Multi-server Marginal Allocation
With CVaR and Abandonment based QoS Measures

Per Enqvist[1] and Göran Svensson[1,2]
[1]*Division of Optimization and Systems Theory, Kungliga Tekniska Högskolan, Lindstedtsv 25, Stockholm, Sweden*
[2]*Teleopti WFM, Teleopti AB, Sweden*

Keywords: Queueing, Queueing Networks, Marginal Allocation, Conditional Value-at-Risk, Abandonments.

Abstract: Two multi-objective minimization problems are posed, one for Erlang-C queues and one for Erlang-A queues. The objectives are to minimize the cost of added agents while also trying to optimize a quality of service measure. For the Erlang-C system we propose using the Conditional Value-at-Risk measure with waiting time as the loss function. We prove that this quality of service measure is integer convex in the number of servers. For the Erlang-A system we use the fraction of abandoning customers and some rate based weighting function as the service measure. Finally, a numerical comparison of the two system types is performed. The numerical results show the similarities between the two systems in terms of optimal points.

1 INTRODUCTION

In this paper we investigate multi-class queueing networks and the optimal allocation of servers with respect to a *Quality of Service* (QoS) measure. We consider two types of queueing systems, one based on the Erlang-C model with *Conditional Value-at-Risk* (CVaR) (Rockafellar and Uryasev, 2000; Rockafellar and Uryasev, 2002) on the waiting time as the QoS measure. In the second system type we look at the Erlang-A model and use the fraction of abandoning customers as the basis of our QoS measure.

The basic model consists of a system of parallel server pools with a corresponding set of agents(servers) that cater to customers(jobs), see Figure 1. Each server pool has a separate and infinite *first-come-first-serve* (FCFS) buffer. The separate and parallel queueing systems are bound together by a common budget constraint.

The optimization problem is formulated and solved in terms of the *marginal allocation* (MA) algorithm (Fox, 1966). When varying the budget constraint the whole efficent front, consisting of efficient solutions, can be found. The MA algorithm depends on the costs per agent and the improvements of the QoS measure, for the different queues, to be separable and (integer) convex functions. In the tradition of (Rolfe, 1971; Dyer and Proll, 1977; Weber, 1980) we proceed to prove that the QoS measure, determined by CVaR, is decreasing and convex in the number of agents. In (Parlar and Sharafali, 2014) the authors summarize other proofs of convexity for different QoS measures. In a similar fashion the system of queues where abandonments are allowed is posed and solved, using a QoS measure based on the fraction of abandoning customers. The convexity of the measure is posed as a conjecture.

2 MODEL DESCRIPTION AND THE QoS MEASURES

We consider a system of $N \in \mathbb{N}$ queues of either $M/M/c$ or $M/M/c+M$ type (using the notation of (Baccelli and Hebuterne, 1981)), i.e., Erlang-C or Erlang-A models, each with its own infinite queueing buffer.

Introduce the index set $I = \{1,\ldots,N\}$. Let c_i, denote the number of servers in queue i, $\mathbf{c} = [c_1 \ \ldots \ c_N]$ and let a_i be the corresponding cost for each agent of type i. The total cost of the agents is a separable and integer convex function in $\mathbf{c}$. Furthermore, assume that the numbers of agents available for assignment to the different demands are also limited. Let $d_i \in \mathbb{N}$, be the maximum number of available agents of type i.

The arrival process to queue i is a homogeneous Poisson process with arrival rate parameter λ_i. The service rate of each server in pool i is denoted by μ_i and service times are exponentially distributed. In the case of the Erlang-A type systems the abandonment

Enqvist, P. and Svensson, G.
Multi-server Marginal Allocation - With CVaR and Abandonment based QoS Measures.
DOI: 10.5220/0006652602970303
In *Proceedings of the 7th International Conference on Operations Research and Enterprise Systems (ICORES 2018)*, pages 297-303
ISBN: 978-989-758-285-1

rate of each customer is included. The time to abandonment of a customer in queue i is exponentially distributed with rate parameter θ_i. Hence, the customer may leave the system due to either service completion or impatience, whichever occurs first. It is assumed that customers being served do not defect.

Let b denote a budget constraint on the system, meaning that the total cost of the assigned servers to the queueing network must be within budget. The model may be extended to include several budget constraints, affecting only a subset of the queues.

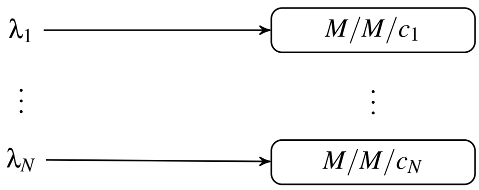

Figure 1: A system of N parallell $M/M/c$ queues.

2.1 The Conditional Value-at-Risk Measure

To promote a positive customer experience the queueing system is endowed with a QoS measure. Typical QoS measures for contact centers include *average speed of answer* (ASA) and the *telephone service factor* (TSF) (Gans et al., 2003), also known as *service level* (SL). ASA measures the average time a customer spends waiting on service while TSF considers the *acceptable waiting time* (AWT) that a certain percentile of the customers have to wait before service. The TSF type of QoS measure is also used in finance, under the name of *Value-at-Risk* (VaR), to describe the risk of (large) losses. The CVaR measure is often preferred to the VaR type measure due to its convexity (Rockafellar and Uryasev, 2000; Rockafellar and Uryasev, 2002) and coherent measure properties (Artzner et al., 1999). In general CVaR offers a means of controlling worst case outcomes, which might be of great importance in fields like healthcare, as pointed out in (Parlar and Sharafali, 2014). We consider the QoS measure corresponding to CVaR where the loss function is taken to be a customers waiting time in the queue. We apply the CVaR measure to the system of $M/M/c$ queues.

First some preliminary results. The waiting time distribution for a $M/M/c$ queue is given by (Kleinrock, 1996)

$$\Pr(W_q > t) = \Pi_q(c,\eta)e^{-(c\mu-\lambda)t} = 1 - F_{W_q}(t), \quad (1)$$

where W_q is the random variable representing the waiting time and $\Pi_q(c,\eta)$ is the probability of delay, i.e., of having to wait when there are c homogeneous agents working under load $\eta = \frac{\lambda}{\mu}$. This is given by the Erlang-C formula (Kleinrock, 1996),

$$\Pi_q(c,\eta) = \frac{\eta^c/c!}{(1-\eta/c)\sum_{i=0}^{c-1}\eta^i/i! + \eta^c/c!}. \quad (2)$$

It can be calculated efficiently via the following recursion (Zeng, 2003)

$$\Pi_q(c+1,\eta) = \frac{\lambda}{\mu}\frac{(c\mu-\lambda)\Pi_q(c,\eta)}{\big((c+1)\mu-\lambda\big)c - \lambda\Pi_q(c,\eta)}. \quad (3)$$

This formula avoids numerical issues. We note that the recursion is initiated by $\Pi_q(1,\eta) = \eta$, and any value larger than one should be interpreted as one, since the larger values corresponds to unstable queues when abandonments are excluded.

The different queues may have different QoS requirements, thus let β_i denote the quantile level for queue i and t_i be the VaR(AWT) value for queue i. The QoS requirement for the VaR type measure is then given by

$$\Pr(W_{q,i} \le t_i) \ge \beta_i, \quad i \in I, \quad (4)$$

and the corresponding β_i-VaR is defined as

$$\begin{aligned} t_i(c_i) &= \min\{t | F_{W_{q,i}}(t) \ge \beta_i, c_i\} \\ &= \frac{\log\Pi_q(c_i,\eta_i) - \log(1-\beta_i)}{c_i\mu_i - \lambda_i}. \end{aligned} \quad (5)$$

We use the notation of t_i to underscore that the VaR value is measured in units of time, in this case.

The corresponding β-CVaR is defined by

$$\phi_{\beta_i}(c_i) = \frac{1}{1-\beta_i}\int_{t_i(c_i)}^{\infty} t\,dF_{W_{q,i}}(t), \quad i \in I, \quad (6)$$

with explicit formulation for $\Pi_q(c_i,\eta_i) \ge 1-\beta_i$:

$$\begin{aligned} \phi_{\beta_i}(c_i) &= \frac{1}{c_i\mu_i - \lambda_i}\left(\log\big(\frac{\Pi_q(c_i,\eta_i)}{1-\beta_i}\big) + 1\right) \\ &= t_i(c_i) + \frac{1}{c_i\mu_i - \lambda_i}, \quad i \in I, \end{aligned} \quad (7)$$

where $t_i(c_i)$ is given by (5).

Remark 1 (Probability atoms and CVaR)**.** *If β is small then a probability atom might have to be handled. To avoid such an atom it is required that $\Pi_q \ge 1-\beta$ holds. For a complete treatment of CVaR in cases where there is a probability atom see (Rockafellar and Uryasev, 2002). However, for most realistic choices of parameters this is not an issue and will thus be ignored throughout the rest of the paper.*

The QoS measure given by this definition of the β-CVaR is integer convex and decreasing in the number of agents, c. The use of the MA algorithm rests on this fact. This is formalized in Proposition 1.

Proposition 1 (Integer Convexity of β-CVaR)**.** *Consider a $M/M/c$ queue with constant rate parameters $\mu, \lambda > 0$ such that $c\mu > \lambda$ and that $K < \beta \leq 1$, where K is large enough that there is no probability atom. Then*

$$\phi_{\beta_i}(c_i) = t_i(c_i) + \frac{1}{c_i\mu_i - \lambda_i}, \quad i \in I,$$

is decreasing and integer convex in $c \in \mathbb{N}$.

The proof is given in the Appendix.

2.2 Probability of Abandonment

For a system of multi-server Erlang-A queues other QoS measures are of interest. In (Garnett et al., 2002; Mandelbaum and Zeltyn, 2007) different measures for queueing systems with abandonments are considered. Erlang-A queues are stable for arbitrary loads while Erlang-C queues are only stable when $c\mu > \lambda$. Here we consider systems in steady state, with constant (random) demand. Perhaps the most obvious measure is the probability that a customer will abandon (defect) the queue before recieving service. This may occur if the arriving customer finds all servers occupied and thus have to wait in queue. Then the customer will abandon if his/her patience runs out before a server becomes available.

Let $\pi_j^{(i)}$ denote the probability of queue i being in state $j \in \mathbb{N}$, where the states are given by the number of customers in queue i. Furthermore, let $E_{1,c_i}^{(i)}$ denote the Erlang blocking formula of the i:th queue with c_i servers. The incomplete Gamma function is defined as

$$\gamma(x,y) = \int_0^y t^{x-1} e^{-t} dt, \tag{8}$$

then, in accordance with (Mandelbaum and Zeltyn, 2007), let

$$A(x,y) = \frac{xe^y}{y^x}\gamma(x,y), \quad x > 0, y \geq 0. \tag{9}$$

The probability of an arrival finding all servers busy is given by

$$\Pr(W_q > 0) = \frac{A(\frac{c\mu}{\theta}, \frac{\lambda}{\theta})E_{1,c}}{1 + \left(A(\frac{c\mu}{\theta}, \frac{\lambda}{\theta}) - 1\right)E_{1,c}}. \tag{10}$$

The fraction of customers abandoning, conditioned on having to wait on arrival, is given by

$$\Pr(\text{Ab}|W_q > 0) = \frac{1}{\rho A(\frac{c\mu}{\theta}, \frac{\lambda}{\theta})} + 1 - \frac{1}{\rho}, \tag{11}$$

where $\rho = \frac{\lambda}{c\mu}$, the offered load per agent. Then using the definition of conditional probability yields the fraction of customers abandoning the queue.

The QoS measure given by the fraction of customers abandoning is decreasing and integer convex in the number of servers according to numerical tests we performed. We are convinced it holds in general and may be known. At this stage we pose it as a conjecture.

Conjecture 1 (Integer Convexity of Pr(Ab))**.** *Consider a $M/M/c+M$ queue with constant rate parameters $\mu, \lambda > 0$. Then*

$$\Pr(Ab) = \Pr(Ab|W_q > 0)\Pr(W_q > 0)$$
$$= \left(\frac{1}{\rho A(\frac{c\mu}{\theta}, \frac{\lambda}{\theta})} + \frac{\rho - 1}{\rho}\right)\left(\frac{A(\frac{c\mu}{\theta}, \frac{\lambda}{\theta})E_{1,c}}{1 + \left(A(\frac{c\mu}{\theta}, \frac{\lambda}{\theta}) - 1\right)E_{1,c}}\right) \tag{12}$$

is decreasing and integer convex in $c \in \mathbb{N}$.

3 OPTIMIZATION FORMULATION

Here we pose optimization formulations based on the systems and QoS measures of Subsections 2.1 and 2.2. Since costs, CVaR and the probability to abandon are integer convex functions in the number of servers, the multi-objective solutions given by the MA algorithm will be in terms of an efficient front.

First we will formulate the multi-objective problem that minimizes the total agent cost and β-CVaR for all $M/M/c$ queues. Then the following Optimization problem can be formulated,

$$\begin{array}{llll} \min\limits_{\mathbf{c}} & \varphi \sum\limits_{i=1}^{N} a_i c_i & +\psi \sum\limits_{i=1}^{N} \phi_{\beta_i}^{(i)}(c_i), & \\ \text{Sub. to} & \sum\limits_{i=1}^{N} a_i c_i & \leq b, & \\ & c_i & \leq d_i, & \forall i \in I, \\ & c_i \in \mathbb{N}, & a_i, b, \varphi, \psi > 0, & \forall i \in I, \end{array} \tag{13}$$

where $\phi_{\beta_i}^{(i)}(c_i)$ denotes the QoS measure, for the i:th queue, based on CVaR with waiting time as the loss function. The weights φ and ψ are constant and determine the relative importance of the two objectives. Note that both functions in the objective are separable functions.

In a similar fashion, we pose an optimization problem minimizing the agent costs and the probabilities to abandon modified by a weighting function for the $M/M/c+M$ queues. The optimization problem for a queue with abandonments can be formulated as

$$\begin{array}{llll}\min_{\mathbf{c}} & \varphi \sum_{i=1}^{N} a_i c_i & +\psi \sum_{i=1}^{N} \omega(\lambda_i,\mu_i)\Pr_i\{Ab|c_i\}, & \\ \text{Sub. to} & \sum_{i=1}^{N} a_i c_i & \leq b, & \\ & c_i & \leq d_i, & \forall i \in I, \\ & c_i \in \mathbb{N}, & a_i,b,\varphi,\psi > 0, & \forall i \in I,\end{array} \tag{14}$$

where $\Pr_i\{Ab|c_i\}$ denotes the probability of abandonment, for the i:th queue, given that there are c_i servers at work and $\omega(\lambda_i,\mu_i)$ is some weight function that correspond to some notion of the population of queue i. We will use the offered load. Note that both functions in the objective are separable functions.

The optimization problem formulated in (14) can be extended to include a waiting time based QoS measure. Ideally we would like said condition to be the CVaR measure with loss function in terms of the waiting time under the condition that the waiting customer gets served. This measure needs to be defined as of yet. Instead we settle for average waiting time(mean delay) under the condition that the demand eventually gets served. This result can be found in (Riordan, 1962, p. 112, eq. 88). This condition forms a lower bound on the number of agents and it is included in the marginal allocation algorithm as the initiating number of agents of each type, thus guaranteeing that this QoS measure will be met.

3.1 The Marginal Allocation Algorithm

An efficient algorithm for finding the efficient points for two integer convex and separable functions is the MA algorithm, described in (Fox, 1966; Svanberg, 2009).

In general when minimizing the multi-objective optimization problem for functions $f,g:\mathbb{N}^N \to \mathbb{R}$ where

$$\begin{array}{ll}\Delta f_j(x_j) \leq \Delta f_j(x_j+1) < 0 & \forall j,x_j \in \mathbb{N}, \\ 0 < \Delta g_j(x_j) \leq \Delta g_j(x_j+1) & \forall j,x_j \in \mathbb{N},\end{array} \tag{15}$$

the optimal vector $\mathbf{x}^* \in \mathbb{N}^N$ minimizes $\varphi g(x) + \psi f(x)$ if and only if the following conditions are satisfied for each $j = 1,\dots,N$:

$$\begin{cases} \frac{-\Delta f_j(x_j^*)}{\Delta g_j(x_j^*)} \leq \frac{\varphi}{\psi} \leq \frac{-\Delta f_j(x_j^*-1)}{\Delta g_j(x_j^*-1)} & \text{if } x_j^* > 0, \\ \frac{-\Delta f_j(0)}{\Delta g_j(0)} \leq \frac{\varphi}{\psi} & \text{if } x_j^* = 0. \end{cases} \tag{16}$$

x_j^* is an efficient solution if and only if there are constants $\varphi,\psi > 0$ such that the conditions (16) are satisfied for each $j = 1,\dots,N$.

Marginal Allocation Algorithm

Step 0: Generate a table with N columns, fill the columns with the quotients $-\Delta f_j(n)/\Delta g_j(n)$ for $n = 0,1,2,\dots$ where n starts at zero. Set $k = 0, x^{(0)} = (0,\dots,0)$, $g(x^{(0)}) = g(\mathbf{0})$ and $f(x^{(0)}) = f(\mathbf{0})$.

Step 1: Select the largest uncancelled quotient in the table. Cancel it and let l be the corresponding column number.

Step 2: Let $k := k+1$ then let $x_l^{(k)} = x_l^{(k-1)}$ and $x_j^{(k)} = x_j^{(k-1)}, \forall j \neq l$.
Let $f(x^{(k)}) = f(x^{(k-1)}) + \Delta f_l(x_l^{k-1})$ and $g(x^{(k)}) = g(x^{(k-1)}) + \Delta g_l(x_l^{k-1})$.
Terminate algorithm if $g(x^{(k)}) \geq g^{\max}$; otherwise go to Step 1.

We identify our budget constraint b to be $g^{\max}$ and the constraint on available agents, $d_i, i \in I$, to be the number of quotients to calculate for column i. This algorithm can now be applied to (13) and (14), where

$$g(x) = \sum_{i=1}^{N} a_i c_i \tag{17}$$

$$f(x) = \begin{cases} \sum_{i=1}^{N} f_{\text{CVaR}}^{(i)}(c_i) & \text{for (13)}, \\ \sum_{i=1}^{N} \omega(\lambda_i,\mu_i)\Pr_i\{Ab|c_i\} & \text{for (14)}. \end{cases} \tag{18}$$

4 NUMERICAL EXAMPLES

The main benefit of using the MA approach is that huge systems can be optimized almost effortlessly. An example of the efficient front of a system of 100 $M/M/c$ queues, with a (maximal) budget constraint of $b = 4500$, is shown in Fig 2. The input parameters in terms of arrivals, service rates and more were randomly generated. The optimal staffing calculation took less than a second to perform on a laptop.

To compare the Erlang-C and the Erlang-A based systems we look at a three class multi-server system (i.e., three queues). First, find the efficient points for the Erlang-C type system with CVaR as the service measure. Then use those points in the solution of the Erlang-A type system with fraction of abandonments as basis for the service measure. These points are then compared to the optimal points, shown in Figure 3.

The parameters used were $\boldsymbol{\beta} = [.95\ \ .95\ \ .95]^T, \boldsymbol{\lambda} = [15\ \ 10\ \ 20]^T, \boldsymbol{\mu} = [0.5\ \ 0.6\ \ 0.7]^T$ and $\mathbf{a} = [12\ \ 15\ \ 18]^T$. The procedure was repeated for two sets of impatience rates, $\boldsymbol{\theta} = [0.25\ \ 0.25\ \ 0.25]^T$ and $\boldsymbol{\theta} = [10\ \ 10\ \ 10]^T$, respectively. The weight function, $\omega(\lambda_i,\mu_i)$, in (14) is defined to be the offered load, η_i.

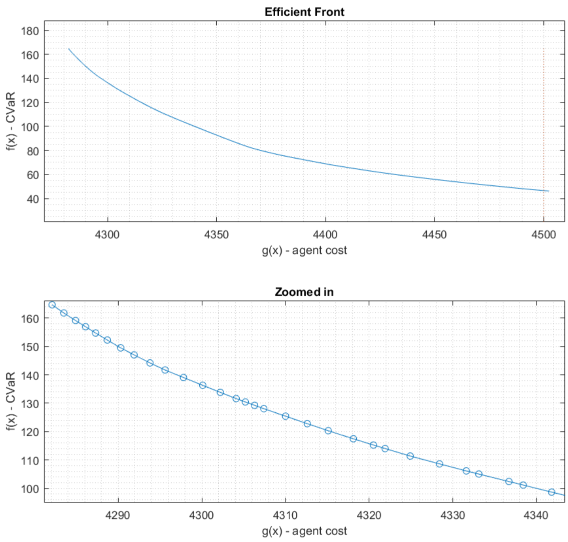

Figure 2: Showing the efficient front for a system of 100 $M/M/c$-queues, using CVaR as QoS measure and with a budget constraint of 4500. The parameters where randomly generated. In the bottom figure some specific efficient points are shown.

With respect to the patience parameter the two solutions do not differ significantly. The given solutions are very similar, which seem to hold for larger systems and for a wide range of parameter values. Even though the probability of abandonment is very similar for the two approaches, the optimal number of agents may be different. Therefore, we consider the optimal distributions in Table 1 to illustrate this for the example seen in Figure 3. With abandonments the first agent pool is prioritized over the second and third, and recieves one or two more agents compared to the CVaR assignment.

5 SUMMARY AND CONCLUSIONS

The main objectives of this paper is to find an optimal distribution of agents between N queues, bound by budget constraints, via the MA algorithm. The QoS measures used are CVaR, with waiting time as the loss function, for the systems of $M/M/c$ queues and a weighted probability of abandonment for the $M/M/c+M$ queues. The advantage of using the MA algorithm is that it can find the efficient points extremely fast. As can be seen from the numerical examples studying two different queueing networks give similar solutions. The two QoS measures have rather different shape, consider the curvature of the graphs in Figure 3, but generate similar optimal points. We believe this is partly a consequence of the exponential distributions of the service times. It may well diverge for other service time distributions, e.g. lognormal distributions.

Table 1: Showing the partial agent distributions for the CVaR solution (left figure) and for the abandonment solution where $\theta = 0.25$ (middle figure).

Agents	CVaR			Ab		
77	31	17	29	32	17	28
78	31	18	29	33	17	28
79	31	18	30	33	17	29
80	32	18	30	33	18	29
81	32	19	30	34	18	29
82	33	19	30	34	18	30
83	33	19	31	34	19	30
84	33	20	31	35	19	30
85	34	20	31	35	19	31
86	34	20	32	36	19	31
87	35	20	32	36	19	32
88	35	21	32	36	20	32
89	36	21	32	36	20	33
90	36	21	33	37	20	33
91	36	22	33	37	21	33

ACKNOWLEDGEMENTS

The authors would like to thank Teleopti AB, Stockholm Sweden, for their support of our work.

REFERENCES

Artzner, P., Delbaen, F., Eber, J., and Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, 9(3):203–228.

Baccelli, F. and Hebuterne, G. (1981). On queues with impatient customers. *Research report, RR-0094, INRIA*.

Dyer, M. and Proll, L. (1977). On the validity of marginal analysis for allocating servers in M / M / c queues. *Management Science*, 23(9).

Fox, B. (1966). Discrete optimization via marginal analysis. *Management Science*, 13(3):210–216.

Gans, N., Koole, G., and Mandelbaum, A. (2003). Telephone call centers: Tutorial, review, and research prospects. *Manufacturing Service Oper. Management*.

Garnett, O., Mandelbaum, A., and M.Reiman (2002). *Manufacturing and Service Operations Management*, 4:208–227.

Kleinrock, L. (1996). *Queueing systems : problems and solutions*. Wiley, New York.

Mandelbaum, A. and Zeltyn, S. (2007). *Service Engineering in Action: The Palm/Erlang-A Queue, with Appli-*

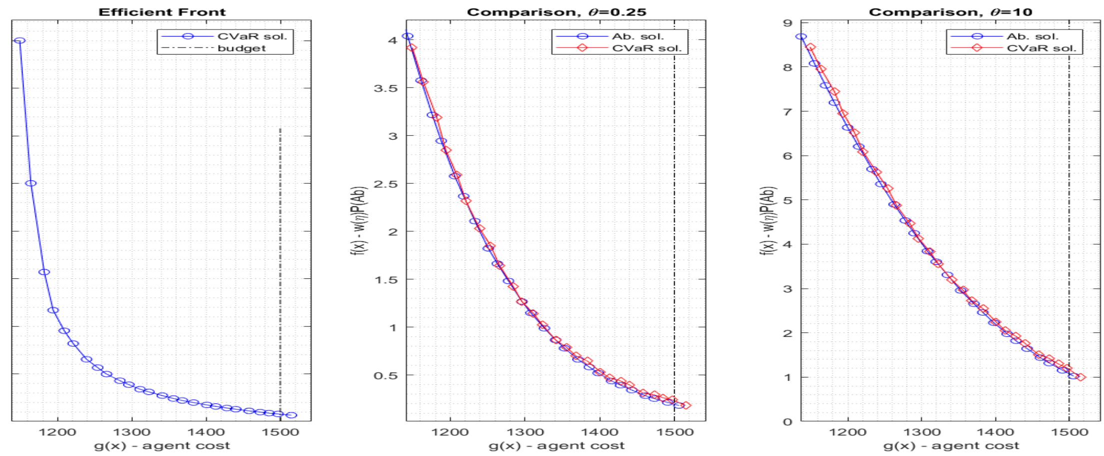

Figure 3: The left figure shows the CVaR solution for a system of three queues. In the middle and right figures the solution under abandonments are shown, for $\theta = 0.25$ and $\theta = 10$. The MA solution for abandonments is compared to the CVaR solution for abandonments.

cations to Call Centers, pages 17–45. Springer Berlin Heidelberg, Berlin, Heidelberg.

Parlar, M. and Sharafali, M. (2014). Optimal design of multi server markovian queues with polynomial waiting and service costs. *Applied Stochastic Models in Business and Industry*.

Riordan, J. (1962). *Stochastic Service Systems*. Wiley.

Rockafellar, R. and Uryasev, S. (2000). Optimization of conditional value-at-risk. *Journal of risk*, 2.

Rockafellar, R. and Uryasev, S. (2002). Conditional value-at-risk for general loss distributions. *Journal of banking & Finance*, 26:1443–1471.

Rolfe, A. J. (1971). A note on marginal allocation in multiple-server service systems. *Management Science*.

Svanberg, K. (2009). *On marginal allocation*. Dept. of mathematics, KTH, Stockholm.

Weber, R. (1980). On the marginal benefit of adding servers to G/GI/m queues. *Management Science*, 26(9).

Zeng, G. (2003). Two common properties of the Erlang-B function, Erlang-C function, and Engset blocking function. *Mathematical and Computer Modelling*, 37(12):1287–1296.

APPENDIX

Proof of Proposition 1

First note that β-CVaR satisfies $\phi_\beta(c,\lambda,\mu) = (1/\lambda)\Phi_\beta(c,1,\mu/\lambda)$, so we can without loss of generality assume that $\lambda = 1$ and $\mu < 1$.
The forward difference $\Delta\phi_\beta(c) = \phi_\beta(c+1) - \phi_\beta(c)$ is given by

$$\frac{1}{c\mu+\mu-1}\left[\log\frac{\Pi_{c+1}}{\Pi_c} - \mu\phi_\beta(c)\right] \le 0 \qquad (19)$$

since Π_c is non-increasing, hence $\phi_\beta(c)$ is non-increasing.

The second forward difference is

$$\Delta^2\phi_\beta(c) = \frac{1}{C_1C_2}\left[2\mu^2\phi_\beta(c) + G\right] \qquad (20)$$

where $C_k = (c+k)\mu - 1$, and

$$\begin{aligned} G \;=\;& C_1\log\left(\frac{C_1}{C_0}\frac{C_1c-\Pi_c}{C_2(c+1)-\Pi_{c+1}}\right) \\ & -2\mu\log\left(\frac{C_0}{(C_1c-\Pi_c)\mu}\right). \end{aligned} \qquad (21)$$

Convexity of β-CVaR follows by showing that $G \ge 0$. Using that $x/(1+x) \le \log(1+x) \le x$ we have that

$$\log\frac{C_1}{C_2(c+1)-\Pi_{c+1}} \ge -\log(c+1) \qquad (22)$$

$$-\frac{1}{c+1}\frac{\mu(c+1)-\Pi_{c+1}}{C_1}, \qquad (23)$$

and

$$\log\frac{C_1c-\Pi_c}{C_0} \ge \log(c) + \frac{c\mu-\Pi_c}{cC_0+\mu c-\Pi_c}. \qquad (24)$$

Then

$$\begin{aligned} G \;\ge\;& C_1\log\frac{c}{c+1} + 2\mu\log(c\mu) - \mu \\ & +\frac{\Pi_{c+1}}{c+1} + \frac{C_3(c\mu-\Pi_c)}{cC_0+(c\mu-\Pi_c)}. \end{aligned} \qquad (25)$$

From (3), it follows that

$$G \ge \frac{C_{-1}}{c} + \frac{\Pi_c\left(\frac{C_0}{(c+1)\mu} - C_2\right) + 2\mu^2c}{(cC_0+(c\mu-\Pi_c))}. \qquad (26)$$

Note that $c(c\mu-1)+c\mu-\Pi_c > 0$ since $c\mu > 1$. Using that $\Pi_c \le \frac{1}{c\mu+1}\frac{1}{(c-1)\mu+1}$, $G(c(c\mu-1)+c\mu-\Pi_c)$ is bounded from below by

$$\begin{aligned} &(c\mu-1)^2 + \mu^2(2c-1) - \frac{1}{c\mu+1}\frac{1}{(c-1)\mu+1}\cdot \\ &\left[(c\mu-1)\left(\frac{c+1}{c} - \frac{1}{(c+1)\mu}\right) + \mu\left(2-\frac{1}{c}\right)\right] \end{aligned} \qquad (27)$$

Introduce $\varepsilon = \frac{c\mu-1}{\mu}$, and eliminate c in (27) to obtain

$$\mu^2(\varepsilon^2+2\varepsilon+\frac{2}{\mu}-1)-\frac{H}{(2+\mu\varepsilon)(2+\mu\varepsilon-\mu)}, \quad (28)$$

where $H=\mu(2+\varepsilon)+\mu^2\frac{\varepsilon-1}{1+\mu\varepsilon}-\frac{1}{1+\mu+\mu\varepsilon}$.
Consider two cases, $\varepsilon \geq 1$ and $\varepsilon < 1$.

If $\varepsilon \geq 1$, then (28) is greater than

$$\begin{aligned} & \mu^2(\varepsilon^2+2\varepsilon+\frac{2}{\mu}-1)-H/4 \quad (29)\\ \geq \quad & (\varepsilon\mu+\frac{7\mu-2}{8})^2+\frac{12+11\mu+97\mu(1-\mu)}{64} \end{aligned}$$

which is non-negative and we used that $\frac{1}{1+\mu\varepsilon} \leq 1$ and $\frac{1}{1+\mu+\mu\varepsilon} \geq 1-\mu-\mu\varepsilon$.

If $\varepsilon < 1$, then we write (28) on common denominator and use $\frac{1}{1+\mu\varepsilon} \geq 1-\mu\varepsilon$ and $\frac{1}{1+\mu+\mu\varepsilon} \geq 1-\mu-\mu\varepsilon$, we can show that the numerator is bounded below by

$$\begin{aligned} (1-\mu)\,\big[1-\varepsilon\mu+\mu^3+5\mu(1-\mu)+\mu(1-\varepsilon) \\ +4\varepsilon\mu^2(13/4-\mu+\varepsilon(1+\mu)))\big] \\ +\mu^3\varepsilon^3(4\varepsilon+\mu\varepsilon^2+7) \geq 0 \quad (30) \end{aligned}$$

for all $\varepsilon, \mu \in (0,1)$.

Then $\Delta^2\phi_\beta(c,1,\mu) \geq \frac{2\mu^2\phi_\beta(c,1,\mu)}{C_2C_1} \geq 0$, and ϕ_β is integer-convex.