Applied Mathematical Modelling 34 (2010) 2074–2088

Contents lists available at ScienceDirect

Applied Mathematical Modelling

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

Modeling a supply chain using a network of queues

Vidhyacharan Bhaskar [a,*], Patrick Lallement [b]

[a] Department of Electronics and Communication Engineering, S.R.M. University, Kattankulathur, Kancheepuram District, Tamilnadu 603203, India
[b] Institut Charles Delaunay, Universite de Technologie de Troyes, 12 Rue Marie Curie, 10000 Troyes, France

ARTICLE INFO

Article history:
Received 12 March 2009
Received in revised form 8 October 2009
Accepted 14 October 2009
Available online 21 October 2009

Keywords:
Average queue lengths
Average response times
Average waiting times
Utilizations
Steady-state probability
Supply chain

ABSTRACT

In this paper, a supply chain is represented as a two-input, three-stage queuing network. An input order to the supply chain is represented by two stochastic variables, one for the occurrence time and the other for the quantity of items to be delivered in each order. The objective of this paper is to compute the minimum response time for the delivery of items to the final destination along the three stages of the network. The average number of items that can be delivered with this minimum response time constitute the optimum capacity of the queuing network. After getting serviced by the last node (a queue and its server) in each stage of the queuing network, a decision is made to route the items to the appropriate node in the next stage which can produce the least response time.

1. Introduction

1.1. Overview

Queuing models have been used to investigate supply chain problems for many years. In the 1940s, queuing models were used to solve a variety of machine interference problems, i.e., how many repair persons are needed to be assigned to properly maintain a system, or how many telephone operators are required to handle traffic calls. Queuing models are used to analyze tradeoffs concerning the number of servers versus the waiting time of the customers. Clearly, if the number of servers is high, the cost of the servers is high, but the waiting time (cost of customer idle time) is low.

Queuing models calculate the optimum number of customer/order service points (servers) to minimize cost for business. It considers the average arrival rate of orders, the average customer service rate, the cost to the business of order waiting time (customer dissatisfaction), and the cost to operate customer service points. Queuing models are used to obtain apriori information not only about important performance measures like queue lengths, response times, and waiting times, but also other performance measures like: (a) probability that any delay will occur, (b) probability that the total delay is greater than a predetermined value, (c) probability that all service facilities will be idle, (d) expected idle time of the total facility, and (e) probability of turnaways due to insufficient waiting accommodation. Some kind of queuing problems involve determining the appropriate number of service facilities to cover expected demand, as well as determining the efficiency of servers and the number of servers of different types at the service facilities [1]. Suri suggested [2] the use of queuing theory to provide quick solutions to supply chain problems.

Current generation enterprizes such as global supply chains, virtual enterprizes and e-businesses are driving research in the area of enterprize modeling framework suitable for a distributed environment. Supply Chain (SC) is a concept which can

* Corresponding author.
E-mail address: meetvidhyacharan@yahoo.com (V. Bhaskar).

0307-904X/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.apm.2009.10.019

be considered analogous to a pipeline of physical and informational flows between suppliers and customers. From an operational point of view, this pipeline works like a process of activities, and these activities are distributed. So, the term "chain" could be replaced by the term "network" [3]. Each company is at the center of a network of suppliers and customers.

The supply chain could be defined as *a network of connected and interdependent organizations mutually and co-operatively working together to control, manage, and improve the flow of materials and information from suppliers to end users* [4]. Since the supply chain management (SCM) is a market-driven concept, it is necessary to adopt the customer's point of view. For this reason, the concept "process" has been introduced in logistics for strategic reasons [5]. The definitions for "process" and "activity" have been specified in [6]. The activities represent the system functionality. They can be scheduled, and they need time and resources. The term "process" represents the global behavior of the industrial system. It is a logical sequence of activities to realize a predefined objective. A process can be planned, but rarely scheduled. The objective is generally expressed in terms of delay, quantity and quality. Among these attributes, the delay is generally the most critical one. In this paper, we refer to the central process of any industrial system as "treatment of orders". It reflects the fact when considering the performance of a SCM system that the inputs are the "orders" and the outputs are the "goods".

A process is composed of activities that use resources which are network-configured. There are other processes in industrial systems, such as supply or maintenance, that can be considered to be collaborative processes of the main process. Fig. 1 illustrates how the logical concept is mapped to the "physical system". In this figure, successive activities represent stages (or steps) in a given process. Each of them is realized in a site with specific resources. In the general case, from a given activity output, there are many possible connections to the next activity. If sites are geographically scattered (sites of the same company, subcontractors), a transport activity must be inserted between two transformation activities. In manufacturing systems, the production nomenclature means "assembling of components" and "convergence of physical flows" to a final point.

When the treatment of orders is made, their evaluation consists in a comparison between the objective and the result. The global challenge of the process approach is: (a) to initialize correctly, each process objective with a realistic delay value. The delay objective to be assigned represents the expected value of the delay (or lead-time) plus a security margin. It can be derived from the statistics (average response time) or from the actual state of the system in terms of waiting times and service times at the nodes. In this case, it is necessary to convert all waiting activities and services to be executed in a global throughput time, and to choose the route that minimizes the lead-time. The same evaluation results may also be useful for (b) a negotiation with a potential client during an e-business or an e-commerce transaction. For the company, it represents the lead-time promised for a given order.

The main challenge of the SCM system is to improve the performance while reducing the costs (generally in terms of trade-offs). The challenge addressed in this paper is to represent a physical network of resources with a queuing model. Each resource is modeled as a server and waiting activities are in a queue. More precisely, an activity is a logical object which contains attributes such as: reference process (order number), quantity to produce, and objective delay. Throughout this paper, we assume that a process will correspond to each order. By the virtualisation induced using process approach, the object process tracks the physical flow. Among classical performance measures obtained by a queuing representation (average queue length, average response time, etc.), this will lead to estimating a minimum lead-time. The computing challenge is to determine the best strategy to setup a process in terms of delay measures. This efficiency is measurable with the number of processes which fulfil their objective.

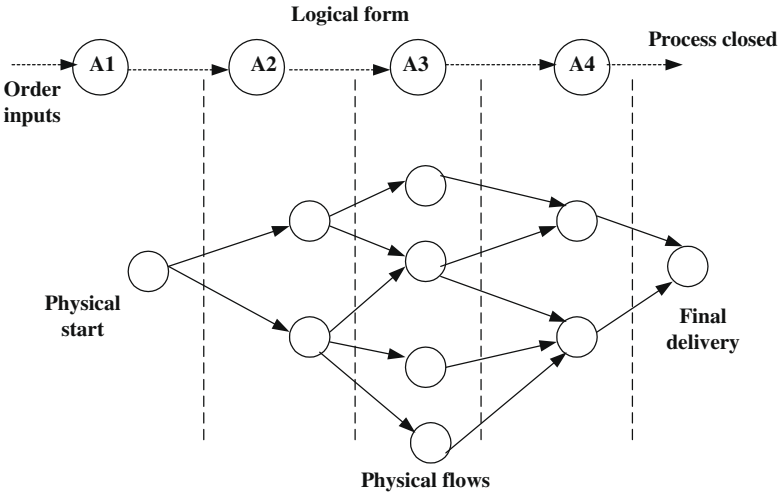

Fig. 1. Physical flow diagram in the supply chain.

All processes are supervised and need to be controlled. The control is generally performance-centric, which means that during the life of a process cycle, a drift situation can be detected between the result and the objective, and corrections can be applied. Since processes are in competition to access resources, and since resources are capacity-limited, the drift situation can be due to breakdown problem of a resource, set-up times, and activities (inventory, transport) of interfaces. One of the correction variables is the possibility to re-route the physical flow from one node to an alternative resource for the next activity (if several ways are possible). This local challenge is similar to point (a), except that the route includes the breakpoint. This is a routing problem and is similar to those that have been addressed in the field of telecommunication networks.

1.2. Organization of the paper

The objective of this paper is to compute the minimum response time for the delivery of an item to the final destination along the three stages of the queuing network. The average number of items that can be delivered with this response time constitute the capacity of the network. Section 2 describes the supply chain (textile manufacturing system) and discusses the literature review in detail. Section 3 presents a queuing network approach to model a textile manufacturing system. Closed-form expressions are derived for utilizations for each node (queue and server) in the network. Section 4 derives, plots and discusses performance measures like average response times, average queue lengths, and average waiting times of individual nodes and different paths in the network. This section also discusses the average queue lengths, average response times, and equivalent service rate of the equivalent single queue, single server network. Section 5 describes the numerical results. Finally, Section 6 presents the conclusions.

2. Supply chain description and literature review

2.1. Textile manufacturing system

The supply chain constitutes some basic activities. They are (i) Knitting, (ii) Making, and (iii) Distribution (central warehousing). The Warehouse corresponds to an European Warehouse. For performance evaluation, the supply chain is modeled by a process with these three activities (three stages). In fact, these activities may be supported by operational resources physically distributed in many sites and interlinked by transportation. Consider Fig. 2.

- *Knitting* locations are in **L1** (France) and **L2** (Morocco).
- *Making* locations are in **L1** and **L2**, **L3** (Morocco), and **L4** (Tunisia). The *warehouse* location is in **L1**.

There are routing choices for the physical flows at two steps of the processes. They are at:

- *Knitting*: From S_0 to **(L1 or L2)**.
- *Making*: From **L1** to **(L1 or L3)**, (or) from **L2** to **(L2 or L4)**.

Each resource is modeled as a queue where batches are waiting to be processed (see Fig. 2). The routing decision may be performed considering the estimated throughput delay from S_0 to S_1. This delay includes the manufacturing delay (depending on the batch quantities to be processed) and the total waiting times in all the downstream queues. Comparing with the routing problem in telecommunication networks (IP networks), the problem is not a hop by hop problem, but we consider the whole route to make the decision.

We can easily separate the global problem to the example given in this paper by providing this example as an illustration of a more general issue. It is interesting to focus on processes which use resources that are network-configured. In other words, those orders which follow a particular distribution on the arrival rate will be configured to the network. We can thus propose an interest to queuing modeling by considering the arrival and departure processes modeled using a particular distribution.

For each output of **L1** and **L2** in Stage I, there are two possible connections (two routes), and this example is like any assembling system. The output of **L1** of Stage I is connected to nodes **L1** and **L3** of Stage II, whereas the output of **L2** of Stage I is connected to nodes **L2** and **L4** of Stage II. Finally, the departures from servers, A11 and A13 arrive at S_1, and the departures from servers $A12$ and $A14$ arrive at S_1. Thus, the nodes in different paths are not cross-linked. It is important to deal with this special case to accommodate the case of "urgent orders" and "regular orders". Urgent orders correspond to orders which require quick processing and regular orders correspond to orders which require normal processing. The orders can be routed appropriately to **L1** and **L2** of Stage I, and subsequently to **L1**, **L2**, **L3**, and **L4** of Stage II for processing.

In order to make a comparison of this special industrial system to a more generic supply chain, the service times and order arrivals can have a different distribution than that considered in this paper. For example, the service time could be modeled as a Lognormal distribution proposed to model supplier delay. A G/M/1 or G/G/1 queue could be used to model a generic supply chain. The global challenge is to be able to estimate an "apriori" performance measure which is necessary to propose a suitable Quality of Service (QoS) (for eg., minimum response time) to the client.

Fig. 1: Queuing formulation of the network of processes

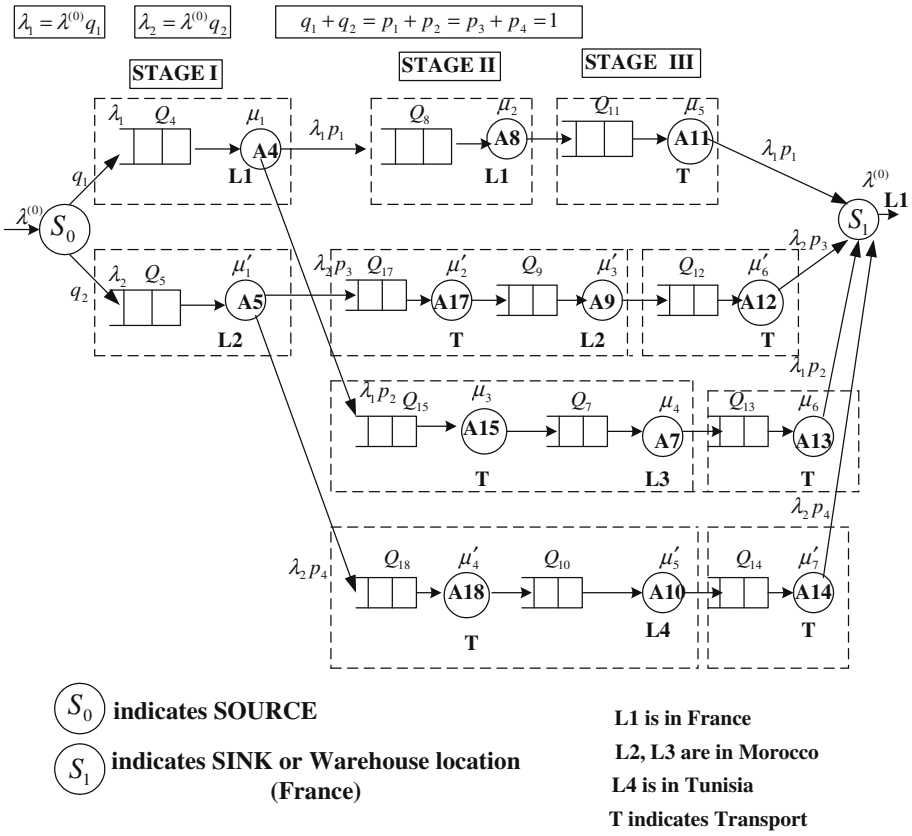

Fig. 2. Block diagram for Queuing formulation of the network of processes (2-input network).

The various kinds of difficulties in modeling the SCM system using queues other than M/M/1 queue are described as follows:

(1) M/M/1 queues have Poisson arrival process and exponentially distributed service times. State description for M/M/1 queuing model is simple as one needs just a number in the system denoting the system state. This is possible because the exponential service-time distribution is memoryless. For M/G/1 queues, where the arrival process is Poisson, but service times have a general (arbitrary) distribution, the general state description would require specifications on both the number in the system and the amount of service already provided to the customer currently being served.
(2) The G/M/1 queue is the dual of M/G/1 queue, where the arrival process is a general one, but the service times are exponentially distributed. The state descriptions are found under equilibrium conditions at the time instants just before job arrivals to the system. The state distributions are also valid for the departure instants (just after a job leaves the system) as Kleinrock's principle is applicable to this system. The state distributions are not valid at arbitrary time instants (or ergodic, time-average results), since Poisson Arrival See Time Averages (PASTA) will not be applicable to the system (i.e., the arrival process is not Poisson).
(3) For $M/E_k/1$ or $E_k/E_k/1$ queues, where E_k is the Erlang distribution with k phases, the probability of packet loss or the probability of packet delay can be determined according to various assumptions made to find if the blocked orders are aborted (Erlang B) or blocked orders are queued until served (Erlang C) (Erlang B and Erlang C formulas are in everyday use for traffic modeling or transportation applications).
(4) For G/G/1 and G/G/m queues (m is the number of servers), only when the offered traffic is high (i.e., utilization, ρ, gets close to 1), the distribution of the waiting time will be approximately exponentially distributed. The waiting times become very large as $\rho \approx 1$. For other values of offered traffic, the distribution of the waiting time has a general distribution, thus making state descriptions not valid at arbitrary time instants. Thus, it is reasonable to use M/M/1 queues to offer a simple and feasible solution to the given SCM problem.

The importance of minimum response time estimation in the supply chain network is given below:

In a distributed hard real-time system, such as the supply chain problem considered in this paper, communication between tasks on different processors must occur in bounded time. The inevitable communication delay is composed of both

the delay in transmitting a message on the communication media, and also the delay in delivering the data to the destination task. A simple delivery approach is considered in this paper, the arrival of an "order" generates an interrupt called "on-demand" approach. As soon as the order arrives at the source node, it is routed to all intermediate nodes leading to the destination node. The objective is to find the path between the source node and the destination node which provides the minimum response time.

The shortest path problem in the dynamic supply chain network considered in this paper is a problem of sending an order from an origin node to a destination node with the least delay over a network that has no perfect, permanent fixed structure, and which is subjected to varying volumes of traffic. The optimal path connecting the origin and destination nodes through several intermediate nodes is called the shortest path since it produces the least response time. Once a shortest path is identified, care should be taken to see that all incoming orders are not dumped onto this path, thereby causing congestion on the shortest path route. So, it is advisable to increase the service rate or reduce the service time of servers on all non-shortest path routes, thereby redistributing the incoming orders to balance the load (offered traffic).

2.2. Literature review

There has been quite a number of research papers published in the area of modeling e-businesses, enterprize systems, assembly and manufacturing systems using queuing networks.

The contributions and applications of queuing theory in the field of discrete part manufacturing is discussed in [7]. Provided are concise, descriptive summaries, rather than detailed mathematical models of the various queuing theory results in the manufacturing context. In [8], a discrete-time Markov chain is developed to model the routing of new emails through a contact center. An open queuing network is used to model the email customer contact center. The fundamental matrix of the absorbing Markov chain developed in [8], is used to obtain the average number of visits an email makes to a particular node before getting resolved. In [9], the joint equilibrium distribution of queue sizes in a network of queues containing N service centers and R classes of customers is derived. Also, the equilibrium state probabilities for both open and closed queuing networks are derived.

Queuing network model is a very useful tool to analyze the performance of a system from an abstract model. In [10], a transformation technique is proposed from Unified Modeling Language (UML) to queuing network model. This approach avoids the need for a prototype implementation since we can determine the overall performance from the architectural design description. A man–machine system is modeled and analyzed in [11] using graphical simulation software package. The validated model is used for evaluating alternate routings to find out the optimum route. The goal of research in [12] is to develop a simulation model for overhead monorail conveyor systems and statistical methods for the analysis and multi-objective optimization of the manufacturing process. Such systems provide connectivity to large area and buffering to streamline the material flow between machinery. In [13], the authors analyze the memory interference caused by several processors simultaneously using several memory modules. The assumptions and results of the simple model are tested against some measurements of program behavior and simulations of systems using memory references from real programs. In [14], the problem of assigning the best service rate to minimize the expected delay under a cost constraint is considered. Also studied are systems with several types of customers, general service-time distributions, stochastic or deterministic routing, and a variety of service regimes.

A model of a closed queuing network within which customer routing between queues depends on the state of the network, is presented in [15]. The routing functions allowed may be rational functions of the queue lengths of various downstream queues which reside within special subnetworks called p-subnetworks. In [16], the focus is on characterizing the average end-to-end delay and maximum achievable per node throughput in random access multihop wireless adhoc networks with stationary nodes. The random access multihop wireless networks can be modeled as a G/G/1 queuing network model, and uses the diffusion approximation to evaluate closed-form expressions for the average end-to-end delay. For closed product-form queuing networks with n customers, the Sevcik–Mitrani arrival theorem in [17] states that an arriving customer would see the network in equilibrium with one less customer. Muppala et al. use the stochastic reward nets (SRN) for the compact specification, automatic generation and solution of large Markov chains in [18]. This allows them to solve large and complex models. Closed-form solutions have been derived for the response time distributions through a particular path in open product-form queuing networks in [19]. A multi-layered queuing network that models a client–server system where clients and servers communicate via synchronous and asynchronous messages is discussed in [20]. The queuing network is approximately analyzed using a decomposition algorithm.

In [21], the authors compute approximations for response time distributions for queuing networks with Poisson or phase-type arrival processes and general service-time distributions. In [22], the performance evaluation of an assembly system with components or sub-assemblies feeding into a kitting and assembly stage framework is studied. In that example, the quantity of the orders were considered as a constant. This could be the case in some supply chains when orders are generated by important applicants (for eg., supermarkets). In [23], Whitt described the Queuing Network Analyzer (QNA), a software package developed at Bell Laboratories to calculate approximate congestion measures for a network of queues. Congestion measures for the network as a whole are obtained by assuming as an approximation that the nodes are stochastically independent given the approximate flow parameters.

A lot of focus on queuing models of manufacturing systems, and many approximations for evaluating the performance of queuing networks are carried out in [24–28]. When the clients are heterogeneous goods (simple customers, shops, stores,

etc.), the quantity of orders may vary largely and they have to be modeled as a stochastic variable. Without any additional information about the economical context, the best way is to assume a uniform distribution for the quantity of orders, thus characterizing the quantity of orders to lie between a minimum and a maximum value. In this case, any input (order) to the queuing system has to be represented by two stochastic variables, one for the time of occurrence, and one for the quantity to deliver. This constitutes the main improvement of this paper over [29].

3. Queuing network description

We shall consider one type of product (tee-shirt) in the supply chain given in Fig. 2. Orders arrive in one portal, but processes can start in two places. There are three stages (i.e., three activities in any process): Knitting, Making and Delivery, which can be realized in four sites. Because of the network structure, different routes are possible depending on the traffic, which implies transport activity is necessary. All physical flows converge to a central warehouse. The analysis of the two-input, three-stage queuing network is made as follows:

There are 2-inputs in the queuing network considered in Fig. 2. The two-input queuing network receives orders from clients, and the orders are waiting to be served. The quantity of items to be delivered in each order is assumed to be uniformly distributed. The arrival rates at the 2-inputs are λ_1 and λ_2, respectively. The arrival rate at source (S_0) is $\lambda^{(0)}$. The probability of arrivals at Q_4 and Q_5 are q_1 and q_2, respectively. It is clear that $q_1 + q_2 = 1$. Let $\lambda_1 = \lambda^{(0)} q_1$ be the arrival rate of the jobs at Q_4, and let $\lambda_2 = \lambda^{(0)} q_2$ be the arrival rate at Q_5. Let the service rates of servers, $A4$ and $A5$ be μ_1 and μ_1', respectively. After getting serviced by server $A4$, the jobs arrive at the queues Q_8 and Q_{15} with probabilities p_1 and p_2, respectively, where $p_1 + p_2 = 1$. So, the arrival rate at Q_8 is $\lambda_1 p_1$ and the arrival rate at Q_{15} is $\lambda_1 p_2$. After getting serviced by server $A8$, the jobs arrive at queue Q_{11} with arrival rate $\lambda_1 p_1$. The service rates of servers $A8$ and $A11$ are μ_2 and μ_5, respectively.

Now, queues Q_{15} and Q_7 are in serial connection. So, the jobs which are serviced by server $A15$ are again serviced by server $A7$, after waiting at Q_7. The service rates of servers $A15$ and $A7$ are μ_3 and μ_4, respectively. The arrival rate at Q_7 is $\lambda_1 p_2$. After getting serviced by server $A7$, the jobs arrive at queue $Q13$ with the same arrival rate, $\lambda_1 p_2$, and get serviced by server $A13$. The service rate of server $A13$ is μ_6. After getting serviced by server $A5$, the jobs arrive at the queues Q_{17} and Q_{18} with probabilities p_3 and p_4, respectively, where $p_3 + p_4 = 1$. So, the arrival rate at Q_{17} is $\lambda_2 p_3$, and the arrival rate at Q_{18} is $\lambda_2 p_4$.

Now, queues Q_{17} and Q_9 are in serial connection. So, the jobs which are serviced by server $A17$ are again serviced by server $A9$, after waiting at Q_9. The service rates of servers $A17$ and $A9$ are μ_2' and μ_3', respectively. The arrival rate at Q_9 is $\lambda_2 p_3$. After getting serviced by server $A9$, the jobs arrive at queue Q_{12} with the same arrival rate, $\lambda_2 p_3$, and get serviced by server $A12$, whose service rate is μ_6'.

Similarly, queues Q_{18} and Q_{10} are in serial connection. So, the jobs which are serviced by server $A18$, are again serviced by server $A10$, after waiting at Q_{10}. The service rates of servers $A18$ and $A10$ are μ_4' and μ_5', respectively. The arrival rate at Q_{10} is $\lambda_2 p_4$. After getting serviced by server $A10$, the jobs arrive at queue Q_{14} with the same arrival rate, $\lambda_2 p_4$, and get serviced by server $A14$, whose service rate is μ_7'. Finally, jobs after service completion at servers, $A11$, $A12$, $A13$ and $A14$, arrive at the sink, S_1, with departure rate $\lambda^{(0)}$. From Fig. 2, we have:

$$\lambda_1 = \lambda^{(0)} q_1, \quad \lambda_2 = \lambda^{(0)} q_2,$$
$$\lambda_1 + \lambda_2 = \lambda^{(0)}, \quad \text{and} \quad p_1 + p_2 = p_3 + p_4 = q_1 + q_2 = 1.$$

For satisfactory management and control requirements, it can be assumed that $\mu_1 = \mu_1'$ and $\mu_2 = \mu_3' = \mu_4 = \mu_5'$. Fig. 3 represents the overall system in terms of (i) Production and (ii) Delivery.

A node is defined by a queue and its corresponding server.

- The nodes in Stage I are $(Q_4, A4)$, $(Q_5, A5)$.
- The nodes in Stage II are $(Q_8, A8)$, $(Q_{17}, A17)$, $(Q_9, A9)$, $(Q_{15}, A15)$, $(Q_7, A7)$, $(Q_{18}, A18)$, $(Q_{10}, A10)$.
- The nodes in Stage III are $(Q_{11}, A11)$, $(Q_{12}, A12)$, $(Q_{13}, A13)$, $(Q_{14}, A14)$.

Each activity belongs to a specific process. Each activity is an object which describes a specific task that the resource has to do. Here, *Ai*'s are the differential activities carried out in this supply chain. The activities, $A4$ and $A5$ are called "Knitting",

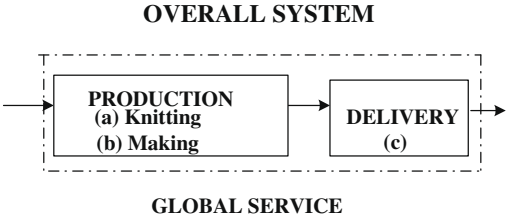

Fig. 3. Overall system.

activities A8, A9, A7 and A10 are called "Making", and activities A17, A15, A18, A11, A12, A13 and A14 are called "Transporting". Let the service rates of A4, A8, A15, A7, A11, A13 be μ_1, μ_2, μ_3, μ_4, μ_5 and μ_6, respectively.

From the client's point of view, a supply chain is equivalent to a queue. The queue is receiving orders. These orders are waiting to be served. The service is a production center and the results are products, items, etc.

Orders are characterized by (i) Occurrence, (ii) Quantity, and (iii) Delay.

(i) Occurrence (λ): This could be stochastic in nature (Poissonian) or deterministic.
(ii) Quantity: This is the quantity of items to be delivered. They are stochastic in nature (uniform distribution).
(iii) Delay: This is the main QoS indicator.

Since the occurrence (λ) is stochastic in nature with Poissonian distribution, and the quantity of jobs to be processed (for each occurrence of λ) is uniformly distributed, we have an equivalent random variable Z that is a function of both X and Y, where

$X \triangleq$ random variable denoting the occurrence time of an order,
$Y \triangleq$ random variable denoting the number of items in each order, and
$Z \triangleq XY$ denotes the occurrence time along with the number of items in each order.

The cumulative distribution function of Z is [30]:

$$F_Z(z) = P(Z \leqslant z) = \int_{A_z} \int f_{XY}(x,y)dxdy, \tag{1}$$

where A_z is a subset of R^2 given by $A_z = \{(x,y)|\Phi(x,y) \leqslant z\}$. Since $f_X(x)$ and $f_Y(y)$ are independent (we assume that with each occurrence, the number of items are random), we have $f_{XY}(x,y) = f_X(x)f_Y(y)$. Hence,

$$F_Z(z) = \int_{A_z} \int f_X(x)f_Y(y)dxdy. \tag{2}$$

Let X be exponentially distributed random variable, and Y be a random variable, uniformly distributed between a and b ($b > a$). So,

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}; & x \geqslant 0 \\ 0; & x < 0 \end{cases},$$

$$\text{and} \quad f_Y(y) = \begin{cases} \frac{1}{b-a}; & a < y < b \\ 0; & \text{otherwise} \end{cases}.$$

The region ΔD_z such that $z < xy < z + dz$ is the portion of the curve lying between the outer boundaries of the two rectangular hyperbolas, $y = \frac{z}{x}$ and $y = \frac{z+dz}{x}$. The coordinates of a point in this region are $\frac{z}{x}, x$ and $|\frac{dz}{dy}| = |x|$. So, $|\frac{dy}{dz}| = \frac{1}{|x|}$. The area of a differential equals $\frac{1}{|x|}dxdz$. Since the random variables X and Y are independent, the probability density function of Z is given by Papoulis [31] and Rohatgi [32]:

$$f_Z(z) = \frac{\lambda}{b-a}\int_{z/d}^{z/c} \frac{e^{-\lambda x}}{|x|}dx \;\forall\; 0 < z < \infty. \tag{3}$$

Making change of variables in the integral of (3) by substituting $t = \lambda x$ and $dt = \lambda dx$, we have:

$$f_Z(z) = \frac{\lambda}{b-a}\left(E_1\left(\frac{\lambda z}{b}\right) - E_1\left(\frac{\lambda z}{a}\right)\right), \tag{4}$$

where $E_1(x) = \int_x^\infty \frac{e^{-u}}{u}du$ is the exponential integral defined by $E_n(x) = \int_1^\infty \frac{e^{-xt}}{t^n}dt$ at $n = 1$ [33]. Now, $E_1(x) = -E_i(-x)$, where $E_i(x) = -\int_{-x}^\infty \frac{e^{-t}}{t}dt$ is the exponential integral function [33]. Substituting the relation between $E_i(x)$ and $E_1(x)$ in (4), we have:

$$f_Z(z) = \frac{\lambda}{b-a}\left(E_i\left(-\frac{\lambda z}{a}\right) - E_i\left(-\frac{\lambda z}{a}\right)\right). \tag{5}$$

The mean value of the occurrence time, $E(Z)$ is given by

$$E(Z) = \int_0^\infty zf_Z(z)dz = \frac{\lambda}{b-a}\int_0^\infty z\left(E_i\left(-\frac{\lambda z}{a}\right) - E_i\left(-\frac{\lambda z}{b}\right)\right)dz. \tag{6}$$

From [34], the integral shown in (6) can be evaluated as

$$E(Z) = \frac{\lambda}{b-a}\left[\frac{z^2}{2}\left(E_i\left(-\frac{\lambda z}{a}\right) - E_i\left(-\frac{\lambda z}{b}\right)\right) + \left(\frac{az}{2\lambda} + \frac{a^2}{2\lambda^2}\right)\exp\left(-\frac{\lambda z}{a}\right) - \left(\frac{bz}{2\lambda} + \frac{b^2}{2\lambda^2}\right)\exp\left(-\frac{\lambda z}{b}\right)\right]_0^\infty = \left(\frac{b+a}{2\lambda}\right). \tag{7}$$

Thus, the inter-arrival times of the orders (occurrence and quantity) are exponentially distributed with mean $E(Z)$. Service times of orders are independent identically distributed random variables, the common distribution being exponential with mean $\frac{1}{\mu}$, where μ is the service rate.

Assume that orders are served in their sequence of arrivals (FCFS scheduling). If the "order" denotes a job arriving into a computer system, then the server represents the computer system. Let $N(t)$ denote the number of orders in the system (those queued plus the one in service) at time t. Then $\{N(t)|t \geqslant 0\}$ is a birth–death process with minimum arrival rate:

$$\Lambda_k = \lambda^{(0)} = \frac{1}{E(z)} = \frac{2\lambda}{b+a}, \tag{8}$$

and service rate $\mu_k = \mu; k \geqslant 1$. The ratio:

$$\rho = \frac{\text{mean service time}}{\text{mean interarrival time}} = \frac{\Lambda_k}{\mu_k} = \frac{1}{\mu E(z)} = \frac{2\lambda}{\mu(b+a)} \ \forall\, a, b > 0, b > a. \tag{9}$$

The quantity, ρ, is an important parameter, called the traffic intensity of the system. Traffic intensity is usually expressed in Erlangs. From the birth–death process for continuous-time homogeneous Markov chains, the steady-state probability of having k jobs in the system with batch arrivals is given by Tran and Do [35]:

$$\Pi_k = (\exp(-(1-\rho)))^k \Pi_0 = \exp(-k(1-\rho))\Pi_0 \ \forall\, a, b > 0, b > a. \tag{10}$$

Summing (10) from 0 to ∞ and equating the result to 1, we have $\Pi_0 = \exp(1-\rho) - 1$ provided $\rho < 1$, i.e., when the traffic intensity is less than unity. The server utilization is $U_0 = 1 - \Pi_0 = 2 - \exp(1-\rho)$. It can be shown that the mean and variance of the number of customers in the system are

$$E[N] = \sum_{k=0}^{\infty} k\Pi_k = \Pi_0 \sum_{k=0}^{\infty} k \exp(-k(1-\rho)) = \frac{1}{1-\exp(-(1-\rho))}, \tag{11}$$

and

$$\sigma_N^2 = \sum_{k=0}^{\infty} (k - E(N))^2 \Pi_k = \frac{\exp(-(1-\rho))}{(1-\exp(-(1-\rho)))^2}, \tag{12}$$

respectively.

Let the random variable R denote the response time (defined as the time elapsed from the instant of job arrival until its completion) in the steady-state. In order to compute the average response time $E[R]$ we use the well-known Little's theorem, which states that the mean number of jobs in a queuing system in the steady-state is equal to the product of the arrival rate and mean response time. When applied to the present case, Little's formula gives us $E[N] = \lambda E[R]$ [30]. Hence:

$$E[R] = \frac{E[N]}{\lambda} = \frac{1}{\lambda(1-\exp(-(1-\rho)))}. \tag{13}$$

Note: Congestion is present in the system, and hence the mean response time, $E[R]$, builds rapidly as the traffic intensity, ρ increases.

Let the random variable W denote the waiting time in the queue. The average waiting time:

$$E[W] = E[R] - \frac{1}{\mu} = \frac{\mu - \lambda + \lambda \exp\left(-\left(1 - \frac{2\lambda}{\mu(b+a)}\right)\right)}{\lambda\mu\left(1 - \exp\left(-\left(1 - \frac{2\lambda}{\mu(b+a)}\right)\right)\right)}, \tag{14}$$

$\forall\, a, b > 0, b > a$. If now, let the random variable Q denote the number of jobs waiting in the queue (excluding those, if any, in service), then, to determine the average number of jobs $E[Q]$ in the queue, we apply Little's formula to the queue excluding the server to obtain:

$$E[Q] = \lambda E[W] = \frac{\mu - \lambda + \lambda \exp\left(-\left(1 - \frac{2\lambda}{\mu(b+a)}\right)\right)}{\mu\left(1 - \exp\left(-\left(1 - \frac{2\lambda}{\mu(b+a)}\right)\right)\right)}. \tag{15}$$

Note that the average number of jobs found in the server is

$$E[N] - E[Q] = \frac{\lambda\mu - \mu + \lambda - \lambda\exp(-(1-\rho))}{\lambda\mu(1-\exp(-(1-\rho)))}. \tag{16}$$

Stage I: $A4, A5 \rightarrow$ Knitting
Stage II: $A7, A8, A9, A10 \rightarrow$ Making$A15, A17, A18 \rightarrow$ Transport
Stage III: $A11, A12, A13, A14 \rightarrow$ Transport.

Case (i): Nodes A4, A8, A15, A7, A11 and A13 constitute the nodes visited from Input 1

The service rates of the servers $A4, A8, A15, A7, A11$, and $A13$ are $\mu_1, \mu_2, \mu_3, \mu_4, \mu_5$, and μ_6, respectively. The arrival rates at the queues of these servers are $\lambda_1, \lambda_1 p_1, \lambda_1 p_2, \lambda_1 p_2, \lambda_1 p_1$, and $\lambda_1 p_2$, respectively. The utilizations of these servers are

$$\rho_i^{(j)} = \frac{\lambda^{(0)} w_i}{\mu_i} = \frac{2\lambda}{b+a}\frac{w_i}{\mu_i}, \tag{17}$$

$\forall j = A4, A8, A15, A7, A11$ and $A13$, corresponding to $i = 1, 2, 3, 4, 5$, and 6, respectively, $w_1 = q_1$, $w_2 = w_5 = q_1 p_1$, $w_3 = w_4 = w_6 = q_1 p_2$.

Case (ii): Nodes A5, A17, A9, A18, A10, A12 and A14 constitute the nodes visited from Input 2

The service rates of the servers $A5, A17, A9, A18, A10, A12, A14$ are $\mu_1', \mu_2', \mu_3', \mu_4', \mu_5', \mu_6'$, and μ_7', respectively. The arrival rates at the queues of these servers are $\lambda_2, \lambda_2 p_3, \lambda_2 p_3, \lambda_2 p_4, \lambda_2 p_4, \lambda_2 p_3$, and $\lambda_2 p_4$, respectively. The utilizations of these servers are

$$\rho_i'^{(j)} = \frac{\lambda^{(0)} h_i}{\mu_i'} = \frac{2\lambda}{b+a}\frac{h_i}{\mu_i'}, \tag{18}$$

$\forall j = A5, A17, A9, A18, A10, A12$ and $A14$ corresponding to $i = 7, 8, 9, 10, 11, 12$, and 13, respectively, $h_1 = q_2$, $h_2 = h_3 = h_6 = q_2 p_3$, $h_4 = h_5 = h_7 = q_2 p_4$.

Note 1: Throughout this paper, we will associate the index j for servers $A4, A8, A15, A7, A11$, and $A13$, corresponding to $i = 1, 2, 3, 4, 5$, and 6 for the average queue lengths, $E[N_i^{(j)}]$, the average response times, $E[R_i^{(j)}]$, and the average waiting times, $E[W_i^{(j)}]$.

Note 2: Also in this paper, we will associate the index j for servers $A5, A17, A9, A18, A10, A12$, and $A14$, corresponding to $i = 1, 2, 3, 4, 5, 6$, and 7 for the average queue lengths, $E[N_i'^{(j)}]$, the average response times, $E[R_i'^{(j)}]$, and the average waiting times, $E[W_i'^{(j)}]$.

The difference between the performance measures in *Note 1* and *Note 2* is the notation for the queuing model performance measures, such as, average queue lengths, average response times, and average waiting times for different sets of servers but having the same subscript index i.

4. Performance measures

The performance of the single-server system is measured by the average queue lengths, average waiting times, average response times, and the average number of orders in the system [30].

4.1. Average queue lengths, average response times, and average waiting times

The average queue lengths, average response times, and average waiting times at the nodes $A4, A8, A15, A7, A11$, and $A13$, are

$$\begin{aligned}
E\left[N_i^{(j)}\right] &= \frac{1}{1-\exp\left(-\left(1-\rho_i^{(j)}\right)\right)},\\
E\left[R_i^{(j)}\right] &= \frac{E\left[N_i^{(j)}\right]}{\lambda^{(0)} w_i} = \frac{b+a}{2\lambda w_i\left(1-\exp\left(-\left(1-\rho_i^{(j)}\right)\right)\right)},\\
E\left[W_i^{(j)}\right] &= E\left[R_i^{(j)}\right] - \frac{1}{\mu_i} = \frac{(b+a)\mu_i - 2\lambda w_i\left(1-\exp\left(-\left(1-\rho_i^{(j)}\right)\right)\right)}{2\lambda w_i \mu_i\left(1-\exp\left(-\left(1-\rho_i^{(j)}\right)\right)\right)}.
\end{aligned} \tag{19}$$

The average queue lengths, average response times, and average waiting times at the nodes $A5, A17, A9, A18, A10, A12$, and $A14$, are

$$\begin{aligned}
E\left[N_i'^{(j)}\right] &= \frac{1}{1-\exp\left(-\left(1-\rho_i'^{(j)}\right)\right)},\\
E\left[R_i'^{(j)}\right] &= \frac{E\left[N_i'^{(j)}\right]}{\lambda^{(0)} h_i} = \frac{b+a}{2\lambda h_i\left(1-\exp\left(-\left(1-\rho_i'^{(j)}\right)\right)\right)},\\
E\left[W_i'^{(j)}\right] &= E\left[R_i'^{(j)}\right] - \frac{1}{\mu_i'} = \frac{(b+a)\mu_i' - 2\lambda h_i\left(1-\exp\left(-\left(1-\rho_i'^{(j)}\right)\right)\right)}{2\lambda h_i \mu_i'\left(1-\exp\left(-\left(1-\rho_i'^{(j)}\right)\right)\right)}.
\end{aligned} \tag{20}$$

4.2. Average queue lengths in different paths

The average number of jobs in path $V_1(A4, A8, A11)$, path $V_2(A4, A15, A7, A13)$, path $V_3(A5, A17, A9, A12)$ and path $V_4(A5, A18, A10, A14)$ are

$$\begin{aligned}
E[N_{V_1}] &= E\left[N_1^{(A4)}\right] + E\left[N_2^{(A8)}\right] + E\left[N_5^{(A11)}\right],\\
E[N_{V_2}] &= E\left[N_1^{(A4)}\right] + E\left[N_3^{(A15)}\right] + E\left[N_4^{(A7)}\right] + E\left[N_6^{(A13)}\right],\\
E[N_{V_3}] &= E\left[N_7'^{(A5)}\right] + E\left[N_8'^{(A17)}\right] + E\left[N_9'^{(A9)}\right] + E\left[N_{12}'^{(A12)}\right],\\
E[N_{V_4}] &= E\left[N_7'^{(A5)}\right] + E\left[N_{10}'^{(A18)}\right] + E\left[N_{11}'^{(A10)}\right] + E\left[N_{13}'^{(A14)}\right],
\end{aligned} \tag{21}$$

respectively.

4.3. Average response times in different paths

The global throughput delay from S_0 to S_1 in Fig. 2 can be chosen to be the minimum of the response times of the four paths shown below. The global throughput delay represents the order's cycle. This can be done by

- considering that orders are independently and equally routed from S_0 to S_1, and
- optimizing the route by taking into account the present state of the network.

The average response times in path $V_1(A4, A8, A11)$, path $V_2(A4, A15, A7, A13)$, path $V_3(A5, A17, A9, A12)$ and path $V_4(A5, A18, A10, A14)$ are

$$\begin{aligned}
E[R_{V_1}] &= E\left[R_1^{(A4)}\right] + E\left[R_2^{(A8)}\right] + E\left[R_5^{(A11)}\right],\\
E[R_{V_2}] &= E\left[R_1^{(A4)}\right] + E\left[R_3^{(A15)}\right] + E\left[R_4^{(A7)}\right] + E\left[R_6^{(A13)}\right],\\
E[R_{V_3}] &= E\left[R_1'^{(A5)}\right] + E\left[R_8'^{(A17)}\right] + E\left[R_9'^{(A9)}\right] + E\left[R_{12}'^{(A12)}\right],\\
E[R_{V_4}] &= E\left[R_1'^{(A5)}\right] + E\left[R_{10}'^{(A18)}\right] + E\left[R_{11}'^{(A10)}\right] + E\left[R_{13}'^{(A14)}\right],
\end{aligned} \tag{22}$$

respectively.

4.4. Equivalent network

From [29], the queue lengths and response times of the equivalent network (Fig. 4) are given by

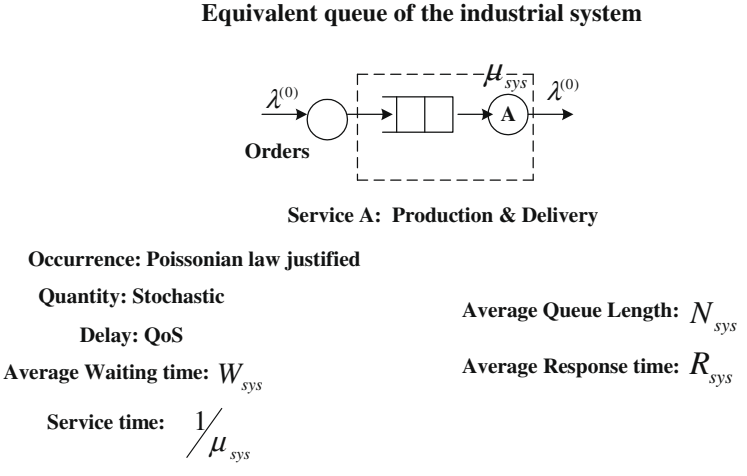

Fig. 4. Equivalent queue of the supply chain.

$$
\begin{aligned}
E[N_{\text{sys}}] &= E\left[N_{\text{eq}}^{(10)}\right] + E\left[N_{\text{eq}}^{(11)}\right] \\
&= \frac{\lambda_1}{\mu_1-\lambda_1} + \frac{\lambda_1 p_1(\mu_2+\mu_5-2\lambda_1 p_1)}{(\mu_2-\lambda_1 p_1)(\mu_5-\lambda_1 p_1)} + \frac{\lambda_2}{\mu_1'-\lambda_2} \\
&\quad + \frac{\lambda_1 p_2\left[(\mu_3+\mu_4-2\lambda_1 p_2)(\mu_6-\lambda_1 p_2)+(\mu_3-\lambda_1 p_2)(\mu_4-\lambda_1 p_2)\right]}{(\mu_3-\lambda_1 p_2)(\mu_4-\lambda_1 p_2)(\mu_6-\lambda_1 p_2)} \\
&\quad + \frac{\lambda_2 p_3\left[(\mu_2'+\mu_3'-2\lambda_2 p_3)(\mu_6'-\lambda_2 p_3)+(\mu_2'-\lambda_2 p_3)(\mu_3'-\lambda_2 p_3)\right]}{(\mu_2'-\lambda_2 p_3)(\mu_3'-\lambda_2 p_3)(\mu_6'-\lambda_2 p_3)} \\
&\quad + \frac{\lambda_2 p_4\left[(\mu_4'+\mu_5'-2\lambda_2 p_4)(\mu_7'-\lambda_2 p_4)+(\mu_4'-\lambda_2 p_4)(\mu_5'-\lambda_2 p_4)\right]}{(\mu_4'-\lambda_2 p_4)(\mu_5'-\lambda_2 p_4)(\mu_7'-\lambda_2 p_4)},
\end{aligned} \tag{23}
$$

and

$$
\begin{aligned}
E[R_{\text{sys}}] &= \frac{E[N_{\text{sys}}]}{\lambda} \\
&= \frac{q_1}{\mu_1-\lambda_1} + \frac{q_1 p_1(\mu_2+\mu_5-2\lambda_1 p_1)}{(\mu_2-\lambda_1 p_1)(\mu_5-\lambda_1 p_1)} + \frac{q_2}{\mu_1'-\lambda_2} \\
&\quad + \frac{q_1 p_2\left[(\mu_3+\mu_4-2\lambda_1 p_2)(\mu_6-\lambda_1 p_2)+(\mu_3-\lambda_1 p_2)(\mu_4-\lambda_1 p_2)\right]}{(\mu_3-\lambda_1 p_2)(\mu_4-\lambda_1 p_2)(\mu_6-\lambda_1 p_2)} \\
&\quad + \frac{q_2 p_3\left[(\mu_2'+\mu_3'-2\lambda_2 p_3)(\mu_6'-\lambda_2 p_3)+(\mu_2'-\lambda_2 p_3)(\mu_3'-\lambda_2 p_3)\right]}{(\mu_2'-\lambda_2 p_3)(\mu_3'-\lambda_2 p_3)(\mu_6'-\lambda_2 p_3)} \\
&\quad + \frac{q_2 p_4\left[(\mu_4'+\mu_5'-2\lambda_2 p_4)(\mu_7'-\lambda_2 p_4)+(\mu_4'-\lambda_2 p_4)(\mu_5'-\lambda_2 p_4)\right]}{(\mu_4'-\lambda_2 p_4)(\mu_5'-\lambda_2 p_4)(\mu_7'-\lambda_2 p_4)},
\end{aligned} \tag{24}
$$

respectively, where $\lambda_1 = \lambda^{(0)} q_1$, $\lambda_2 = \lambda^{(0)} q_2$, and $\lambda^{(0)}$ is as shown in (8). From the results of Bhaskar and Lallement [29], the service rate of the equivalent server is given by

$$
\mu_{\text{sys}} = \lambda + \frac{1}{D12a + D12b + D12c + D12d + D12e + D12f}, \tag{25}
$$

where

- $D12a = \frac{q_1}{\mu_1-\lambda_1}$,
- $D12b = \frac{q_1 p_1(\mu_2+\mu_5-2\lambda_1 p_1)}{(\mu_2-\lambda_1 p_1)(\mu_5-\lambda_1 p_1)}$,
- $D12c = \frac{q_2}{\mu_1'-\lambda_2}$,
- $D12d = \frac{q_1 p_2[(\mu_3+\mu_4-2\lambda_1 p_2)(\mu_6-\lambda_1 p_2)+(\mu_3-\lambda_1 p_2)(\mu_4-\lambda_1 p_2)]}{(\mu_3-\lambda_1 p_2)(\mu_4-\lambda_1 p_2)(\mu_6-\lambda_1 p_2)}$,
- $D12e = \frac{q_2 p_3[(\mu_2'+\mu_3'-2\lambda_2 p_3)(\mu_6'-\lambda_2 p_3)+(\mu_2'-\lambda_2 p_3)(\mu_3'-\lambda_2 p_3)]}{(\mu_2'-\lambda_2 p_3)(\mu_3'-\lambda_2 p_3)(\mu_6'-\lambda_2 p_3)}$, and
- $D12f = \frac{q_2 p_4[(\mu_4'+\mu_5'-2\lambda_2 p_4)(\mu_7'-\lambda_2 p_4)+(\mu_4'-\lambda_2 p_4)(\mu_5'-\lambda_2 p_4)]}{(\mu_4'-\lambda_2 p_4)(\mu_5'-\lambda_2 p_4)(\mu_7'-\lambda_2 p_4)}$.

5. Numerical results

5.1. Queue lengths with and without weights for the most optimal path in the 2-input network

5.1.1. No weights

Let λ be the total number of arrivals in the 2-input queuing network. In the example considered in this section, the arrival rate, $\lambda = 2, 4, \ldots, 20$. The values of a and b are 2 and 10, respectively. The other specifications include:

(a) Probability of arrivals at queues Q_4 and Q_5 are $(q_1, q_2) = (0.5, 0.5)$, respectively.
(b) The service rate specifications of different servers in the network are $\mu_2 = \mu_3' = \mu_4 = \mu_5' = \mu_c = 9$, $\mu_1 = \mu_1' = 15$, $\mu_2' = 7$, $\mu_3 = 6$, $\mu_4' = 5$, $\mu_5 = 11$, $\mu_6' = 5$, $\mu_6 = 8$, and $\mu_7' = 9$.
(c) The probabilities (p_1, p_2) and (p_3, p_4) are $(0.3, 0.7)$ and $(0.4, 0.6)$, respectively.

For each value of λ, the utilizations, average queue lengths, average response times, and average waiting times in all the nodes of the 2-input queuing network are computed. The average queue lengths in paths V_1, V_2, V_3, and V_4, respectively, are computed from (21). The average response times in paths V_1, V_2, V_3 and V_4, respectively, are computed from (22).

The minimum of the average response times of the nodes in all the paths is computed. It is found that for all arrival rates, the minimum response time corresponds to path V_2. Consequently, path V_2 is declared as the "optimal path". The optimal path is the path in which the sum of the average response times in each node is the minimum as compared to those of the other paths. The nodes in path V_2 are $(Q_4, A4)$, $(Q_{15}, A15)$, $(Q_7, A7)$ and $(Q_{13}, A13)$. The average queue length corresponding to path V_2 is noted for arrival rates, $\lambda = 2, 4, \ldots, 20$.

5.1.2. Including weights

When weights are incorporated, the service rates of $A4$, $A5$, $A8$, $A9$, $A7$ and $A10$ are halved from their original values given in part (1) (no weights section). All other specifications remain unchanged. The arrival rate, $\lambda = 2, 4, \ldots, 20$. The values of a and b are 2 and 10, respectively. The utilizations, average queue lengths, average response times and average waiting times in all the nodes are computed for each value of λ. It is found that for all arrival rates, the minimum response time corresponds to path V_2. The nodes in path V_2 are $(Q_4, A4)$, $(Q_{15}, A15)$, $(Q_7, A7)$, and $(Q_{13}, A13)$. It is found that as the arrival rate increases, the queue length also increases.

The queue lengths for nodes in path V_2 for both cases (with and without weights) in the 2-input network are plotted in Fig. 5. From the figure, it is clear that the average queue lengths for the case when weights are included, is larger than that for the no weight case for a particular arrival rate. This is because, when weights are included, the service rates of some servers are halved, which means that the service time is doubled. Because of this reason, lesser number of customers are served for a particular arrival rate.

5.1.3. Analysis of bottleneck servers in the queuing network

Bottleneck analysis using queuing network models is an important technique for the performance analysis and capacity planning of computer and communication systems. For the set of specifications considered in Section 5.1.1 on the probabilities of entering the individual branches of the queuing network, total arrival rate, arrival rates in individual branches, and service rates of different servers in the network for the no weight case, the utilizations of servers $A4$, $A8$, $A15$, $A7$, $A11$, $A13$, $A5$, $A17$, $A9$, $A18$, $A10$, $A12$, and $A14$ as a function of the arrival rates are shown in Table 1. From Table 1, the maximum utilization among the servers, $A4$, $A8$, $A15$, $A7$, $A11$ and $A13$ occurs for server $A15$ when $\lambda = 20$, and that value is $\rho_3^{(A15)} = \frac{7}{36} < 1$. Similarly, the maximum utilization among the servers, $A5$, $A17$, $A9$, $A18$, $A10$, $A12$, and $A14$ occurs for server $A18$ when $\lambda = 20$, and that value is $\rho_4^{\prime(A18)} = 0.2 < 1$.

For the case including weights, the service rates of servers, $A4$, $A8$, $A7$, $A5$, $A9$, $A10$ are halved from their original values. Thus, the utilizations of these servers are doubled from their corresponding previous values. The maximum utilization among the servers, $A4$, $A8$, $A15$, $A7$, $A11$, and $A13$ occurs for server $A7$ when $\lambda = 20$, and that value is $\rho_4^{(A7)} = \frac{7}{27} < 1$. Similarly, the maximum utilization among the servers, $A5$, $A17$, $A9$, $A18$, $A10$, $A12$, and $A14$ occur for server $A10$ when $\lambda = 20$, and that value is $\rho_{11}^{\prime(A10)} = \frac{2}{9} < 1$. Since the respective maximum utilizations for the case including weights are much lesser than unity, the queuing model with the given set of specifications is feasible to implement.

For the severs in the no weight case as well as in the case incorporating weights, to act as a bottleneck, the utilizations of these servers must be high and close to 1. Since none of the servers have utilizations close to unity for arrival rates,

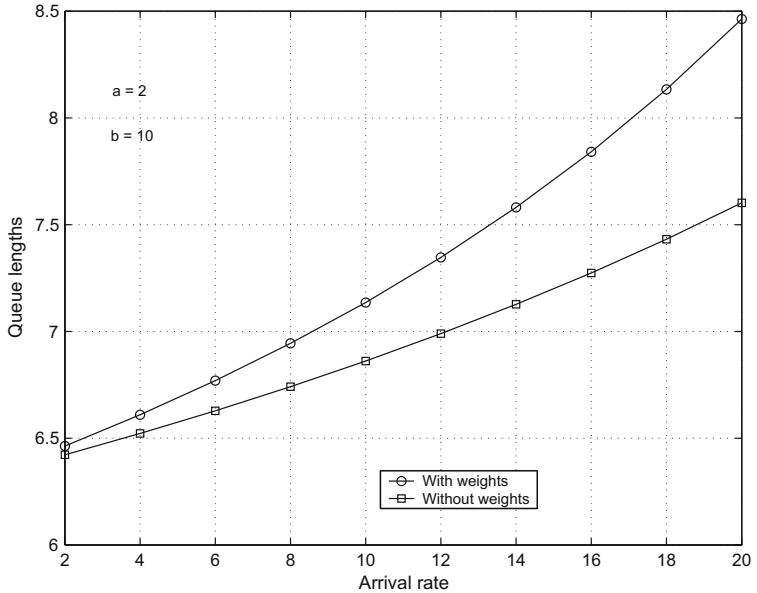

Fig. 5. Queue lengths for nodes in path V_2 in the 2-input network with and without weights.

Table 1
Utilization of different servers in the network.

Utilizations	Arrival rate ($\lambda = 2, 4, \ldots, 20$)
$\rho_1^{(A4)}$	$\frac{\lambda}{180}$
$\rho_2^{(A8)}$	$\frac{\lambda}{360}$
$\rho_3^{(A15)}$	$\frac{7\lambda}{720}$
$\rho_4^{(A7)}$	$\frac{7\lambda}{1080}$
$\rho_5^{(A11)}$	$\frac{3\lambda}{1320}$
$\rho_6^{(A13)}$	$\frac{7\lambda}{960}$
$\rho_7'^{(A5)}$	$\frac{\lambda}{180}$
$\rho_8'^{(A17)}$	$\frac{\lambda}{210}$
$\rho_9'^{(A9)}$	$\frac{\lambda}{270}$
$\rho_{10}'^{(A18)}$	$\frac{\lambda}{100}$
$\rho_{11}'^{(A10)}$	$\frac{\lambda}{180}$
$\rho_{12}'^{(A12)}$	$\frac{\lambda}{150}$
$\rho_{13}'^{(A14)}$	$\frac{\lambda}{180}$

$\lambda = 2, 4, \ldots, 20$, the queuing model with the given specifications is feasible to implement. The limitations of our data formulation occur only when any or some of the utilizations become high and get closer to unity. This can occur when

(a) the arrival rates become much higher than those considered, or
(b) when specifications like probabilities and service rates are initialized to values such that the utilizations get close to unity.

5.2. Queue lengths with and without weights in the equivalent single queue–single server network

5.2.1. No weights

Let λ be the total number of arrivals at the equivalent single-queue single-server system as shown in Fig. 4. In the example considered in this section, the arrival rate is $\lambda = 2, 4, \ldots, 20$. The service rate specifications of different servers in the network are $\mu_2 = \mu_3' = \mu_4 = \mu_5' = \mu_c = 9$, $\mu_1 = \mu_1' = 15$, $\mu_2' = 7$, $\mu_3 = 6$, $\mu_4' = 5$, $\mu_5 = 11$, $\mu_6' = 5$, $\mu_6 = 8$, and $\mu_7' = 9$. The probabilities, (p_1, p_2), (p_3, p_4), and (q_1, q_2) are (0.3, 0.7), (0.4, 0.6) and (0.5, 0.5), respectively.

For each value of λ, the average queue length of the equivalent queue, and the average response time of the equivalent queue are computed from (23) and (24), respectively. The service rate of the equivalent server is computed from (25). The average queue length of the equivalent queue increases as the arrival rate increases.

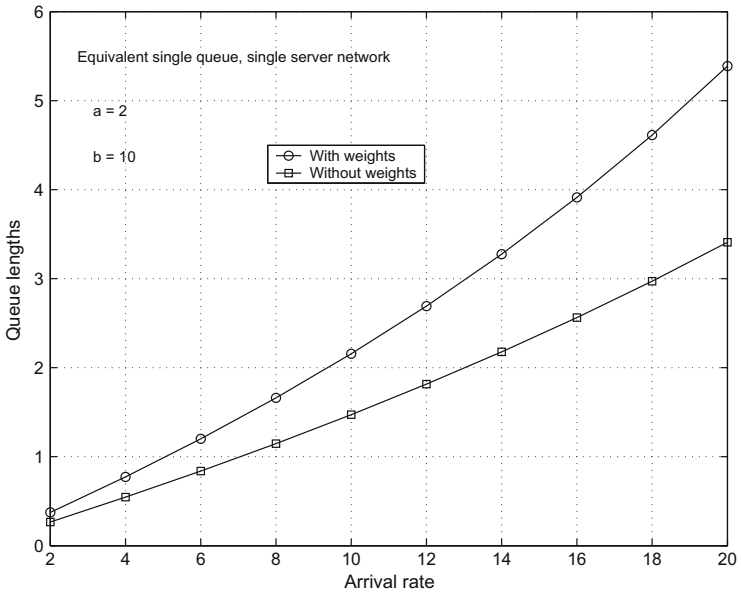

Fig. 6. Queue lengths in the equivalent single queue–single server network with and without weights.

5.2.2. Including weights

When weights are incorporated, the service rates of $A4$, $A5$, $A8$, $A9$, $A7$ and $A10$ are halved from their original values in part (1) (no weights section). All other specifications remain unchanged. The arrival rate is $\lambda = 2, 4, \ldots, 20$. The average queue length and average response time of the equivalent queue are computed from (23) and (24), respectively for each value of λ. The service rate of the equivalent server is computed from (25).

The queue lengths in the equivalent single queue–single server network for both cases (with and without weights) are plotted in Fig. 6. Again, it can be observed that the average queue lengths for the case when weights are included, is larger than that for the no weight case for a particular arrival rate. Lesser number of customers are served in a system with a lower service rate as compared to a system with a higher arrival rate. The equivalent service rate for the system when weights are not included is $\mu_{sys} = 8.4735$ and for the case when weights are included, it is $\mu_{sys} = 5.5024$, which also explains the behavior of the curves in Fig. 6.

6. Conclusions

In this paper, the most optimal path for routing items is path V_2 because it produces the least response time for the given set of specifications (probability of entering a new path, arrival rate and service rates). The nodes in the optimal path, V_2, are $(Q_4, A4)$, $(Q_{15}, A15)$, $(Q_7, A7)$, and $(Q_{13}, A13)$. The choice of the optimal path depends on the arrival rate at each queue, service rate of each server, and the probability of entering a particular node. The corresponding total number of items in all the nodes of the most optimal path constitutes the capacity of the 2-input network. Decision for routing is made at the last node in each stage of the network as to which path to choose to obtain the least response time. Performance measures such as average queue lengths are derived and plotted. Performance measures such as average response times, average waiting times and steady-state probabilities are also derived.

The supply chain is modeled as an equivalent single queue–single server system. Performance measures such as average queue lengths and average response times are derived and plotted for the equivalent single queue–single server network. The service rates of the equivalent server with and without weights are also derived and computed numerically.

Appendix

In this section, the expressions for the *steady-state probability* of having a certain number of jobs in the system for each of the two cases is presented.

Case (i): The steady-state probability of having k_i jobs at node i is $\Pi_i(k_i) = (1 - \rho_i)\rho_i^{k_i}$ [30]. The steady-state probabilities at the nodes ($A4$, $A8$, $A15$, $A7$, $A11$, and $A13$) in the queuing network are

$$\Pi_i^{(j)}(k_i) = \left(1 - \rho_i^{(j)}\right)\left(\rho_i^{(j)}\right)^{k_i} = \left(1 - \frac{2\lambda}{b+a}\frac{w_i}{\mu_i}\right)\left(\frac{2\lambda}{b+a}\frac{w_i}{\mu_i}\right)^{k_i}, \tag{26}$$

$\forall\, j = A4, A8, A15, A7, A11$, and $A13$ corresponding to $i = 1, 2, 3, 4, 5$, and 6, respectively, $w_1 = q_1$, $w_2 = w_5 = q_1p_1$, and $w_3 = w_4 = w_6 = q_1p_2$.

Case (ii): The steady-state probabilities at the nodes ($A5$, $A17$, $A9$, $A18$, $A10$, $A12$ and $A14$) in the queuing network are

$$\Pi_i'^{(j)}(k_i') = \left(1 - \rho_i'^{(j)}\right)\left(\rho_i'^{(j)}\right)^{k_i'} = \left(1 - \frac{2\lambda}{b+a}\frac{h_i}{\mu_i'}\right)\left(\frac{2\lambda}{b+a}\frac{h_i}{\mu_i'}\right)^{k_i'}, \tag{27}$$

$\forall j = A17, A5, A9, A18, A10, A12$, and $A14$ corresponding to $i = 1, 2, 3, 4, 5, 6$, and 7, respectively, $h_1 = q_2$, $h_2 = h_3 = h_6 = q_2p_3$, and $h_4 = h_5 = h_7 = q_2p_4$.

References

[1] F. Hillier, G. Lieberman, Introduction to Operation Research, eighth ed., McGraw Hill, NY, USA, 2005.
[2] R. Suri, Quick Response Manufacturing, Productivity Press, Portland, OR, USA, 1998.
[3] M. Christopher, Logistics and Supply Chain Management, third ed., Prentice Hall Inc., NJ, USA, 2005.
[4] J. Aitken, Supply Chain Integration Within the Context of a Supplier Association, Cranfield University, Ph.D. Thesis, 1998.
[5] J.L. Heskett, Logistics: essential to strategy, Harv. Bus. Rev. 85 (6) (1977) 85–96.
[6] F.B. Vernadat, Enterprise Modeling and Integration: Principles and Applications, Chapman and Hall, London, 1996.
[7] Manish K. Govil, Michael C. Fu, Queuing theory in manufacturing: a survey, J. Manuf. Syst. 18 (1999) 3.
[8] M. Chinnaswamy, M. Kamath, On queuing network models of service systems, in: IE Research Conference, Atlanta, GA, May 2005.
[9] F. Baskett, K. Chandy, R. Muntz, F. Palacios, Open, closed, and mixed network of queues with different classes of customers, JACM 22 (2) (1975) 248–260.
[10] H. Youn, S. Jang, E. Lee, Deriving queuing network model for UML for software performance prediction, in: Fifth International Conference on Software Engineering Research Management and Applications, August 2007, pp. 125–131.
[11] R. Raja, K. Suryaprakash Rao, Performance evaluation through simulation modeling in a cotton spinning system, J. Simul. Model. Pract. Theory 15 (9) (2007) 1163–1172.

[12] S.V. Serguyevich, M.G. Ortega Rosales, J. Marcos Garcia, L.A. Zamora Quintana, R. Pena Lopez, Chain conveyor system simulation and optimization, in: 17th IASTED International Conference on Modeling and Simulation, 2006, pp. 172–177
[13] F. Baskett, A. Smith, Interference in multiprocessor computer systems with interleaved memory, Commun. ACM 19 (6) (1976).
[14] P. Pollett, Resource Allocation in General Queuing Networks with Applications to Data Networks, Technical Report, Department of Mathematics, University of Queensland, Queensland 4072, Australia, 1998.
[15] D. Towsley, Queuing network models with state-dependent routing, JACM 27 (2) (1980) 323–337.
[16] N. Bisnik, A. Abouzeid, Queuing network models for delay analysis of multihop wireless ad hoc networks, in: International Conference on Communications and Mobile Computing, 2006, pp. 773–778.
[17] K. Sevcik, J. Mitrani, The distribution of queuing network states at input and output instants, JACM 28 (1981) 358–371.
[18] G. Ciardo, J. Muppala, K. Trivedi, SRNP: Stochastic Petri nets and Performance Models, IEEE Computer Society Press, Los Alamitos, CA, 1989. pp. 142–150.
[19] R. Schassberger, H. Daduna, Sojourn times in queuing networks with multiserver models, J. Appl. Probab. 24 (1987) 511–521.
[20] S. Ramesh, H.G. Perros, A multi-layer client–server queuing network model with synchronous and asynchronous messages, IEEE Trans. Softw. Eng. 26 (11) (2000) 1086–1100.
[21] V. Mainkar, Solutions of Large and Non-Markovian Performance Models, Ph.D Dissertation, Department of Computer Science, Duke University, Durham, NC, 1994.
[22] Y. Leung, M. Kamath, Performance analysis of a single-stage assembly system, in: INFORMS Annual Meeting, Detroit, MI, October 1994.
[23] W. Whitt, The queuing network analyzer, Bell Syst. Tech. J. 62 (9) (1983).
[24] J.A. Buzacott, J.G. Shanthikumar, Queuing models of manufacturing and service systems, in: HandBook of Industrial Engineering, third ed., Wiley InterScience, 2001, pp. 1627–1668.
[25] J.A. Buzacott, J.G. Shanthikumar, D.D. Yao, Jackson network models of manufacturing systems, in: Stochastic Models and Analysis of Manufacturing Systems, Springer, 1994, pp. 1–45.
[26] D. Gupta, J.A. Buzacott, A framework for understanding flexibility of manufacturing systems, J. Manuf. Syst. 8 (1989) 89–97.
[27] J.A. Buzacott, Y. Kahyaoglu, Flexibility and robustness in manufacturing, Int. J. Manuf. Technol. Manage. 2 (2000) 546–558.
[28] J.A. Buzacott, R.Q. Zhang, Financial flows and material flows, in: Stochastic Modeling and Optimization of Manufacturing Systems and Supply Chains, Kluwer, 2003, pp. 375–404.
[29] V. Bhaskar, P. Lallement, Activity routing in a distributed supply chain: performance evaluation with two inputs, J. Netw. Comput. Appl. (2008), doi:10.1016/j.jnca.2008.02.001.
[30] K. Trivedi, Probability and Statistics with Reliability, Queuing and Computer Science Applications, Prentice Hall, New Jersey, 1982.
[31] A. Papoulis, Probability, Random Variables and Stochastic Processes, third ed., McGraw Hill, NewYork, 1991.
[32] V.K. Rohatgi, An introduction to probability theory, Mathematical Statistics, Wiley, NewYork, 1976.
[33] I.S. Gradshteyn, I.M. Ryzhik, Table of Integrals, Series, and Products, fifth ed., Academic Press, San Diego, CA, 1994.
[34] Wolfram, Wolfram Mathematics Online Integrator, Wolfram Research Inc., 2008. <http://integrals.wolfram.com/index.jsp>.
[35] H.T. Tran, T.V. Do, An iterative method for queuing systems with batch arrivals and batch departures, in: Proceedings of the 8th IFIP Workshop on Performance Modeling and Evaluation of ATM and IP Networks, 2000.