rates are class dependent, and denoted by λ_i , μ_i and θ_i , respectively, for class i . The individual holding cost per unit time for class i is denoted by c_i . We study this model in a many-server fluid regime. The formal scaling limit leads to a deterministic ordinary differential equation, and the solution to the corresponding control problem is a lower bound on the limiting cost of the stochastic queueing problem, under any policy. The solution to the deterministic problem turns out to be particularly simple and explicitly solvable. When translated back to the queueing model, this solution formally corresponds to assigning preemptive priority to the classes in agreement with the order of the indices $c_i\mu_i/\theta_i$. This priority rule indeed achieves in the limit the lower bound alluded to above, hence it is asymptotically optimal for the queueing system. Since an often more realistic model is one where service to customers cannot be interrupted, we are mainly interested in a nonpreemptive version of this rule. The scaling is taken in such a way that the service times are not accelerated, and therefore it is not obvious that the non-preemptive policy should behave similarly. Our results show that a non-preemptive version of the priority rule alluded to above is, in fact, asymptotically optimal in the scaling limit.

Other recent contributions to queueing models in fluid regimes, motivated by applications to call centers, include Bassamboo et. al. [3], introducing a two-scale parameter regime and developing a linear program based approach to dynamic routing, and Whitt [9] in an approach that emphasizes the role played by abandonment. See also references cited in [1] for related queueing models in diffusion regime.

In the next section we describe the model and state the main result. We discuss the result in Section 3, and give a rough sketch of the proof in Section 4.

2. MODEL AND MAIN RESULT

The model consists of I customer classes and a pool with n identical servers. Customers of class i arrive according to a Poisson process with rate λ_i , for $i = 1, \dots, I$. Customers who cannot be served immediately upon arrival, are kept in an infinite-capacity queue that is dedicated to their specific class. A customer who has been held in queue may lose her patience and abandon the system. Customer “impatience” is assumed to be exponentially distributed with mean $1/\theta_i$ for class- i customers. Once admitted to service, a customer

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ValueTools 2008, October 21–23, 2008, Athens, GREECE.
Copyright ©2008 ICST ISBN # 978-963-9799-31-8 .

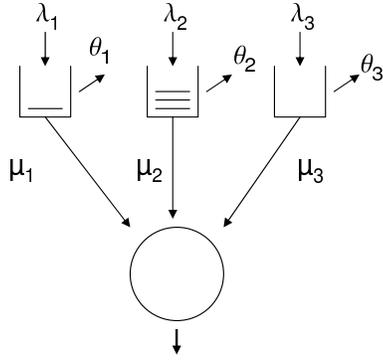


Figure 1: A model with three classes

of class i is served with exponentially distributed time duration, with mean $1/\mu_i$. The control involves choosing the customer class to be served next when a server becomes free.

The stochastic processes involved in the model will be defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The parameter n , denoting the number of servers in the pool, will appear in the notation of the stochastic processes.

For $i \in \mathcal{I} := \{1, \dots, I\}$, denote by $X_i^n(t)$ the total headcount of class i customers in the system, by $Q_i^n(t)$ the queue length of class- i customers, and by $Z_i^n(t)$ the number of servers occupied by customers of class i at time t . Clearly, for every $t \geq 0$,

$$\sum_{i \in \mathcal{I}} Z_i^n(t) \leq n \quad (1)$$

$$X_i^n(t) - Z_i^n(t) = Q_i^n(t) \geq 0, \quad i \in \mathcal{I}. \quad (2)$$

The arrival processes A_i^n are modeled as Poisson with rate λ_i^n . We use $D_i^n(t)$ and $R_i^n(t)$ to denote the number of class- i service completions and number of class- i abandonments, by time t , respectively. By the exponential assumption, these are given as

$$D_i^n = \tilde{D}_i^n \left(\int_0^\cdot Z_i^n(s) ds \right), \quad R_i^n = \tilde{R}_i^n \left(\int_0^\cdot Q_i^n(s) ds \right), \quad (3)$$

for some Poisson processes \tilde{D}_i^n and \tilde{R}_i^n with rates μ_i^n and θ_i^n , respectively. The 3I processes A_i^n , D_i^n , R_i^n , referred to as the *primitive processes*, are further assumed to be mutually independent (for each n). Finally, we have

$$X_i^n(t) = X_i^n(0) + A_i^n(t) - D_i^n(t) - R_i^n(t), \quad i \in \mathcal{I}, \quad t \geq 0. \quad (4)$$

We will use bold font for vector notation, as in \mathbf{X}^n for the vector whose i -th element is X_i^n .

Equations (1)–(4) do not fully describe the model, since the routing mechanism has not been specified. We will use a very elaborate definition of the term ‘policy’, that will only rely on the above equations and the primitive processes. Any process

$$\pi^n = (\mathbf{D}^n, \mathbf{R}^n, \mathbf{X}^n, \mathbf{Q}^n, \mathbf{Z}^n)$$

will be referred to as a policy, provided that equations (1)–(4) hold, and that the primitive processes satisfy our probabilistic assumptions mentioned earlier. The collection of

all policies π^n will be denoted by Π^n . Note that we do not require a policy to satisfy any work-conservation condition.

Let constants $c_i \geq 0$ be given, denoting holding cost per unit time for class- i customers. For any policy π_n define the cost as

$$C_{n,T}(\pi_n) = \frac{1}{nT} \mathbb{E}^{\pi_n} \left[\int_0^T \mathbf{c} \cdot \mathbf{Q}^n(t) dt \right]. \quad (5)$$

Let the corresponding value be defined by

$$V_{n,T} = \inf_{\pi_n \in \Pi_n} C_{n,T}(\pi_n). \quad (6)$$

We consider a sequence of queueing systems as above where now the number of servers $n \in \mathbb{N}$ is used as an index to the sequence. The parameters are assumed to satisfy the following. There are positive constants $\lambda_i, \mu_i, \theta_i$, such that, as $n \rightarrow \infty$,

$$\frac{\lambda_i^n}{n} \rightarrow \lambda_i, \quad \mu_i^n \rightarrow \mu_i, \quad \theta_i^n \rightarrow \theta_i. \quad (7)$$

Note that the system may be overloaded in the sense that the workload exceeds the service capacity. However, since the abandonment rates are non zero stability holds automatically.

To motivate the proposed policy via a heuristic discussion, consider quantities \mathbf{x} , \mathbf{q} and \mathbf{z} , that formally represent the steady state values of \mathbf{X}^n , \mathbf{Q}^n and \mathbf{Z}^n respectively, for large values of n and T . These quantities must satisfy $\mathbf{x} = \mathbf{q} + \mathbf{z}$, and in addition

$$\begin{cases} z_i, q_i \geq 0 \\ \lambda_i = \mu_i z_i + \theta_i q_i \\ \sum_{i \in \mathcal{I}} z_i \leq 1. \end{cases} \quad (8)$$

Consider the problem

$$\text{minimize } \mathbf{c} \cdot \mathbf{q} \text{ such that } (\mathbf{q}, \mathbf{z}) \text{ satisfies (8)}. \quad (9)$$

We denote by $(\mathbf{x}, \mathbf{q}, \mathbf{z})$ the solution to this linear program, and by V the corresponding minimum value. The solution to this problem (that we do not provide here) has the property that $q_i = 0$ for all i whose indices $c_i \mu_i / \theta_i$ are the largest (provided that $q_i = 0$ for at least one i). This suggests that priority should be given according to this index.

Motivated by the foregoing discussion, we define π_n^* to be the routing policy that assigns nonpreemptive priority according to this index. More precisely, under this policy, after each service completion the next customer to be admitted to service is chosen according to its class’s index, (where the higher the index is, the higher is the priority). By assumption, this policy is non-preemptive, and work-conserving. We refer to it as the **$c\mu/\theta$ non-preemptive priority rule**, or simply the **$c\mu/\theta$ rule**.

Our result shows that in the many-server fluid regime, the $c\mu/\theta$ rule achieve asymptotic optimality with respect to the cost criterion alluded to above, among all policies.

THEOREM 2.1. *Assume $n^{-1} \mathbf{X}^n(0) \rightarrow \mathbf{x}$, as $n \rightarrow \infty$. Then*

$$\liminf_{T \rightarrow \infty} \liminf_{n \rightarrow \infty} V_{n,T} = \limsup_{T \rightarrow \infty} \limsup_{n \rightarrow \infty} C_{n,T}(\pi_n^*). \quad (10)$$

The proof of this result will appear in [2]. The main idea is provided below in Section 4.

3. DISCUSSION

The result presented in the previous section addresses a long run average cost. The following example discusses informally the short term performance of this policy. It indicates that one should not expect the $c\mu/\theta$ rule to be optimal on a short time horizon.

EXAMPLE 3.1. Consider a queueing system with $\mathcal{I} = \{1, 2\}$ and let the model parameters take the values: $\mu_1 = \mu_2 = 1$, $\lambda_1 = \lambda_2 = 3/4$ and $\theta_1 = \epsilon$, $\theta_2 = 1$ where $0 < \epsilon \ll 1$ such that

$$\frac{c_1}{\theta_1} > \frac{c_2}{\theta_2},$$

but $c_2 > c_1$. The initial state is $x_1(0) = z_1(0) = 1/2$, $x_2(0) = z_2(0) = 1/2$. Consider the cumulative holding cost till time $T \ll 1$. Applying the $c\mu/\theta$ priority rule, we prioritize the first class, and, a simple calculation yields that the holding cost is given by

$$C_T(\text{pr. 1}) = \frac{c_2}{2} \cdot T - \frac{c_2}{2} (1 - \exp\{-T\}) \approx \frac{1}{4} c_2 T^2,$$

for small T . On the other hand, if we prioritize the second class then

$$C_T(\text{pr. 2}) = \frac{c_1}{2\epsilon} \cdot T - \frac{c_1}{2\epsilon^2} (1 - \exp\{-\epsilon T\}) \approx \frac{1}{4} c_1 T^2.$$

Hence, $C_T(\text{pr. 2}) < C_T(\text{pr. 1})$ for T small enough.

Next, note that even if we start with initial state having positive queues at time $t = 0$, the phenomenon may repeat itself. Choose $x_1(0) = x_2(0) = 3/4$. According to $c\mu/\theta$ rule, we set $z_1(0) = 3/4$, $q_1(0) = 0$, $z_2(0) = 1/4$ and $q_2(0) = 1/2$. Since this is the desirable steady state, there is no initial change in either of the processes, and therefore,

$$C_T(\text{pr. 1}) = \frac{c_2}{2} \cdot T.$$

If we apply the opposite priority rule, in which we prioritize the second class right from the beginning, we set $z_1(0) = 1/4$, $q_1(0) = 1/2$, $z_2(0) = 3/4$ and $q_2(0) = 0$. Then, $q_1(t) = 1/2\epsilon - (1 - \epsilon)/2\epsilon \cdot \exp\{-\epsilon t\}$ and we obtain

$$\begin{aligned} C_T(\text{pr. 2}) &= \frac{c_1}{2\epsilon} \cdot T - \frac{c_1(1 - \epsilon)}{2\epsilon^2} (1 - \exp\{-\epsilon T\}) \\ &\approx \frac{c_1}{2} T - \frac{c_1}{4} (1 - \epsilon) T^2. \end{aligned}$$

Since $c_1 < c_2$, we again conclude $C_T(\text{pr. 2}) < C_T(\text{pr. 1})$.

A heuristic explanation of the first example (with empty initial queues) is that when the queues are nearly zero, there is no abandonment, and then it is the $c\mu$ rule that should count.

We can conclude from this example that the optimality of the $c\mu/\theta$ rule is sensitive to the form of the cost (particularly, to the time horizon). However, it is possible that this policy features other modes of robustness. In particular, note that no information on the arrival processes enters the definition of the policy. It is plausible that this policy has good performance, with respect to long run average holding cost, under far more general assumptions on the arrivals, and it will be interesting to address, in future work, this and other robustness aspects.

We would like to point out that one can incorporate into the model an abandonment-count cost, a performance measure that is important in applications. Indeed, the expected

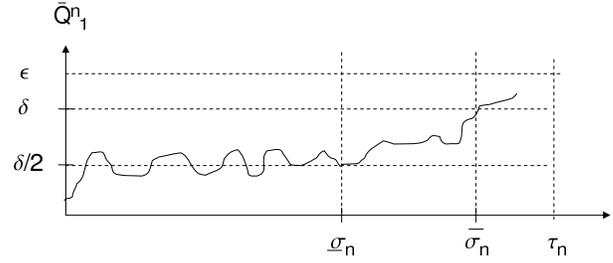


Figure 2: The random times $\bar{\sigma}_n$ and $\underline{\sigma}_n$

rate at which customers abandon from queue i is $\theta_i \mathbb{E}[Q_i]$ (where we remove n from the notation). Letting $b_i \geq 0$ denote penalty per class- i customer abandonment, we obtain that the total customer abandonment cost is given by $\sum_i b_i \theta_i \mathbb{E}[Q_i]$. It is reasonable to take into account both holding cost and abandonment cost, and thus to consider in place of (5) the cost

$$\frac{1}{nT} \mathbb{E}^{\pi^n} \left[\int_0^T \sum_i (b_i \theta_i + c_i) Q^n(t) dt \right]. \quad (11)$$

It is clear that with $\tilde{c}_i = b_i \theta_i + c_i$, the above can be represented in the form (5), and thus this wider apparatus is, in fact, covered by our analysis. Interestingly, in case we are interested only in abandonment count, namely when $c_i = 0$, the index $\tilde{c}_i \mu_i / \theta_i$ is equal to $b_i \mu_i$, and the policy obtained is a version of the $c\mu$ rule (with b_i replacing c_i).

4. PROOF SKETCH

We give here a rough sketch of how Theorem 2.1 is proved in [2].

Recall that V denotes the minimum value for the linear program associated with the fluid model. The first part of the argument is that V is a lower bound on the large n , large T limit of $V_{n,T}$ (i.e., on the l.h.s. of (10)). Taking expectation in equations (1)–(4), dividing by T and sending T to infinity results in a deterministic model of the form (8). It follows that the l.h.s. of (10) is bounded below by V .

In the second part of the proof we analyze the policy π_n^* . Recall that $\mathbf{s} := (\mathbf{x}, \mathbf{q}, \mathbf{z})$ denotes the solution to the linear program, and that by assumption, $n^{-1} \mathbf{X}^n(0)$ is close to \mathbf{x} . Denote $\mathbf{S}^n = (\mathbf{X}^n, \mathbf{Q}^n, \mathbf{Z}^n)$, and $\bar{\mathbf{S}}^n = n^{-1} \mathbf{S}^n$, and, given $T > 0$ and $\epsilon > 0$ let

$$\tau_n = \tau_n(T, \epsilon) = \inf\{t \geq 0 : \|\bar{\mathbf{S}}^n(t) - \mathbf{s}\| \geq \epsilon\} \wedge T.$$

In what follows, write $\|\mathbf{x}\|_u^*$ for $\sup_{t \leq u} \|\mathbf{x}(t)\|$. Uniform integrability (in n) of the random variables $\|\mathbf{Q}^n\|_T^*$ is not hard to obtain by straightforward estimates on the arrival processes. As a result, to prove that the r.h.s. of (10) is equal to V , it suffices to show that

$$\mathbb{P}(\tau_n < T) \rightarrow 0 \text{ as } n \rightarrow \infty, \quad (12)$$

for every T and ϵ , because this will show, in particular, that $\bar{\mathbf{Q}}^n := n^{-1} \mathbf{Q}^n$ is uniformly close to \mathbf{q} .

Since the state of one class influences that of the other classes, the proof of (12) proceeds in several steps, starting

with the class of highest priority, in which we show that $\|\bar{S}_i^n - s_i\|_{\tau_n}^* \rightarrow 0$ in probability, as $n \rightarrow \infty$. We demonstrate here only the part involving \bar{Q}_1^n , since the other parts are based on similar considerations. To this end, consider the event $E_n = \{\|\bar{Q}_1^n - q_1\|_{\tau_n}^* > \delta\}$, for some $\delta < \varepsilon$. Define

$$\bar{\sigma}_n = \inf\{t \in [0, T] : \bar{Q}_1^n(t) > \delta\},$$

$$\underline{\sigma}_n = \sup\{t \in [0, \bar{\sigma}_n] : \bar{Q}_1^n(t) \leq \delta/2\}.$$

Then on the event E_n , we have $0 \leq \underline{\sigma}_n \leq \bar{\sigma}_n \leq \tau_n$, and for every t in the interval $I_n := [\underline{\sigma}_n, \bar{\sigma}_n]$, $\bar{Q}_1^n(t) > 0$. The policy we analyze gives highest priority to class 1. As a result, within I_n , every server that becomes available immediately starts to serve a class-1 customer. The total service rate available to serve class-1 customers quickly rises to a quantity that is significantly larger than the rate of arrivals of customers from this class. This makes it very unlikely for the queue length \bar{Q}_1^n to increase over the interval I_n by a large quantity. However, due to the way $\underline{\sigma}_n$ and $\bar{\sigma}_n$ are defined, the queue does grow by $\delta n/2$ over this interval. We conclude that E_n is unlikely for large values of n .

Acknowledgments: Nahum Shimkin's research was supported by Grant No. 2008480 from the United States–Israel Binational Science Foundation (BSF).

References

- Z. Aksin, M. Armony, and V. Mehrotra. The Modern Call-Center: A Multi-Disciplinary Perspective on Operations Management Research. *Production and Operations Management*, 16(6):665–688, 2007.
- R. Atar, C. Giat, and N. Shimkin. Asymptotic optimality of the $c\mu/\theta$ rule. In preparation, Technion, Israel, 2008.
- A. Bassamboo, J. Harrison, and A. Zeevi. Design and Control of a Large Call Center: Asymptotic Analysis of an LP-Based Method. *Oper. Res.*, 54(3), 2006.
- N. Gans, G. Koole, and A. Mandelbaum. Commissioned Paper: Telephone Call Centers: Tutorial, Review, and Research Prospects. *Manufacturing & Service Operations Management*, 5(2):79–141, 2003.
- A. Mandelbaum and A. Stolyar. Scheduling Flexible Servers with Convex Delay Costs: Heavy-Traffic Optimality of the Generalized $c\mu$ -Rule. *Oper. Res.*, 52(6):836–855, 2004.
- J. Van Mieghem. Dynamic Scheduling with Convex Delay Costs: The Generalized $c\mu$ Rule. *Ann. Appl. Probab.*, 5(3):809–833, 1995.
- J. Van Mieghem. Due-Date Scheduling: Asymptotic Optimality of Generalized Longest Queue and Generalized Largest Delay Rules. *Oper. Res.*, 51(1):113–122, 2003.
- J. Walrand. *An Introduction to Queueing Networks*. Englewood Cliffs, N.J. : Prentice-Hall, 1988.
- W. Whitt. Fluid Models for Multiserver Queues with Abandonments. *Oper. Res.*, 54(1):37–54, 2006.