Optimal Buffers Allocation in an Inter-facility Material Transfer System Modelled as a Closed Queueing Network and Analysed Through a Simulation-Optimisation Approach

Mohamed Amjath
PhD Student
Division of Engineering Management and Decision Sciences
College of Science and Engineering
Hamad Bin Khalifa University, Doha, Qatar.
mamjath@hbku.edu.qa

Laoucine Kerbache
Professor
Division of Engineering Management and Decision Sciences
College of Science and Engineering
Hamad Bin Khalifa University, Doha, Qatar; and HEC Paris.
lakerbache@hbku.edu.qa

Adel Elomri
Assistant Professor
Division of Engineering Management and Decision Sciences
College of Science and Engineering
Hamad Bin Khalifa University, Doha, Qatar.
aelomri@hbku.edu.qa

James MacGregor Smith
Professor
Department of Mechanical and Industrial Engineering
University of Massachusetts
Amherst, MA 01003, USA
jmgsmith@acad.umass.edu

Abstract

Queueing is a common phenomenon in manufacturing and service environments. Practitioners often use queueing network models to analyse these systems for decision-making processes. However, finite queueing systems with blocking are used to a lesser extent due to the complexity of solving the systems. Generally, closed queueing networks (CQN) with finite capacities do not lend themselves to product-form solutions and must be solved using approximation algorithms. This study considers the problem of determining the optimal buffers allocation for an inter-facility material transportation system in a manufacturing facility considering the blocking after service (BAS) phenomenon. Under BAS, a job can only reach the next node until there is space available in the queue of the next node. If the queue capacity of the destination node is full, the job in the preceding node, even after being serviced, will block that server. In a typical material flow, trucks transfer materials between the facility and undergo several sub-processes such as gate entry, weighing, loading, and unloading. This material flow process is modelled using a CQN, and each sub-process is designed as a node with a finite queueing capacity. An analytical model is developed within this context to determine the optimal buffer sizes to maximise the system's throughput. More specifically, a simulation model is developed using the Anylogic software, and the optimisation problem is incorporated into the model to determine the optimal buffer allocations. The Anylogic simulation-optimisation experiment engine uses various metaheuristics through the OptQuest tool to determine the non-dominated solutions of the corresponding CQN.
Keywords
Closed queueing network, Finite queues, Blocking After Service (BAS), Buffers allocation, Optimisation, and Simulation.

1. Introduction
Material handling constitutes a significant portion of operating costs in the manufacturing sector, with estimates indicating that metallurgical industries’ material handling operations can account for up to 50% of total operating costs (Jaoua et al. 2012). The planning of material handling and facility layout are interrelated functions in the manufacturing environment aimed at achieving operational strategies (Perez-Gosende et al. 2021). Implementing a Material Handling System (MHS) in conjunction with an efficient layout plan can result in lower operational costs, improved space and asset utilisation, reduced occupational and health hazards, and increased productivity. The achievement of these benefits can be facilitated by strategies such as reducing the distance travelled in material movement (Burinskeine 2019; Bhawsar and Yadav 2016) and minimising waiting and idle time (Cannas et al. 2020; Zhou and Zhou 2018). Conversely, the absence of an appropriate facility layout design can create system bottlenecks, poor resource utilisation, and disrupted material flows. Azevedo et al. (2017) present Figure 1, which summarises studies related to facility layout problems from an MHS perspective, including the type of problem, layout type, objective, and constraints employed in the literature.

![Figure 1. Characteristics of Facility layout problems](image)

The use of queueing network theory as a mathematical modeling tool has been widely employed to emulate real-world complex and stochastic systems in various industries, such as manufacturing, communication, and service-based processes, for analysis purposes (Martins et al. 2019; Amjath et al. 2022). Various types of queueing networks have been utilised to model these systems, as determined by the nature of the problem and the feasibility of solutions considering time and cost (Smith 2018). Figure 2, presented by Smith (2018), illustrates the different types of queueing networks that have been widely used for this purpose. Among the queueing network models, blocking networks are considered one of the most complex types to analyse. Typically, decomposition techniques are employed to analyse the network and obtain exact outputs, assuming the network has product-form solutions. However, for finite systems such as blocking networks, product-form solutions are not available and approximation algorithms must be employed to solve the model.

Martins et al. (2019) proposed that numerical methods can effectively solve networks that possess a Markovian property, but the computational cost and time associated with these methods tend to grow exponentially with the size of the network. As a result, approximation methods have been adopted to provide efficient approximations in the analysis of these finite networks. Kerbache and Smith (1988) and (1987) presented, respectively, two approximation methods known as the “Expansion Method (EM)” and the “Generalized Expansion Method (GEM)” used, through decomposition approaches, to solve very complex and mathematically intractable open finite queueing networks. The expansion method is used for Jacksonian finite queue networks, whereas the GEM is used for more general finite queue networks. Many researchers still use these two decomposition methods to solve various problems modelled as finite queueing networks.

According to Smith (2018), blocking can be classified into three types in a two-stage queueing network: blocking after service (BAS), blocking before service (BBS), and repetitive service blocking (RSB). In the case of BAS, a customer or agent will only leave the server after being serviced and a vacant space (buffer) is available in the downstream block.
The research questions addressed in this study pertain to the factors that influence the optimal buffer allocation strategies in a finite CQN framework. Specifically, this study aims to investigate how the network population, server utilisation, and arrival rates impact the buffer allocation and throughput of the system. Additionally, the study seeks to examine the combined effect of these factors on various network performance measures, such as waiting times, response times, and utilisation rates.

1.1 Objectives of the study
The transportation of diverse raw materials from storage to an intermediate production facility is facilitated by a fleet of trucks, with the objective of meeting the demand. The inter-facility material transfer operation comprises several sub-processes that are interconnected with resources. In this regard, the operation is modelled as a finite CQN, where each sub-process along with its associated resource is represented as a single server node equipped with a finite buffer. This study aims to determine the optimal allocation of buffers for each service station in order to maximise the system throughput. Additionally, the optimisation problem incorporates a non-binding constraint that limits the utilisation rates of all servers within predefined upper bounds.

2. Literature Review
In the study conducted by Kerbache and Smith (2000), a manufacturing system was modelled as an OQN with finite buffer capacities. The GEM algorithm was used to determine the optimal network performance measures. The authors presented a methodology for determining the optimal topological design of a network with the objective of enhancing the overall throughput. Daskalaki and Smith (2004) formulated a multi-objective optimisation problem encompassing optimal routing and buffer allocation in a series-parallel queueing network and used the EM algorithm for network decomposition and Powell's optimisation procedure for optimisation. Cruz et al. (2005) presented a study which analysed the optimal buffer allocation patterns with varying network topologies and arrival rates in state-dependent queueing networks. As an extension, Cruz et al. (2008) analysed the optimal buffer allocation problem in single-server queueing networks by applying a two-moment approximation formula. The authors presented a novel approximation approach to generate optimal buffer allocation patterns, taking into consideration different network topologies and was validated using a simulation model.

In 2016, Li et al. conducted a study in which they modelled a production line of diesel engine cylinder blocks using a finite queueing network to identify the optimal buffer allocations. The study's objective was to reduce congestion and improve the utilisation rates of handling equipment. Smith (2016) modelled a material handling system using a finite CQN and developed a multi-criteria optimisation problem to maximise the system's throughput. The optimal network population and buffer allocation were determined through the use of a branch and bound approach. Zhang et al. (2016) modelled a four-stage flexible flow shop using a finite OQN to evaluate the optimal buffer allocations from a perspective of resource planning and capacity configuration. The authors used the GEM approach to estimate the network performance measures. In 2017, Pedrielli et al. presented an integrated simulation-optimisation approach for a production line task and buffer allocation problem using a discrete event optimisation (DEO) methodology.
Smith (2018) modelled a material handling system using a finite OQN to determine the optimal service rate and buffer allocation. The study considered various network topologies and employed a combination of decomposition and sequential quadratic programming approaches. In 2022, Yu et al. modelled a flow shop environment with batch transport using a finite OQN. The authors employed an approximation method to estimate the performance measures and an integrated iteration optimisation algorithm to determine the optimal buffer allocations.

This research modelled an inter-facility material transfer operation as a multi-class CQN with limited queue capacities. The service stations are represented using M/M/1/C queue models, characterised by exponentially distributed service times and first-come-first-serve queue discipline. The movements between the service stations are modelled as M/M/∞/∞ queues. The model incorporates the BAS phenomenon in resource nodes in case of full occupancy of buffers. The study proposes a simulation-based optimisation model to identify the optimal buffer allocations for each service station in order to maximise the overall system throughput for a fixed population within a specified time frame. Additionally, the model takes into consideration a constraint relating to the upper limits of utilisation rates for all service stations.

3. Methodology
This section briefly explains the employed modelling and optimisation approach of the buffers allocation problem in the study.

3.1. A Discrete Event Simulation (DES) model using a finite CQN approach to investigate the inter-facility material transfer operations
The trucks transporting different types of materials served as the agents in the simulation model. The subprocesses of the material transfer operations, including loading, unloading, weighing and gate processes, were identified as service stations, while the resources in these stations were represented as servers. The model was constructed using the Anylogic Simulation software, where "Service" blocks represented service stations, and the material flow between these stations was modelled using "Delay" blocks. Multiple sources were utilised to differentiate the various agent (product) types. The service blocks in the simulation software integrated the actions of seizure of resource (server), delay time (service time), and agent release upon completion of service. The agents' behaviour was modelled such that, upon completion of service, they attempt to enter the next block. If the next block is unable to accept the agent due to full capacity, the agent remains in a waiting state within the delay block, resulting in a BAS phenomenon. The simulation was configured by utilising parameters, variables, and functions to specify the agents' characteristics and actions related to the study. An optimisation experiment was performed in the Anylogic software to perform the integrated simulation and optimisation process. Figure 3 shows the service block configurations used in the present study.

The validity of the developed simulation model was confirmed through comparison of its outputs with the actual system setup. The reproducibility of the model outputs ensured its stability when using fixed seed settings. To account for any transient behavior, the simulation was run for a total of 10,000 time units, with a warm-up period of 1,000 time units.

3.2. Problem formulation
In a finite queueing system, the relationship between buffer size, throughput, utilisation rates, and cost is a highly nuanced and complex phenomenon. From an optimisation perspective, the buffer size of a system has a direct impact on its throughput. In general, larger buffer sizes are expected to result in higher throughput, but they also come at a cost, which can increase the overall cost of the system. Furthermore, the utilisation rate of a system is closely correlated with the buffer size and the throughput. These relationships demonstrate the importance of considering all relevant factors when optimising buffer size in a queueing system.
Figure 3. Service block properties and configurations

Notations

✓ 3.2.1. Indices

\( i \)  
Index for service stations  
\( i = 1, \ldots, n \)

\( j \)  
Index for product classes  
\( j = 1, \ldots, m \)

✓ 3.2.2. Parameters

\( n \)  
Number of service stations

\( m \)  
Number of product classes

\( N_j \)  
Number of trucks used by product class \( j \)

\( \lambda_i \)  
Arrival rate of trucks to the node \( i \)

\( \mu_i \)  
Mean service rate at node \( i \)

\( C_i \)  
Capacity of node \( i \) (including in service)

\( L_i \)  
Mean queue length in service station \( i \) (excluding in service)

\( U_i \)  
Mean utilisation rate for service station \( i \)

\( UL_i \)  
Upper limits of utilisation rates for service station \( i \)

\( \theta(N_j) \)  
Throughput of class \( j \) as a function of population \( N \)

\( R(N_i^j) \)  
Mean response time for product class \( j \) in service station \( i \) with \( N \) population

\( P \)  
Cost of a buffer space

\( \beta \)  
Budget for buffer space allocation

\( Z \)  
System throughput

3.3. Optimisation problem

This section delves into the modelling aspect of optimising buffer allocations in inter-facility material transfer operations. A mathematical model is developed by utilising the notations presented in the previous section. This model aims to determine the optimal allocation of buffers to maximise the throughput of material transfer operations between facilities.

Decision variable

\( X_i \)  
Number of buffers allocated to service station \( i \) (excluding in service)

Objective function

\[
\text{Maximise } Z = \sum_{j=1}^{m} \theta(N_j)
\]  
(1)

Constraints

\[
\sum_{i=1}^{n} P_i X_i \leq \beta
\]  
(2)

\( X_i \leq L_i \quad \forall \ i \in (1, n) \)  
(3)

\( U_i \leq UL_i \quad \forall \ i \in (1, n) \)  
(4)

\( X_i \in \mathbb{Z}^+ \)  
(5)

The presented study involves a maximisation problem, the objective of which is captured by the objective function (Equation 1), exhibiting the prominent property of concavity (Smith 2016). The first constraint, represented by Equation 2, serves as a knapsack constraint, ensuring that the cost associated with utilising the total number of buffers does not exceed the predetermined budget. Equation 3 imposes a constraint, ensuring that the utilisation of buffers does not surpass the mean queue length in all service stations. Equation 4 stipulates that the utilisation rates of all service stations must be lower than or equal to the pre-established upper limit. Finally, the last constraint (Equation 5) demonstrates that the decision variable is a positive integer-type variable.
3.4. Solution approach
The buffer allocation problem is a well-known NP-hard problem, which can be formulated using conventional mathematical programming techniques, such as linear programming, mixed-integer linear programming, and non-linear programming, to derive an optimal solution. However, these techniques are often limited by the complex trade-offs between different objectives inherent in NP-hard buffer allocation problems, making it difficult to find a globally optimal solution. This study explores the use of simulation-based approaches to model the system's behaviour and evaluate the performance of different buffer allocation scenarios. The discrete event simulation (DES) model is developed using the AnyLogic simulation software and optimised using its OptQuest engine, leveraging metaheuristics and tabu search approach.

4. Case Study – Numerical experiments
Numerical experiments play a critical role in assessing proposed methodologies' comprehensive performance and efficiency. This study focuses on a steel manufacturing facility where selected raw materials, S, R, and T, must be transported from their storage locations to intermediate demand points. The inter-facility material transfer operations are executed by a fleet of trucks and involve several identified sub-processes. The process flow map, routing, and mean service times for each product class are depicted in Figure 4. The full truckload capacities for materials S, R, and T are 4, 6, and 3 tons, respectively. Truck capacities are assumed to be fully utilised during inter-facility material transfer operations.

This study developed a Discrete Event Simulation (DES) model (Figure 5) using AnyLogic software to represent the above inter-facility material transfer operations. The model was used to determine the optimal buffer allocation for all service stations (1-9) through the implementation of a simulation optimisation approach. This approach was applied to several distinct scenarios, as outlined in Table 1.

Table 1. Identified different truck assignment scenarios for optimisation experiments

<table>
<thead>
<tr>
<th>Scenario #</th>
<th>Number of trucks assigned</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Class S</td>
</tr>
<tr>
<td>Scenario 1</td>
<td>5</td>
</tr>
<tr>
<td>Scenario 2</td>
<td>4</td>
</tr>
<tr>
<td>Scenario 3</td>
<td>6</td>
</tr>
</tbody>
</table>

Figure 4. Inter-facility material flows, routing and average service times
The calculations performed in this study were based on a single shift session, which was assumed to be 8 hours in duration. The results of the simulation optimisation approach are discussed in section 5. For the purposes of the study, the following parameter values (Table 2) were used in the calculations.

Table 2. Parameter values for the case study numerical experiments

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upper limit of utilisation rate (UL_i)</td>
<td>0.95</td>
</tr>
<tr>
<td>Cost of a buffer space (P)</td>
<td>100 $</td>
</tr>
<tr>
<td>Budget ((\beta))</td>
<td>4000 $</td>
</tr>
</tbody>
</table>

5. Analysis of Results

The optimal buffer allocations for the network population scenarios identified in Table 1 were determined through a series of optimisation experiments. The results of these experiments are presented in Table 2, which includes the optimal buffer allocations, the corresponding aggregate material transfers between facilities and the total number of utilised buffers for each scenario. The experiments were performed using Anylogic software, with a setup of 10,000 iterations and a memory allocation of 512 MB. The results indicate that the optimal buffer allocations among all service stations were different across all scenarios. Additionally, it was observed that an increase in buffer size could result in higher overall throughput. Figure 6 provides a visual representation of the optimisation result window for scenario 3.

Table 3. Optimal buffer allocations in each scenario

<table>
<thead>
<tr>
<th>Scenario #</th>
<th>Buffer allocation S. Station (1-9)</th>
<th>Total population (trucks)</th>
<th>Aggregate material transfer (tons)</th>
<th>Total # buffers used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenario 1</td>
<td>(1,3,2,1,1,2,8,1,7)</td>
<td>15</td>
<td>777</td>
<td>26</td>
</tr>
<tr>
<td>Scenario 2</td>
<td>(3,3,3,3,3,3,3,3,3)</td>
<td>12</td>
<td>771</td>
<td>27</td>
</tr>
<tr>
<td>Scenario 3</td>
<td>(3,3,3,3,3,3,3,3,3)</td>
<td>18</td>
<td>806</td>
<td>27</td>
</tr>
</tbody>
</table>

Furthermore, Table 3 reveals that while the total number of trucks is higher in scenario 2 compared to scenario 1, the number of buffers used in scenario 2 is also higher. This is due to the fact that the allocation of trucks varies between product classes, leading to differences in buffer allocations for each scenario. Despite the higher number of trucks in scenario 3 compared to scenario 2, the buffer allocations for both scenarios remain the same. In conclusion, it is imperative to thoroughly analyse each scenario to determine the optimal buffer allocation.

The response time is a crucial metric in the evaluation of the performance of a queueing network, as it provides insight into the waiting time experienced by customers before they receive service. This metric is influenced by a range of factors, including the queue length, the number of servers, the service rate, and the arrival rate of customers at the node. In addition, blocking effects can significantly impact the response time in a finite buffer queueing network. This study considered the BAS phenomena in analysing the response times. Figure 7 presents the response time for materials S, R, and T in scenario 3 for the loading, unloading, gate, and weighing sub-processes.
6. Conclusion
This study considers the problem of determining the optimal buffers allocation for an inter-facility material transportation system in a manufacturing facility considering the blocking after-service (BAS) phenomenon. The buffer allocation problem constitutes a pivotal aspect in the design of material transfer operations, as it directly influences the system's overall efficiency and performance. The buffer allocation problem for a finite CQN is a complex and challenging task that requires careful consideration of various factors, such as the service rate, arrival rate, and network structure. The results of this analysis demonstrate the importance of considering buffer allocation in the design and optimisation of queueing networks. By answering the identified research questions, this study contributes to the existing literature on queueing systems by providing insights into how practitioners and researchers can allocate buffers more efficiently in CQN frameworks. The findings of this study have significant implications for various industries, including manufacturing and service sectors, where efficient management of queueing systems is critical for overall system performance. The results of this study show that organisations can make informed decisions about allocating resources, such as buffer sizes, that will lead to improved performance and reduced costs. Furthermore, utilising a robust tool such as finite CQN models for modelling inter-facility material transfer operations proves substantial value to decision-making at various managerial levels. The study has also showcased the advantages and effectiveness of integrating simulation optimisation techniques in addressing complex buffer allocation problems.

Acknowledgments
The authors are grateful to Hamad Bin Khalifa University, Doha, Qatar, for supporting this study.
References


Biographies

**Mohamed Amjath** is currently reading for his PhD. in Logistics and Supply Chain Management at Hamad Bin Khalifa University, Doha, Qatar. He holds an MSc degree in Maritime Affairs (Shipping Management & Logistics) from World Maritime University, Malmo (Sweden) and a BSc degree in Transport & Logistics Management from the University of Moratuwa (with a first-class). Currently, Mohamed Amjath is working on queueing network applications in material handling system optimization.

**Laoucine Kerbache** is currently serving as a Full Professor and a founding faculty member of the Engineering Management and Decision Sciences Division at Hamad Bin Khalifa University in Doha, Qatar. Further, he has
been Full Professor of Operations and Supply Chain Management at HEC Paris in France for the last twenty-two years. During this tenure period, beside managing many international academic programs, he has also served as Associate Dean of the HEC Paris PhD Program (5 years) and then as Dean and CEO at HEC Paris Qatar (4 years). For over thirty years of academic, research, and consulting activities, he has been very active in his areas of expertise, Operations and Supply Chain Management. He has published over 120 papers in international journals and has been guest speakers at numerous international events. He holds a PhD, an MSc, and a BSc in Industrial Engineering and Operations Research (IEOR) from the Mechanical and Industrial Engineering Department, University of Massachusetts, Amherst, USA. Further, he has an “Habilitation to Direct Research (HDR)” from the University of Nantes (France) and a certification from the “International Teachers Program” from IMD Lausanne (Switzerland).

Adel Elomri is an Assistant Professor of Logistics and Supply Chain Management in the division of Engineering Management and Decision Sciences at the College of Science and Engineering at Hamad Bin Khalifa University (CSE-HBKU). He holds Ph.D. and MSc degrees in Operations Management (with Highest Honors) from CentraleSupélec Paris (France) and a BSc degree in Industrial Engineering from the National Engineering School of Tunisia (with distinction). Dr. Elomri has more than 15 years of international experience as lecturer and researcher in Supply Chain and Operations Management with research interests lying at the interface of operations research, economics, and engineering, with a special focus in modeling and analyzing supply chain networks. Currently Dr. Elomri has projects underway in the areas of healthcare operations management, sustainable supply chain management, smart logistics, and production and operations management.

James MacGregor Smith graduated with a B.Arch. and M. Arch. from the University of California at Berkeley and a Ph.D. in Operations Research from the University of Illinois in Champaign–Urbana. Professor Smith conducts research on topological network design, stochastic network design and analysis, and facility layout and location problems. In particular, he is doing research on Steiner minimal trees in 3D applications of Steiner Trees to Minimum Energy Configurations) and protein modeling. He is also working on state-dependent queueing network analysis and finite buffer queueing network models, quadratic assignment, and set packing problems. Applications include the design and layout of manufacturing plants, health care facilities, and many other systems. He has published in many of the industrial engineering and operations research journals concerned with optimization and stochastic processes.