European Journal of Operational Research 310 (2023) 1101–1117

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Stochastics and Statistics

Design of fork-join networks of First-In-First-Out and infinite-server queues applied to clinical chemistry laboratories

Eline R. Tsai[a,b,c], Derya Demirtas[a], Andrei N. Tintu[b], Robert de Jonge[c], Yolanda B. de Rijke[b,1], Richard J. Boucherie[a,*]

[a] Center for Healthcare Operations Improvement and Research, University of Twente, P.O. box 217, 7500 AE, Enschede, The Netherlands
[b] Erasmus MC, University Medical Center Rotterdam, Department of Clinical Chemistry, P.O. box 2040, 3000 CA, Rotterdam, The Netherlands
[c] Amsterdam University Medical Center, VU University Medical Center, Department of Clinical Chemistry, P.O. box 7057, 1007 MB, Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 22 March 2022
Accepted 3 April 2023
Available online 13 April 2023

Keywords:
Queueing
Optimal design
Queueing network analyzer
Simulated annealing
Laboratory design

ABSTRACT

This paper considers optimal design of queueing networks in which each node consists of a single-server FIFO queue and an infinite-server queue, which is referred to as incubation queue. Upon service completion at a FIFO queue, a job splits (forks) into two parts: the first part is routed to the next node on its route, and the second part is placed in the incubation queue. Routing of the jobs of multiple types is governed by a central decision maker that decides on the routes for each job type and aims to minimize the mean turnaround time of the jobs, i.e., the time spent in the system until service completion at the FIFO queue in the last node, and at all incubation queues on the job's route, which may be viewed as a join operation. We provide explicit results for the turnaround time when all service and inter-arrival time distributions are exponential and invoke the Queueing Network Analyzer when these distributions are general. We then develop a Simulated Annealing approach to find the optimal routing configuration. We apply our approach to determine the optimal routing configuration in a chemistry analyzer line.

1. Introduction

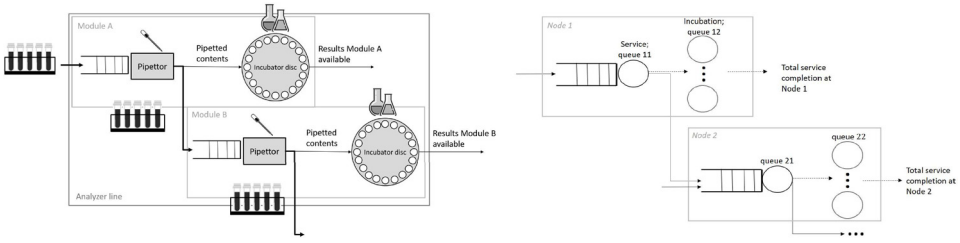
Clinical chemistry laboratories perform various tests on body liquids, thus playing an important role in diagnostics, monitoring and prediction of diseases. Analysis of clinical chemistry samples is performed on so-called analyzer lines, that consist of distinct analyzer modules. Fig. 1 a depicts a small part of the laboratory consisting of two modules. Typically, a small to medium sized laboratory contains several analyzer lines that each contain around four modules. Racks, containing several tubes with samples to be tested, arrive in random order and visit the modules. At each module, a rack joins the queue of the pipettor that handles racks in order of arrival and transfers a defined sample volume from the sample tubes into processing cells. These cells are located on the ample capacity incubator disc, where, according to a test-specific schedule, dilution and reagent fluids are added, the fluid is mixed, incubated and measured, after which the test results become available. At the same time (after sample pipetting is completed), the rack is routed to the next module on its route to join the queue of the pipettor of that module. The modules to be visited are determined by the test mix of the tubes in the rack. The order in which the racks visit the modules mostly does not affect the quality of the test results. The turnaround time (TAT) or end-to-end sojourn time between arrival of samples in the lab and availability of results is the most important performance indicator as this determines the time until the medical doctor receives the patient's test results (Tsai et al., 2019). This paper introduces a queueing network and optimization approach to design a chemistry analyzer line that minimizes TAT.

A pipettor handles racks one by one in order of arrival and may be modeled as a single server First-In-First-Out (FIFO) queue, where the service duration is determined by the number of pipetting operations for the required tests at the module. The service times of different samples on the incubator disc are independent and the incubator has ample positions, so that the incubator disc may be modeled as an infinite-server queue. A natural model

* Corresponding author.
E-mail addresses: e.r.tsai@utwente.nl (E.R. Tsai), d.demirtas@utwente.nl (D. Demirtas), a.tintu@erasmusmc.nl (A.N. Tintu), r.dejonge1@amsterdamumc.nl (R. de Jonge), y.derijke@me.com (Y.B. de Rijke), r.j.boucherie@utwente.nl (R.J. Boucherie).

[1] Y.B. de Rijke received a personal grant from Roche Diagnostics Nederland B.V. to support this research.

https://doi.org/10.1016/j.ejor.2023.04.003



(a) Route of a rack visiting two modules. (b) Queueing network model.

Fig. 1. Laboratory configuration and queueing network model.

for an analyzer line is therefore a network of nodes in which jobs of different classes follow a class dependent route along the nodes that consist of two parts: a FIFO single-server queue and an infinite-server incubation queue (Figure 1 b). Upon service completion at a FIFO queue, a job splits into two parts: one part routes to the incubation queue, and the other part to the next FIFO queue on its route. Upon service completion at the incubation queue, this part leaves the network. A job's TAT ends when service at all FIFO queues along its route *and* all its incubation times are completed. Routing of jobs in the network is governed by a central decision maker that aims to design the lab such that the mean TAT of all jobs is minimized. To this end, the decision maker determines a static routing configuration by dividing the job classes into types, and assigning fixed routes to each type.

Our network has general inter-arrival, service and incubation times characterized via their mean and variance. For some special cases we provide explicit results. For the network with exponential inter-arrival, service, and incubation times we provide explicit results for the joint queue length distribution at the FIFO and incubation queues as well as the Laplace–Stieltjes Transform (LST) of the TAT for a given routing configuration that specifies the route for each job type. In case of a single job type, these results enable closed form evaluation of the mean TAT. For the network with general inter-arrival, service, and incubation times, we develop an approximation to obtain the mean TAT. This approximation involves the Queueing Network Analyzer (QNA) (Whitt, 1983b) to obtain the mean sojourn times at the FIFO queues and the mean TAT for a given routing configuration. The general optimization problem for the decision maker is non-convex. Therefore, we develop a Simulated Annealing (SA) approach to determine the near-optimal routing configuration. Our approach includes several approximation steps of which accuracy is investigated in detail in numerical experiments. We apply our approach to the design of a four module analyzer line in the clinical chemistry laboratory of Erasmus MC, Rotterdam, the Netherlands. For the current load, our optimal routing configuration routes jobs along the nodes from high to low incubation time, which we find to be a good heuristic for general parameter settings. For a 60% load increase, our optimization approach yields roughly 5% reduction in mean TAT compared to this heuristic, which illustrates the quality of the heuristic. In light of the number of tests performed on a chemistry analyzer line, optimization of the routing configuration may result in a substantial improvement of laboratory performance.

Literature: Sample routing in laboratories has been analyzed using operations research techniques. Sample routing between blood collection sites and laboratories was studied using discrete-event simulation (Lote, Williams, & Ülgen, 2009) and modeled as a vehicle routing problem (Grasas et al., 2014; Zabinsky et al., 2020). Lean principles have been applied inside the laboratory, resulting in improved sample routing (Persoon, Zaleski, & Frerichs, 2006; Rutledge, Xu, & Simpson, 2010).

We will use notation as introduced in Kelly (1979), that provides a general description of open and closed product-form queueing networks. For a tandem of FIFO single-server exponential queues the sojourn times at the queues are independent and the distribution of the TAT is available in closed form (Kelly, 1979, Theorem 2.2). Ordering of tandem queues is studied in Suresh & Whitt (1988), where it is shown that arranging the queues in increasing order of service time variability is an effective heuristic to decrease the average sojourn time. For a network of FIFO queues with multiple job types, fixed routes, and exponential service times that do not depend on the job type, the marginal distribution of the sojourn time at each FIFO queue is known (Kelly, 1979, p. 63), from which the mean sojourn time in the network may be obtained. Optimal design of networks of multi-server FIFO and infinite-server queues with fixed routes is studied in Kerbache & MacGregor Smith (2000), where an artificial holding queue is introduced for each finite waiting room FIFO queue to register blocked customers. Such overflow queues differ from our incubation queues. Exact results for the design of networks with exponential service times and concave utility functions are presented in Kameda & Zhang (1995); Shaler (2009). Our utility function TAT is non-convex, see Appendix B. Like the references above, we consider a centralized decision maker. Design of networks with selfish customers or decentralized decision makers are presented in, e.g., (Ghosh & Hassin, 2021; Laan, Timmer, Boucherie, & Nl, 2021). Optimal design of static routes is geared towards optimization of the configuration of the laboratory. Dynamic routing is geared towards the optimization of the operational process, see Shaler (2009) for networks with concave utility function. Our network may be seen as a fork-join network with a fork operation after each FIFO queue, and one join operation when service at all queues along a job's route is completed. The generating function for the queue lengths in fork-join queues are available for the M/M/1 system with two parallel queues (Flatto & Hahn, 2006). Fork-join queues appear, e.g., in parallel or distributed storage and computing (Fidler, Walker, & Bora, 2020; Zubeldia, 2020), proactive coordination between predicted ED patient admissions and inpatient bed management (Lee, Chinnam, Dalkiran, Krupp, & Nauss, 2021), manufacturing systems (Krishnamurthy, Suri, & Vernon, 2003), and container terminals (Kumawat, Roy, De Koster, & Adan, 2021). Our network may also be viewed as a network with regular jobs and positive signals, where a job exits the FIFO queue as positive signal, then visits the incubation queue to increase the number of jobs in that queue by 1, after which it immediately departs the incubation queue as regular job to visit the next FIFO queue on its route. A product-form stochastic upper bound for the stationary distribution of the number of jobs in the network with random routing and exponential single server queues is provided in Huisman & Boucherie (2011).

For our general network the TAT distribution is not available. We approximate the mean TAT via mean sojourn time approximations for the FIFO queues using the QNA. The QNA, originally developed by Whitt (1983b), uses several heavy traffic results, such as Kingman's approximation for the mean waiting time (Kingman, 1961). Accuracy of the QNA depends on the quality of the approximation of non-renewal arrival and departure processes by re-

newal processes (Caldentey, 2001; Whitt, 1983a). The approximation quality of the QNA is studied for several multi-class single-server queueing networks, e.g., Bitran & Tirupati (1988); Fendick, Saksena, & Whitt (1991); Harrison & Nguyen (1990); Whitt (1983a). The QNA has been used to compare system configurations, e.g., (Bai & Menon, 2013; Yu & De Koster, 2008; Zonderland, Boer, Boucherie, De Roode, & Van Kleef, 2009).

The general optimization problem, using the QNA, for the decision maker is non-convex. Global non-convex optimization is NP-hard (Danilova et al., 2022). We develop an SA algorithm to solve the optimization problem of the decision maker. Combining QNA with mathematical optimization is used in Van Nyen, Bertrand, Van Ooijen, & Vandaele (2006); Zhou, Wang, He, & Goh (2017). Zhou et al. (2017) use QNA to obtain the lead time in a manufacturing system to determine the optimal batch size for different product classes. The problem is formulated as a mixed integer program that is solved using a traversal algorithm. Van Nyen et al. (2006) develop a heuristic method for near-optimal production and inventory control decisions, where unimodality of the objective function in the review periods is postulated after extensive tests, which motivated the use of a simple greedy search algorithm called univariant search parallel to the axes. This greedy search algorithm was shown to outperform SA. SA originates from the analogy between combinatorial optimization and the annealing process of solids (Kirkpatrick, Gelatt, & Vecchi, 1983). Convergence theorems for global continuous optimization using SA for real-valued functions are studied in Bélisle (1992); Locatelli (2000). SA has been applied to continuous optimization problems in healthcare (Shepard, Cao, Afghan, & Earl, 2007; Wason & Jaki, 2012). In our network decision variables are coupled. Our combined QNA and SA approach yields a fast approach (run-time of several minutes for the cases in Section 5.3) that is amenable for use in the design phase of a laboratory. We have used discrete-event simulation (DES) to evaluate the accuracy of our approach. The run-time of DES (around 15 hours for each scenario and parameter setting at load 0.45 in Section 5.3) prohibits its direct use in optimal laboratory design.

Statement of contribution: The contribution of this paper is fourfold. First, for the network with one job type, Poisson arrivals, and exponential service and incubation times, we obtain the LST of the TAT distribution. Second, for a given routing configuration in the general network we develop a QNA-based approximation of the mean TAT of each job type. We characterize the quality of our approximations and show their accuracy for our purpose of optimal route selection. Third, we develop an SA numerical optimization approach that provides near-optimal routing configurations and can handle large real-life instances such as the clinical chemistry laboratory case. Fourth, we show that our approach allows optimization of static routes in a real-world clinical chemistry laboratory.

This paper is organized as follows. Section 2 introduces our queueing network model for chemistry analyzer lines. Section 3 considers Poisson arrivals, exponential service and incubation times. Section 4 introduces our QNA and SA approach to obtain near-optimal routing configurations for the general network. Section 5 numerically supports the accuracy of our approximations and applies our approach to a clinical chemistry laboratory. Section 6 concludes our paper.

2. Queueing network model

Consider an open network of nodes $j = 1, \ldots, J$. We will use notation for networks of queues as introduced in Kelly (1979, Chapter 3), that provides a general description of queueing networks. Jobs of class $c = 1, \ldots, C$ arrive to the network with inter-arrival times $A(c)$, with mean $1/\lambda_0(c)$, variance $\sigma_0^2(c)$, and squared coefficient of variation (SCV) $\text{scv}_0(c) = \sigma_0^2(c)\lambda_0^2(c)$. Jobs require service from a subset $R(c) \subseteq \{1, \ldots, J\}$ of the nodes, where each node may be visited only once. Each node j consists of two parts: a FIFO single-server queue $j1$ and an infinite-server incubation queue $j2$, see Fig. 1 b. The operation of the node is as follows. Upon arrival, a job joins the tail of the FIFO queue. Upon service completion at the FIFO queue, the job splits into two parts: the first part routes to the next node on its route, and the second part routes to the incubation queue. Upon service completion of the part at the incubation queue, this part leaves the network. The service time of jobs of class c in queue jk, $B_{jk}(c)$, has mean $1/\mu_{jk}(c)$, variance $\sigma_{jk}^2(c)$, and SCV $\text{scv}_{sjk}(c)$, $c = 1, \ldots, C$, $j = 1, \ldots, J$, $k = 1, 2$. A job departs from the network when the service at the last FIFO node along its route, *and* all incubation times at the nodes are completed. All random variables for service and inter-arrival times are independent. Let $\widetilde{Z}(s) = \mathbb{E}[e^{-sZ}]$, $\text{Re}(s) \geq 0$, denote the LST of a non-negative random variable Z.

Jobs of class c must visit all nodes in $R(c)$, but the order in which these nodes are visited may be different. To this end, jobs of class c may be divided into types c_i, $i = 1, \ldots, I(c)$, where each type corresponds to a fixed route $n(c_i, 1), n(c_i, 2), \ldots, n(c_i, |R(c)|)$ that is available for class c along the nodes in $R(c)$, with $n(c_i, j)$ the j-th node visited by a job of type c_i, $i = 1, \ldots, I(c)$. Clearly, $I(c) \leq |R(c)|!$ as the routes per type are distinct and at most all permutations of nodes in $R(c)$ may be used to assign a route to a job of class c. The TAT of a job is determined by the arrival time, the service completion time at the FIFO queue in the last node on its route, *and* the incubation times at all nodes on the job's route. Routing of the jobs in the network is governed by a central decision maker that aims to minimize the mean time jobs spend in the network. To achieve this goal, the decision maker divides the jobs of class c into types by deciding what fraction $p(c_i)$ of jobs of class c are of type c_i, $i = 1, \ldots, I(c)$, $\sum_{i=1}^{I(c)} p(c_i) = 1$, $c = 1, \ldots, C$.

The tuple $\boldsymbol{p} = \{(p(1_1), \ldots, p(C_{I(C)})\}$ is called a routing configuration. Let $\text{TAT}(c_i)$ denote the TAT of a job of type c_i in the network, $i = 1, \ldots, I(c)$, $c = 1, \ldots, C$. The decision maker aims to find the static routing configuration $\boldsymbol{p}^*$ that results in minimum mean TAT for the jobs:

$$\boldsymbol{p}^* = arg\,min_{\boldsymbol{p}}\ f_{\text{TAT}}(\boldsymbol{p}), \quad \text{with}$$

$$f_{\text{TAT}}(\boldsymbol{p}) = \sum_{c=1}^{C}\sum_{i=1}^{I(c)} \frac{p(c_i)\lambda_0(c)}{\lambda_0}\mathbb{E}[\text{TAT}(c_i)]$$

$$\text{s.t.} \sum_{i=1}^{I(c)} p(c_i) = 1, \qquad c = 1, \ldots, C,$$

$$0 \leq p(c_i) \leq 1, i = 1, \ldots, I(c), \qquad c = 1, \ldots, C, \tag{1}$$

with $\lambda_0 = \sum_{c=1}^{C} \lambda_0(c)$ the total arrival rate of jobs to the network, $p(c_i)\lambda_0(c)/\lambda_0$ the fraction of jobs of type c_i, and $\mathbb{E}[\text{TAT}(c_i)]$ the mean TAT of jobs of type c_i, $i = 1, \ldots, I(c)$, $c = 1, \ldots, C$.

We are mainly interested in the TAT. Let $S_{j1}(c_i)$ and $S_{j2}(c_i)$ denote the sojourn time of a job of type c_i in FIFO queue $j1$ and incubation queue $j2$. Let $\text{TAT}_j(c_i)$ denote the TAT from arrival to node j up to and including the final node a job of type c_i visits in the network. The following result can readily be obtained.

Lemma 1. *The TAT of job type c_i can recursively be obtained as follows:*

$$\text{TAT}_{n(c_i,|R(c)|)}(c_i) = S_{n(c_i,|R(c)|)1}(c_i) + S_{n(c_i,|R(c)|)2}(c_i), \tag{2a}$$

$$\text{TAT}_j(c_i) = S_{j1}(c_i) + \max\left\{S_{j2}(c_i), \text{TAT}_{j+1}(c_i)\right\}, \quad j = n(c_i, |R(c)| - 1), \ldots, n(c_i, 1), \tag{2b}$$

$$\text{TAT}(c_i) = \text{TAT}_{n(c_i,1)}(c_i). \tag{2c}$$

Remark 2 (Sojourn time at the incubation queues; sojourn times at the FIFO queues)**.** The incubation queues are infinite-server queues. The sojourn time of a job of type c_i at incubation queue $j2$ equals its service time $B_{j2}(c)$, $j = 1, \ldots, J$, and is independent of the sojourn times at all other queues in the network.

The sojourn times at the FIFO queues are not affected by the incubation queues. Hence, to analyze the sojourn time of a job along its path through the FIFO queues we may consider the network without the incubation queues.

3. Exponential inter-arrival, service and incubation times

Section 3.1 introduces the network under the assumption of exponential service, incubation and inter-arrival times. Section 3.2 considers a tandem network with one job type. Section 3.3 studies the general exponential network, and optimal job routing is considered in Section 3.4.

3.1. Markov chain

Jobs of type c_i arrive to the network, at node $n(c_i, 1)$, according to a Poisson process with rate $\lambda_0(c_i)$. As jobs follow a fixed route, jobs of type c_i arrive with rate $\lambda_0(c_i)$ at each queue on their route. The service requirement of jobs of type c_i is exponential with rate $\mu_{jk}(c)$. Let $\rho_{jk}(c_i) = \lambda_0(c_i)/\mu_{jk}(c)$ and $\rho_{jk} := \sum_{c=1}^{C}\sum_{i=1}^{I(c)} \rho_{jk}(c_i)$. Assume that $\rho_{j1} < 1$, $j = 1, \ldots, J$.

Characterization of a queue with job types requires a description of the position of the jobs as well as rules for the state change upon arrival of a new job or a service completion, see Kelly (1979, Sec. 3.1). Suppose m_{jk} jobs are present at queue jk. Consider state $\boldsymbol{x}_{jk} = (x_{jk}(1), \ldots, x_{jk}(m_{jk}))$, where $x_{jk}(a)$ records the type of the job in position a. In FIFO queue $j1$, the job in position 1 is in service. If a job of type c_i arrives it is added to the tail of the queue, and the new state is $\boldsymbol{x}'_{j1} = (x_{j1}(1), \ldots, x_{j1}(m_{j1}), c_i)$. If the job in position 1 completes service, the new state is $\boldsymbol{x}'_{j1} = (x_{j1}(2), \ldots, x_{j1}(m_{j1}))$. In the incubation queue $j2$ all jobs are in service. If the job in position a completes service, the new state is $\boldsymbol{x}'_{j2} = (x_{j2}(1), \ldots, x_{j2}(a-1), x_{j2}(a+1), \ldots, x_{j2}(m_{j2}))$. A new job arriving in state $\boldsymbol{x}_{j2} = (x_{j2}(1), \ldots, x_{j2}(m_{j2}))$ moves into position a with probability $1/(m_{j2}+1)$; jobs previously in positions $a, \ldots, m_{j2}$ move to positions $a+1, \ldots, m_{j2}+1$.

The evolution of the number of jobs in the queues is recorded by the Markov chain $\{\mathcal{X}(t),\ t \in \mathbb{R}\}$ at state space $\boldsymbol{X} = \{\boldsymbol{x} = (\boldsymbol{x}_{11}, \boldsymbol{x}_{12}, \ldots, \boldsymbol{x}_{J1}, \boldsymbol{x}_{J2}) : \boldsymbol{x}_{jk} = (x_{jk}(1), \ldots, x_{jk}(m_{jk})),\ x_{jk}(a) \in \{c_i,\ i = 1, \ldots, I(c),\ c = 1, \ldots, C\},\ a = 1, \ldots, m_{jk},\ m_{jk} \in \mathbb{N}_0,\ j = 1, \ldots, J,\ k = 1, 2\}$. The description of the evolution of the queues is provided in Section 2. The transition rates for $\boldsymbol{x} \neq \boldsymbol{x}'$ are:

$$
q(\boldsymbol{x}, \boldsymbol{x}') = \begin{cases}
\lambda_0(c_i), & \text{if } \boldsymbol{x}'_{j1} = (\boldsymbol{x}_{j1}, c_i), \boldsymbol{x}'_{j2} = \boldsymbol{x}_{j2}, \boldsymbol{x}'_{\ell k} = \boldsymbol{x}_{\ell k}, \\
& \ell \neq j, k = 1, 2, j = n(c_i, 1), \\
\frac{1}{m_{j2}+1}\mu_{j1}(c), & \text{if } \boldsymbol{x}'_{j1} = (x_{j1}(2), \ldots, x_{j1}(m_{j1})), x_{j1}(1) = c_i, \\
& \boldsymbol{x}'_{j2} = (x_{j2}(1), \ldots, x_{j2}(a), c_i, \\
& x_{j2}(a+1), \ldots, x_{j2}(m_{j2})), \\
& a = 0, \ldots, m_{j2}, \boldsymbol{x}'_{\ell k} = \boldsymbol{x}_{\ell k}, \ell \neq j, \\
& k = 1, 2, j = n(c_i, |R(c)|), \\
\frac{1}{m_{j2}+1}\mu_{j1}(c), & \text{if } \boldsymbol{x}'_{j1} = (x_{j1}(2), \ldots, x_{j1}(m_{j1})), \\
& x_{j1}(1) = c_i, \boldsymbol{x}'_{\ell 1} = (\boldsymbol{x}_{\ell 1}, c_i) \\
& \boldsymbol{x}'_{j2} = (x_{j2}(1), \ldots, x_{j2}(a), c_i, \\
& x_{j2}(a+1), \ldots, x_{j2}(m_{j2})), \\
& a = 0, \ldots, m_{j2}, \boldsymbol{x}'_{hk} = \boldsymbol{x}_{hk}, h \neq j, \ell, k = 1, 2, \\
& j = n(c_i, r), \ell = n(c_i, r+1), r < |R(c)|, \\
\mu_{j2}(c), & \text{if } \boldsymbol{x}'_{j2} = (x_{j2}(1), \ldots, x_{j2}(a-1), \\
& x_{j2}(a+1), \ldots, x_{j2}(m_{j2})), \\
& x_{j2}(a) = c_i, a = 1, \ldots, m_{j2}, \boldsymbol{x}'_{\ell k} = \boldsymbol{x}_{\ell k}, \\
& \ell \neq j, k = 1, 2, j = 1, \ldots, J.
\end{cases}
$$

3.2. A tandem network with one job type

In this section, we assume there is only one job type that arrives with rate λ and follows route $1, 2, \ldots, J$. Let M_{jk} record the number of jobs in queue jk, $1, \ldots, J$, $k = 1, 2$. First we consider two special cases: zero incubation times and zero service times. Then we proceed with the general tandem network with one job type.

Consider the case with zero incubation times. As a consequence $M_{j2} = 0$, $j = 1, \ldots, J$. The random variables M_{j1}, $j = 1, \ldots, J$, are independent and the equilibrium distribution of the number of jobs in the queues is (Kelly, 1979, p. 37):

$$
\mathbb{P}(M_{11} = m_{11}, \ldots, M_{j1} = m_{j1}) = \prod_{j=1}^{J}\left(1 - \frac{\lambda}{\mu_{j1}}\right)\left(\frac{\lambda}{\mu_{j1}}\right)^{m_{j1}}, \quad \text{with } m_{j1} \in \mathbb{N}_0,\ j = 1, \ldots, J. \tag{3}
$$

The TAT is the sum of the sojourn times S_{j1} in the FIFO queues, which are independent exponential random variables with rate $\mu_{j1} - \lambda$, see (Kelly, 1979, Theorem 2.2). The LST is:

$$
\widetilde{\mathrm{TAT}}(s) = \prod_{j=1}^{J}\frac{\mu_{j1} - \lambda}{\mu_{j1} - \lambda + s}, \quad \mathrm{Re}(s) \geq 0.
$$

Now consider the case with zero service times at the FIFO queues. As a consequence $M_{j1} = 0$, $j = 1, \ldots, J$. The resulting network can now be viewed as a network of J $M/M/\infty$ queues in parallel with simultaneous Poisson arrivals with rate λ. We have $\mathrm{TAT} = \max_{j=1,\ldots,J}\{S_{j2}\}$. As the incubation times are independent random variables:

$$
\mathbb{P}(\mathrm{TAT} \leq x) = \prod_{j=1}^{J}\mathbb{P}(S_{j2} \leq x) = \prod_{j=1}^{J}\left(1 - \mathrm{e}^{-\mu_{j2}x}\right), \quad x \geq 0. \tag{4}
$$

We have simultaneous Poisson arrivals to the incubation queues. The random variables M_{j2}, $j = 1, \ldots, J$, are not independent. The joint generating function of the number of jobs in the queues is, for $|z_{j2}| \leq 1$, $j = 1, \ldots, J$, (Choi & Park, 1992):

$$
\begin{aligned}
&G_{M_{12},\ldots,M_{J2}}(z_{12}, \ldots, z_{J2}) \\
&\quad = \exp\left(\sum_{j=1}^{J}(z_{j2} - 1)\rho_{j2}\right) \\
&\qquad \times \prod_{j=2}^{J}\exp\left(\sum_{\{\ell_1,\ldots,\ell_j\}\subseteq\{1,\ldots,J\}}\frac{\prod_{m=1}^{j}(z_{\ell_m 2} - 1)\rho_{\ell_m 2}}{\prod_{m=1}^{j}\prod_{\substack{n=1\\ n\neq m}}^{j}\rho_{\ell_n 2}}\right).
\end{aligned} \tag{5}
$$

Now consider the general tandem network with one job type. The arrival processes to and the departure processes from the FIFO queues coincide with those processes for the network with zero incubation times. From Burke's theorem (Kelly, 1979, Theorem 2.1) we obtain that the arrival process to each queue jk, $j = 1, \ldots, J$, $k = 1, 2$, in the tandem with one job type, is a Poisson process with rate λ. Thus, the marginal distributions of the number of jobs in the queues are:

$$
\begin{aligned}
&\mathbb{P}(M_{j1} = m_{j1}) = \left(1 - \frac{\lambda}{\mu_{j1}}\right)\left(\frac{\lambda}{\mu_{j1}}\right)^{m_{j1}} \text{ and} \\
&\mathbb{P}(M_{j2} = m_{j2}) = \frac{1}{m_{j2}!}\left(\frac{\lambda}{\mu_{j2}}\right)^{m_{j2}}\mathrm{e}^{-\frac{\lambda}{\mu_{j2}}},
\end{aligned} \tag{6}
$$

with $m_{jk} \in \mathbb{N}_0$, $j = 1, \ldots, J$, $k = 1, 2$. Observe that M_{j2} and $M_{(j+1)1}$, the queue lengths of queues $j2$ and $(j+1)1$, are not independent as the queues have simultaneous arrivals, which prohibits a product form expression for the joint probability of the number of jobs in the queues.

We have the following results for the sojourn times.

Lemma 3. *In the tandem network with one job type, the sojourn times of the jobs in the queues in the network are independent exponential random variables with rate $\mu_{j1}-\lambda$ for queues $j1$ and rate μ_{j2} for queues $j2$, $j=1,\ldots,J$.*

Proof. The sojourn times at queue $j2$, $j=1,\ldots,J$, are exponential random variables that are independent of the sojourn times of the jobs at the FIFO queues. Also observe that the arrival processes to the FIFO queues in the network with incubation queues and in the network without these incubation queues coincide. Thus, the sojourn times of the jobs at each of these J FIFO queues are independent exponential random variables (Kelly, 1979, Theorem 2.2). As a consequence the sojourn times at all the queues are independent exponential random variables. □

Theorem 4. *For the tandem network with one job type, the LST of the TAT can recursively be obtained as follows:*

$$\widetilde{\mathrm{TAT}}_J(s)=\widetilde{S}_{j1}(s)\widetilde{S}_{j2}(s), \tag{7a}$$

$$\widetilde{\mathrm{TAT}}_j(s)=\widetilde{S}_{j1}(s)\widetilde{S}_{j2}(s)\left(\frac{s}{s-\mu_{j2}}\widetilde{\mathrm{TAT}}_{j+1}(\mu_{j2})-\frac{\mu_{j2}}{s-\mu_{j2}}\widetilde{\mathrm{TAT}}_{j+1}(s)\right), \quad j=J-1,\ldots,1, \tag{7b}$$

$$\widetilde{\mathrm{TAT}}(s)=\widetilde{\mathrm{TAT}}_1(s), \tag{7c}$$

where $\widetilde{S}_{j1}(s)=\frac{\mu_{j1}-\lambda}{\mu_{j1}-\lambda+s}$ *and* $\widetilde{S}_{j2}(s)=\widetilde{B}_{j2}(s)=\frac{\mu_{j2}}{\mu_{j2}+s}$, $j=1,\ldots,J$.

Proof. From Lemma 1 we obtain

$$\widetilde{\mathrm{TAT}}_J(s)=\widetilde{S}_{j1}(s)\widetilde{S}_{j2}(s), \tag{8}$$

$$\mathrm{TAT}_j=S_{j1}+S_{j2}+\max\left\{0,\mathrm{TAT}_{j+1}-S_{j2}\right\}, \quad j=J-1,\ldots,1. \tag{9}$$

If A is exponentially distributed with rate μ, and A and S are independent random variables, then the LST of $W=\max\{S-A,0\}$ is:

$$\widetilde{W}(s)=\frac{s}{s-\mu}\widetilde{S}(\mu)-\frac{\mu}{s-\mu}\widetilde{S}(s), \tag{10}$$

see, e.g., (Adan & Resing, 2015, Section 7.5), where this result is derived for Lindley's equation. Lemma 3 implies that the sojourn times S_{j2} in the incubation queues are exponential random variables, independent of TAT_{j+1}, which allow us to use (10) to obtain (7b). □

Corollary 5. *The mean TAT is obtained as follows:*

$$\mathbb{E}[\mathrm{TAT}_J]=\mathbb{E}[S_{J1}]+\mathbb{E}[S_{J2}], \tag{11a}$$

$$\mathbb{E}[\mathrm{TAT}_j]=\mathbb{E}[S_{j1}]+\mathbb{E}[S_{j2}]\mathbb{P}\left(S_{j2}>\mathrm{TAT}_{j+1}\right)+\mathbb{E}[\mathrm{TAT}_{j+1}], \quad j=J-1,\ldots,1, \tag{11b}$$

$$\mathbb{E}[\mathrm{TAT}]=\mathbb{E}[\mathrm{TAT}_1], \tag{11c}$$

where $\mathbb{E}[S_{j1}]=\frac{1}{\mu_{j1}-\lambda}$ *and* $\mathbb{E}[S_{j2}]=\frac{1}{\mu_{j2}}$, $j=1,\ldots,J$.

Proof. We may obtain (11b) from (7b) by differentiation, or from (2b) by taking expectations as follows. If X, Y are independent random variables, and X is exponentially distributed, then

$$\mathbb{E}[\max\{X,Y\}]=\mathbb{E}[Y]+\mathbb{E}[\max\{X-Y,0\}]=\mathbb{E}[Y]+\mathbb{E}[X]\mathbb{P}(X>Y), \tag{12}$$

where, for X exponentially distributed with rate μ, we have used that

$$\begin{aligned}\mathbb{E}[\max\{X-Y,0\}]&=\int_0^\infty \mathrm{d}F_Y(y)\int_y^\infty (x-y)\mu e^{-\mu x}\mathrm{d}x\\&=\int_0^\infty e^{-\mu y}\mathrm{d}F_Y(y)\int_0^\infty x\mu e^{-\mu x}\mathrm{d}x=\mathbb{E}[X]\,\tilde{Y}(\mu),\end{aligned}$$

$$\begin{aligned}\mathbb{P}(X>Y)&=\int_0^\infty \mathrm{d}F_Y(y)\int_y^\infty \mu e^{-\mu x}\mathrm{d}x\\&=\int_0^\infty e^{-\mu y}\mathrm{d}F_Y(y)=\tilde{Y}(\mu).\end{aligned} \tag{13}$$

Inserting $X=S_{j2}$ and $Y=\mathrm{TAT}_{j+1}$ in (12) yields (11b). □

Remark 6. Observe from (13) that $\mathbb{P}\left(S_{j2}>\mathrm{TAT}_{j+1}\right)=\widetilde{\mathrm{TAT}}_{j+1}(\mu_{j2})$. This LST is obtained in Theorem 4, providing an explicit method to calculate Corollary 5.

Remark 7 (Generally distributed incubation times)**.** Observe that Theorem 4 and Corollary 5 require the incubation times at all nodes to be exponentially distributed. We may extend the results of the equilibrium distribution in (6) and the independence result in Lemma 3 to the network with generally distributed incubation times. For (6) observe that the infinite-server queue is insensitive to the distribution of the service time except for its mean (Taylor, 2011). For Lemma 3 observe that the service times at the queues are independent random variables.

3.3. Multiple job types and fixed routes

This section considers the general exponential network with multiple job types, fixed routes, exponential service times and Poisson arrivals, under the assumption that the service rates at the FIFO queues do not depend on the job types: $\mu_{j1}(c)=\mu_{j1}$, $j=1,\ldots,J$, $c=1,\ldots,C$. We first consider the special cases with zero incubation and zero service times.

Consider the case with zero incubation times. From (Kelly, 1979, Theorem 3.1) we obtain that $\{\mathcal{X}(t),\ t\in\mathbb{R}\}$ has unique product form equilibrium distribution

$$\boldsymbol{\pi}(\boldsymbol{x})=\prod_{j=1}^{J}\pi_{j1}(\boldsymbol{x}_{j1}), \quad \boldsymbol{x}=(\boldsymbol{x}_{11},\ldots,\boldsymbol{x}_{J1}), \tag{14}$$

$$\begin{aligned}\pi_{j1}(\boldsymbol{x}_{j1})&=(1-\rho_{j1})\prod_{l=1}^{m_{j1}}\frac{\lambda_0(x_{j1}(\ell))}{\mu_{j1}},\\ \boldsymbol{x}_{j1}&=(x_{j1}(1),\ldots,x_{j1}(m_{j1})), \quad j=1,\ldots,J.\end{aligned} \tag{15}$$

The Arrival Theorem (Kelly, 1979, p. 63) gives that the marginal distribution of the sojourn time in queue $j1$ is equal to the sojourn time as if it were an isolated M/M/1 queue with Poisson arrivals with rate λ_{j1}, $j=1,\ldots,J$. Hence, this marginal distribution is exponential with rate $\mu_{j1}-\lambda_{j1}$. The sojourn times in the queues are not independent in general, for example due to overtaking (Melamed, 1982). The TAT of a job of type c_i is

$$\mathrm{TAT}(c_i)=\sum_{j\in R(c)}S_{j1},$$

$$\text{so the mean TAT is}\quad \mathbb{E}[\mathrm{TAT}(c_i)]=\sum_{j\in R(c)}\frac{1}{\mu_{j1}-\lambda_{j1}}. \tag{16}$$

Now consider the case with zero service times at the FIFO queues. The route, and therefore its type, of a job c_i does not influence its TAT as its arrival results in $|R(c)|$ simultaneous arrivals to the incubation queues visited by class c jobs. Hence $\mathrm{TAT}(c_i)=\max_{j\in R(c)}\{S_{j2}(c)\}$, $i=1,\ldots,I(c)$, $c=1,\ldots,C$. As the incubation times are independent random variables:

$$\mathbb{P}(\mathrm{TAT}(c_i)\le x)=\mathbb{P}(\mathrm{TAT}(c)\le x)=\prod_{j\in R(c)}\left(1-\mathrm{e}^{-\mu_{j2}(c)x}\right), \quad x\ge 0. \tag{17}$$

The mean TAT of a job of type c_i is readily obtained as:

$$\mathbb{E}[\mathrm{TAT}(c_i)]=\int_0^\infty\left(1-\prod_{j\in R(c)}\left(1-\mathrm{e}^{-\mu_{j2}(c)x}\right)\right)\mathrm{d}x$$

$$= \int_0^\infty \sum_{r=1}^{|R(c)|} \sum_{\{\ell_1,\ldots,\ell_r\}\subseteq\{1,\ldots,J\}} (-1)^{r+1} \mathrm{e}^{-\left(\sum_{m=1}^r \mu_{\ell_m 2}(c)\right)x} \mathrm{d}x$$
$$= \sum_{r=1}^{|R(c)|} \sum_{\{\ell_1,\ldots,\ell_r\}\subseteq\{1,\ldots,J\}} \frac{(-1)^{r+1}}{\sum_{m=1}^r \mu_{\ell_m 2}(c)}. \quad (18)$$

The result from Choi & Park (1992) for the joint probability generating function of the queue lengths does not extend to this case as it requires that there is only one job class.

Now consider the general exponential network with multiple job types and fixed routes. Observe again that the arrival process to the FIFO queues in the network with incubation queues and in the network without these incubation queues coincide. As a consequence, the marginal distribution of the number of jobs in the FIFO queues is given in (14) and (15).

Remark 8 (Generally distributed incubation times)**.** For generally distributed incubation times, the marginal distribution of the number of jobs in the FIFO queues is given in (14) and (15).

For this general exponential network we do not have explicit results for the LST of the TAT. However, from the Arrival Theorem (Kelly, 1979, p. 63) we readily obtain that for job class c the marginal distribution of the sojourn at queue $j1$ is exponentially distributed with rate $\mu_{j1} - \lambda_{j1}$. As the incubation times are independent random variables, the sojourn time at queue $j2$ is exponentially distributed with rate $\mu_{j2}(c)$. The mean sojourn time at the queues is:

$$\mathbb{E}[S_{j1}(c)] = \frac{1}{\mu_{j1} - \lambda_{j1}} \quad \text{and} \quad \mathbb{E}[S_{j2}(c)] = \frac{1}{\mu_{j2}(c)}, \quad j = 1,\ldots,J. \quad (19)$$

Each job type proceeds along its fixed route through the nodes in $R(c)$ as if this route is a tandem network. This gives the following result.

Corollary 9. *The mean TAT for job type c_i is obtained as follows:*

$$\mathbb{E}[\mathrm{TAT}_{n(c_i,|R(c)|)}(c_i)] = \mathbb{E}[S_{n(c_i,|R(c)|)1}(c)] + \mathbb{E}[S_{n(c_i,|R(c)|)2}(c)], \quad (20a)$$

$$\mathbb{E}[\mathrm{TAT}_j(c_i)] = \mathbb{E}[S_{j1}(c)] + \mathbb{E}[S_{j2}(c)]\mathbb{P}\big(S_{j2}(c) > \mathrm{TAT}_{j+1}(c_i)\big) + \mathbb{E}[\mathrm{TAT}_{j+1}(c_i)], \quad j = n(c_i,|R(c)|-1),\ldots,n(c_i,1), \quad (20b)$$

$$\mathbb{E}[\mathrm{TAT}(c_i)] = \mathbb{E}[\mathrm{TAT}_{n(c_i,1)}(c_i)], \quad (20c)$$

with $i = 1,\ldots,I(c)$, $c = 1,\ldots,C$.

Corollary 9 is obtained from Lemma 1 by considering $\mathbb{E}[\max\{S_{j2}(c), \mathrm{TAT}_{j+1}(c_i)\}]$, $j = 1,\ldots,J$, by analogy with the result of Corollary 5. In contrast with the result of Corollary 5, we do not have explicit results for $\mathbb{P}\big(S_{j2}(c) > \mathrm{TAT}_{j+1}(c_i)\big)$ or the distribution of TAT_j, $j = 1,\ldots,J$. Therefore, Corollary 9 does not enable us to explicitly evaluate the mean TAT.

3.4. Optimal routing configuration

Let $\mu_{j1}(c) = \mu_{j1}$, $c = 1,\ldots,C$, $j = 1,\ldots,J$. Consider the case with zero incubation times. Combining (16) and (1):

$$f_{\mathrm{TAT}}(\boldsymbol{p}) = \sum_{c=1}^{C}\sum_{i=1}^{I(c)} \frac{p(c_i)\lambda_0(c)}{\lambda_0} \sum_{j\in R(c)} \frac{1}{\mu_{j1} - \lambda_{j1}} = \sum_{c=1}^{C} \frac{\lambda_0(c)}{\lambda_0} \sum_{j\in R(c)} \frac{1}{\mu_{j1} - \lambda_{j1}}. \quad (21)$$

For zero incubation times, the objective function does not depend on the routing configuration as each job of class c must visit all nodes in $R(c)$ and the load of the nodes is determined only by the arrival rate of jobs to the nodes. Optimal design of static routes is considered in Kameda & Zhang (1995); Shaler (2009).

For zero service times at the FIFO queues, the type c_i does not influence its TAT as its arrival results in $|R(c)|$ simultaneous arrivals to the incubation queues visited by class c jobs. As a consequence, the mean TAT of the jobs is the same for each routing configuration.

4. General inter-arrival, service, and incubation times

This section considers the mean TAT for the network with multiple job types, fixed routes and generally distributed inter-arrival, service and incubation times. The TAT distribution is not available in closed form. Thus, we do not have explicit results for the term $\mathbb{E}[\max\{S_{j2}(c), \mathrm{TAT}_{j+1}(c_i)\}]$ in (2b), which also prohibits explicit evaluation of the mean TAT. We propose a two step approach to approximate the mean TAT. First, in Section 4.1, we approximate $\mathbb{E}[\max\{S_{j2}(c), \mathrm{TAT}_{j+1}(c_i)\}]$, and subsequently in Section 4.2 we invoke the Queueing Network Analyzer (QNA). Section 4.3 considers optimization of the routing configuration via a Simulated Annealing (SA) approach.

4.1. Approximation of the mean TAT

Evaluation of the mean TAT is cumbersome for generally distributed service and incubation times since we do not have an explicit expression for $\mathbb{E}[\max\{S_{j2}(c), \mathrm{TAT}_{j+1}(c_i)\}]$ in (2b).

We first elaborate on bounds for the expectation $\mathbb{E}[\max\{X,Y\}]$ for independent and non-negative random variables X, Y. We have

$$\max\{\mathbb{E}[X], \mathbb{E}[Y]\} \le \mathbb{E}[\max\{X,Y\}], \quad (22)$$

$$\mathbb{E}[\max\{X,Y\}] \le \mathbb{E}[X] + \mathbb{E}[Y], \quad (23)$$

where (22) follows by Jensen's inequality since $f(X,Y) = \max(X,Y)$ is convex, and (23) follows from $\max\{X,Y\} = X + Y - \min\{X,Y\}$. If we further assume that X and Y are independent and that X is exponentially distributed we may evaluate the error in the lower bound (22) as

$$\mathbb{E}[\max\{X,Y\}] - \max\{\mathbb{E}[X], \mathbb{E}[Y]\} = \begin{cases} \mathbb{E}[X]\mathbb{P}(X > Y), & \text{if } \mathbb{E}[Y] \ge \mathbb{E}[X], \\ \mathbb{E}[Y] - \mathbb{E}[X]\mathbb{P}(X \le Y), & \text{if } \mathbb{E}[Y] < \mathbb{E}[X]. \end{cases} \quad (24)$$

Assuming that X and Y are independent and both exponentially distributed, we may sharpen the upper bound in (23) to

$$\mathbb{E}[\max\{X,Y\}] = \mathbb{E}[X] + \mathbb{E}[Y] - \frac{\mathbb{E}[X]\mathbb{E}[Y]}{\mathbb{E}[X] + \mathbb{E}[Y]}. \quad (25)$$

If $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ differ considerably in magnitude, then the upper and lower bound, (23) and (22), tend to be close to $\mathbb{E}[X]$ if $\mathbb{E}[X] >> \mathbb{E}[Y]$, and close to $\mathbb{E}[Y]$, otherwise, so that the lower bound $\max\{\mathbb{E}[X], \mathbb{E}[Y]\}$ may be a good approximation of $\mathbb{E}[\max\{X,Y\}]$.

To approximate $\mathbb{E}[\max\{S_{j2}(c), \mathrm{TAT}_{j+1}(c_i)\}]$, observe that, often, the expected incubation time at queue j is smaller than the expected TAT from the subsequent nodes on the route of a job: $\mathbb{E}[S_{j2}(c_i)] < \mathbb{E}[\mathrm{TAT}_\ell(c_i)]$, $j = n(c_i,r)$, $\ell = n(c_i,r+1)$, $i = 1,\ldots,I(c)$, $r = 1,\ldots,|R(c)|-1$, $c = 1,\ldots,C$. The longer the residual route of the job from node j, the larger the difference between these two expectations. In contrast, towards the end of a job's route the incubation time may outweigh the residual TAT. This supports the approximation $\mathbb{E}[\max\{S_{j2}(c), \mathrm{TAT}_{j+1}(c_i)\}] \approx \max\{\mathbb{E}[S_{j2}(c)], \mathbb{E}[\mathrm{TAT}_{j+1}(c_i)]\}$. In our numerical results we will also consider the upper bound in (23), and the approximation resulting from (25), see Section 5.2 and Section 5.3.

Approximation 10 (Approximation of the mean TAT)**.** For $i = 1, \ldots, I(c)$, $c = 1, \ldots, C$, let $\widehat{\overline{\text{TAT}}}(c_i)$ be determined as

$$\begin{aligned}\widehat{\overline{\text{TAT}}}_{n(c_i,|R(c)|)}(c_i) &= \mathbb{E}[S_{n(c_i,|R(c)|)1}(c)] + \mathbb{E}[S_{n(c_i,|R(c)|)2}(c)],\\ \widehat{\overline{\text{TAT}}}_j(c_i) &= \mathbb{E}[S_{j1}(c)] + \max\{\mathbb{E}[S_{j2}(c)], \widehat{\overline{\text{TAT}}}_{j+1}(c_i)\},\\ j &= n(c_i, |R(c)| - 1), \ldots, n(c_i, 1),\\ \widehat{\overline{\text{TAT}}}(c_i) &= \widehat{\overline{\text{TAT}}}_{n(c_i,1)}(c_i).\end{aligned}$$

4.2. Queueing Network Analyzer

Approximation 10 requires the mean sojourn times at the FIFO queues. The sojourn times at the FIFO queues are not affected by the incubation queues, see Remark 2. We will use the QNA to approximate the mean sojourn times at the FIFO queues in our network.

The QNA basically assumes that all FIFO queues are independent $GI/G/1$ queues for which the mean waiting time is approximated using a modified version of Kingman's approximation

$$\begin{aligned}\overline{W}_{j1} &= \frac{\tau_{j1}\rho_{j1}(\text{scv}_{aj1} + \text{scv}_{sj1})g_{j1}}{2(1-\rho_{j1})},\\ g_{j1} &= \begin{cases}\exp\left[-\frac{2(1-\rho_{j1})}{3\rho_{j1}}\frac{(1-\text{scv}_{aj1})}{\text{scv}_{aj1}+\text{scv}_{sj1}}\right], & \text{scv}_{aj1} < 1,\\ 1, & \text{scv}_{aj1} \geq 1,\end{cases}\end{aligned} \tag{27}$$

with scv_{aj1} the SCV of the aggregated inter-arrival times of the arrival process to queue $j1$, scv_{sj1} the SCV of the aggregated service time distribution at queue $j1$, τ_{j1} the mean service time of a random job at queue $j1$, $j = 1, \ldots, J$, see Whitt (1983b). For $\text{scv}_{aj1} < 1$, we obtain the Kraemer and Langenbach-Belz approximation (Krämer & Langenbach-Belz, 1976), and for $\text{scv}_{aj1} \geq 1$, (27) reduces to Kingman's approximation (Kingman, 1961). The QNA is included in Appendix A. We obtain the approximate mean sojourn time, $\overline{S}_{j1}$, of job type c_i at queue $j1$ as

$$\overline{S}_{j1}(c) = \overline{W}_{j1} + \mathbb{E}[B_{j1}(c)], \quad j = 1, \ldots, J, \ c = 1, \ldots, C. \tag{28}$$

We propose the following approximation for the mean TAT of jobs of type c_i that is obtained using the QNA for the mean sojourn times in the FIFO queues in Approximation 10.

Approximation 11 (Approximation of the mean TAT using the QNA)**.** For $i = 1, \ldots, I(c)$, $c = 1, \ldots, C$, let $\overline{\text{TAT}}(c_i)$ be determined as

$$\overline{\text{TAT}}_{n(c_i,|R(c)|)}(c_i) = \overline{S}_{n(c_i,|R(c)|)1}(c) + \mathbb{E}[S_{n(c_i,|R(c)|)2}(c)], \tag{29a}$$

$$\begin{aligned}\overline{\text{TAT}}_j(c_i) &= \overline{S}_{j1}(c) + \max\{\mathbb{E}[S_{j2}(c)], \overline{\text{TAT}}_{j+1}(c_i)\},\\ j &= n(c_i, |R(c)| - 1), \ldots, n(c_i, 1),\end{aligned} \tag{29b}$$

$$\overline{\text{TAT}}(c_i) = \overline{\text{TAT}}_{n(c_i,1)}(c_i). \tag{29c}$$

The QNA is exact for a network with one job type, Poisson arrivals and exponential service times (Whitt, 1983b). In Section 5.3, we investigate the quality of Approximation 11 using both interchanging of mean and max and the QNA for the relevant range of parameters in our network.

4.3. Simulated Annealing for optimal routing configuration in the QNA

This section introduces a Simulated Annealing (SA) approach to obtain a near-optimal routing configuration in the QNA, and an approximate upper bound on the optimality gap, for the network with general inter-arrival, service and incubation times and multiple job types.

For real life instances, such as the clinical chemistry laboratory case, determining the optimal routing configuration from the mathematical program (1) with TAT replaced by $\overline{\text{TAT}}$ is infeasible considering the size of its solution space and its non-convex objective function. In Appendix B we provide a counterexample for convexity of the objective function.

We follow the general SA setting, which requires a feasible initial solution, neighborhood, acceptance probabilities, and a cooling scheme (van Laarhoven & Aarts, 1987). An initial solution may be obtained by letting all job types visit their required nodes in decreasing order of the mean incubation times. Our numerical experiments show that this is a good rule of thumb that is often close to the optimal routing configuration. Alternatively, if an initial network design is already in place, we may select its routing configuration as the initial solution. The neighboring solutions are constructed such that they lie close to the current solution and that it is possible to reach each possible valid routing configuration. Neighboring solutions are obtained by adding a uniformly distributed value to the current fractions: $p(c_i) := p(c_i) + \text{Unif}(-0.01, 0.01)$, where values larger than 1 and smaller than 0 are rounded off to 1 and 0. The resulting fractions are normalized such that for each job class these fractions sum up to 1:

$$p(c_i) := \frac{p(c_i)}{\sum_{i=1}^{I(c)} p(c_i)}, \qquad i = 1, \ldots, I(c), \ c = 1, \ldots, C.$$

Acceptance of a neighboring solution depends on the acceptance probabilities which are a function of the value of the current solution (f_{current}), the value of the neighboring solution (f_{neighb}) and the current cooling parameter (d):

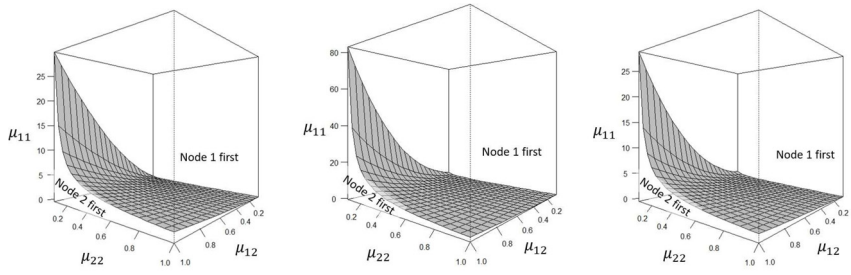
$$\mathbb{P}_{\text{accept}}(f_{\text{neighb}}, f_{\text{current}}, d) = \begin{cases}1, & \text{if } f_{\text{neighb}} \leq f_{\text{current}},\\ e^{(f_{\text{current}} - f_{\text{neighb}})/d}, & \text{if } f_{\text{neighb}} > f_{\text{current}}.\end{cases}$$

The SA algorithm can accept routing configurations that result in a higher objective value to avoid getting stuck in a local minimum. Closer to the stopping value for d, the algorithm is less likely to accept a routing configuration that is worse. The cooling scheme is chosen such that the fraction of accepted transitions for the initial value of the cooling parameter d is approximately equal to 1. After a fixed number of k steps (in literature referred to as fixed Markov chain length), the cooling parameter will be decreased by a fixed factor. Both the decrease factor and the number of steps k will depend on the problem instance, see Sections 5.3 and 5.4.2.

An approximate upper bound on the optimality gap of a near-optimalsolution is obtained via comparison with the lower bound of the objective function value in the QNA. A lower bound on the objective value in the QNA is obtained as follows. Observe that τ_{j1}, ρ_{j1} and scv_{sj1} do not depend on the routing configuration as the total arrival rate and service duration at the modules are the same for each configuration. Observe from (27) that minimum waiting times are then obtained by minimizing scv_{aj1} and g_{j1}, i.e., by assuming deterministic inter-arrival times of the jobs at the nodes. Letting jobs route through the nodes from highest to lowest incubation time and using these minimum waiting times at the FIFO queues yields a lower bound on the objective function value and, hence, an approximate upper bound on the optimality gap.

5. Optimal routing configurations

We start this section with an illustration of the relation between service times and incubation times for the network with one job type, Poisson arrivals, exponential service and incubation times in Section 5.1. Section 5.2 investigates the impact of Approximation 10 on the mean TAT and the routing configuration. Section 5.3 provides numerical results on the accuracy of our numerical procedure using the QNA and SA for general networks.



(a) Two nodes (b) Three nodes, $\mu_{32} = 1$ (c) Three nodes, $\mu_{32} = 3$

Fig. 2. Switching curve for the network of 2 and 3 nodes with $\lambda = 0.5$ and $\mu_{21} = 1$, and $\mu_{31} = 1$.

Section 5.4 applies our method to design the optimal routing configuration for one analyzer line of the laboratory of the Erasmus MC.

5.1. Impact of service and incubation times on the routing configuration

The impact of the service and incubation times on the optimal routing configuration is considered for a network of two and of three nodes with one job type, Poisson arrivals, exponential service and incubation times.

First, consider a network of two nodes, node 1 and node 2. Let TAT_{12} and TAT_{21} denote the TAT when the jobs visit the nodes in the sequence 12 and 21. From Corollary 5,

$$\mathbb{E}[\text{TAT}_{12}] = \mathbb{E}[S_{11}] + \mathbb{E}[S_{12}]\mathbb{P}(S_{12} > S_{21} + S_{22}) + \mathbb{E}[S_{21}] + \mathbb{E}[S_{22}],$$

with similar expression for $\mathbb{E}[\text{TAT}_{21}]$. As all random variables are exponential, we readily obtain

$$\mathbb{P}(S_{12} \leq S_{21} + S_{22}) = \frac{\mu_{12}(\mu_{12} + \mu_{21} + \mu_{22} - \lambda)}{(\mu_{12} + \mu_{21} - \lambda)(\mu_{12} + \mu_{22})}, \tag{30}$$

so that

$$\mathbb{E}[\text{TAT}_{12}] \leq \mathbb{E}[\text{TAT}_{21}] \iff \mu_{22}(\mu_{22} + \mu_{11} - \lambda) - \mu_{12}(\mu_{12} + \mu_{21} - \lambda) \geq 0. \tag{31}$$

The switching curve $\mathbb{E}[\text{TAT}_{12}] = \mathbb{E}[\text{TAT}_{21}]$ is depicted in Fig. 2 a for $\lambda = 0.5$ and $\mu_{21} = 1$. Above the curve it is optimal to first visit node 1. The switching curve is symmetrical in the nodes. The seemingly larger volume where first visiting node 1 is preferred is due to setting $\mu_{21} = 1$. As a rule of thumb, if μ_{11} does not deviate much from μ_{21}, then it is optimal to first visit the node with the largest incubation time, which is in agreement with intuition.

Now consider a network of 3 nodes. We will compare $\mathbb{E}[\text{TAT}_{123}]$ and $\mathbb{E}[\text{TAT}_{213}]$ denoting the mean TAT when the jobs visit the nodes in the sequence 123 and 213. All expressions in Corollary 5 may be explicitly evaluated to obtain a switching curve $\mathbb{E}[\text{TAT}_{123}] = \mathbb{E}[\text{TAT}_{213}]$ that is depicted for $\lambda = 0.5$ and $\mu_{21} = \mu_{31} = 1$ and two values for μ_{32} in Figs. 2 b and 2 c. Observe that these curves resemble the switching curve for the network of two nodes. Visiting node 1 first becomes more favorable when μ_{11} increases. Again, in agreement with intuition, among nodes 1 and 2, it seems optimal to first visit the node with the largest incubation time.

The last two nodes in the network of three nodes have Poisson arrivals (Burke's theorem (Kelly, 1979, Theorem 2.1)), and may therefore be considered as a network of two nodes that should be arranged such that the node with the largest incubation time is visited first. As a consequence, as a rule of thumb, also in the network of three nodes *the nodes should be visited in decreasing order of the incubation times.*

5.2. Accuracy of approximation 10

We first investigate the error in Approximation 10 for a tandem network of five nodes. Then, we investigate the effects of the error in this approximation on the routing configuration.

Section 4.1 indicates that the error $\mathbb{E}[\text{TAT}(c_i)] - \widehat{\text{TAT}}(c_i)$ in the approximation is mainly incurred in the last nodes along a route when the incubation times are identically and exponentially distributed. Following up on this claim, for a tandem of length 5 in which all jobs visit the nodes in the sequence 12345, Table 1 presents $\mathbb{P}(S_{j2} > \text{TAT}_{j+1})$, $\widehat{\text{TAT}}_j$, and the error in Approximation 10, $j = 1, \ldots, 4$, for $\lambda = 0.5$, and 7 scenarios for the exponential service duration in the queues. We omitted $\widehat{\text{TAT}}_5$ as Approximation 10 is exact for the last node in the tandem. In scenarios 1, 2, and 3, the incubation times are all exponentially distributed with the same rate. Observe that $\mathbb{P}(S_{j2} > \text{TAT}_{j+1})$ increases in j. The error in the approximation $\widehat{\text{TAT}}_j$ for scenario 1 is equal to 0.17 for node 4; at node 3 the error increases by 0.06 to 0.23; at node 2 the error increases by 0.02 to 0.25; and at node 1 by 0.02 to 0.27. This shows that for a tandem with identically distributed incubation times the error in $\widehat{\text{TAT}}_j$ is mainly incurred in the last nodes on the route. This result seems to extend to tandems in which the incubation rates μ_{j2} are similar, as illustrated in scenarios 4 and 5. These results support Approximation 10 that interchanges max and expectation. In scenarios 6 and 7, jobs visit the nodes in order of substantially increasing incubation rates μ_{j2}. The monotonicity in $\mathbb{P}(S_{j2} > \text{TAT}_{j+1})$ breaks down as the long incubation times at the initial nodes of the tandem extend beyond the mean sojourn time along the FIFO queues.

For the networks of Section 5.1 we now compare the switching curves from the exact mean TAT and Approximation 10. For the network with two nodes, let $\widehat{\text{TAT}}_{ij}$ denote the approximated TAT when node i is visited first. Approximation 10 gives

$$\widehat{\text{TAT}}_{12} = \frac{1}{\mu_{11} - \lambda} + \max\left\{\frac{1}{\mu_{12}}, \frac{1}{\mu_{21} - \lambda} + \frac{1}{\mu_{22}}\right\},$$

with similar expression for $\widehat{\text{TAT}}_{21}$. The switching curve under Approximation 10 is $\mu_{12} = \mu_{22}$. Fig. 3 a depicts the switching curves for multiple values of μ_{11}. The approximate switching curve lies in the area where node 2 should be visited first under the exact solution. Hence, the error between the switching curves is due to visiting node 1 first instead of node 2 first as indicated by the exact switching curves. The error in the TAT approximation is maximum at the curve $\mu_{12} = \mu_{22}$. On the switching curve $\mu_{12} = \mu_{22}$ for $\lambda = 0.5$, $\mu_{11} = 0.6$ and $\mu_{21} = 1$, Table 2 a presents the exact values $\mathbb{E}[\text{TAT}_{12}]$, $\mathbb{E}[\text{TAT}_{21}]$, as well as lower bound $\widehat{\text{TAT}}$ obtained from (22), $\widehat{\text{TAT}}_{12,\text{exp}}$, and $\widehat{\text{TAT}}_{21,\text{exp}}$ under approximation (25), and the upper bound $\widehat{\text{TAT}}_{\text{UB}}$ obtained from (23). The approximation $\widehat{\text{TAT}}$ is better for $\mathbb{E}[\text{TAT}_{21}]$ than for $\mathbb{E}[\text{TAT}_{12}]$ as $\mu_{12} = \mu_{22}$ lies in the area where node 2 must be visited first. Observe that the difference between $\mathbb{E}[\text{TAT}_{12}]$, $\mathbb{E}[\text{TAT}_{21}]$ and $\widehat{\text{TAT}}$ decreases in μ_{12}. To

Table 1
The error $\mathbb{E}[\text{TAT}_j] - \widehat{\text{TAT}}_j$. Scenarios 1–3: $\mu_{j2} = 1$, scenarios 4–7: $\mu_{j1} = 1$, $j = 1, \ldots, 5$.

Scen.:	(μ_{j1}, μ_{j2})					$\mathbb{P}(S_{j2} > \text{TAT}_{j+1})$				$\widehat{\text{TAT}}_j$ + error			
	j=1	2	3	4	5	1	2	3	4	1	2	3	4
1	(1,1)	(1,1)	(1,1)	(1,1)	(1,1)	0.013	0.027	0.060	0.167	11.00+0.27	9.00+0.25	7.00+0.23	5.00+0.17
2	(0.8,1)	(0.9,1)	(1,1)	(1.1,1)	(1.2,1)	0.014	0.034	0.081	0.206	11.93+0.33	8.60+0.32	6.10+0.29	4.10+0.21
3	(1.2,1)	(1.1,1)	(1,1)	(0.9,1)	(0.8,1)	0.010	0.017	0.037	0.115	11.93+0.18	10.50+0.17	8.83+0.15	6.83+0.12
4	(1,1.2)	(1,1.1)	(1,1)	(1,0.9)	(1,0.8)	0.008	0.020	0.055	0.168	11.25+0.27	9.25+0.26	7.25+0.24	5.25+0.19
5	(1,0.8)	(1,0.9)	(1,1)	(1,1.1)	(1,1.2)	0.022	0.034	0.065	0.163	10.83+0.28	8.83+0.25	6.83+0.21	4.83+0.15
6	(1,0.2)	(1,0.4)	(1,0.6)	(1,0.8)	(1,1)	0.222	0.129	0.129	0.214	11.00+1.91	9.00+0.8	7.00+0.48	5.00+0.27
7	(1,0.1)	(1,0.3)	(1,0.5)	(1,0.7)	(1,0.9)	0.413	0.183	0.157	0.234	12.00+4.50	9.11+1.26	7.11+0.65	5.11+0.33

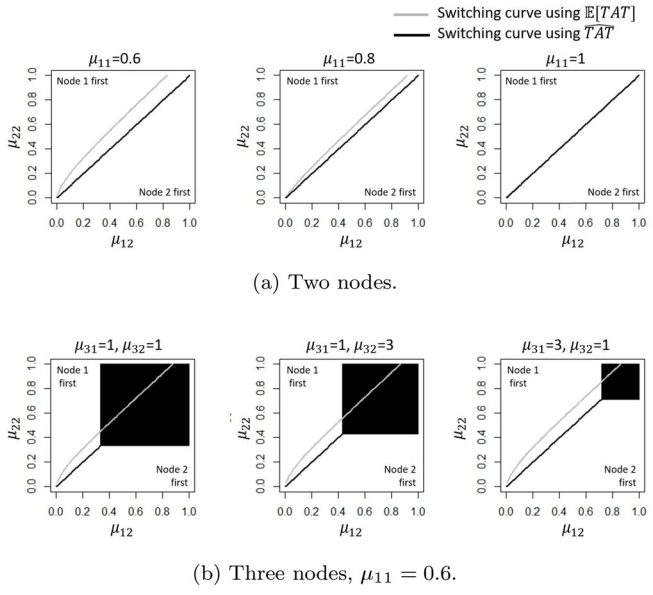
Table 2
TAT values on the switching curve $\mu_{12} = \mu_{22}$ for $\lambda = 0.5$, $\mu_{11} = 0.6$ and $\mu_{21} = \mu_{31} = \mu_{32} = 1$.

(a) Network with two nodes.

μ_{12}	$\mathbb{E}[\text{TAT}_{12}]$	$\mathbb{E}[\text{TAT}_{21}]$	$\widehat{\text{TAT}}$	$\widehat{\text{TAT}}_{12,\text{exp}}$	$\widehat{\text{TAT}}_{21,\text{exp}}$	$\widehat{\text{TAT}}_{\text{UB}}$
0.2	18.79	17.83	17.00	19.08	18.25	22.00
0.4	15.19	14.75	14.50	15.39	14.92	17.00
0.6	14.05	13.79	13.67	14.19	13.88	15.33
0.8	13.49	13.32	13.25	13.60	13.38	14.50
1	13.17	13.05	13.00	13.25	13.08	14.00

(b) Network with three nodes.

μ_{12}	$\mathbb{E}[\text{TAT}_{123}]$	$\mathbb{E}[\text{TAT}_{213}]$	$\widehat{\text{TAT}}$	$\widehat{\text{TAT}}_{123,\text{exp}}$	$\widehat{\text{TAT}}_{213,\text{exp}}$	$\widehat{\text{TAT}}_{\text{UB}}$
0.2	19.27	18.58	17.00	20.03	19.31	25.00
0.4	16.59	16.21	15.00	16.86	16.51	20.00
0.6	15.84	15.59	15.00	15.98	15.78	18.33
0.8	15.51	15.34	15.00	15.60	15.47	17.50
1	15.35	15.22	15.00	15.41	15.32	17.00



(a) Two nodes.

(b) Three nodes, $\mu_{11} = 0.6$.

Fig. 3. Switching curves for the network with $\lambda = 0.5$ and $\mu_{21} = 1$.

understand this, observe that the term $\mathbb{E}[S_{12}]\mathbb{P}(S_{12} \leq S_{21} + S_{22})$ in (30) is decreasing in μ_{12}. The upper bound $\widehat{\text{TAT}}_{\text{UB}}$ seems less accurate, while $\widehat{\text{TAT}}_{12,\text{exp}}$, and $\widehat{\text{TAT}}_{21,\text{exp}}$ seem to overestimate the true values and are a good approximation of these values.

We now consider TAT_{123} and TAT_{213} for the network with three nodes. The switching curve under Approximation 10 is $\mu_{12} = \mu_{22}$ if $\max\{1/\mu_{12}, 1/\mu_{22}\} \geq 1/(\mu_{31} - \lambda) + 1/\mu_{32}$, otherwise the order of visiting nodes 1 and 2 is irrelevant. Fig. 3 b depicts these switching curves for three choices of μ_{31} and μ_{32}. The black square indicates that for these parameters it is irrelevant in which order nodes 1 and node 2 are visited. The line $\mu_{12} = \mu_{22}$ is a good approximation of the true switching curve. Table 2 b presents the values of $\mathbb{E}[\text{TAT}_{123}]$, $\mathbb{E}[\text{TAT}_{213}]$, $\widehat{\text{TAT}}$, $\widehat{\text{TAT}}_{123,\text{exp}}$, $\widehat{\text{TAT}}_{213,\text{exp}}$ and $\widehat{\text{TAT}}_{\text{UB}}$ on $\mu_{12} = \mu_{22}$ for $\lambda = 0.5$, $\mu_{11} = 0.6$ and $\mu_{21} = \mu_{31} = \mu_{32} = 1$. As for two nodes, the difference between $\mathbb{E}[\text{TAT}_{123}]$, $\mathbb{E}[\text{TAT}_{213}]$, and $\widehat{\text{TAT}}$ decreases in μ_{12}. Conclusions on the accuracy of $\widehat{\text{TAT}}_{\text{UB}}$ $\widehat{\text{TAT}}_{123,\text{exp}}$, and $\widehat{\text{TAT}}_{213,\text{exp}}$ coincide with the two node case.

Observe from Tables 2 a, and 2 b that it seems optimal *for equal incubation times $\mu_{12} = \mu_{22}$ to visit the nodes in increasing order of the service times at the FIFO queues*, which is intuitively clear as this results in an earlier start of the first incubation time.

5.3. Accuracy of approximation 11 and optimal routing configurations

This section considers the accuracy of Approximation 11 using objective function (1) with TAT replaced by $\widehat{\text{TAT}}$ to obtain optimal routes by comparison with discrete-event simulation (DES), where we have used the replication/deletion approach using Welch's graphical method (Law, 2015). Subsequently, we will study performance of our SA approach to determine near-optimal routes. Approximations using (23) or (25) are considered for the scenarios with non-zero incubation times.

Consider a network of three nodes with two job classes and twelve possible job types, along with six scenarios for inter-arrival, service and incubation times as displayed in Table 3. We will focus on the impact of the SCV of the inter-arrival and service times and the mean of the incubation times. Therefore, in our experiments, the incubation times are deterministic with different values for their means, the service times have mean $\mathbb{E}[B_{j1}(c)] = 1$, $c = 1, 2$, $j = 1, 2, 3$, and different SCV, the inter-arrival times have different SCV, where these SCVs are chosen as deterministic (SCV = 0), exponential (SCV = 1), or log-normal (SCV $\neq$ 0, 1). For example, Scenario 4 refers to the network in which both job classes have Poisson arrivals, service times have SCV 0, 1, and 2 at the three FIFO queues, and incubation times are 8, 4, 1, 1, 4, and 8 time units. In our experiments we vary the load at the FIFO queues via four cases of the arrival rates, ranging from A: $\lambda_0(1) = \lambda_0(2) = 0.2$ (low–moderate load) to D: $\lambda_0(1) = \lambda_0(2) = 0.45$ (high load).

Motivation for the selection of the scenarios is as follows. Scenarios 1 and 2 contain the FIFO queues only and are included to zoom in on the accuracy of the QNA in the relevant range of system parameters for the FIFO queues. The QNA is exact for the mean waiting times in Scenario 3 with exponential service times (Whitt, 1983b). This scenario focuses on the impact of the incubation times on the optimal routes. Scenarios 3, 4 and 5 consider the impact of variability of the inter-arrival and service times. Scenarios 1, 4 and 6 focus on the impact of the incubation times. Note that the sojourn time of a job along its path through the FIFO queues is identical in Scenarios 1, 4, and 6, as well as in Scenarios 2 and 5, see Remark 2.

Table 4 presents a comparison of the mean TAT along the FIFO queues from our QNA approximation and DES including 95% confidence interval (CI) for the case in which job type 1 passes

Table 3
Job classes, job types and scenarios in the numerical experiments.

(a) Job classes.

c	$R(c)$
1	{1,2,3}
2	{1,2,3}

(b) Job types.

c_i	Route	c_i	Route
1_1	1,2,3	2_1	1,2,3
1_2	1,3,2	2_2	1,3,2
1_3	2,1,3	2_3	2,1,3
1_4	2,3,1	2_4	2,3,1
1_5	3,1,2	2_5	3,1,2
1_6	3,2,1	2_6	3,2,1

(c) Scenarios.

Scenario:	1	2	3	4	5	6
scv(1)	1	0	1	1	0	1
scv(2)	1	2	1	1	2	1
scv_{s11}	0	1	1	0	1	0
scv_{s21}	1	1	1	1	1	1
scv_{s31}	2	1	1	2	1	2
$\mathbb{E}[B_{12}(1)]$	0	0	8	8	8	0
$\mathbb{E}[B_{22}(1)]$	0	0	4	4	4	8
$\mathbb{E}[B_{32}(1)]$	0	0	1	1	1	0
$\mathbb{E}[B_{12}(2)]$	0	0	1	1	1	0
$\mathbb{E}[B_{22}(2)]$	0	0	4	4	4	8
$\mathbb{E}[B_{32}(2)]$	0	0	8	8	8	0

the queues in the sequence 123 and type 2 in the sequence 123 ($p(1_1) = p(2_1) = 1$) or 321 ($p(1_1) = p(2_6) = 1$) for scenarios 1, 4 and 6 (top part) and scenarios 2 and 5 (bottom part). These tables also list the mean waiting times at the queues to provide more detailed insight into the origin of the approximation error. Note that the mean waiting times at the FIFO queues in Scenarios 2 and 5 for route $p(1_1) = 1$ and $p(2_1) = 1$ coincide since the service times at these queues are identically distributed and all the jobs follow the same route through these three queues.

Accuracy of the QNA depends on the quality of the approximation of non-renewal arrival and departure processes by renewal processes (Whitt, 1983a). The QNA uses several heavy traffic results (Whitt, 1983b), which carries over in the observation that the QNA performs better for load D than load A. In all considered cases, the QNA approximation of $f_{\text{TAT}}(\boldsymbol{p})$ lies within 10% of the DES value. The error in the QNA for Scenarios 1, 4 and 6 seems to a large extent be due to the waiting time errors for FIFO queue 2, which is explained by the error in the SCV of the departure process approximation from queue 1. Observe that larger relative errors in the waiting times are mainly incurred at queues with low–moderate load. As the waiting times for low–moderate load are relatively small, the influence of this approximation error on the TAT is very limited. We conclude that the QNA provides a sufficiently accurate approximation of the mean TAT.

We proceed to determine optimal routing configurations for the network with job types and scenarios listed in Table 3 under loads A–D. In the experiments, for the SA algorithm the initial cooling parameter was set to 4; the stopping value to 0.00005; the decrease factor to 0.995 and the Markov chain length to 75. Experiments were conducted on an Intel Core i7-7700HQ 2.80GHz processing system with 16GB of RAM. The run-time of this cooling scheme was about 9 minutes.

Table 5 presents the near-optimal routing configurations from the QNA for the 24 scenario/load combinations, the corresponding value function $f_{\text{TAT}}(\boldsymbol{p})$, and its lower bound (LB) as described in Section 4.3. Observe that the optimal routing configuration may be deterministic, e.g., Scenario 4A–C, may randomize jobs over different routes with fixed probabilities, e.g., Scenario 2, or may have these probabilities in an interval, e.g., Scenario 3A. Randomized routes are typically obtained for scenarios with identically distributed service times, where the level of randomization increases with the load, see Scenarios 2 and 5. Routing in increasing order of service time variability is preferred for Scenario 1C–D, which is in accordance with the heuristic of Suresh & Whitt (1988), as Scenario 1 has incubation times equal to 0. Routing from high to low incubation times is optimal for scenarios with identically distributed inter-arrival times and varying service times at the FIFO queues, with some randomization for high loads, see Scenarios 3 and 4.

Table 5 includes the lower bound of $f_{\text{TAT}}(\boldsymbol{p})$ assuming deterministic inter-arrival times, see Section 4.3. This lower bound

Table 4
Comparison of TAT and waiting times from QNA and DES including 95% CI.

	QNA	DES	QNA	DES	QNA	DES	QNA	DES
$\lambda_0(c)$	A: 0.2		B: 0.3		C: 0.4		D: 0.45	
Scenarios 1, 4 and 6: $p(1_1) = 1$ and $p(2_1) = 1$								
$f_{\text{TAT}}(\boldsymbol{p})$	4.91	4.79±0.035	7.11	6.93±0.077	13.51	13.75±0.255	26.39	27.92±0.792
$\mathbb{E}[W_{11}]$	0.33	0.34±0.004	0.75	0.75±0.007	2.00	2.00±0.031	4.50	4.58±0.217
$\mathbb{E}[W_{21}]$	0.62	0.49±0.009	1.26	1.10±0.035	2.89	2.97±0.090	5.95	6.75±0.345
$\mathbb{E}[W_{31}]$	0.96	0.96±0.042	2.10	2.09±0.071	5.62	5.78±0.205	12.94	13.59±0.657
Scenarios 1, 4 and 6: $p(1_1) = 1$ and $p(2_6) = 1$								
$f_{\text{TAT}}(\boldsymbol{p})$	5.00	4.92±0.046	7.51	7.38±0.045	15.06	14.75±0.281	30.19	30.38±0.776
$\mathbb{E}[W_{11}]$	0.33	0.33±0.005	0.75	0.75±0.022	2.00	1.94±0.077	4.51	4.37±0.070
$\mathbb{E}[W_{21}]$	0.67	0.63±0.034	1.51	1.46±0.052	4.06	3.84±0.180	9.18	9.05±0.323
$\mathbb{E}[W_{31}]$	1.00	0.97±0.027	2.25	2.17±0.040	6.00	5.98±0.383	13.51	13.95±0.925
Scenarios 2 and 5: $p(1_1) = 1$ and $p(2_1) = 1$								
$f_{\text{TAT}}(\boldsymbol{p})$	5.00	4.56±0.012	7.50	6.92±0.050	15.00	14.25±0.330	30.00	28.80±0.852
$\mathbb{E}[W_{11}]$	0.67	0.46±0.004	1.50	1.17±0.028	4.00	3.52±0.114	9.00	8.24±0.378
$\mathbb{E}[W_{21}]$	0.67	0.53±0.006	1.50	1.34±0.020	4.00	3.82±0.116	9.00	8.72±0.441
$\mathbb{E}[W_{31}]$	0.67	0.56±0.008	1.50	1.41±0.013	4.00	3.91±0.135	9.00	8.85±0.272
Scenarios 2 and 5: $p(1_1) = 1$ and $p(2_6) = 1$								
$f_{\text{TAT}}(\boldsymbol{p})$	4.99	4.52±0.018	7.47	6.82±0.084	14.94	14.08±0.417	29.92	29.21±0.640
$\mathbb{E}[W_{11}]$	0.58	0.50±0.007	1.24	1.20±0.029	3.08	3.39±0.129	6.67	7.97±0.146
$\mathbb{E}[W_{21}]$	0.67	0.52±0.015	1.50	1.31±0.038	4.00	3.77±0.149	9.00	8.87±0.509
$\mathbb{E}[W_{31}]$	0.73	0.50±0.010	1.73	1.31±0.026	4.86	3.93±0.202	11.16	9.38±0.832

Table 5
Routing configurations for the network with 6 job types.

Scenario	1	2	3	4	5	6	Scenario	1	2	3	4	5	6
$\lambda_0(c)$	A: 0.2						$\lambda_0(c)$	B: 0.3					
$p(1_1)$	0	1	$p(1_1)$	1	1	0	$p(1_1)$	0	1	1	1	1	0
$p(1_2)$	0.28	0	$p(1_2)$	0	0	0	$p(1_2)$	0	0	0	0	0	0
$p(1_3)$	0.72	0	0	0	0	$p(1_3)$	$p(1_3)$	1	0	0	0	0	$p(1_3)$
$p(1_4)$	0	0	0	0	0	$p(1_4)$	$p(1_4)$	0	0	0	0	0	$p(1_4)$
$p(1_5)$	0	0	0	0	0	0	$p(1_5)$	0	0	0	0	0	0
$p(1_6)$	0	0	0	0	0	0	$p(1_6)$	0	0	0	0	0	0
$p(2_1)$	0	0	0	0	0	0	$p(2_1)$	0	0	0	0	0	0
$p(2_2)$	0.30	0	0	0	0	0	$p(2_2)$	0	0	0	0	0	0
$p(2_3)$	0.70	0.46	0	0	0	$p(2_3)$	$p(2_3)$	1	0.06	0	0	0	$p(2_3)$
$p(2_4)$	0	0.01	0	0	0	$p(2_4)$	$p(2_4)$	0	0.20	0	0	0	$p(2_4)$
$p(2_5)$	0	0.02	$p(2_5)$	0	0	0	$p(2_5)$	0	0.61	0	0	0	0
$p(2_6)$	0	0.51	$p(2_6)$	1	1	0	$p(2_6)$	0	0.13	1	1	1	0
$f_{TAT}(\boldsymbol{p})$	4.95	4.94	9.67	9.67	9.66	9.67	$f_{TAT}(\boldsymbol{p})$	7.26	7.29	10.5	10.50	10.48	10.50
LB	3.53	3.37	9.12	9.20	9.12	9.12	LB	4.68	4.44	9.48	9.60	9.48	9.48
$\lambda_0(c)$	C: 0.4						$\lambda_0(c)$	D: 0.45					
$p(1_1)$	1	1	1	1	1	0	$p(1_1)$	1	1	$p(1_1)$	1	1	0
$p(1_2)$	0	0	0	0	0	0	$p(1_2)$	0	0	0	0	0	0
$p(1_3)$	0	0	0	0	0	1	$p(1_3)$	0	0	$p(1_3)$	0	0	1
$p(1_4)$	0	0	0	0	0	0	$p(1_4)$	0	0	0	0	0	0
$p(1_5)$	0	0	0	0	0	0	$p(1_5)$	0	0	0	0	0	0
$p(1_6)$	0	0	0	0	0	0	$p(1_6)$	0	0	0	0	0	0
$p(2_1)$	1	0.22	0	0	0	0	$p(2_1)$	1	0.19	0	0	0	0
$p(2_2)$	0	0.09	0	0	0	0	$p(2_2)$	0	0.14	0	0	0.07	0
$p(2_3)$	0	0.12	0	0	0	1	$p(2_3)$	0	0.15	0	0	0	1
$p(2_4)$	0	0.13	0	0	0	0	$p(2_4)$	0	0.14	$p(2_4)$	0	0.38	0
$p(2_5)$	0	0.29	0	0	0.14	0	$p(2_5)$	0	0.22	0	0.85	0.14	0
$p(2_6)$	0	0.15	1	1	0.86	0	$p(2_6)$	0	0.16	$p(2_6)$	0.15	0.41	0
$f_{TAT}(\boldsymbol{p})$	13.51	14.17	16.00	16.06	15.89	13.89	$f_{TAT}(\boldsymbol{p})$	26.39	27.99	31.00	30.34	29.76	26.94
LB	8.37	8.08	10.69	11.03	10.69	10.69	LB	15.85	15.54	16.54	17.85	16.54	15.85

Table 6
$f_{TAT}(\boldsymbol{p})$ using QNA for the current lab load (100%) and load increase to 120%, 140% and 160%.

Load	Historic	Optimal	High to low	Low to high
100%	857.93 s	841.42 s	841.42 s	903.75 s
120%	884.78 s	867.87 s	867.87 s	941.21 s
140%	938.98 s	922.65 s	923.15 s	995.86 s
160%	1152.52 s	1097.22 s	1149.97 s	1270.07 s

underestimates the waiting times and clearly is more accurate with reducing load. The lower bound provides a good benchmark for loads A and B for scenarios with high incubation times (Scenarios 3–6). Table C.7 provides further details on the QNA and DES objective values for three routes under loads B and D, the near-optimal routing configuration displayed in Table 5 (O), routing from high to low incubation time (H), and the deterministic version of O in which the largest $p(c_i)$ per class i is set to 1 (D). Table C.7 underpins the accuracy of our approach using Approximation 11 as for all routing configurations (O, H, D) the approximation for $f_{TAT}(\boldsymbol{p})$ lies within 10% of the DES value. This table also reveals that the optimal routing configuration often routes from high to low incubation times. Furthermore, the deterministic routing configuration (D) is a good approximation of our near-optimal routing configuration.

Table C.8 provides a comparison of the optimal routing configurations and objective function values using the QNA under three approximations, where O refers to the lower bound approximation (22), O_{exp} to approximation (25), and O_{UB} to approximation (23), as well as routing from high to low incubation time, H. Comparison of the rows for $f_{TAT}(\boldsymbol{p})$ shows that approximation (22) yields the best approximation of the DES values for load 0.3 and that for high load 0.45 approximations (25) and (22) show similar performance; in all cases the upper bound (23) shows poor performance. For load 0.3 the optimal strategies under (25) and (22) are similar as the optimal values $f_{TAT}(\boldsymbol{p})$ do not show a significant difference in the DES. For load 0.45 the average values for DES are different, but the confidence intervals overlap. Loads in the clinical chemistry labs, which is our main application, are low to moderate. Therefore, we conclude that for the loads of interest approximation (22) is preferred over (25).

5.4. Case study: Clinical chemistry laboratory at Erasmus MC

This section applies our method to the optimal design, minimizing mean TAT, of a clinical chemistry laboratory analyzer line using data obtained from Erasmus MC, Rotterdam, the Netherlands. A description of an analyzer line is included in Section 1 and Fig. 1.

5.4.1. Input data

The analyzer line of the clinical chemistry laboratory of Erasmus MC has 4 analyzer modules and is operational 24/7. Technical details of these modules affect the TAT only via the pipetting duration and incubation times. For lab design, the TAT during the busiest time of day is of main importance, as the physician has to wait the longest for test results in this time frame. Therefore, we consider samples arriving between 9 AM and 2 PM on weekdays in March 2019. The data contained 2232 racks, which in total contained 6708 clinical chemistry samples. This results in 150 tests per hour at peak load on this analyzer line. We obtained 15 job classes characterized by the modules visited by a rack, resulting in 64 possible job types (Table D.10). A sample may require multiple tests on a single module, that are processed in parallel on the incubator disc, in which case the incubation time is the maximum of these times as this is the time when the test results for the module become available. From the laboratory information system (LIS) and analyzer log files, for each job class we obtained the mean

and SCV of the service duration at the pipettor and incubator disc (Table D.9), the historical lab routes and the mean TAT that was found to be 17.86 minutes. Our model does not incorporate delays caused by transport and scanning of the samples, with the largest deviations as compared to reality observed between arrival at an analyzer line and arrival at the initial node, that is estimated from the LIS to be 2 minutes. Smaller transport delays that cannot be estimated accurately from the LIS occur between analyzer modules.

5.4.2. Optimal routing configuration

We used (1) with TAT replaced by $\overline{TAT}$ to compare the historical mean TAT with the mean TAT from our model using the historical laboratory routes, which was found to be 14.30 minutes (857.93 seconds, see Table 6). This result should be compared to the historical value of 17.86 minutes, minus the initial delay of 2 minutes and additional minor transport delays between modules. This indicates that (1) using $\overline{TAT}$ provides a good estimate of the historical mean TAT.

We invoked our optimization method to determine optimal allocation of jobs to routes. Calculating the mean waiting time of a specific analyzer line routing configuration using the QNA takes 0.01 to 0.05 seconds. For SA, the initial value of the cooling parameter was set to 40, the stopping value was set to 0.0005, with a decrease factor of 0.999 and Markov chain length of 100. The run-time of the algorithm is approximately 7 hours, which is acceptable considering we are interested in optimal laboratory design that typically has a duration of several weeks or months.

Table 6 compares the performance of four routing configurations: historical lab route, best route found using our optimization method, jobs routed from highest to lowest incubation time, and jobs routed from lowest to highest incubation time that is included to compare performance of the other routes with a route that intuitively does not perform well. Results are presented for the current lab load and scenarios in which the lab load is increased to 120%, 140% and 160% of the current lab load. The routing configurations used in Table 6 are detailed in Table D.10.

The current load at the four FIFO queues is 0.23, 0.57, 0.48, and 0.13. The optimal routing configuration for the current lab load has a mean TAT of 841.42 seconds. The lower bound on the objective function value is 801.10 seconds, see Section 4.3, which is expected to be a good benchmark for the current load, see Section 5.3. The gap of 31.32 seconds (5.03%) between our optimum and this minimum value indeed may be completely allocated to the additional waiting times at the FIFO queues since the inter-arrival times are not deterministic as assumed in the lower bound calculations, which indicates that our method closely approximates the optimal value.

The optimal routing configuration for the current lab load routes jobs from highest to lowest incubation time and provides a 2% decrease in mean TAT compared to the historic laboratory route. As the load increases, the heuristic routing configuration high to low continues to perform well, but above 140% load our optimal routing configuration (Table D.10) outperforms this heuristic. The possible improvement in mean TAT increases with the load. When increasing the load to 160%, resulting in loads 0.38, 0.92, 0.76 and 0.20 at the four FIFO queues, our proposed route results in a 5% reduction of the mean TAT.

Our results support that routing jobs according to the heuristic that routes jobs from high to low incubation times is optimal for the current load, as well as for increased load up to 140% of the current load. For a 60% load increase, our optimization approach yields roughly 5% reduction in mean TAT compared to this heuristic, which illustrates the quality of the heuristic, as well as the gain that may be achieved by optimization. In light of the number of tests performed on a chemistry analyzer line (150 per hour at current peak load in Erasmus MC), optimization of the routing configuration may result in a substantial improvement of laboratory performance.

6. Discussion and conclusion

Motivated by chemistry analyzer lines, we have considered optimal design of queueing networks in which each node consists of a single-server FIFO queue and an infinite-server incubation queue. A job departing from a single-server queue forks to the incubation queue and the next FIFO queue on its route. We have provided exact results for the queue length distribution and TAT as well as a QNA and SA optimization approach to determine the optimal routing configuration for a central decision maker that aims to minimize the mean TAT for all jobs.

Generalizations: Our results may be extended to incubation queues of different types. This is clear in Lemma 1, as we only require the maximum of the sojourn time at the incubation queue and remaining part of the route. Theorem 5 requires the sojourn time at the incubation queue to be exponentially distributed, which is also the case if the incubation queue is a single-server FIFO queue with exponential service requirement. Our exact results may be extended to more general queues $j1$, such as those modeled using the (ϕ, γ, δ)-protocol (Kelly, 1979, Section 3.1). Approximation 11 that uses the QNA requires the queues $j1$ to be FIFO queues, but allows for multi-server queues and general incubation queues as long as the mean sojourn time for these queues is known. Setting all incubation times to zero shows that our novel QNA and SA approach may also be used for the optimal design of networks of single-server FIFO queues.

Our results may be extended to include several parallel incubation queues as this requires evaluation of the maximum sojourn time over these queues. By setting the service times at the first FIFO queue to zero, this shows that we may also approximate the TAT distribution of a fork-join queue with an arbitrary number of parallel queues. We assume that the nodes in the network are distinct and therefore the set of nodes visited by a job are uniquely defined. An interesting extension is to allow for duplicate nodes, where a job should visit one of these duplicate nodes. The QNA also allows for approximation of the variance of the sojourn times per node assuming independence of the sojourn times at the nodes. Among our aims for further research is exploring how such results may be used to approximate the variance of the TAT and the fraction of samples that completes TAT before the due date.

Limitations and further research: Our QNA and SA approach shows good performance in terms of accuracy and computation time for realistic size network design challenges. A possible direction for future research is to improve the QNA approximation (Caldentey, 2001; Harrison & Nguyen, 1990), and the SA algorithm using approaches such as improvement of the cooling scheme, the stopping value and the Markov chain length (van Laarhoven & Aarts, 1987).

In highly congested analyzer lines, blocking of jobs in between modules might occur, which is not included in our model. A possible extension of our model is to further develop the approach of Kerbache & MacGregor Smith (2000) to include incubation queues.

Our results consider optimal design of a laboratory via static route allocation. Dynamic routes are of interest to avoid congestion in an operational setting. Extension of the results for networks of FIFO queues with exponential service times and concave utility functions (Shaler, 2009) to include incubation queues is of considerable interest for daily operation of laboratories.

Conclusion: We have illustrated the intricate relation between TAT, service and incubation times. A heuristic routing configuration supported by our theoretical results routes jobs along the nodes in decreasing order of the incubation times. Our numerical results reveal that this heuristic is close to optimum for realistic network

parameters, which make the results amenable for inclusion in the lab ICT system that routes samples in analyzer lines. The accuracy of our approach supports its use in design of clinical chemistry laboratories.

Appendix A. Queueing Network Analyzer

The Queueing Network Analyzer (QNA) is developed in Whitt (1983b) to approximate the mean sojourn times at the queues of a network of multi-server FIFO queues with multiple job types and general inter-arrival and service time distributions. Below, we present the QNA for a network of single-server FIFO queues and the expressions for the mean sojourn times in the FIFO queues in our network used in Section 4.2. The mean sojourn times at the FIFO queues in our network are not affected by the incubation queues, so that the QNA yields an approximation of the sojourn times in the FIFO queues in the network with incubation queues, recall Remark 2.

The QNA uses as input the mean and the variance of the inter-arrival and service times of each job type. The arrival rate of type c_i jobs is $\lambda_0(c_i) = p(c_i)\lambda_0(c)$. The squared coefficient of variation (SCV) of the arrival process of job type c_i is (Whitt, 1983b):

$$\text{scv}(c_i) = p(c_i)\sigma_0^2(c)\lambda_0^2(c) + 1 - p(c_i), \quad i = 1, \ldots, I(c), \ c = 1, \ldots, C.$$

The service requirements of a job at queue jk on its route depend on its class and therefore $\mu_{jk}(c_i) = \mu_{jk}(c)$. The SCV of the service time of job type c_i at queue jk on its route is

$$\text{scv}_{sjk}(c_i) = \sigma_{jk}^2(c)\mu_{jk}(c)^2, \quad i = 1, \ldots, I(c), \ c = 1, \ldots, \ j = 1, \ldots, J, \ k = 1, 2.$$

The aggregated external arrival rate to queue $j1$ is:

$$\lambda_{0j1} = \sum_{c=1}^{C} \sum_{i=1}^{I(c)} \lambda_0(c_i) \mathbb{1}\{n(c_i, 1) = j\}.$$

The aggregated internal flow rate from queue $j1$ to queue $j'1$:

$$\lambda_{j1j'1} = \sum_{c=1}^{C} \sum_{i=1}^{I(c)} \sum_{\ell=1}^{|R(c)|-1} \lambda_0(c_i) \mathbb{1}\{n(c_i, \ell) = j, n(c_i, \ell+1) = j'\}$$

The departure rate from queue $j1$ out of the network is:

$$\lambda_{j10} = \sum_{c=1}^{C} \sum_{i=1}^{I(c)} \lambda_0(c_i) \mathbb{1}\{n(c_i, |R(c)|) = j\}.$$

The total aggregated arrival flow rate to queue $j1$ is:

$$\lambda_{j1} = \sum_{j'=0}^{J} \lambda_{j'1j1}.$$

The routing matrix Q has elements $q_{j1j'1}$ equal to the proportion of jobs that go from queue $j1$ to queue $j'1$. The element $q_{jj'}$ can be seen as the probability that a job exiting queue $j1$ will then go to queue $j'1$:

$$q_{j1j'1} = \frac{\lambda_{j1j'1}}{\sum_{\ell=0}^{J} \lambda_{j1\ell1}}, \quad \text{with} \ \sum_{j'=1}^{J} q_{j1j'1} = 1.$$

Service time parameters are obtained by averaging the service times of jobs that visit queue $j1$:

$$\tau_{j1} = \frac{\sum_{c=1}^{C} \sum_{i=1}^{I(c)} \sum_{\ell=1}^{|R(c)|} \lambda_0(c_i) \mathbb{E}\left[B_{n(c_i,\ell)1}(c)\right] \mathbb{1}\{n(c_i, \ell) = j\}}{\sum_{c=1}^{C} \sum_{i=1}^{I(c)} \sum_{\ell=1}^{|R(c)|} \lambda_0(c_i) \mathbb{1}\{n(c_i, \ell) = j\}}.$$

The SCV of the service time at queue $j1$ is calculated as follows:

$$\text{scv}_{sj1} = \frac{\sum_{c=1}^{C} \sum_{i=1}^{I(c)} \sum_{\ell=1}^{|R(c)|} \lambda_0(c_i) \mathbb{E}\left[B_{n(c_i,\ell)1}(c)\right]^2 \left(\text{scv}_{sn(c_i,\ell)1}(c_i) + 1\right) \mathbb{1}\{n(c_i, \ell) = j\}}{\tau_{j1}^2 \sum_{c=1}^{C} \sum_{i=1}^{I(c)} \sum_{\ell=1}^{|R(c)|} \lambda_0(c_i) \mathbb{1}\{n(c_i, \ell) = j\}} - 1.$$

The utilization of queue $j1$ is given by:

$$\rho_{j1} = \lambda_{j1}\tau_{j1}.$$

The proportion of arrivals to $j'1$ that came from $j1$ is equal to:

$$p_{j1j'1} = \frac{\lambda_{j1j'1}}{\lambda_{j1}}, \quad \text{with:} \ \sum_{j=1}^{J} p_{j1j'1} = 1.$$

The SCV of the external arrival process to queue $j1$ is:

$$\text{scv}_{0j1} = (1 - u_{j1}) + u_{j1}\left[\sum_{c=1}^{C} \sum_{i=1}^{I(c)} \text{scv}(c_i) \left(\frac{\lambda_0(c_i) \mathbb{1}\{n(c_i, 1) = j\}}{\sum_{c=1}^{C} \sum_{i=1}^{I(c)} \lambda_0(c_i) \mathbb{1}\{n(c_i, 1) = j\}} \right) \right],$$

where

$$u_{j1} = u_{j1}(\rho_{j1}, v_{j1}) = \frac{1}{1 + 4(1 - \rho_{j1})^2(v_{j1} - 1)}$$

and

$$v_{j1} = \left[\sum_{c=1}^{C} \sum_{i=1}^{I(c)} \left(\frac{\lambda_0(c_i) \mathbb{1}\{n(c_i, 1) = j\}}{\sum_{c=1}^{C} \sum_{i=1}^{I(c)} \lambda_0(c_i) \mathbb{1}\{n(c_i, 1) = j\}} \right)^2 \right]^{-1}.$$

The approximation of the SCV of the arrival process at each queue is calculated as follows:

$$\text{scv}_{aj1} = a_{j1} + \sum_{j'=1}^{J} \text{scv}_{aj1} b_{j'1j1} \quad \Leftrightarrow \text{scv}_a = (I - B^T)^{-1} a.$$

The a_{j1} and $b_{j1j'1}$ are constants depending on the input data:

$$a_{j1} = 1 + u_{j1}\left(p_{0j1}\text{scv}_{0j1} - 1 + \sum_{j'=1}^{N} p_{j'1j1}\left(\left(1 - q_{j'1j1}\right) + q_{j'1j1}\rho_{j'1}^2 x_{j'1} \right) \right)$$

and

$$b_{j1j'1} = u_{j1} p_{j1j'1} q_{j1j'1}(1 - \rho_{j1}^2),$$

with

$$x_{j1} = 1 + m_{j1}^{-0.5}(\max\{\text{scv}_{sj1}, 0.2\} - 1),$$

$$u_{j1} = \frac{1}{1 + 4(1 - \rho_{j1})^2(v_{j1} - 1)} \quad \text{and} \quad v_{j1} = \left[\sum_{j'=0}^{J} p_{j'1j1}^2 \right]^{-1}.$$

Given τ_{j1}, ρ_{j1}, scv_{aj1} and scv_{sj1}, we can calculate the approximate mean waiting times, and thus the approximate mean sojourn times at the FIFO queues using (27) and (28).

Appendix B. Example non-convexity of the QNA objective function

This section gives an example of a small network for which the objective function of the QNA is non-convex. We will follow the steps of the QNA as outlined in Appendix A.

Consider a network of two nodes with one job class and 2 possible job types. Type 1_1 visits the queues in order 12, while type 1_2 visits them in order 21. We consider zero incubation times and the service discipline at the FIFO queues only depends on the queues $\mu_{j1}(c_i) = \mu_{j1}$, $j = 1, 2$. The arrival rate of type 1_i jobs is $\lambda_0(1_i) = p(1_i)\lambda_0$, $i = 1, 2$, $p(1_1) + p(1_2) = 1$. The SCV of the arrival process of job type 1_i is $\text{scv}(1_i) = p(1_i)\sigma_0^2\lambda_0^2 + 1 - p(1_i)$, $i = 1, 2$.

The mean service time at queue $j1$ is $\tau_{j1} = \frac{1}{\mu_{j1}}$. The SCV of the service time of job type 1_i at queue $j1$ on its route is $\text{scv}_{sj1} = \sigma_{j1}^2 \mu_{j1}^2$, $i = 1, 2,\ j = 1, 2$. The total aggregated arrival flow rate to queue $j1$ is $\lambda_{j1} = \lambda_0$. The routing matrix Q is

$$Q = \begin{bmatrix} 0 & p(1_1) & p(1_2) \\ p(1_2) & 0 & p(1_1) \\ p(1_1) & p(1_2) & 0 \end{bmatrix}$$

The utilization of queue $j1$ is $\rho_{j1} = \lambda_0 \tau_{j1}$. The proportion of arrivals to $j'1$ that came from $j1$ is captured in the following matrix:

$$P = \begin{bmatrix} 0 & p(1_1) & p(1_2) \\ p(1_2) & 0 & p(1_1) \\ p(1_1) & p(1_2) & 0 \end{bmatrix}$$

The SCV of the external arrival process to queue $j1$ is $\text{scv}_{0j1} = p(1_j)\sigma_0^2\lambda_0^2 + 1 - p(1_j)$, $j = 1, 2$. The approximation of the SCV of the arrival process at each queue is:

$$\text{scv}_a = (I - B^T)^{-1} a,$$

with

$$a = \begin{bmatrix} 1 + u_{11}\big(p(1_1)[p(1_1)\sigma_0^2\lambda_0^2 + 1 - p(1_1)] - 1 \\ \quad + p(1_2)(1 - p(1_2) + p(1_2)\rho_{21}^2 x_{21})\big) \\ 1 + u_{21}\big(p(1_2)[p(1_2)\sigma_0^2\lambda_0^2 + 1 - p(1_2)] - 1 \\ \quad + p(1_1)(1 - p(1_1) + p(1_1)\rho_{11}^2 x_{11})\big) \end{bmatrix},$$

$$B = \begin{bmatrix} 0 & u_{11}p(1_1)^2(1 - \lambda_0^2\tau_{11}^2) \\ u_{21}p(1_2)^2(1 - \lambda_0^2\tau_{21}^2) & 0 \end{bmatrix},$$

$$u_{j1} = \frac{1}{1 + 4(1 - \lambda_0\tau_{j1})^2([p(1_1)^2 + p(1_2)^2]^{-1} - 1)}, \quad j = 1, 2,$$

$$x_{j1} = 1 + (\max\{\text{scv}_{sj1}, 0.2\} - 1), \quad j = 1, 2.$$

Mathematica (version 12.3) was used to find scv_{aj1}, $j = 1, 2$, and to determine whether these functions are convex or not. We find:

$$\text{scv}_{a11} = \frac{\frac{p(1_2)\left(\lambda_0^2 p(1_2)\tau_{21}^2 x_{21} - p(1_2) + 1\right) + p(1_1)\left(\lambda_0^2 p(1_1)\sigma_0^2 - p(1_1) + 1\right) - 1}{4(1-\lambda_0\tau_{11})^2(\nu-1)+1} + 1}{1 - \frac{p(1_1)^2 p(1_2)^2\left(1-\lambda_0^2\tau_{11}^2\right)\left(1-\lambda_0^2\tau_{21}^2\right)}{\left(4(1-\lambda_0\tau_{11})^2(\nu-1)+1\right)\left(4(1-\lambda_0\tau_{21})^2(\nu-1)+1\right)}}$$
$$+ \frac{p(1_2)^2\left(1-\lambda_0^2\tau_{21}^2\right)\left(\frac{p(1_1)\left(\lambda_0^2 p(1_1)\tau_{11}^2 x_{11} - p(1_1) + 1\right) + p(1_2)\left(\lambda_0^2 p(1_2)\sigma_0^2 - p(1_2) + 1\right) - 1}{4(1-\lambda_0\tau_{21})^2(\nu-1)+1} + 1\right)}{\left(4(1-\lambda_0\tau_{21})^2(\nu-1)+1\right)\left(1 - \frac{p(1_1)^2 p(1_2)^2\left(1-\lambda_0^2\tau_{11}^2\right)\left(1-\lambda_0^2\tau_{21}^2\right)}{\left(4(1-\lambda_0\tau_{11})^2(\nu-1)+1\right)\left(4(1-\lambda_0\tau_{21})^2(\nu-1)+1\right)}\right)},$$

$$\text{scv}_{a21} = \frac{1 + \frac{-1 + p(1_2)(1-p(1_2)+\lambda_0^2 p(1_2)\sigma_0^2) + p(1_1)(1-p(1_1)+\lambda_0^2 p(1_1)\tau_{11}^2 x_{11})}{1+4(-1+\nu)(1-\lambda_0\tau_{21})^2}}{1 - \frac{p(1_1)^2 p(1_2)^2(1-\lambda_0^2\tau_{11}^2)(1-\lambda_0^2\tau_{21}^2)}{(1+4(-1+\nu)(1-\lambda_0\tau_{11})^2)(1+4(-1+\nu)(1-\lambda_0\tau_{21})^2)}}$$
$$+ \frac{p(1_1)^2(1-\lambda_0^2\tau_{11}^2)\left(1 + \frac{-1 + p(1_1)(1-p(1_1)+\lambda_0^2 p(1_1)\sigma_0^2) + p(1_2)(1-p(1_2)+\lambda_0^2 p(1_2)\tau_{21}^2 x_{21})}{1+4(-1+\nu)(1-\lambda_0\tau_{11})^2}\right)}{(1+4(-1+\nu)(1-\lambda_0\tau_{11})^2)\left(1 - \frac{p(1_1)^2 p(1_2)^2(1-\lambda_0^2\tau_{11}^2)(1-\lambda_0^2\tau_{21}^2)}{(1+4(-1+\nu)(1-\lambda_0\tau_{11})^2)(1+4(-1+\nu)(1-\lambda_0\tau_{21})^2)}\right)},$$

with

$$\nu = \frac{1}{p(1_1)^2 + p(1_2)^2}.$$

Table C1
Objective function values for a selection of routing configurations under loads B and D.

Scenario	1			2			3			4			5			6		
Route	O	H	D	O	H	D	O	H	D	O	H	D	O	H	D	O	H	D
$\lambda_0(c)$									B: 0.3									
$p(1_1)$	0	$\frac{1}{6}$	0	1	$\frac{1}{6}$	1	1	1	1	1	1	1	1	1	1	0	0	0
$p(1_2)$	0	$\frac{1}{6}$	0	0	$\frac{1}{6}$	0	0	0	0	0	0	0	0	0	0	0	0	0
$p(1_3)$	1	$\frac{1}{6}$	1	0	$\frac{1}{6}$	0	0	0	0	0	0	0	0	0	0	1	0.5	1
$p(1_4)$	0	$\frac{1}{6}$	0	0	$\frac{1}{6}$	0	0	0	0	0	0	0	0	0	0	0	0.5	0
$p(1_5)$	0	$\frac{1}{6}$	0	0	$\frac{1}{6}$	0	0	0	0	0	0	0	0	0	0	0	0	0
$p(1_6)$	0	$\frac{1}{6}$	0	0	$\frac{1}{6}$	0	0	0	0	0	0	0	0	0	0	0	0	0
$p(2_1)$	0	$\frac{1}{6}$	0	0	$\frac{1}{6}$	0	0	0	0	0	0	0	0	0	0	0	0	0
$p(2_2)$	0	$\frac{1}{6}$	0	0	$\frac{1}{6}$	0	0	0	0	0	0	0	0	0	0	0	0	0
$p(2_3)$	1	$\frac{1}{6}$	1	0.06	$\frac{1}{6}$	0	0	0	0	0	0	0	0	0	0	1	0.5	1
$p(2_4)$	0	$\frac{1}{6}$	0	0.20	$\frac{1}{6}$	0	0	0	0	0	0	0	0	0	0	0	0.5	0
$p(2_5)$	0	$\frac{1}{6}$	0	0.61	1	0	0	0	0	0	0	0	0	0	0	0	0	0
$p(2_6)$	0	$\frac{1}{6}$	0	0.13	$\frac{1}{6}$	0	1	1	1	1	1	1	1	1	1	0	0	0
$f_{\text{TAT}}(\boldsymbol{p})$ QNA	7.26	7.51	7.26	7.29	7.50	7.34	10.50	10.50	10.50	10.50	10.50	10.50	10.48	10.48	10.48	10.50	10.50	10.50
$f_{\text{TAT}}(\boldsymbol{p})$ DES	7.10 ±0.07	7.46 ±0.14	7.10 ±0.07	6.84 ±0.06	7.07 ±0.05	6.81 ±0.05	11.40 ±0.07	11.40 ±0.07	11.40 ±0.07	11.52 ±0.05	11.52 ±0.05	11.52 ±0.05	10.91 ±0.05	10.91 ±0.05	10.91 ±0.05	11.23 ±0.05	11.31 ±0.03	11.23 ±0.05
$\lambda_0(c)$									D: 0.45									
$p(1_1)$	1	$\frac{1}{6}$	1	1	$\frac{1}{6}$	1	1	1	1	1	1	1	1	1	1	0	0	0
$p(1_2)$	0	$\frac{1}{6}$	0	0	$\frac{1}{6}$	0	0	0	0	0	0	0	0	0	0	0	0	0
$p(1_3)$	0	$\frac{1}{6}$	0	0	$\frac{1}{6}$	0	0	0	0	0	0	0	0	0	0	1	0.5	1
$p(1_4)$	0	$\frac{1}{6}$	0	0	$\frac{1}{6}$	0	0	0	0	0	0	0	0	0	0	0	0.5	0
$p(1_5)$	0	$\frac{1}{6}$	0	0	$\frac{1}{6}$	0	0	0	0	0	0	0	0	0	0	0	0	0
$p(1_6)$	0	$\frac{1}{6}$	0	0	$\frac{1}{6}$	0	0	0	0	0	0	0	0	0	0	0	0	0
$p(2_1)$	1	$\frac{1}{6}$	1	0.19	$\frac{1}{6}$	0	0	0	0	0	0	0	0	0	0	0	0	0
$p(2_2)$	0	$\frac{1}{6}$	0	0.14	$\frac{1}{6}$	0	0	0	0	0	0	0	0.07	0	0	0	0	0
$p(2_3)$	0	$\frac{1}{6}$	0	0.15	$\frac{1}{6}$	0	0	0	0	0	0	0	0	0	0	1	0.5	1
$p(2_4)$	0	$\frac{1}{6}$	0	0.14	$\frac{1}{6}$	0	0	0	0	0	0	0	0.38	0	0	0	0.5	0
$p(2_5)$	0	$\frac{1}{6}$	0	0.22	1	0	0	0	0	0.85	0	1	0.14	0	0	0	0	0
$p(2_6)$	0	$\frac{1}{6}$	0	0.16	$\frac{1}{6}$	0	1	1	1	0.15	1	0	0.41	1	1	0	0	0
$f_{\text{TAT}}(\boldsymbol{p})$ QNA	26.39	30.15	26.39	27.99	30.00	29.62	31.00	31.00	31.00	30.34	31.19	30.37	29.76	30.92	30.92	26.94	30.18	26.94
$f_{\text{TAT}}(\boldsymbol{p})$ DES	27.92 ±0.79	30.37 ±0.91	27.92 ±0.79	28.60 ±0.41	29.39 ±0.79	28.59 ±0.62	31.44 ±0.17	31.44 ±0.17	31.44 ±0.17	32.01 ±1.67	32.17 ±0.77	32.08 ±0.60	30.91 ±0.32	30.90 ±1.10	30.90 ±1.10	29.40 ±0.81	30.90 ±0.76	29.40 ±0.81

Both scv_{a11} and scv_{a21} are neither convex nor concave in $p(1_1)$ and $p(2_1)$. Recall (27):

$$\overline{W}_{j1} = \frac{\tau_{j1}\rho_{j1}(\text{scv}_{aj1} + \text{scv}_{sj1})g_{j1}}{2(1-\rho_{j1})},$$
$$g_{j1} = \begin{cases} \exp\left[-\frac{2(1-\rho_{j1})}{3\rho_{j1}}\frac{(1-\text{scv}_{aj1})}{\text{scv}_{aj1}+\text{scv}_{sj1}}\right], & \text{scv}_{aj1} < 1, \\ 1, & \text{scv}_{aj1} \geq 1. \end{cases} \quad \text{(B.1)}$$

Observe that scv_{sj1} does not depend on $p(1_i)$. For $\text{scv}_{aj1} \geq 1$ we have $g_{j1} = 1$, and $\overline{W}_{j1}$ is linear in scv_{sj1} and scv_{aj1}, and therefore not convex in $p(1_1)$ and $p(2_1)$. For $\text{scv}_{aj1} < 1$, g_{j1} is not convex in $p(1_1)$ and $p(2_1)$. The objective function is a linear combination of $\overline{\text{TAT}}(c_i)$, which in turn consists of summing and maximizing over a non-convex part ($\overline{W}_{j1}$) and a constant ($\mathbb{E}[B_{j1}]$ or $\mathbb{E}[S_{j2}]$). Therefore, the objective function for this small example is non-convex.

Appendix C. Objective values for selected loads and accuracy of approximations

Appendix D. Input parameters and routing configurations for case study

Table D.9 gives the input parameters for the arrival and service times of the job classes for the case study in Section 5.4. There is no incubation queue at node 1, i.e., $B_{12}(c) = 0$, $c = 1, \ldots, C$. The incubation time at node 3 is deterministic and equal to 10 minutes, i.e., $B_{32}(c) = 600\text{s}$, $\text{scv}_{s32} = 0$, $c = 1, \ldots, C$.

Table D.10 provides the details of the routing configuration and job types in Section 5.4.

Table C2
Objective function values for a selection of routing configurations under loads B and D, including three objective function approximations. The columns give the corresponding routing configuration. For each routing configuration the value of $f_{\text{TAT}}(\boldsymbol{p})$ is then determined for each approximation (22), (25), (23) and via DES in the corresponding rows, i.e., in the row (22) the value 13.75 indicates the value of $f_{\text{TAT}}(\boldsymbol{p})$ for the listed optimal routing configuration under O_{UB} evaluated using approximation (22) in the QNA.

Scenario			4				5			
Route			O	O_{exp}	O_{UB}	H	O	O_{exp}	O_{UB}	H
	$\lambda_0(c)$					0.3				
	$p(1_1)$		1	1	0	1	1	1	1	1
	$p(1_2)$		0	0	0	0	0	0	0	0
	$p(1_3)$		0	0	1	0	0	0	0	0
	$p(1_4)$		0	0	0	0	0	0	0	0
	$p(1_5)$		0	0	0	0	0	0	0	0
	$p(1_6)$		0	0	0	0	0	0	0	0
	$p(2_1)$		0	0	0	0	0	0	0	0
	$p(2_2)$		0	0	0	0	0	0	0	0
	$p(2_3)$		0	0	1	0	0	0	0.06	0
	$p(2_4)$		0	0	0	0	0	0	0.19	0
	$p(2_5)$		0	0	0	0	0	0.11	0.61	0
	$p(2_6)$		1	1	0	1	1	0.89	0.14	1
$f_{\text{TAT}}(\boldsymbol{p})$		(22)	10.50	10.50	13.75	10.50	10.48	10.50	11.03	10.48
		(25)	14.63	14.63	15.47	14.63	14.58	14.57	14.76	14.58
		(23)	20.51	20.51	20.26	20.51	20.47	20.41	20.29	20.47
		DES	11.52	11.52	13.94	11.52	10.91	10.95	11.47	10.91
			±0.05	±0.05	±0.06	±0.05	±0.05	±0.03	±0.04	±0.05
	$\lambda_0(c)$					0.45				
	$p(1_1)$		1	1	0	1	1	1	0	1
	$p(1_2)$		0	0	0	0	0	0	0	0
	$p(1_3)$		0	0	1	0	0	0	0	0
	$p(1_4)$		0	0	0	0	0	0	1	0
	$p(1_5)$		0	0	0	0	0	0	0	0
	$p(1_6)$		0	0	0	0	0	0	0	0
	$p(2_1)$		0	0.76	0	0	0	0	0.22	0
	$p(2_2)$		0	0	0	0	0.07	0.09	0.16	0
	$p(2_3)$		0	0	1	0	0	0	0.14	0
	$p(2_4)$		0	0	0	0	0.38	0.26	0.19	0
	$p(2_5)$		0.85	0.24	0	0	0.14	0.26	0.14	0
	$p(2_6)$		0.15	0	0	1	0.41	0.39	0.15	1
$f_{\text{TAT}}(\boldsymbol{p})$		(22)	30.34	30.60	31.44	31.19	29.76	29.83	34.20	30.92
		(25)	33.15	32.62	33.63	34.51	33.13	33.07	35.63	34.16
		(23)	41.06	39.58	39.94	43.19	41.44	41.3	40.99	42.92
		DES	32.01	32.14	33.82	32.17	30.91	31.40	35.30	30.90
			±1.67	±1.46	±0.74	±0.77	±0.32	±0.15	±0.69	±1.1

Table D1
Arrival and service time input parameters for the laboratory case study. The time unit is seconds.

			1		2				3		4			
c	$\lambda_0(c) \times 10^3$	scv(c)	$\mathbb{E}[B_{11}(c)]$	scv$_{s11}$	$\mathbb{E}[B_{21}(c)]$	scv$_{s21}$	$\mathbb{E}[B_{22}(c)]/60$	scv$_{s22}$	$\mathbb{E}[B_{31}(c)]$	scv$_{s31}$	$\mathbb{E}[B_{41}(c)]$	scv$_{s41}$	$\mathbb{E}[B_{42}(c)]/60$	scv$_{s42}$
1	1.957	1.58			26.32	0.74	9.80	0.01						
2	0.759	1.87							47.06	0.28				
3	0.457	2.49									26.70	0.36	16.89	0.05
4	0.395	0.94	21.21	0.59										
5	0.451	1.06			23.51	0.33	7.03	0.25			28.01	0.46	20.72	0.09
6	0.975	1.40			36.30	0.47	10.00	0.00	50.78	0.22				
7	1.062	1.54	19.78	0.21	32.16	0.24	9.72	0.02						
8	4.228	1.94	27.77	0.25	59.56	0.32	10.00	0.00	50.82	0.16				
9	0.142	3.39	26.40	1.19					43.63	0.22				
10	2.537	1.71	29.34	0.21	62.87	0.26	10.00	0.00	57.81	0.28	30.92	0.30	18.41	0.09
11	0.056	1.37	25.03	0.27							40.63	0.41	18	0.13
12	0.321	0.90			41.39	0.28	9.87	0.01	47.75	0.32	26.22	0.20	20.00	0.08
13	0.327	1.01	23.62	0.19	55.65	1.17	9.74	0.02			32.21	0.22	16.56	0.14
14	0.056	2.01							65.39	0.43	28.63	0.12	18.14	0.07
15	0.056	2.22	30.37	0.26					61.61	0.30	33.67	0.22	15.56	0.22

Table D2
Routing configurations discussed in Section 5.4.

				Routing configuration						
$n(c_i,1)$	$n(c_i,2)$	$n(c_i,3)$	$n(c_i,4)$	Historic	Optimal route				High to	Low to
				lab route	lab load (100%)	120% load	140% load	160% load	low B_{j2}	high B_{j2}
1				1	1	1	1	1	1	1
2				1	1	1	1	1	1	1
3				1	1	1	1	1	1	1
4				1	1	1	1	1	1	1
1	2			0.90	0	0	0.16	0.72	0	1
2	1			0.10	1	1	0.84	0.28	1	0
1	3			0.74	0	0	0.29	0.20	0	1
3	1			0.26	1	1	0.71	0.80	1	0
1	4			0.11	0	0	0.78	0.82	0	1
4	1			0.89	1	1	0.22	0.18	1	0
2	3			0.68	0.5	0.5	0.05	0.01	0.5	0.5
3	2			0.32	0.5	0.5	0.95	0.99	0.5	0.5
2	4			0.08	0	0	0.01	0	0	1
4	2			0.92	1	1	0.99	1	1	0
3	4			0.44	0	0	0.24	0.28	0	1
4	3			0.56	1	1	0.76	0.72	1	0
1	2	3		0.50	0	0	0	0	0	0.5
1	3	2		0.32	0	0	0.02	0.74	0	0.5
2	1	3		0.08	0	0	0.02	0	0	0
2	3	1		0.01	0.5	0.5	0.33	0	0.5	0
3	1	2		0.08	0	0	0	0	0	0
3	2	1		0.01	0.5	0.5	0.62	0.26	0.5	0
1	2	4		0	0	0	0	0.01	0	1
1	4	2		0.25	0	0	0.01	0.35	0	0
2	1	4		0	0	0	0	0	0	0
2	4	1		0	0	0	0.01	0	0	0
4	1	2		0.70	0	0	0.47	0.29	0	0
4	2	1		0.06	1	1	0.51	0.36	1	0
1	3	4		0.11	0	0	0.28	0.11	0	1
1	4	3		0.44	0	0	0.46	0.05	0	0
3	1	4		0	0	0	0.01	0.06	0	0
3	4	1		0	0	0	0.10	0.18	0	0
4	1	3		0.44	0	0	0.05	0.30	0	0
4	3	1		0	1	1	0.11	0.30	1	0
2	3	4		0.02	0	0	0	0.03	0	0.5
2	4	3		0.06	0	0	0.02	0	0	0
3	2	4		0	0	0	0	0	0	0.5
3	4	2		0.06	0	0	0.34	0.01	0	0
4	2	3		0.58	0.5	0.5	0.51	0.19	0.5	0
4	3	2		0.29	0.5	0.5	0.14	0.77	0.5	0
1	2	3	4	0.01	0	0	0	0	0	0.5
1	2	4	3	0.03	0	0	0	0	0	0
1	3	4	2	0.02	0	0	0	0	0	0
1	3	2	4	0	0	0	0	0	0	0.5
1	4	2	3	0.11	0	0	0	0	0	0
1	4	3	2	0.05	0	0	0	0.32	0	0
2	1	3	4	0	0	0	0	0	0	0
2	1	4	3	0.02	0	0	0	0	0	0
2	3	4	1	0	0	0	0	0	0	0
2	3	1	4	0	0	0	0	0	0	0
2	4	1	3	0.01	0	0	0	0	0	0
2	4	3	1	0	0	0	0	0	0	0
3	1	2	4	0	0	0	0	0	0	0
3	1	4	2	0.01	0	0	0	0	0	0
3	2	1	4	0	0	0	0	0	0	0
3	2	4	1	0	0	0	0	0	0	0
3	4	1	2	0.02	0	0	0	0	0	0
3	4	2	1	0	0	0	0	0	0	0
4	1	2	3	0.38	0	0	0.16	0	0	0
4	1	3	2	0.20	0	0	0.49	0.21	0	0
4	2	1	3	0.06	0	0	0.03	0	0	0
4	2	3	1	0.01	0.5	0.5	0.01	0	0.5	0
4	3	1	2	0.05	0	0	0.10	0	0	0
4	3	2	1	0	0.5	0.5	0.21	0.46	0.5	0

References

Adan, I. J. B. F., & Resing, J. A. C. (2015). *Queueing systems: Lecture notes*. Eindhoven University of Technology.

Bai, X., & Menon, P. K. (2013). Decision support for optimal runway reconfiguration. *2013 Aviation Technology, Integration, and Operations Conference*, 1–15.

Bélisle, C. J. P. (1992). Convergence theorems for a class of simulated annealing algorithms on $\mathbb{R}^d$. *Journal of Applied Probability, 29*(4), 885–895.

Bitran, G. R., & Tirupati, D. (1988). Multiproduct queueing networks with deterministic routing: Decomposition approach and the notion of interference. *Management Science, 34*(1), 75–100.

Caldentey, R. (2001). Approximations for multi-class departure processes. *Queueing Systems, 38*(2), 205–212.

Choi, B. D., & Park, K. K. (1992). The $m^k/M/\infty$ queue with heterogeneous customers in a batch. *Journal of Applied Probability, 29*(2), 477–481.

Danilova, M., Dvurechensky, P., Gasnikov, A., Gorbunov, E., Guminov, S., Kamzolov, D., & Shibaev, I. (2022). Recent theoretical advances in non-convex optimization. *High-Dimensional Optimization and Probability*, 79–163.

Fendick, K. W., Saksena, V. R., & Whitt, W. (1991). Investigating dependence in packet queues with the index of dispersion for work. *IEEE Transactions on Communications, 39*(8), 1231–1244.

Fidler, M., Walker, B., & Bora, S. (2020). Tiny tasks - A remedy for synchronization constraints in multi-server systems. *Proceedings - IEEE INFOCOM, 2020-July*, 1063–1072.

Flatto, L., & Hahn, S. (2006). Two parallel queues created by arrivals with two demands i. *SIAM Journal on Applied Mathematics, 44*(5), 1041–1053.

Ghosh, S., & Hassin, R. (2021). Inefficiency in stochastic queueing systems with strategic customers. *European Journal of Operational Research, 295*(1), 1–11.

Grasas, A., Ramalhinho, H., Pessoa, L. S., Resende, M. G., Caballé, I., & Barba, N. (2014). On the improvement of blood sample collection at clinical laboratories. *BMC Health Services Research, 14*(1), 1–9.

Harrison, J. M., & Nguyen, V. (1990). The QNET method for two-moment analysis of open queueing networks. *Queueing Systems 1990 6:1, 6*(1), 1–32.

Huisman, T., & Boucherie, R. J. (2011). Decomposition and aggregation in queueing networks. *International Series in Operations Research and Management Science, 154*, 313–344.

Kameda, H., & Zhang, Y. (1995). Uniqueness of the solution for optimal static routing in open BCMP queueing networks. *Mathematical and Computer Modelling, 22*(10–12), 119–130.

Kelly, F. P. (1979). *Reversibility and stochastic networks*. Cambridge University Press.

Kerbache, L., & MacGregor Smith, J. (2000). Multi-objective routing within large scale facilities using open finite queueing networks. *European Journal of Operational Research, 121*(1), 105–123.

Kingman, J. F. C. (1961). The single server queue in heavy traffic. *Mathematical Proceedings of the Cambridge Philosophical Society, 57*(4), 902–904.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science, 220*(4598), 671–680.

Krämer, W., & Langenbach-Belz, M. (1976). Approximate formulae for the delay in the queueing system GI/G/l. *Congressbook, 8th ITC, Melbourne, 235*(1), 1–8.

Krishnamurthy, A., Suri, R., & Vernon, M. (2003). Two-moment approximations for throughput and mean queue length of a fork/join station with general inputs from finite populations. In *Stochastic modeling and optimization of manufacturing systems and supply chains* (pp. 87–126). Springer, Boston, MA.

Kumawat, G. L., Roy, D., De Koster, R., & Adan, I. (2021). Stochastic modeling of parallel process flows in intra-logistics systems: Applications in container terminals and compact storage systems. *European Journal of Operational Research, 290*(1), 159–176.

Laan, C. M., Timmer, J., Boucherie, R. J., Nl, J. B., et al., (2021). Non-cooperative queueing games on a network of single server queues. *Queueing Systems, 97*, 279–301.

van Laarhoven, P. J. M., Aarts, E. H. L., et al., (1987). *Simulated annealing: Theory and applications*. Springer Netherlands.

Law, A. M. (2015). *Simulation modeling and analysis* (5th). Mcgraw-Hill Education.

Lee, S. Y., Chinnam, R. B., Dalkiran, E., Krupp, S., & Nauss, M. (2021). Proactive coordination of inpatient bed management to reduce emergency department patient boarding. *International Journal of Production Economics, 231*, 107842.

Locatelli, M. (2000). Simulated annealing algorithms for continuous global optimization: Convergence conditions. *Journal of Optimization Theory and Applications, 104*(1), 121–133.

Lote, R., Williams, E. J., & Ülgen, O. M. (2009). Simulation of medical laboratory operations to achieve optimal resource allocation. In *Proceedings - 23rd european conference on modelling and simulation, ECMS 2009* (pp. 249–255).

Melamed, B. (1982). Sojourn times in queueing networks. *Mathematics of Operations Research, 7*(2), 223–244.

Persoon, T. J., Zaleski, S., & Frerichs, J. (2006). Improving preanalytic processes using the principles of lean production (toyota production system). *American Journal of Clinical Pathology, 125*(1), 16–25.

Rutledge, J., Xu, M., & Simpson, J. (2010). Application of the toyota production system improves core laboratory operations. *American Journal of Clinical Pathology, 133*(1), 24–31.

Shaler, S. J. (2009). *Optimal design of queueing systems*. Chapman and Hall/CRC.

Shepard, D. M., Cao, D., Afghan, M. K. N., & Earl, M. A. (2007). An arc-sequencing algorithm for intensity modulated arc therapy. *Medical Physics, 34*(2), 464–470.

Suresh, S., & Whitt, W. (1988). Arranging queues in series. *Technical Report*. Murray Hill, NY

Taylor, P. G. (2011). Insensitivity in stochastic models. *International Series in Operations Research and Management Science, 154*, 121–140.

Tsai, E. R., Tintu, A. N., Demirtas, D., Boucherie, R. J., de Jonge, R., & de Rijke, Y. B. (2019). A critical review of laboratory performance indicators. *Critical Reviews in Clinical Laboratory Sciences, 56*(7), 458–471.

Van Nyen, P. L. M., Bertrand, J. W. M., Van Ooijen, H. P. G., & Vandaele, N. J. (2006). A heuristic to control integrated multi-product multi-machine production-inventory systems with job shop routings and stochastic arrival, set-up and processing times. *Stochastic Modeling of Manufacturing Systems: Advances in Design, Performance Evaluation, and Control Issues*, 253–288.

Wason, J. M. S., & Jaki, T. (2012). Optimal design of multi-arm multi-stage trials. *Statistics in Medicine, 31*(30), 4269–4279.

Whitt, W. (1983a). Performance of the queueing network analyzer. *Bell System Technical Journal, 62*(9), 2817–2843.

Whitt, W. (1983b). The queueing network analyzer. *Bell System Technical Journal, 62*(9), 2779–2815.

Yu, M., & De Koster, R. (2008). Performance approximation and design of pick-and-pass order picking systems. *IIE Transactions, 40*(11), 1054–1069.

Zabinsky, Z. B., Dulyakupt, P., Zangeneh-Khamooshi, S., Xiao, C., Zhang, P., Kiatsupaibul, S., & Heim, J. A. (2020). Optimal collection of medical specimens and delivery to central laboratory. *Annals of Operations Research, 287*(1), 537–564.

Zhou, F. L., Wang, X., He, Y. D., & Goh, M. (2017). Production lot-sizing decision making considering bottleneck drift in multi-stage manufacturing system. *Advances in Production Engineering & Management, 12*(3).

Zonderland, M. E., Boer, F., Boucherie, R. J., De Roode, A., & Van Kleef, J. W. (2009). Redesign of a university hospital preanesthesia evaluation clinic using a queuing theory approach. *Anesthesia and Analgesia, 109*(5), 1612–1621.

Zubeldia, M. (2020). Delay-optimal policies in partial fork-join systems with redundancy and random slowdowns. *Proceedings of the ACM on Measurement and Analysis of Computing Systems, 4*(1), 1–49.