Mohammadi, M., Dauzère-pérès, S., Yugma, C., & Karimi-Mamaghan, M. (2020). A queue-based aggregation approach for performance evaluation of a production system with an AMHS. *Computers & Operations Research*, *115*, 104838.

https://doi.org/10.1016/j.cor.2019.104838

A queue-based aggregation approach for performance evaluation of a production system with an AMHS

Mehrdad Mohammadi[a*], Stéphane Dauzère-pérès[b,c], Claude Yugma[b], Maryam Karimi-Mamaghan[a]

[a]*IMT Atlantique, Lab-STICC, UBL, F-29238 Brest, France*
[b]*Mines Saint-Etienne, Univ. Clermont Auvergne, CNRS, UMR 6158 LIMOS, CMP, Department of Manufacturing Sciences and Logistics, F-13541 Gardanne, France*
[c]*Department of Accounting, Auditing and Business Analytics, BI Norwegian Business School, 0484 Oslo, Norway*

Abstract

Production planning optimization remains a major challenge in almost all industries, particularly in high-tech manufacturing. A critical task to support such optimization is performance evaluation, wherein an accurate estimation of the cycle time as a function of the throughput rate plays a key role. This paper develops a novel aggregation model based on a queueing network approach, so-called queue-based aggregation (QAG) model, to estimate the cycle time of a job-shop production system that consists of several processing workstations, and in which products are transferred via an Automated Material Handling System (AMHS). The proposed model aggregates both production and automated material handling systems and provides an accurate and fast estimation of the overall cycle time. The performance and superiority of the proposed model is validated by comparing its results with those of a detailed simulation model. Numerous sensitivity analyses are performed to provide valuable managerial insights on both the production and automated material handling systems.

Keywords: Performance evaluation; Aggregation model; Queuing network; Cycle time estimation; Automated material handling system

1. Introduction

Processing a wide range of products in response to the dynamic nature of the customers' demand has become one of the greatest challenges in almost all manufacturing systems [1]. This challenge becomes more severe as manufacturing systems have to meet customers' orders in shorter time and at minimal cost to remain competitive. This makes production planning of such manufacturing systems more and more important and complex to improve the performance of factories. Several indicators are being used to analyze the performance of manufacturing systems such as the cycle time and the throughput. The cycle time is the average time spent by a product unit (i.e., waiting time plus processing time) in the system and the throughput, also referred to as productivity, is the number of product units processed per time unit [2].

Managers aim at designing and planning highly productive production systems, in particular in capital intensive industries, by keeping equipment throughput as high as possible. At the same time, they should respect the due dates of customers' orders by keeping product cycle times as short as possible. Therefore, there is usually a trade-off between cycle time minimization and throughput maximization in production systems. As it is well known, increasing the throughput results in a large Work-in-Process (WIP) and, consequently, significant waiting times in the production system. The longer the waiting time, the longer the cycle time of products. The increased level of complexity in the manufacturing systems alongside with the pressure on low prices of the final products require a scientific approach to manage this trade-off between productivity and due date satisfaction.

Accordingly, an accurate prediction of the cycle time as a function of the throughput is necessary. A prediction model has to incorporate all the contributing components of the production system. Processing workstations have a major effect on the cycle time. The consideration of processing workstations in cycle time prediction models has been well studied in the literature [3]–[7]. Another important component that highly impacts the cycle time is the Material Handling System (MHS). A MHS performs as an interconnector for workstations and should deliver the right amount of materials, to the right place, at the right time and at a minimal cost [8]–[10]. The MHS impacts the performance of production systems, mainly by impacting the WIP level. An efficient MHS can decrease the cycle time by reducing the WIP level and consequently the waiting time. The contribution of the MHS to the total operating cost ranges between 15% and 70%, and this value is around

[*] Corresponding author: Mehrdad MOHAMMADI (mehrdad.mohammadi@imt-atlantique.fr). Tel: +33 (0)2 29 00 10 30.

20% of the product cycle time [11], [12]. Therefore, it is highly important to consider the MHS when developing a cycle time prediction model, when the MHS's capacity is limited and/or the transportation time is long. However, when the MHS has infinite (very high) capacity or the transportation time is very short, this consideration is not necessary. Different factors could be incorporated in the prediction model including processing several products at the same time, predicted/unpredicted breakdowns of workstations and of the MHS, vehicle blockage in the MHS, and dispatching rules. Although incorporating a large number of shop floor factors makes the prediction model close to real settings, it also makes the model more complex. Broadly speaking, a performance evaluation model should not be too complex and should require limited development and maintenance effort and the model execution should be computationally cheap.

There are two common modeling approaches to predict cycle times in production systems: I) "Simple" analytical models (i.e., models that require only few easily-used and easily-estimated parameters) and II) (mainly discrete-event) Simulation models. Commonly-used simple analytical models are closed form G/G/m queueing models [3]–[5], [7], [10], [11], [13]–[27]. Despite the promptness and easiness of analytical models to provide solutions and performance evaluation, they cannot model complex shop-floor behavior such as workstation/MHS breakdowns, and they suffer from the lack of accuracy for complex workstations/equipment as well. Alternatively, simulation models are used to model complex behavior and to accurately represent the processing operations in workstations [28]–[30]. Detailed simulation modeling allows all relevant shop-floor details to be considered [31], [32]. An accurate simulation model requires all required input data related to processing workstations and MHS. Consequently, it may become computationally expensive and require significant development time for experimentation with a large set of scenarios. This issue renders simulation model impractical for quick cycle time estimations as well as computationally heavy for designing or redesigning a system [33].

One way to overcome this time inefficiency and to abstract a detailed simulation model is to perform simulation runs according to a design of experiments, and use the responses to generate a metamodel [34], [35]. Re-running the simulation model after only making a small change is one of the drawbacks of these experiments. Another approach to abstract a detailed simulation model is aggregation [4]. Aggregation can be done by simplifying the complex components/assumptions of the system such as replacing non-bottleneck workstations by a constant delay [36], [37]. An aggregation approach aggregates the raw processing times and all shop-floor realities and processing disturbances at different workstations into a single process time distribution. The inclusion of multiple phenomena into a single distribution is referred to as aggregation.

In this article, a new aggregation model based on a queueing network approach, so-called queue-based aggregation (QAG) model, is proposed to estimate the cycle time of a job-shop (*JS*) production system that consists of several processing workstations and where products are transferred via a MHS. The proposed QAG aggregates a set of operating workstations in a job-shop frame into a single workstation. Open and closed queueing networks with finite/infinite buffers have been typically used for modeling and analyzing discrete event systems such as manufacturing systems, computer systems, communication systems, and MHS [38]. Comparing to a detailed simulation, the proposed QAG model provides a prompt and reasonably accurate estimation of a *JS*'s cycle time and throughput. The QAG model is proposed to support the early stages of design by exploring and evaluating different scenarios before committing to the time and cost of a detailed simulation model.

The remainder of the paper is organized as follows. Section 2 reviews the most relevant articles and outlines the literature gaps. Section 3 develops the QAG model. The proposed model is validated through comprehensive experimental results in Section 4. The paper is concluded in Section 5.

2. Literature review

This section reviews what we believe are the most relevant articles in the literature that have studied the performance evaluation of production systems. To the best of our knowledge, most of the papers have studied processing workstations and MHS separately. Neglecting one of them may lead to a poor approximation of the *JS* performance.

2.1. Performance evaluation of processing workstations

Brooks and Tobias [39] and Johnson et al. [36] provide a simple aggregation model by reducing non-bottleneck processing workstations to a constant delay to estimate long-term throughput rates, but their model does not estimate processing cycle times. They propose an eight-stage procedure for the reductions and study two manufacturing cases. Similarly, Rose [37] aggregates all processing workstations, except the bottleneck workstation, by replacing them with a constant delay; however, for certain scenarios, these simple aggregation models fail to predict the cycle time distributions accurately.

Hopp and Spearman [40], [41] and Jacobs et al. [38], [42] propose an algorithm to aggregate the process time distributions and impose delays (i.e., failures, maintenance, operator unavailability, etc.) in processing workstations. Hopp and Spearman [40] define the aggregation process time as the Effective Process Time (EPT). EPT is "the process time seen by a product at a workstation from a logistical point of view". The Expected value and variance of the EPTs are approximated from the raw process times, and the preemptive and non-preemptive delays. After approximating the mean and the variance of the EPTs, a closed-form G/G/m queuing model is used to estimate the mean cycle time. Due to the dependency between the EPT distribution parameters and the WIP, Kock et al. [5] approximate a new WIP-based EPT and propose a G/G/m-based aggregate simulation model to estimate the mean cycle time as a function of the processing throughput. The proposed model does not necessarily lead to accurate cycle time estimation, due to the First-Come-First-Served (FCFS) rule in the aggregate model.

To cope with this issue, Veeger et al. [6] develop an aggregate model using a WIP-dependent EPT distribution that considers the order in which products are processed. Accordingly, each product that arrives in the aggregate model has a probability of overtaking a number of other products already in the system. This number is determined by a WIP-dependent overtaking distribution measured from the arrival and departure times of the products. The authors in [5] and [6] first build a detailed simulation model of the processing workstations and then the aggregate model is trained using arrival and departure data measured in the simulated workstation model. Numerous replications (i.e., 10^5 runs) were required in a specific level of throughput rate (e.g., 80%) for the simulated model to obtain significantly enough arrival and departure events to get a reliable EPT estimation. This approach is computationally expensive due to the need for creating detailed simulation models with numerous replications. Furthermore, the models are built for a specific throughput rate and particular details of the processing workstation (e.g., number of parallel servers, number of processing steps, etc.). Making a small change in the models requires the detailed simulation model to be re-built and re-ran, making this approach relatively heavy.

Unlike simulation-based aggregation models, Morrison and Martin [4] and Morrison [43] use analytical methods to estimate the cycle time in single-product clustered photolithography tools in semiconductor manufacturing. Morrison and Martin [4] model each processing workstation as a general G/G/m queue and propose a closed deterministic formulation to approximate the cycle time of the workstations subject to server failures and cycle time offsets (e.g., events such as travel and hold upstream of a workstation). Morrison [43] develops several analytical flow line models to calculate the cycle time of each processing workstation. In the flow line models, cycle time, start time, and the process time of each processing workstation are recursively calculated from the completion times of products. Although the proposed analytical models of [4] and [43] are computationally cheaper than EPT-based aggregation models, they neither work for multiple-product workstations nor are able to approximate the cycle time of integrated workstations including serial/non-serial configurations of processing workstations.

Various studies have aimed at evaluating the performance of serial/non-serial configured production systems through approximating queue length and waiting time at each processing workstation [7], [44]–[48]. As a sample, Satyam and Krishnamurthy [45] and Satyam et al. [47] propose a new approach based on parametric decomposition. Two-moment approximations are used to estimate the performance of individual processing workstations. Subsequently, the traffic process parameters at different workstations are linked using stochastic transformation equations. The resulting set of non-linear equations is solved using an iterative algorithm to obtain estimates of key performance indicators (KPIs) such as the throughput and mean queue lengths.

Apart from queuing network and simulation techniques, Hillion and Proth [49] utilize timed event-graphs, as a special class of timed Petri nets, for modelling and analyzing job-shop systems. Their proposed model allows the steady-state performance of the system to be evaluated under a deterministic and cyclic production process. The authors explain that, by fixing processing times, the production throughput can be determined from the initial state. In addition, given any desired product mix, it is possible to start the system with enough WIPs in front of bottleneck machines. In that case, the system works at the maximal rate and the throughput is optimal. Suri et al. [50] discuss different approaches for performance evaluation of discrete manufacturing systems through different performance indicators such as the utilization of each server, the throughput, the flow time of a job through the system, the queue length at a station, and the WIP in the system. These approaches fall into three categories as: Static (allocation) models, aggregate dynamic models (ADMs), and detailed dynamic models.

Silva and Morabito [51] apply performance evaluation and capacity allocation models to support decisions in the design (or redesign) and planning of a job-shop queueing network in a metallurgical plant. In this work, approximate parametric decomposition methods are used to evaluate system performance measures, such as the expected WIP and production lead times. Based on these methods, optimization models are then applied for the allocation (or reallocation) of capacity to the stations of the job-shop network. Abdi and Labib [52] develop holonic architecture for reconfigurable manufacturing systems (RMS), which are capable of adapting to unpredictable changes in demands. RMS are designed to produce various products grouped into families in a short time and at low cost. A holonic structure reflecting basic holons for RMS is developed and then linked to an analytical network process (ANP) model, as a multi-criteria approach, to evaluate the system performance. The proposed generic model provides flexibility for holons and facilitates the evaluation of RMS considering economical and operational aspects as the main performance objectives.

To the best of our knowledge, almost all of these works have decomposed the production system into its contributing workstations and have evaluated each workstation separately to estimate its queue length and waiting time. None of these works propose an overall approximation model for the whole production system.

Recently, we developed a new aggregation model [53] based on queuing network, so-called queue-based aggregation (QAG) model, to estimate the cycle time in a production system. Multiple workstations in serial and job-shop configurations were aggregated into a single-step workstation. The parameters of the aggregated workstation are approximated based on the parameters of the original workstations. Finally, the numerical experiments indicate that the proposed QAG model is computationally efficient and yields fairly accurate results when compared to other aggregation approaches in the literature. Our recent model does not take the MHS into account, while the MHS has a significant effect on the overall performance of the production system.

2.2. Performance evaluation of MHS

Bedell and Smith [54] categorize material handling devices into three groups:

I. Transporters (vehicles, carts, trucks, people, etc.),
II. Conveyors (belt, chain, roller, overhead conveyors, escalators, etc.),
III. Lifts (hoists, elevators, cranes, etc.).

Simulation models have been widely employed to model MHS in manufacturing and service systems. There are few analytical models and their wide-spread use is not significant. Analytical models have been proposed for both discrete part transfer and continuous part material handling flow [54]. The discrete part transfer is mainly concerned with carts, transporters, or people moving the products (i.e., group I). Continuous material handling flows correspond to material handling devices of groups II and III. This section focuses on analytical models for the performance evaluation of MHS with discrete part transfer.

The first analytical approach for MHS's performance evaluation is the queuing-based approach proposed by Benson and Gregory [55] who consider closed cyclic systems with exponentially distributed transit times between successive processing workstations. As an extension, Posner and Bernholtz [56] model a closed queueing network of two processing workstations and model the transfer time between these stations with a general distribution through a supplementary variable technique. It has been shown in the literature that finite buffer, single server queues with general service times along with state-dependent queues are robust with respect to the general flow processes in MHS [56]–[58].

Bedell and Smith [54] examine the integration of multi-server and state-dependent queues and the robustness properties in these queueing network models. Raman et al. [10] develop a two-step analytical approach to determine the quantity of material handling equipment (MHE) required for effective handling of products among facilities. In the first step, the authors obtain a preliminary solution by considering the time required for loading and unloading products, loaded traveling, empty traveling and breakdown of the MHE. A detailed model, which integrates both operational and cost performance factors such as the utilization of the MHE, the work-in-process at the MHS, and the life-cycle cost is then utilized to rank alternatives that are generated from the preliminary solution. They finally use queuing theory in order to cope with the stochasticity of the MHS.

Nazzal and McGinnis [59] propose a queue-based analytical approach to evaluate the performance of a simple closed loop automated MHS (AMHS), which is typical in 300 mm wafer fab bays. Due to the significant impact of vehicle blocking, a straightforward queueing network model which treats the material handling system as a central server can be inaccurate. Therefore, the authors propose an alternative model to estimate the MHS performance considering the possibility of vehicle blocking. Nazzal [33] models a multi-vehicle material handling system as a closed-loop queueing network with finite buffers and general service times, where the vehicles model the jobs in the network. In this work, the vehicles' residence times on track segments (servers) depend on the number of jobs (vehicles) in circulation. A new iterative approximation algorithm is developed that estimates throughput capacity and decomposes the network consisting of S servers into S separate $G/G/1$ systems.

Tu et al. [60] study the AMHS capacity determination model in order to maintain the originally designed optimal production throughput or cycle time of products. A GI/G/m queuing model is applied based on the FCFS dispatching rule of the AMHS to determine the required number of vehicles. In this model, products should be transported to the specific workstation before the workstation completes the existing process; therefore, sufficient WIP in front of this specific workstation should be kept. In order to improve the performance of the AMHS, Johnson [61] presents analytical models to predict empty vehicle travel under two popular vehicle dispatching rules. It is shown that using analytical models to estimate empty vehicle traffic in the design process can significantly improve the AMHS performance.

2.3. Paper contributions

The main contributions of the paper to the literature are detailed below:

- Queuing systems have been widely used to evaluate the performance of manufacturing systems [3]–[5], [7], [10], [11], [13]–[27], [62], [63]. However, it is worth mentioning that the existing papers evaluate the production system workstation per workstation. To the best of our knowledge, there is no aggregation model that approximates the overall performance of the system through closed form analytical formulations as the ones proposed in this paper.
- Moreover, there is no work in the literature that explicitly considers the AMHS in the performance evaluation of the production system, although it can be very important when developing cycle-time prediction models for highly automated manufacturing systems. The previous studies discussed in Section 2.2 focus on the optimal performance of AMHS. However, the AMHS is only an auxiliary system for production, and the optimal performance of the AMHS does not guarantee a better performance of the overall production system. Accordingly, this paper develops an aggregation model for a job-shop production system with an AMHS.
- Furthermore, two important factors affecting the performance of complex production systems are unexpected failures and the priorities of products. These factors have not been taken into account in the performance evaluation of job-shop production systems. This paper investigates the impact of these factors on the overall performance of the production system.
- In terms of computational complexity, the recent simulation-based aggregation models in [5], [6] and [62] are not timely-efficient for quick and real time decisions; while the aggregation model proposed in this paper is solved within less than one second, even for complex production systems with numerous workstations and products. Also, the proposed model has very small gaps when compared to the detailed simulation models.
- Besides to comparison between the proposed QAG model and detailed simulation, a remarkable sensitivity analysis is performed to investigate the sensitivity of the KPIs to the input parameters. Sensitivity analysis using detailed simulation model is difficult in practice since it is timely expensive. Performing sensitivity analysis provides knowledge on the behavior of important parameters. Having this knowledge helps experts to achieve the desired level of KPIs by correctly adjusting these parameters.
- As the last effort, a capacity planning is conducted for the AMHS system by identifying the required number of vehicles to reach a high throughput rate of the production system.

To summarize, we propose a new aggregation model based on a queueing network approach, so-called queue-based aggregation (QAG) model, to estimate the cycle time in a *JS* production system that consists of several processing workstations and in which products are transferred via a MHS. This work is actually an extension of our previous work [53], but with the aggregation of the AMHS that makes the model closer to real settings. Promising and more interesting results are provided. In our model, the whole *JS* system is aggregated into a single-step workstation. The parameters of the single-step aggregated workstation are approximated based on the parameters of the original processing workstations and the MHS. This paper not only helps to provide an abstraction of detailed simulation models for complex manufacturing systems, but also proposes an accurate and timely efficient estimation of the cycle time as a function of the throughput rate.

3. Proposed QAG Model

This section develops new aggregation models based on queuing networks to estimate the cycle time in complex manufacturing systems accurately and efficiently. This is done by modeling manufacturing systems as single/multiple-class open queuing networks, wherein the nodes in the network are the processing workstations of the manufacturing system. A queueing network is said to be open when external flow units (products) can enter the network at every node, and the internal flow can leave the network from any node. A queueing network is called closed when flow units can neither enter nor leave the network. Accordingly, the number of flow units in a closed queue network is constant. The queueing network can also be called single class or multiclass, depending on whether the queueing network serves a single type or multiple types of products [45].

Queueing networks with several service workstations are more suitable than simple queue models with only a single service station for modeling the structure of many manufacturing systems with a wide range of resources. Note that an underlying assumption in a queueing network is that at least two workstations are connected to each other.

3.1. Job-shop manufacturing system

Figure 1 shows an example of a multi-class *JS* manufacturing system with six workstations and three products. Each product requires a set of workstations, and the processing routes of the products are different, e.g. the processing routes of products P_1, P_2 and P_3 are $E_1 = \{1, 2, 4\}$, $E_2 = \{4, 5, 3\}$, and $E_3 = \{5, 6, 1\}$, respectively. In addition, the subsets R_i of products requiring workstation *i* are $R_1 = \{1,3\}$, $R_2 = \{1\}$, $R_3 = \{2\}$, $R_4 = \{1,2\}$, $R_5 = \{2,3\}$, and $R_6 = \{3\}$. Figure 2 depicts the product-focused structure of the multi-class *JS* system and shows the serial processing route of each product [53].

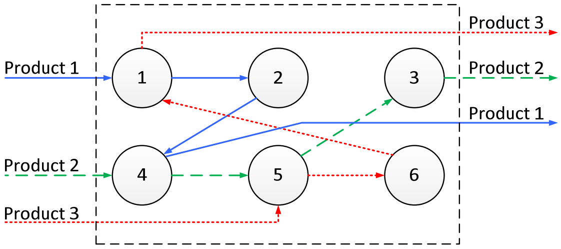

Figure 1. Process-focused structure of a multi-class *JS* manufacturing system

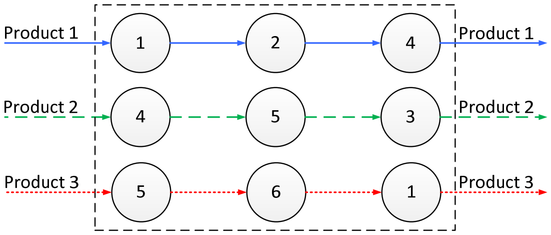

Figure 2. Product-focused structure of a multi-class *JS* manufacturing system

The products are transferred between each pair of workstations using the AMHS and the FCFS dispatching rule. Figure 3 illustrates the integration of the AMHS in the *JS* manufacturing system of Figure 2. In the following, analytical models are provided based on queuing theory to analyze the performance of the workstations and the AMHS and finally, to evaluate the performance of the *JS* manufacturing system in terms of the product cycle time. Figure 4 depicts how two workstations and the in-between AMHS are modeled by queuing systems. The product cycle time is the time spent by the products in the workstations and the AMHS.

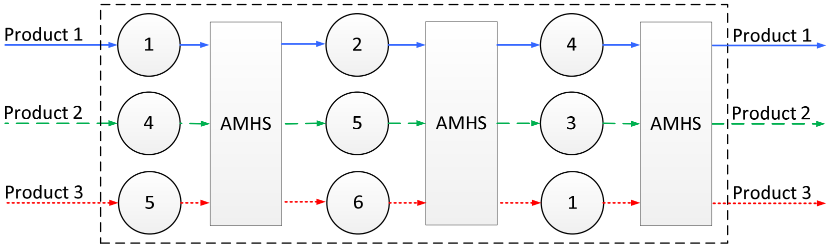

Figure 3. Integration of the AMHS in the *JS* manufacturing system

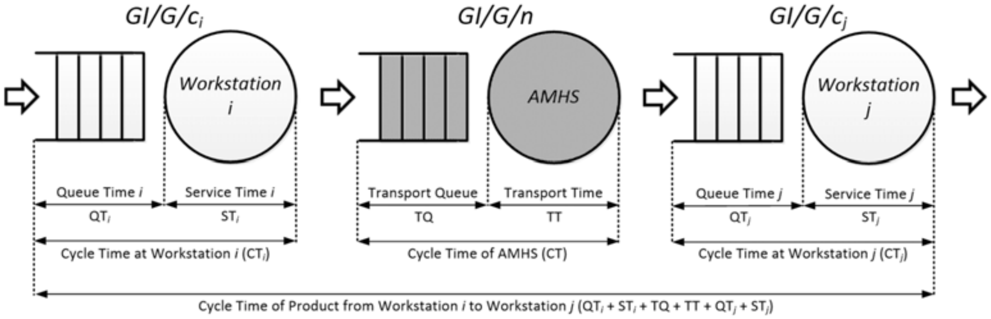

Figure 4. Two workstations and the in-between AMHS modeled by the queuing system

3.2. Processing workstation characterization

3.2.1. Notations

Before characterizing the processing workstation and developing the QAG model for the *JS* system, the required notations are provided in this section.

Sets and Indices:

i, j Indices of workstations; $i, j \in \{1, \dots, WS\}$, where *WS* is the number of workstations,
p Index of products; $p \in \{1, \dots, P\}$, where *P* is the number of products,
E_p Processing route of product p ($E_p = \{e_{1p}, e_{2p}, \dots, e_{ip}, \dots, e_{Np}\}$),
R_i Subset of products requiring workstation i.

Parameters:

$\lambda_{A,pi}$ Arrival rate of product *p* to workstation i ($\lambda_{A,pi} = \sum_{j=0}^{W} \lambda_{A,pji}$),
μ_{pi} Service rate of product *p* at workstation i,
μ_{pB} Service rate of bottleneck workstation *B* corresponding to product *p,*
$\tau_{A,pi}$ Inter-arrival time of product p at workstation i,
$\tau_{S,pi}$ Service (process) time of product p at workstation i including downtime ($\tau_{S,pi} = 1/c_i\mu_{pi}$),
c_i Number of servers (machines) at workstation i,
$\sigma^2_{A,pi}$ Variance of the inter-arrival time of product *p* at workstation i,
$\sigma^2_{S,pi}$ Variance of the service time of product *p* at workstation i,
$c^2_{A,pi}$ Squared coefficient of variation (SCV) of the inter-arrival time of product *p* at workstation i,
$c^2_{S,pi}$ SCV of the service time of product *p* at workstation i,
ρ_{pi} Throughput rate of product *p* at workstation i,
w_{pi} Weight of product *p* at workstation i.

In the following, variables are defined as either related to the "*Original products*" or related to the "*Aggregated product*". The original products are the real products $p \in \{1, \dots, P\}$ that enter the production system, while the *aggregated product* corresponds to the aggregation of the *original products* into a dummy single product.

Variables:

- **Original products**

$\lambda_{D,pi}$ Departure rate of product *p* from workstation i,
$c^2_{D,pi}$ SCV of the inter-departure time of product *p* at workstation i,
W_{pi} Waiting time of product *p* in the queue at workstation i,
L_{pi} WIP of product *p* in the queue at workstation i,
CT_{pi} Cycle time of product *p* in the queue at workstation i.

- **Aggregated product**

$\lambda_{A,\mathbb{P}i}$ Arrival rate of the aggregated product to workstation i,
$\lambda_{D,\mathbb{P}i}$ Departure rate of the aggregated product from workstation i,
$\mu_{\mathbb{P}i}$ Service rate of the aggregated product at workstation i,
$\rho_{\mathbb{P}i}$ Utilization rate of the aggregated product at workstation i ($\rho_{\mathbb{P}i} = \sum_{p \in R_i} \rho_{pi} = \lambda_{A,\mathbb{P}i}/c_i\mu_{\mathbb{P}i}$),

$c^2_{A,\mathbb{P}i}$ SCV of the inter-arrival time of the aggregated product at workstation i,

$c^2_{S,\mathbb{P}i}$ SCV of the service time of the aggregated product at workstation i,

$c^2_{D,\mathbb{P}i}$ SCV of the inter-departure time of the aggregated product at workstation i,

$\tau_{A,\mathbb{P}i}$ Inter-arrival time of the aggregated product at workstation i ($\tau_{A,\mathbb{P}i} = 1/\lambda_{\mathbb{P}i}$),

$\tau_{S,\mathbb{P}i}$ Service time of the aggregated product at workstation i including downtime ($\tau_{S,\mathbb{P}i} = 1/c_i\mu_{i\mathbb{P}}$),

$W_{\mathbb{P}i}$ Waiting time of the aggregated product at workstation i,

$L_{\mathbb{P}i}$ WIP of the aggregated product at workstation i,

$CT_{\mathbb{P}i}$ Cycle time of the aggregated product at workstation i,

CT_{pi} Cycle time of product *p* at workstation i.

3.2.2. The QAG formulations for the processing workstations

In general, the products arriving for service in a workstation have to wait in the queue if the workstation is busy. This waiting process mainly depends on the rate at which the processed products arrive at the workstation and the rate at which the workstation services the products. Neither the inter-arrival time between two arrivals nor the service time of the workstation are deterministic but distributed around a mean value. Accordingly, it is essential to include the variability in time and its distribution while evaluating the performance indicators of the workstation (i.e., waiting time, queue length, throughput rate). Therefore, each workstation is modeled as a multi-server GI/G/c queue, wherein the inter-arrival times between flow units of each product *p* are given by a random variable with general distribution and mean of $1/\lambda_{A,pi}$. The service times are imposed by a random variable with general distribution and mean of $1/\mu_{pi}$. Before each workstation *i*, there exists an infinite buffer and the products are served in a first-come-first-served (FCFS) dispatching rule. Each workstation *i* has *c* parallel servers and each server processes only one unit of each product at a time and devotes all of its resources to complete the transaction. The throughput rate of product *p* at workstation *i* is $\rho_{pi} = \lambda_{A,pi}/c_i\mu_{pi}$ (i.e., $\rho_{pi} = \tau_{S,pi}/c_i\tau_{A,pi}$).

Obtaining an exact value for a performance indicator of workstations modeled by GI/G/c queues is complicated. Therefore, an approximation technique is necessary. The fork-join analysis [45] is adopted to evaluate the performance of each workstation separately. In this analysis, the arrival products are aggregated into a single product (called aggregated product ℙ) and the performance of the workstation is evaluated for the aggregated product. The arrival, service and departure parameters for the aggregated product are approximated based on the parameters of the original products at each workstation. Hereafter, the characterization of the aggregated product and the inter-departure expressions are approximated. Figure 5 illustrates a workstation *i* that processes a given set of products R_i [53].

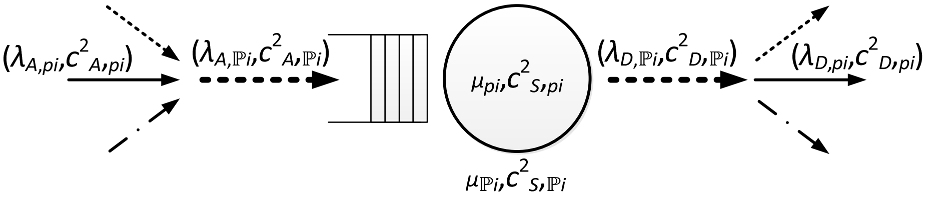
$(\lambda_{A,pi},c^2_{A,pi})$
$(\lambda_{A,\mathbb{P}i},c^2_{A,\mathbb{P}i})$
$\mu_{pi},c^2_{S,pi}$
$(\lambda_{D,\mathbb{P}i},c^2_{D,\mathbb{P}i})$
$(\lambda_{D,pi},c^2_{D,pi})$
$\mu_{\mathbb{P}i},c^2_{S,\mathbb{P}i}$

Figure 5. Workstation *i* and aggregated product

Based on the notations in Section 3.2, Equations (1) to (4) are proposed to estimate the parameters of the aggregate product [65], [66].

$$\lambda_{A,\mathbb{P}i} = \sum_{p\in R_i} \lambda_{A,pi} \qquad \forall i \qquad (1)$$

$$\tau_{S,\mathbb{P}i} = \sum_{p\in R_i} \left(\frac{\lambda_{A,pi}}{\lambda_{A,\mathbb{P}i}}\right) \tau_{S,pi} \qquad \forall i \qquad (2)$$

$$c^2_{A,\mathbb{P}i} = \theta_i \delta_i^2 + (1-\theta_i) \qquad \forall i \qquad (3)$$

$$c^2_{S,\mathbb{P}i} = \frac{\left\{\sum_{p\in R_i} \left(\frac{\lambda_{A,pi}}{\lambda_{A,\mathbb{P}i}}\right) \left(c^2_{S,pi}+1\right)\left(\tau_{S,pi}\right)^2\right\} - \left(\tau_{S,\mathbb{P}i}\right)^2}{\left(\tau_{S,\mathbb{P}i}\right)^2} \qquad \forall i \qquad (4)$$

where $\theta_i = [1 + 4(1 - \rho_{\mathbb{P}i})^2(v_i - 1)]^{-1}$, $\delta_i^2 = \sum_p \left(\frac{\lambda_{A,pi}}{\lambda_{A,\mathbb{P}i}}\right) c_{A,pi}^2$ and $v_i{}^{-1} = \sum_p \left(\frac{\lambda_{A,pi}}{\lambda_{A,\mathbb{P}i}}\right)^2$. In Equation (1), the arrival rate of the aggregated product to a workstation is the sum of the arrival rates of the original products arriving to that workstation. Equation (2) expresses that the service time of the aggregated product in workstation i, $\tau_{S,\mathbb{P}i}$, is equal to the sum of the expected service times of the original products in that workstation. The ratio $\frac{\lambda_{A,pi}}{\lambda_{A,\mathbb{P}i}}$ is the probability of processing product *p* at workstation *i*. In Equation (3), the SCV of the inter-arrival time of the aggregated product is a function of the inter-arrival time of the original products. Equation (4) states that the SCV of the service time of the aggregated product at workstation *i*, $c_{S,\mathbb{P}i}^2$, is equal to the deviation of the expected maximum service time at workstation *i* (i.e., $\sum_{p \in R_i} \left(\frac{\lambda_{A,pi}}{\lambda_{A,\mathbb{P}i}}\right) \left(c_{S,pi}^2 + 1\right)\left(\tau_{S,pi}\right)$) from the service time of the aggregated product, $\tau_{S,\mathbb{P}i}$. This deviation is expressed as a squared value in Equation (4).

Since the workstations are connected via the AMHS, the SCV of the inter-arrival time of product *p* at workstation i, $c_{A,pi}^2$, depends on the SCV of the inter-departure time of the AMHS, c_{DV}^2. In such situations, $c_{A,pi}^2$ is calculated as Equation (5) ([63] and [64]).

$$c_{A,pi}^2 = \pi_i c_{DV}^2 + (1 - \pi_i) \qquad \forall i, p \tag{5}$$

where $\pi_i = \sum_{p=1}^{P} \frac{\lambda_{A,pi}}{\lambda_{AT}}$ and the SCV of the inter-departure time of the AMHS, c_{DV}^2, is calculated in Section 3.4. Based on the estimated input parameters of the aggregated product determined through Equations (1) to (4), the average departure rate, $\lambda_{D,\mathbb{P}i}$, and the SCV of the inter-departure time, $c_{D,\mathbb{P}i}^2$, of the aggregated product at workstation *i* can be found in Equations (6) and (7), respectively [45]. Since the throughput rate remains constant through the production system and no waste is considered at workstations, the departure rate of the aggregated product is equal to its arrival rate as stated in Equation (6). In Equation (7), the SCV of the inter-departure time for the aggregated product is a linear combination of two terms: The SCV of the inter-arrival time and the SCV of the service time.

$$\lambda_{D,\mathbb{P}i} = \lambda_{A,\mathbb{P}i} \qquad \forall i \tag{6}$$

$$c_{D,\mathbb{P}i}^2 = (1 - \rho_{\mathbb{P}i}^2) c_{A,\mathbb{P}i}^2 + \rho_{\mathbb{P}i}^2 c_{S,\mathbb{P}i}^2 \qquad \forall i \tag{7}$$

Finally, the average departure rate, $\lambda_{D,pi}$, and the SCV of the inter-departure time, $c_{D,pi}^2$, of product *p* at processing step *i* can be written as Equations (8) and (9) respectively [45]. Equation (8) follows the same rule as Equation (6), where the throughput rate remains constant through the production system and no waste is considered at workstations for the original products. In Equation (9), the SCV of inter-departure time for each original product *i* is a function of three terms: I) The SCV of the service time of *i*, II) The SCV of the inter-arrival time of *i*, and III) The impact of the other original products that are using the same workstation.

$$\lambda_{D,pi} = \lambda_{A,pi} \qquad \forall i, p \tag{8}$$

$$c_{D,pi}^2 = \rho_{pi}{}^2 c_{S,pi}^2 + \left(1 - 2\rho_{pi}\rho_{\mathbb{P}i} + \rho_{pi}{}^2\right) c_{A,pi}^2 + \left(\frac{\lambda_{A,pi}}{\lambda_{A,\mathbb{P}i}}\right) \sum_{\substack{r \neq p \\ p \in R_i}} \frac{\lambda_{A,\mathbb{P}i}\rho_{ri}^2}{\lambda_{A,ri}} \left(c_{A,ri}^2 + c_{S,ri}^2\right) \qquad \forall i, p \tag{9}$$

After calculating the parameters of the aggregated product at workstation *i*, the waiting time of product *p* in the queue at workstation *i*, W_{pi}, the WIP of product *p* at workstation *i*, L_{pi}, and the WIP of the aggregated product at workstation *i*, $L_{\mathbb{P}i}$, are estimated as Equations (10) to (12), respectively.

$$W_{pi} \approx W_{\mathbb{P}i} \approx \frac{\phi_{\mathbb{P}i}\tau_{S,\mathbb{P}i}}{1 - \rho_{\mathbb{P}i}} \times \frac{c_{A,\mathbb{P}i}^2 + c_{S,\mathbb{P}i}^2}{2} \times G_{\mathbb{P}i} \qquad \forall i, p \in R_i \tag{10}$$

$$L_{pi} = \lambda_{A,pi} W_{pi} \qquad \forall i, p \in R_i \tag{11}$$

$$L_{\mathbb{P}i} = \sum_{p \in R_i} L_{pi} \qquad \forall i \tag{12}$$

where $\phi_{\mathbb{P}i}$ and $G_{\mathbb{P}i}$ are calculated as Equations (13) and (14):

$$\phi_{\mathbb{P}i} = \frac{(c_i\rho_{\mathbb{P}i})^{c_i}}{c_i!\,(1 - \rho_{\mathbb{P}i})} \left[\sum_{k=0}^{c_i - 1} \frac{(c_i\rho_{\mathbb{P}i})^k}{k!} + \frac{(c_i\rho_{\mathbb{P}i})^{c_i}}{c_i!} \frac{1}{1 - \rho_{\mathbb{P}i}}\right]^{-1} \qquad \forall i \tag{13}$$

$$G_{\mathbb{P}i} = \begin{cases} \exp\left(-\frac{2}{3} \times \frac{1-\rho_{\mathbb{P}i}}{\rho_{\mathbb{P}i}} \times \frac{\left(1-c_{A,\mathbb{P}i}^2\right)^2}{c_{A,\mathbb{P}i}^2 + c_{S,\mathbb{P}i}^2}\right), & 0 \leq c_{A,\mathbb{P}i} \leq 1 \\ \exp\left(-(1-\rho_{\mathbb{P}i}) \times \frac{c_{A,\mathbb{P}i}^2 - 1}{c_{A,\mathbb{P}i}^2 + c_{S,\mathbb{P}i}^2}\right), & c_{A,\mathbb{P}i} > 1 \end{cases} \qquad \forall i \tag{14}$$

Finally, the cycle time of the aggregated product and product *p* at workstation *i* are calculated as Equations (15) and (16), respectively.

$$CT_{\mathbb{P}i} = W_{\mathbb{P}i} + \tau_{S,\mathbb{P}i} \qquad \forall i \tag{15}$$
$$CT_{pi} = W_{pi} + \tau_{S,pi} \qquad \forall i, p \tag{16}$$

As a complementary study, we investigate the impact of disruption [67] of the processing workstations on the overall performance of the production system. In this regard, we consider that each workstation *i* is subject to random failures. Once functional, the failures occur at workstation *i* stochastically with rate of f_i and consequently the time until a failure occurs is exponentially distributed with mean of $1/f_i$. Once there is a failure, workstation *i* becomes unavailable and it is stochastically retrieved with rate of r_i. Accordingly, the time until the workstation is repaired is generally distributed with mean of $1/r_i$, standard deviation of σ_{ri} and SCV of c_{Ri}^2 (i.e., $c_{Ri}^2 = r_i\sigma_{ri}$). Hence, the mean availability of workstation *i* when failures are considered is $AV_i \ = \frac{r_i}{r_i+f_i}$ [4]. In this case, the service rate of product *p* at workstation *i* (μ_{pi}^*) and the SCV of the service time of product *p* at workstation *i* ($c_{S,pi}^{2*}$) are modified as Equations (17) and (18). Equations (1) to (16) are modified accordingly (if necessary).

$$\mu_{pi}^* \equiv AV_i\mu_{pi} = \frac{r_i\mu_{pi}}{r_i + f_i} \qquad \forall i, p \in R_i \tag{17}$$
$$c_{S,pi}^{2*} = c_{S,pi}^2 + (1 + c_{Ri}^2)\left(1 - \frac{r_i}{r_i + f_i}\right)\frac{\mu_{pi}}{(r_i + f_i)} \qquad \forall i, p \in R_i \tag{18}$$

3.3. AMHS characterization

3.3.1. Notations

Before characterizing the processing workstation and developing the QAG model for the AMHS, the required notations are provided in this section.

Parameters:

$\lambda_{A,pij}$	Arrival rate of product *p* at workstation *j* from workstation i,
λ_{AT}	Total arrival rate of products in the whole system ($\lambda_{AT} = \sum_{p=1}^{P}\sum_{i=0}^{W}\sum_{j=1}^{W+1}\lambda_{A,pij}$),
c_{RV}^2	SCV of the repair time of vehicles in the AMHS,
d_{ij}	Distance between workstation i and workstation j,
V	Speed of vehicles in the AMHS,
t_{ij}	Transportation time from workstation i to workstation j,
π_i	Probability of transferring products from workstation i,
f_V	Failure rate of vehicles in the AMHS,
r_V	Retrieval rate of vehicles in the AMHS,
LT_V	Loading time of vehicles in the AMHS,
UT_V	Unloading time of vehicles in the AMHS.

Variables:

c_{AV}^2	SCV of the inter-arrival time of transport request for the AMHS,
c_{SV}^2	SCV of the service time of the AMHS,
τ_{SV}	Expected service time of vehicles in the AMHS,
μ_V	Service rate of vehicles in the AMHS,
ρ_V	Utilization rate of vehicles in the AMHS,
n	Number of vehicles in the AMHS,
n_{min}	Minimum number of vehicles to keep the AMHS functional,
c_{DV}^2	SCV of the inter-departure time of the AMHS,
W_V	Waiting time for vehicles at the AMHS.

3.3.2. The QAG formulations for AMHS

Similar to processing workstations, in an AMHS, the products processed in workstations have to wait in a buffer for a free vehicle. This waiting process mainly depends on the rate at which the processed products (transfer requests) arrive at the vehicles and the rate at which the AMHS services the requests. The AMHS is modeled as a multi-server queue (GI/G/*n*), wherein the inter-arrival times between transfer requests are given by a random variable with general distribution and mean of $1/\lambda_{AT}$. The service time is imposed by a random variable with general distribution and mean of $1/\mu_V$. After each workstation, there is an infinite buffer and the products wait until the first vehicle becomes available. The AMHS has *n* vehicles and each vehicle transfers only one unit of each product at a time and devotes all of its resources to complete the travel. The utilization rate of AMHS is $\rho_V = \lambda_{AT}/n\mu_V$. The following assumptions are considered for the AMHS:

- Products are transferred using a FCFS rule.
- The product-independent travel time (t_{ij}) between any two workstations *i* and *j* is deterministic and is given by $t_{ij} = d_{ij}/V$.
- Each vehicle is subject to random failures. Once functional, the failures arrive stochastically with rate of f_v and consequently the time until a failure occurs is exponentially distributed with mean of $1/f_v$.
- Once a failure occurs, the vehicle becomes unavailable and it is stochastically retrieved with rate of r_v. Accordingly, the time until the vehicles is repaired is generally distributed with mean of $1/r_v$, standard deviation of σ_r and SCV of c_{RV}^2 (i.e., $c_{RV}^2 = r_v\sigma_r$).
- The mean availability of the AMHS in case of failure is determined as $AV = \frac{r_v}{r_v+f_v}$.

The first step when analyzing the AMHS is to calculate the mean service time (τ_{SV}) of vehicles. The vehicle service time is the sum of the empty travel time (ETT), the loading time (LT), the loaded travel time (LTT), and the unloading time (UT). The loading time and unloading time are the average time required for loading and unloading one unit of product, and they are usually given and are product-independent. Once requesting a transfer from workstation *i* to workstation *j*, the AMHS spends an empty travel time (ETT) from its current workstation *k*, which is the location of its last delivery, to workstation *i* and spends LT to load the product. Next, it transfers the product from workstation *i* to workstation *j* (i.e., LTT) followed by the unloading time (UT) to unload the product. Although LTT is deterministic as the origin and destination of a request are given based on the product routing information, the ETT is non deterministic since the AMHS can be located at any workstation. These time components are illustrated in Figure 6. Another time component that can happen at any moment is the repair time since the AMHS is subject to failures. This time component is included in the waiting time that the AMHS imposes to the products.

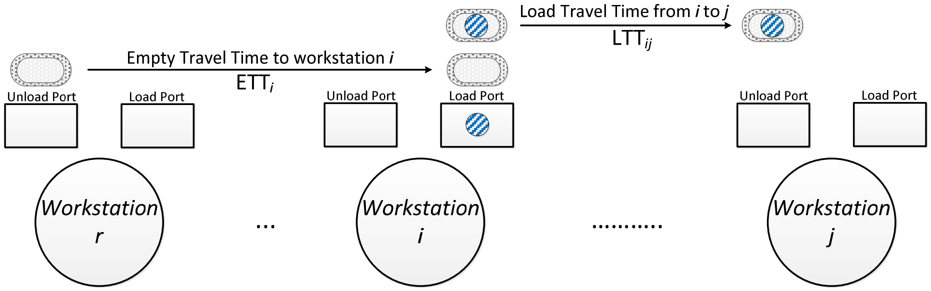

Figure 6. Time components of a transfer request

The ETT can be defined as the sum of two factors: 1) The probability that the vehicle is located in *k* and 2) The travel time from location *k* to the origin location *i* [60]. The EET to workstation *i* is estimated as Equation (19).

$$ETT_i = \sum_{j=1}^{W+1} \pi_j t_{ji} \qquad \forall i \tag{19}$$

The LTT is deterministic as the origin and destination of a request is given. LTT_{ij} is calculated as Equation (20).

$$LTT_{ij} = t_{ij} = \frac{d_{ij}}{V} \qquad \forall i, j \tag{20}$$

Finally, the mean service time of a vehicle (τ_{SV}) is derived from ETT, LTT, LT, and UT, as Equation (21). The concept of expected value is also used to estimate τ_{SV}. $\frac{\lambda_{A,pij}}{\lambda_{AT}}$ is the probability of transferring product *p* from workstation *i* to workstation *j*.

$$\tau_{SV} = \sum_{i=0}^{W} \sum_{j=1}^{W+1} \sum_{p=1}^{P} \frac{\lambda_{A,pij}}{\lambda_{AT}} \left(ETT_i + LT_V + LTT_{ij} + UT_V \right) \tag{21}$$

After calculating the mean service time of the AMHS, we determine the minimum number of vehicles (*n_{min}*) that can meet the basic requirements of the system. In order to keep the system functional at steady state, the throughput of the AMHS (ρ_v) must be smaller than one. Therefore, the initial capacity of the AMHS should be the smallest integer that is larger than the arrival rate divided by the service rate, as shown in Equations (22) and (23).

$$\rho_V = \frac{\lambda_{AT}}{n\mu_V} = \frac{\lambda_{AT}\tau_{SV}}{n} < 1 \tag{22}$$

$$n_{min} = \lfloor \lambda_{AT}\tau_{SV} \rfloor + 1 \tag{23}$$

Based on Whitt's approximation for super-positioning the arrival process in general queue systems [65], the SCV of the inter-arrival time of transport requests for the AMHS (c_{AV}^2) is determined as Equation (24). The remaining equations are derived in the same way than the equations in Section 3.2.2 for the products on the processing workstations.

$$c_{AV}^2 = \theta_V \delta_V^2 + (1 - \theta_V) \tag{24}$$

where $\theta_V = [1 + 4(1 - \rho_V)^2(v_V - 1)]^{-1}$, $\delta_V^2 = \sum_i \pi_i c_{D,pi}^2$ and $v_V{}^{-1} = \sum_i \pi_i{}^2$. The SCV of the inter-departure time of the transport requests for the AMHS (c_{DV}^2) and the SCV of the service time of AMHS (c_{SV}^2) are calculated as Equations (25) and (26).

$$c_{DV}^2 = (1 - \rho_V^2)c_{AV}^2 + \rho_V^2 c_{SV}^2 \tag{25}$$

$$c_{SV}^2 = \frac{\sum_i \left\{ \left(\frac{\lambda_{A,\mathbb{P}i}}{\lambda_{AT}} \right) \left(c_{S,\mathbb{P}i}^2 + 1 \right) \left(\tau_{S,\mathbb{P}i} \right)^2 \right\} - (\tau_{SV})^2}{(\tau_{SV})^2} \tag{26}$$

Finally, c_{AV}^2 can be directly derived as Equation (27).

$$c_{AV}^2 = \frac{(1 - \theta_V) + \theta_V \left(\frac{\sum_p \left[\lambda_{A,p,e_{1p}} c_{A,p,e_{1p}}^2 \right]}{\lambda_{AT}} \right) + \theta_V \sum_i \pi_i \left\{ 1 + \rho_{\mathbb{P}i}^2 \left(c_{S,\mathbb{P}i}^2 - 1 \right) + \pi_i (1 - \rho_{\mathbb{P}i}^2)[\rho_V^2 c_{SV}^2 - 1] \right\}}{1 - \theta_V \sum_i \{ \pi_i{}^2 (1 - \rho_V^2)(1 - \rho_{\mathbb{P}i}^2) \}} \tag{27}$$

After calculating the parameters of the AMHS, the waiting time of the AMHS, W_V, and the WIP at the AMHS, L_V, are estimated as Equations (28) and (29), respectively.

$$W_V \approx \frac{\phi_V \tau_{SV}}{1 - \rho_V} \times \frac{c_{AV}^2 + c_{SV}^2}{2} \times G_V \tag{28}$$

$$L_V = \lambda_{AT} W_V \tag{29}$$

where ϕ_V and G_V are calculated as Equations (30) and (31):

$$\phi_V = \frac{(n\rho_V)^n}{n!\,(1 - \rho_V)} \left[\sum_{k=0}^{n-1} \frac{(n\rho_V)^k}{k!} + \frac{(n\rho_V)^n}{n!} \frac{1}{1 - \rho_V} \right]^{-1} \tag{30}$$

$$G_V = \begin{cases} \exp\left(-\frac{2}{3} \times \frac{1-\rho_V}{\rho_V} \times \frac{(1-c_{AV}^2)^2}{c_{AV}^2 + c_{SV}^2}\right), & 0 \le c_{AV} \le 1 \\ \exp\left(-(1-\rho_V) \times \frac{c_{AV}^2 - 1}{c_{AV}^2 + c_{SV}^2}\right), & c_{AV} > 1 \end{cases} \tag{31}$$

As a complementary study, we consider that the AMHS is subject to random failures. Once functional, the AMHS stochastically fails with rate of f_V and consequently the time until a failure occurs is exponentially distributed with mean of $1/f_V$. Once a failure occurs, the AMHS becomes unavailable and is stochastically repaired with rate of r_V. Accordingly, the time until the repair is completed is generally distributed with mean of $1/r_V$, standard deviation of σ_{rV} and SCV of c_{RV}^2 (i.e., $c_{RV}^2 = r_V\sigma_{rV}$). Hence, the mean availability of the AMHS in this case is $AV_V = \frac{r_V}{r_V+f_V}$ [4]. Also, the service rate (μ_V^*) and the SCV of the service time of the AMHS (c_{SV}^{2*}) are modified as Equations (32) and (33). Equations (1) to (31) are modified accordingly (if necessary).

$$\mu_V^* \equiv AV_V\mu_V = \frac{r_V\mu_V}{r_V + f_V} \tag{32}$$

$$c_{SV}^{2*} = c_{SV}^2 + (1 + c_{RV}^2)\left(1 - \frac{r_V}{r_V + f_V}\right)\frac{\mu_V}{(r_V + f_V)} \tag{33}$$

3.3.3. Capacity planning for AMHS

Two major KPIs for any production system are production throughput and product cycle time. Since the AMHS is an auxiliary system in any production system, the aim of capacity planning for the AMHS is to determine the right number of vehicles of the AMHS such that the AMHS has no impact on the KPIs [60]. Therefore, the AMHS should feed all workstations at the right moment and should not let the workstations become idle. Two important workstations in any production system are the troublesome workstation (*TW*) and the fastest workstation (*FW*). The overall performance of a production system is easily affected by *TW* and *FW*.

In manufacturing systems, *TW* is usually expensive, it is hard to increase its capacity, it has a high failure rate and a large unavailability, and it is generally hard to plan its production as well as its capacity. Therefore, managers attempt not to allow the AMHS to decrease the throughput of *TW*. The throughput of a production system is usually strongly affected by its *TW*. Therefore, the best way to achieve the highest throughput is to prevent *TW* from being idle, i.e. the AMHS should always be ready to serve *TW* when necessary.

Delp et al. [68], [69] and Tu et al. [60] propose the concept of *X*-factor contribution (*XFC*) to determine the contribution of each workstation to the overall system performance. The *XFC* of workstation *i* and of *TW* are defined as Equations (34) and (35).

$$XFC_i = \frac{\tau_{S,\mathbb{P}i}\left(1 + \frac{\phi_{\mathbb{P}i}}{1-\rho_{\mathbb{P}i}} \times \frac{c_{A,\mathbb{P}i}^2 + c_{S,\mathbb{P}i}^2}{2} \times G_{\mathbb{P}i}\right)}{\sum_{j=1}^{W} \tau_{S,\mathbb{P}j}} \qquad \forall j \tag{34}$$

$$TW = \left\{ i | XFC_i = \max_j\left\{XFC_j\right\}\right\} \tag{35}$$

FW is the workstation with the highest service rate or the lowest service time, and is determined as (36).

$$FW = \left\{ i | \tau_{S,\mathbb{P}i} = \min_j\left\{\tau_{S,\mathbb{P}j}\right\}\right\} \tag{36}$$

The main goal of AMHS capacity planning is to determine the number of vehicles that leads to an optimal performance of the whole production system. Let us define the optimal performance of the whole production system in two cases: 1) No workstation is idle and 2) The production system operates at its highest throughput. In order to study these two cases, we adopted two policies: **Policy 1**, "Preventing *FW* from being idle" or **Policy 2**, "Preventing *TW* from being idle". **Policy 1** aims at ensuring that all workstations are always working, i.e. that they are never idle, and **Policy 2** pushes the production system to operate at its highest throughput. The goal is to keep the WIP strictly positive in front of *FW* and *TW*. In **Policy 1**, no workstation is idle and the WIP in front of all workstations is larger than zero because if the AMHS can feed *FW* on time, then all workstations are also fed on time. This policy is expensive and may need significant storage space. Consequently, managers may adopt the second policy by only preventing *TW* from being idle. This section proposes the minimum number of vehicles to optimize the throughput for the two policies above.

In Figure 4, consider that workstation *j* is *FW* (*TW*); hence, to avoid the negative impact of the AMHS and ensure the highest performance of the *JS*, the WIP in front of *FW* (*TW*) should always be strictly positive ($L_{\mathbb{P}j}$ > 0; *j* = *FW* or *TW*). By considering the AMHS, *FW,* and *TW* as queuing systems, the sum of the product queue time TQ at the AMHS (W_V) and the mean transportation time TT of the AMHS (τ_{SV}/n) should be lower than the mean service time of *FW*, $\tau_{S,\mathbb{P}}^{FW}$ or lower than the mean service time of *TW*, $\tau_{S,\mathbb{P}}^{TW}$. Conditions (37) and (38) are proposed to satisfy **Policies 1** and **2**, respectively.

$$W_V+\frac{\tau_{SV}}{n}<\tau_{S,\mathbb{P}}^{FW} \tag{37}$$

$$W_V+\frac{\tau_{SV}}{n}<\tau_{S,\mathbb{P}}^{TW} \tag{38}$$

Since it is hard in practice to strictly satisfy Inequalities (37) and (38), two chance constraints (39) and (40) are proposed that correspond to the probability that ensures the highest performance of *JS* under **policies 1** and **2** is larger than the target probabilities α_{FW} and α_{TW}, respectively.

$$P\left\{W_V+\frac{\tau_{SV}}{n}<\tau_{S,\mathbb{P}}^{FW}\right\}>\alpha_{FW} \tag{39}$$

$$P\left\{W_V+\frac{\tau_{SV}}{n}<\tau_{S,\mathbb{P}}^{TW}\right\}>\alpha_{TW} \tag{40}$$

or we have

$$P\left\{W_V<\tau_{S,\mathbb{P}}^{FW}-\frac{\tau_{SV}}{n}\right\}>\alpha_{FW} \tag{41}$$

$$P\left\{W_V<\tau_{S,\mathbb{P}}^{TW}-\frac{\tau_{SV}}{n}\right\}>\alpha_{TW} \tag{42}$$

Using the waiting time distribution function of Whitt [70], the left-hand-side (LHS) of Inequalities (41) and (42) can be rewritten as:

$$P\left\{W_V<\tau_{S,\mathbb{P}}^{FW}-\frac{\tau_{SV}}{n}\right\}\approx 1-\omega_V e^{-\eta_V\left(\tau_{S,\mathbb{P}}^{FW}-\frac{\tau_{SV}}{n}\right)} \tag{43}$$

$$P\left\{W_V<\tau_{S,\mathbb{P}}^{TW}-\frac{\tau_{SV}}{n}\right\}\approx 1-\omega_V e^{-\eta_V\left(\tau_{S,\mathbb{P}}^{TW}-\frac{\tau_{SV}}{n}\right)} \tag{44}$$

Where

$$\omega_V\approx\eta_V\times W_V \tag{45}$$

$$\eta_V=\frac{2n\times(1-\rho_V)}{c_{AV}^2+c_{SV}^2} \tag{46}$$

Decision makers should set a target probability whereby the highest performance for the production system is kept. Since proposing a close formulation for the number of vehicles *n* in (37) to (46) is complicated, the value of *n* is found for each policy by trial and error. The minimum number of vehicles that satisfies Inequalities (41) and (42) is the best capacity of the AMHS under policies 1 and 2, respectively.

Hence, by considering a target probability α_{FW}, a minimal WIP is kept in front of *FW*. It is then possible to absorb any fluctuation of the production system such as transportation delays of the AMHS. For optimal throughput, a normal and acceptable value of the target probability α_{FW} can be defined around 95%. Increasing too much the target probability α_{FW} will increase the number of vehicles; however there is no guarantee that the throughput of the production system increases as well. Therefore, the AMHS capacity may be overestimated.

On the other hand, setting the target probability α_{FW} around 100% and keeping the WIP (>0) in front of *FW* results in an unnecessary WIP in front of slower workstations. In addition, considering a target probability α_{FW} around 100% and preventing cycle time losses by the AMHS is not practical from a queuing theory point of view.

3.4. Proposed QAG model for the *AJS* system

According to Equation (10), the mean waiting time at workstation *i* is the same for all products in R_i. Since the processing route of each product is known as E_p, the *JS* system can be aggregated into a single-step aggregated JS (*AJS*) system for each product *p*. The *AJS* system for each product *p* is illustrated in Figure 7 with the corresponding parameters. We model the *AJS* system as a GI/G/1 queue system. Before proposing the approximation formulations, the required notations are provided below:

$\lambda_{A,p}^{AJS}$	Arrival rate of product *p* to the *AJS* system,
$\lambda_{D,p}^{AJS}$	Departure rate of product *p* from the *AJS* system,
μ_{p}^{AJS}	Service rate of product *p* in the *AJS* system,
ρ_{p}^{AJS}	Utilization rate of product *p* in the *AJS* system,
$c_{A,p}^{2,AJS}$	SCV of the inter-arrival time of product *p* at the *AJS* system,
$c_{S,p}^{2,AJS}$	SCV of the service time of product *p* in the *AJS* system,
$c_{D,p}^{2,AJS}$	Coefficient of variation of the inter-departure time of product *p* from the *AJS* system,
$\tau_{A,p}^{AJS}$	Inter-arrival time of product *p* to the *AJS* system,
$\tau_{S,p}^{AJS}$	Service time of product *p* in the *AJS* system,
W_{p}^{AJS}	Waiting time of product *p* in the queue in the *AJS* system,
L_{p}^{AJS}	Queue length of product *p* in the *AJS* system,
CT_{p}^{AJS}	Cycle time of product *p* in the *AJS* system.

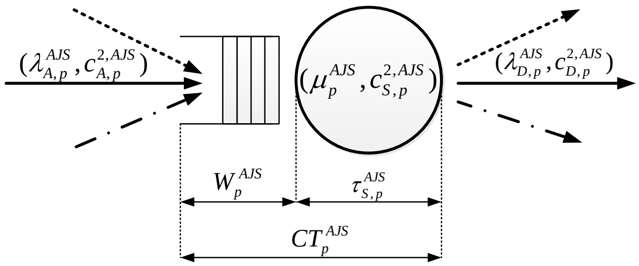
$(\lambda_{A,p}^{AJS}, c_{A,p}^{2,AJS})$
$(\mu_{p}^{AJS}, c_{S,p}^{2,AJS})$
$(\lambda_{D,p}^{AJS}, c_{D,p}^{2,AJS})$
W_{p}^{AJS}
$\tau_{S,p}^{AJS}$
CT_{p}^{AJS}

Figure 7. *AJS* system

This section aims at approximating the variables of the *AJS* system based on the parameters of the original *JS* system. Based on its process route, the average cycle time of product *p* through the entire *JS* system, CT_{p}^{JS}, can be expressed as Equation (47). CT_{p}^{JS} is the sum of the cycle times at all corresponding workstations of E_p plus the time spent in the AMHS that is the sum of the waiting times in the AMHS (W_V) and the mean service time of the AMHS (τ_{SV}).

$$CT_{p}^{JS} = \sum_{i \in E_p} CT_{pi} + |E_p|(W_V + \tau_{SV}) \tag{47}$$

where $|E_p|$ is the number of workstations that serve product *p*. The idea of aggregation in this paper is that the average service time of the *AJS* system for product *p,* $\tau_{S,p}^{AJS}$, is the total time that each product spends in the system after departing the queue of the first workstation (i.e., *i*=1). This time is equal to the sum of the service times for all workstations and the in-between waiting times. By "in-between" waiting time, we mean the sum of the waiting times of workstations *i*; $i = 2, \dots, e_{Np}$. Accordingly, let us express $\tau_{S,p}^{AJS}$ as Equation (48). Note that unlike the literature, $\tau_{S,p}^{AJS}$ in Equation (48) is WIP-dependent. When the WIP increases, the waiting time increases and the service time of the *APS* system increases as well for all products.

$$\tau_{S,p}^{AJS} = \sum_{i=2}^{e_{Np}} W_{pi} + \sum_{i=1}^{e_{Np}} \tau_{S,\mathbb{P}i} + |E_p|(W_V + \tau_{SV}) \tag{48}$$

An underlying assumption in heavy traffic conditions for serial queue networks is that the waiting time of the network is dominated by the waiting time at the bottleneck workstation. Knowing that $\lambda_{A,pi} = \lambda_1$ for $i = 2, \dots, e_{Np}$, the bottleneck workstation *BW* is expressed as the workstation with minimum service rate or longest service time. Because the AMHS is considered in the production system, the service rate of product *p* in the *AJS* system μ_p^{AJS} is the minimum between the bottleneck's service rate and the service rate of the AMHS as expressed in Equation (49).

$$\mu_p^{AJS} = \min\left\{\min_{i \in E_p} \frac{c_i}{\tau_{S,\mathbb{P}i}}, \mu_V\right\} \qquad \forall p \tag{49}$$

To calculate the SCV of the service time of the *AJS* system for product *p*, $c_{S,p}^{2,AJS}$, we first need to calculate the variance of the service time in the *AJS* system, $\sigma_{S,p}^{2,AJS}$, as Equation (50), wherein Var(*X*) is the variance of variable *X*.

$$\sigma_{S,p}^{2,AJS} = \mathrm{Var}\left(\sum_{i=2}^{e_{Np}} W_{pi} + \sum_{i=1}^{e_{Np}} \tau_{S,\mathbb{P}i} + \left|E_p\right|(W_V + \tau_{SV})\right) \qquad \forall p \tag{50}$$

Since the waiting times as well as the service times of the workstations are identical and independent, Equation (50) can be re-written as Equations (51) and (52).

$$\sigma_{S,p}^{2,AJS} = \sum_{i=2}^{e_{Np}} \mathrm{Var}(W_{pi}) + \sum_{i=1}^{e_{Np}} \mathrm{Var}(\tau_{S,pi}) + \left|E_p\right|^2 [\mathrm{Var}(W_V) + \mathrm{Var}(\tau_{SV})] \qquad \forall p \tag{51}$$

$$\sigma_{S,p}^{2,AJS} = \sum_{i=2}^{e_{Np}} \mathrm{Var}(W_{pi}) + \sum_{i=1}^{e_{Np}} \sigma_{S,\mathbb{P}i}^2 + \left|E_p\right|^2 [\mathrm{Var}(W_V) + c_{SV}^2 \tau_{SV}^2] \qquad \forall p \tag{52}$$

Applying the Kingman-Kollerstrom approximation [71], the cumulative distribution function of the waiting time of product *p* at workstation*,* $F_{W_{pi}}$, and the cumulative distribution function of the waiting time at the AMHS, F_{W_V}, can be well approximated by Equations (53) and (54).

$$F_{W_{pi}}(x) = 1 - \omega_{pi} e^{(-\eta_{pi} x)} \qquad \forall p, i \in E_p \tag{53}$$

$$F_{W_V}(x) = 1 - \omega_V e^{(-\eta_V x)} \tag{54}$$

where ω_{pi} and η_{pi} are approximated as (55) and (56).

$$\omega_{pi} = \eta_{pi} W_{pi} \qquad \forall p, i \in E_p \tag{55}$$

$$\eta_{pi} = \frac{2c_i(1 - \rho_{\mathbb{P}i})}{c_{A,\mathbb{P}i}^2 + c_{S,\mathbb{P}i}^2} \tag{56}$$

Accordingly, $\mathrm{Var}(W_{pi})$ and $\mathrm{Var}(W_V)$ can be expressed as:

$$\mathrm{Var}(W_{pi}) = \frac{\omega_{pi}}{{\eta_{pi}}^2} \qquad \forall p, i \in E_p \tag{57}$$

$$\mathrm{Var}(W_V) = \frac{\omega_V}{{\eta_V}^2} \tag{58}$$

Finally, $c_{S,p}^{2,AJS}$ is approximated as (59):

$$c_{S,p}^{2,AJS} \approx \frac{\sum_{i=2}^{e_{Np}} \left[\frac{\omega_{pi}}{{\eta_{pi}}^2}\right] + \sum_{i=1}^{e_{Np}} c_{S,\mathbb{P}i}^2 \tau_{S,\mathbb{P}i}^2 + \left|E_p\right|^2 \left[\frac{\omega_V}{{\eta_V}^2} + c_{SV}^2 \tau_{SV}^2\right]}{\left(\sum_{i=2}^{e_{Np}} W_{pi} + \sum_{i=1}^{e_{Np}} \tau_{S,pi} + \left|E_p\right|(W_V + \tau_{SV})\right)^2} \qquad \forall p, i \in E_p \tag{59}$$

Other parameters of the *AJS* system in Figure 7 are provided in Equations (60) to (67).

$$\lambda_{A,p}^{AJS} = \lambda_{A,pi} \qquad \forall p, i = e_{p1} \tag{60}$$

$$c_{A,p}^{2,AJS} = c_{A,pi}^2 \qquad \forall p, i = e_{p1} \tag{61}$$

$$c_{D,p}^{2,AJS} = c_{D,pi}^{2} \qquad \forall p, i = e_{Np} \tag{62}$$

$$\rho_p^{AJS} = \frac{\lambda_{A,p}^{AJS}}{\mu_p^{AJS}} \qquad \forall p \tag{63}$$

$$W_p^{AJS} \approx \frac{\rho_p^{AJS}/\mu_p^{AJS}}{1-\rho_p} \times \frac{c_{A,p}^{2,AJS} + c_{S,p}^{2,AJS}}{2} \times G_p^{AJS} \qquad \forall p \tag{64}$$

where G_p^{AJS} is calculated as Equation (66).

$$G_p^{AJS} = \begin{cases} \exp\left(-\frac{2}{3} \times \frac{1-\rho_p^{AJS}}{\rho_p^{AJS}} \times \frac{\left(1-c_{A,p}^{2,AJS}\right)^2}{c_{A,p}^{2,AJS} + c_{S,p}^{2,AJS}}\right), & 0 \le c_{A,p}^{AJS} \le 1 \\ \exp\left(-\left(1-\rho_p^{AJS}\right) \times \frac{c_{A,p}^{2,AJS} - 1}{c_{A,p}^{2,AJS} + c_{S,p}^{2,AJS}}\right), & c_{A,p}^{AJS} > 1 \end{cases} \qquad \forall p \tag{65}$$

Finally, the cycle time and the WIP of each product in the *AJS* system are calculated as Equations (67) and (68), respectively.

$$CT_p^{AJS} = W_p^{AJS} + \tau_{S,p}^{AJS} \qquad \forall p \tag{66}$$

$$L_p^{AJS} = \lambda_{A,p}^{AJS} CT_p^{AJS} \qquad \forall p \tag{67}$$

3.5. *AJS* with priorities on products

In order to stay competitive and to satisfy customers' demand with high variety and significant variability, companies often need to handle customers and customer orders differently, and thus products need to have different priorities. Priorities usually can be divided into three levels: Hot, rush and normal [72]. A product with a higher priority requires a shorter cycle time and should be processed before the lower priority products, whereas products with lower priority need to wait until higher priority products are completed and the workstation becomes available. Because of longer waiting times, the cycle time of lower priority products is longer.

Accordingly, we consider products with multiple priorities in the *AJS* system and investigate the effect of considering priorities on the cycle time of products. This is done by developing a priority queuing system for the *AJS* system. All products keep their priority on all their corresponding workstations. In a priority queueing system, we assume that an arriving product belongs to a priority class *r* (*r* = 1, 2, ..., *R*). The next product to be served is the customer with the highest priority *r*. Among products with the same priority, the queueing discipline is FCFS. We consider priority queue without preemption, wherein a product already in service is not preempted by an arriving product with higher priority. The mean waiting time for an arriving product *p* with priority level *r* consists of three components as in [73]:

I. The mean remaining service time τ_0^{AJS} of the product in service (if any) in the *AJS* system,
II. The mean service time of products in the queue with the same or higher priority as the tagged product,
III. The mean service time of products with higher priority that arrive at the queue while the tagged product is in the queue and are served before the tagged product.

The necessary notations are first defined below and only for the *AJS* system.

P_r Set of products with priority *r*.

$\mathbb{Q}_{r'r}^{AJS}$ Mean number of products with priority *r'* found in the *AJS* queue with the tagged product of priority *r* and receiving service before it

$\mathbb{Z}_{r'r}^{AJS}$ Mean number of products with priority *r'* who arrive during the waiting time of the tagged product of priority *r* and receive service before it

$\lambda_{A,p}^{AJS,r}$ Arrival rate of product *p* with priority *r* to the *AJS* system

$\mu_p^{AJS,r}$ Service rate of product *p* with priority *r* in the *AJS* system

$c_{A,p}^{2,AJS,r}$ SCV of the inter-arrival time of product *p* with priority *r* at the *AJS* system

$c_{S,p}^{2,AJS,r}$ SCV of the service time of product *p* with priority *r* in the *AJS* system

$\tau_{S,p}^{AJS,r}$ Service time of product *p* with priority *r* in the *AJS* system

$W_p^{AJS,r}$ Waiting time of product *p* with priority *r* in the queue in the *AJS* system
$L_p^{AJS,r}$ Queue length of product *p* with priority *r* in the *AJS* system
$CT_p^{AJS,r}$ Cycle time of product *p* with priority *r* in the *AJS* system

The mean waiting time of products with priority *r* in the AJS system, $W_p^{AJS,r}$, can be then calculated as the sum of three above-mentioned components:

$$W_p^{AJS,r} = \tau_0^{AJS} + \sum_{r'=r}^{R} \frac{\mathbb{Q}_{r'r}^{AJS}}{\mu_p^{AJS,r'}} + \sum_{r'=r+1}^{R} \frac{\mathbb{Z}_{r'r}^{AJS}}{\mu_p^{AJS,r'}} \qquad \forall r, p \in P_r \tag{69}$$

where

$$\tau_0^{AJS} \approx \sum_{r'=1}^{R} \sum_{\substack{p=1 \\ p\in P_{r'}}}^{P} \frac{\lambda_{A,p}^{AJS,r'}}{\lambda_{A,\mathbb{P}1}} \frac{c_{A,p}^{2,AJS,r'} + c_{S,p}^{2,AJS,r'}}{2\mu_p^{AJS,r'}} \tag{70}$$

$$\mathbb{Q}_{r'r}^{AJS} = \sum_{\substack{p=1 \\ p\in P_{r'}}}^{P} \lambda_{A,p}^{AJS,r'} W_p^{AJS,r'} \qquad \forall r', r; r' \geq r \tag{71}$$

$$\mathbb{Z}_{r'r}^{AJS} = \sum_{\substack{p=1 \\ p\in P_{r'}}}^{P} \lambda_{A,p}^{AJS,r'} W_p^{AJS,r} \qquad \forall r', r; r' > r \tag{72}$$

The cycle time and the WIP of product *p* with priority *r* in the *AJS* system are calculated as Equations (75) and (76), respectively.

$$CT_p^{AJS,r} = W_p^{AJS,r} + \tau_{S,p}^{AJS,r} \qquad \forall p \tag{75}$$

$$L_p^{AJS,r} = \lambda_{A,p}^{AJS,r} CT_p^{AJS,r} \qquad \forall p \tag{76}$$

4. Computational results

This section provides a comprehensive analysis of computational results to validate the correctness and the performance of the proposed QAG model. Numerous experiments with different parameter settings are first designed in Section 4.1. Then, Section 4.2 compares a detailed discrete event simulation approach and the proposed QAG model. A comprehensive sensitivity analysis is performed in Section 4.3 using the QAG model to provide valuable managerial insights on the production system and the AMHS. Finally, an analysis of the capacity planning of the AMHS is provided in Section 4.4.

4.1. Design of experiments

In order to compare the performance of a discrete event simulation approach and the proposed QAG model, the job-shop production system in Figure 8 is considered. Figure 8 illustrates a (part of a) production system with six workstations to manufacture three products (i.e., P_1, P_2, and P_3). The processing routes E_p of the products are $E_1 = \{1 \to 3 \to 4 \to 5 \to 6\}$, $E_2 = \{1 \to 2 \to 4 \to 6\}$, and $E_3 = \{2 \to 3 \to 5\}$. Accordingly, the subsets of products in workstations are $R_1 = \{P_1, P_2\}$, $R_2 = \{P_2, P_3\}$, $R_3 = \{P_1, P_3\}$, $R_4 = \{P_1, P_2\}$, $R_5 = \{P_1, P_3\}$ and $R_6 = \{P_1, P_2\}$. The performance of the proposed QAG model is tested through different instances labeled I1 to I20. Table 1 provides the settings of the production system and of the AMHS for each instance. The processing times of the workstations and the AMHS are specifically determined to respect the condition $\rho = \lambda/c\mu < 1$. On the other hand, *cμ* is considered as close as possible to *λ* in some scenarios to evaluate the accuracy of the QAG model in case of high congestion. The throughput (utilization) rate is actually calculated as λ/λ_{max} for each scenario, wherein λ for each product is generated from Table 1, and λ_{max} is actually the maximum arrival rate of a product that all workstations in the production system can process, such that values larger than λ_{max} will saturate the system in terms of WIPs. Accordingly, λ_{max} is considered close to *cμ* (i.e., $\lambda_{max} \sim c\mu$) for each product.

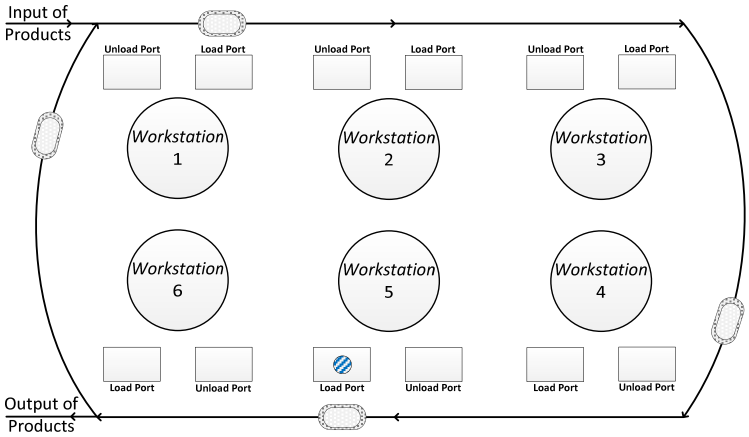

Figure 8. Production system with six workstations

The simulation experiments for the smallest instance I1 took at least 4 hours on average. The simulation results were recorded based on 20 independent simulation runs with a 95% confidence interval. This ensured that the standard deviation of the throughput value from different replications is within ±0.5% of the mean. Each run corresponds to the production of at least 15,000 units of products. The statistics corresponding to the first 1,500 units were neglected to account for the transient start-up effects.

The KPI recorded in all experiments is the total cycle time of each product. To determine the numerical accuracy of the proposed QAG model, the percentage errors between the QAG model and the simulation (SIM) in terms of cycle time are computed as in Equation (77).

$$\Delta_{KPI}= \frac{KPI^{(SIM)} - KPI^{(QAG)}}{KPI^{(SIM)}} \times 100 \tag{77}$$

Table 1. Parameter settings of experimental instances

Instance	Production system						AMHS				
	λ (*unit/h*)		c	f (#/h)	r		V (m/s)	LT&UT	f (#/h)	r	
	mean	c_A^2			*mean*	c_R^2				*mean*	c_{RV}^2
I1	[50,150]	[2,5]	1	[0.1,0.5]	[1.0,2.0]	[0.1,0.4]	0.25	[20,30]	1	2	[0.2,0.4]
I2	[150,200]	[2,5]	1	[0.5,1.0]	[2.0,5.0]	[0.2,0.6]	0.25	[25,35]	1.5	3	[0.4,0.8]
I3	[200,250]	[2,5]	1	[1.0,1.5]	[6.0,9.0]	[0.4,0.8]	0.25	[30,40]	2	5	[1.0,2.0]
I4	[250,300]	[2,5]	1	[1.5,2.0]	[8.0,18]	[0.6,1.0]	0.25	[35,50]	3	7	[2.0,3.0]
I5	[50,150]	[5,10]	2	[0.1,0.5]	[1.0,2.0]	[0.1,0.4]	0.50	[20,30]	1	2	[0.2,0.4]
I6	[150,200]	[5,10]	2	[0.5,1.0]	[2.0,5.0]	[0.2,0.6]	0.50	[25,35]	1.5	3	[0.4,0.8]
I7	[200,250]	[5,10]	2	[1.0,1.5]	[6.0,9.0]	[0.4,0.8]	0.50	[30,40]	2	5	[1.0,2.0]
I8	[250,300]	[5,10]	2	[1.5,2.0]	[8.0,18]	[0.6,1.0]	0.50	[35,50]	3	7	[2.0,3.0]
I9	[50,150]	[10,15]	3	[0.1,0.5]	[1.0,2.0]	[0.1,0.4]	0.75	[20,30]	1	2	[0.2,0.4]
I10	[150,200]	[10,15]	3	[0.5,1.0]	[2.0,5.0]	[0.2,0.6]	0.75	[25,35]	1.5	3	[0.4,0.8]
I11	[200,250]	[10,15]	3	[1.0,1.5]	[6.0,9.0]	[0.4,0.8]	0.75	[30,40]	2	5	[1.0,2.0]
I12	[250,300]	[10,15]	3	[1.5,2.0]	[8.0,18]	[0.6,1.0]	0.75	[35,50]	3	7	[2.0,3.0]
I13	[50,150]	[15,20]	4	[0.1,0.5]	[1.0,2.0]	[0.1,0.4]	1.00	[20,30]	1	2	[0.2,0.4]
I14	[150,200]	[15,20]	4	[0.5,1.0]	[2.0,5.0]	[0.2,0.6]	1.00	[25,35]	1.5	3	[0.4,0.8]
I15	[200,250]	[15,20]	4	[1.0,1.5]	[6.0,9.0]	[0.4,0.8]	1.00	[30,40]	2	5	[1.0,2.0]
I16	[250,300]	[15,20]	4	[1.5,2.0]	[8.0,18]	[0.6,1.0]	1.00	[35,50]	3	7	[2.0,3.0]
I17	[50,150]	[20,25]	5	[0.1,0.5]	[1.0,2.0]	[0.1,0.4]	1.25	[20,30]	1	2	[0.2,0.4]
I18	[150,200]	[20,25]	5	[0.5,1.0]	[2.0,5.0]	[0.2,0.6]	1.25	[25,35]	1.5	3	[0.4,0.8]
I19	[200,250]	[20,25]	5	[1.0,1.5]	[6.0,9.0]	[0.4,0.8]	1.25	[30,40]	2	5	[1.0,2.0]
I20	[250,300]	[20,25]	5	[1.5,2.0]	[8.0,18]	[0.6,1.0]	1.25	[35,50]	3	7	[2.0,3.0]

4.2. Comparing the QAG model and a simulation model

This section compares the results from the analytical QAG model with those obtained from the discrete event simulation model built using AnyLogic (www.anylogic.com) on an Intel Pentium IV PC. The QAG model is implemented in MATLAB 2014 and executed in less than one second for each problem instance on the same PC.

Table 2 shows the corresponding results of the QAG and detailed simulation (SIM) models under the setting of each instance. The results of Table 2 provide insights regarding the impact of different input parameters on the cycle time of each product. We observed that the mean gap between the proposed QAG model and the detailed simulation (SIM) is 2.23%, 2.69% and 1.87% for products P_1 to P_3, respectively. The proposed QAG model takes less than one second to execute any instance. These acceptable gaps and low computational times validate the effectiveness and the high performance of the proposed QAG model. Accordingly, the proposed QAG model can replace time-consuming simulation methods with the guarantee of providing acceptable results.

Another observation is that the QAG performance degrades when the arrival rates increase. The increase of arrival rates leads to higher congestion in the system, which degrades the accuracy of the QAG model in approximating the waiting time (i.e., Equation (65)) and, consequently, the cycle time (i.e., Equation (67)) is approximated with higher gaps. In order to identify the limit of the proposed QAG model, Figure 9 depicts the mean gap value in terms of throughput rates (i.e., λ/λ_{max}).

Figure 9 shows that the proposed QAG model performs quite well for throughput rates lower than 80%, and that the QAG model can replace simulation models, although the performance degrades for higher values of λ/λ_{max} ($\geq$ 80%). Accordingly, the utilization of the proposed QAG model with throughput rates higher than 80% might not be recommended; however, it depends on the tolerance of the experts to the accuracy of the cycle time approximation. One may agree to sacrifice the accuracy to quickly obtain approximate results.

Table 2. Comparison between QAG and SIM in terms of cycle time

Instance	QAG			Vehicle min	SIM			Δ (%)		
	P_1	P_2	P_3		P_1	P_2	P_3	P_1	P_2	P_3
I1	20.68	25.76	7.94	2	20.76	25.87	7.96	0.38	0.43	0.32
I2	104.25	134.17	40.00	2	106.09	136.57	40.60	1.76	1.79	1.52
I3	31.42	46.45	11.39	3	31.60	46.78	11.45	0.58	0.71	0.44
I4	106.17	209.74	35.56	4	108.09	214.48	36.07	1.81	2.26	1.43
I5	147.46	178.59	57.69	1	151.78	185.07	59.12	2.93	3.63	2.47
I6	24.81	32.94	9.37	2	24.94	33.14	9.41	0.49	0.59	0.39
I7	171.86	239.98	63.93	2	179.97	252.91	66.21	4.72	5.39	3.58
I8	244.79	434.23	84.48	3	257.89	465.88	88.50	5.35	7.29	4.76
I9	97.93	119.07	38.25	1	99.53	121.10	38.82	1.63	1.70	1.48
I10	18.86	25.39	7.08	2	18.91	25.48	7.09	0.29	0.34	0.25
I11	81.11	115.93	29.95	2	82.23	117.80	30.43	1.38	1.61	1.59
I12	158.43	299.63	53.89	3	164.42	313.90	55.54	3.78	4.76	3.05
I13	83.55	101.85	32.61	1	84.79	103.39	33.05	1.48	1.51	1.37
I14	16.70	22.70	6.25	2	16.73	22.75	6.26	0.18	0.22	0.15
I15	62.62	90.77	23.03	2	63.16	91.74	23.22	0.87	1.06	0.79
I16	139.66	270.73	47.26	3	143.27	278.96	48.26	2.58	3.04	2.12
I17	76.74	93.74	29.94	1	77.60	94.98	30.20	1.11	1.32	0.87
I18	968.11	1247.22	373.34	1	1064.43	1400.01	404.06	9.95	12.25	8.23
I19	54.82	80.28	20.12	2	55.29	81.04	20.27	0.86	0.94	0.71
I20	131.69	258.87	44.45	3	135.00	266.35	45.30	2.51	2.89	1.93
							Min	0.18	0.22	0.15
							Mean	2.23	2.69	1.87
							Max	9.95	12.25	8.23

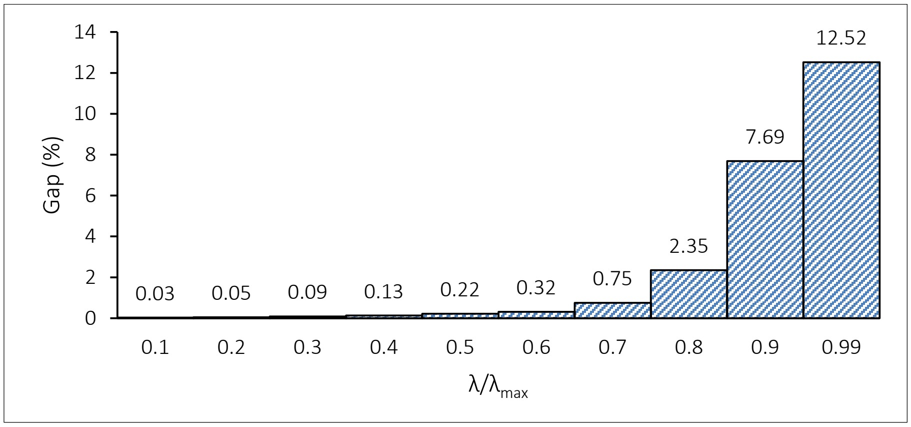

Figure 9. Gap vs. Throughput rate

In order to validate the performance of the proposed QAG model on larger instances, five new instances I21-I25 with more than 10 products and 20 workstations are generated based on the instances I4, I8, I12, I16, and I20, respectively. The results can be found in Table 3, where the first column shows the instance number, and the second and third columns show the number of products *P* and the number of workstations *WS*. For example, instance I21 takes the same parameter setting as instance I4 but with *P*=10 products and *WS*=15 workstations.

Note that, when the size of the production system is increasing, the accuracy of the proposed method slightly decreases compared to the detailed simulation model but still with small gaps.

Table 3. Gap between QAG and detailed simulation model on larger instances

Instance	Parameters		Δ (%)		Min # of vehicles
	P	*WS*	Δ_{min}	Δ_{max}	
I21	10	15	1.98	2.54	5
I22	15	25	2.05	2.86	9
I23	30	30	2.42	3.31	12
I24	50	40	3.12	3.69	17
I25	80	100	4.35	6.78	26

4.3. Sensitivity analyses

A remarkable advantage of the proposed QAG model is the possibility of analyzing the sensitivity of the KPIs to the input parameters. Sensitivity analysis using detailed simulation model is difficult in practice since it is timely expensive. Performing sensitivity analysis provides knowledge on the behavior of important parameters. Having this knowledge helps experts to achieve the desired level of KPIs by correctly adjusting these parameters.

4.3.1. Impact of AMHS on the cycle time

The main aim of this section is to do a sensitivity on the input parameters of both production system and AMHS to finally show the impact of AMHS on the cycle time of the production system.

In this regard, Table 4 shows the impact of changing the value of parameters of both production system and AMHS on the cycle time. All values are in percentage and, for each analysis, a base case has been defined to calculate the increase or decrease of the cycle time. In Table 4, the second column shows the base case for each analysis. The final column reports changes on the number of vehicles for each analysis. Negative values correspond to decrease of the cycle time; e.g. increasing the number of servers (c_i) decreases the cycle time. In addition, NaN means no feasible analysis possible because of system saturation (i.e., queue length goes to infinity).

Table 4 shows that the arrival rate (λ) has the highest impact on the cycle time increase, since increasing the arrival rate increases the cycle time polynomially. When analyzing the positive impact, the corresponding parameters can be ordered as follows: λ >> f_i > f_V >> c_A^2. From this analysis, it can be understood that the failure rate of the AMHS highly impact the cycle time as the third impacting parameter. The reason is that any disruption in the AMHS will interrupt the transport of products through the production system and this interruption increases the waiting time of products and consequently the cycle time of the products in the production system. When analyzing the negative impact, the corresponding parameters can be ordered as follows: V > r_V > r_i > c_i..

Therefore, it can be observed that the speed of vehicles has the highest impact on the cycle time decrease. This shows the importance of the AMHS on the production performance.

Regarding the number of vehicles, it can be seen as well that increasing λ, r_i, and f_V requires a larger number of vehicles while increasing f_i, r_V, and V decreases the number of vehicles. When λ increases, higher number of products demand to be handled by AMHS. Consequently, higher number of vehicles are required to response this demand. In addition, when r_i increases, workstations become quickly available in case of disruption; therefore, higher number of vehicles are necessary to transfer the products between workstations as well as to avoid the workstations' idleness. In a special case, when f_i increases, workstations become less available; therefore, fewer products need to be transferred and consequently less vehicles are required. Finally, When r_V and V increase, vehicles become more available and faster. Accordingly, lower number of vehicles are required to handle the products' transportation through the production system.

Table 4. Sensitivity analysis results

Parameter	Base-case	Parameter increase (%)	Cycle Time Increase (%)			No. of Vehicles
			P_1	P_2	P_3	
λ/λ_{max}	λ/λ_{max} = 0.1	300	1429	1719	1169	1
		500	4991	6611	3944	2
		700	15664	23620	11989	2
		880	49409	86813	37357	3
c_A^2	c_A^2 = 2	100	0,21	0,39	0,21	2
		300	0,64	1,18	0,65	2
		500	1,07	1,97	1,11	2
		700	1,51	2,76	1,58	2
c_i	c_i = 1	100	-78	-86	-75	2
		200	-82	-89	-79	2
		300	-83	-90	-81	2
		400	-84	-91	-82	2
f_i	f_i = 0.25	100	124	131	112	2
		300	473	528	421	2
		700	1665	2084	1460	1
r_i	r_i = 1	150	-50	-52	-48	1
		500	-75	-77	-74	2
		700	-80	-82	-79	2
f_V	f_V = 1	50	50	50	50	2
		100	112	112	112	2
		200	291	292	291	3
r_V	r_V = 2	50	NaN*	NaN	NaN	4
		150	-74	-71	-69	3
		250	-88	-82	-74	2
LT&UT	LT&UT = 20	25	61	61	61	2
		50	153	154	152	2
		100	550	551	547	3
V	V = 0.25	100	-28	-33	-27	3
		200	-69	-71	-68	3
		300	-77	-79	-77	2
		400	-81	-82	-81	2

* System saturation with no feasible solution

4.3.2. Joint impact of arrival rate and number of vehicles

In the following, comprehensive sensitivity analyses are done based on instance I20 to investigate the joint impact of the arrival rate and the number of products on the KPIs of the production system: cycle time and queue length.

As the first experiment, Figure 10 shows the cycle time of the products for different throughput rates as well as the impact of the number of vehicles on the cycle times. Figure 10 illustrates that system saturation (i.e., infinite cycle time) happens twice for throughput rates around 45% and 85%. This multiple-saturation event shows the impact of the AMHS on the production system performance. Actually, the system operates with the minimum number of vehicles and once the saturation happens, an extra vehicle is added to the AMHS. Consequently, the production system returns to a stable state with lower cycle times. The next saturation happens when two vehicles are unavailable to respond to the increase of the throughput rate. Similarly, the third vehicle is injected to the AMHS and so on. Ignoring the cost of vehicles in the AMHS, this addition can be continued in order to find the minimum number of vehicles so that the AMHS has no impact on the performance of the production system. This value helps the manufacturers to estimate the capability of their AMHS to support the production system. However, in real settings, reaching such a minimum value is expensive and there is a trade-off between the production system and the AMHS. Therefore, the proposed QAG model provides such a capacity planning opportunity for the manufacturer in a very short computational time (i.e., less than one second).

Another important point in Figure 10 is that the system saturations happen at the same throughput ranges for all products. This is because all products have the same priority and they are similarly impacted by the congestion of the production system.

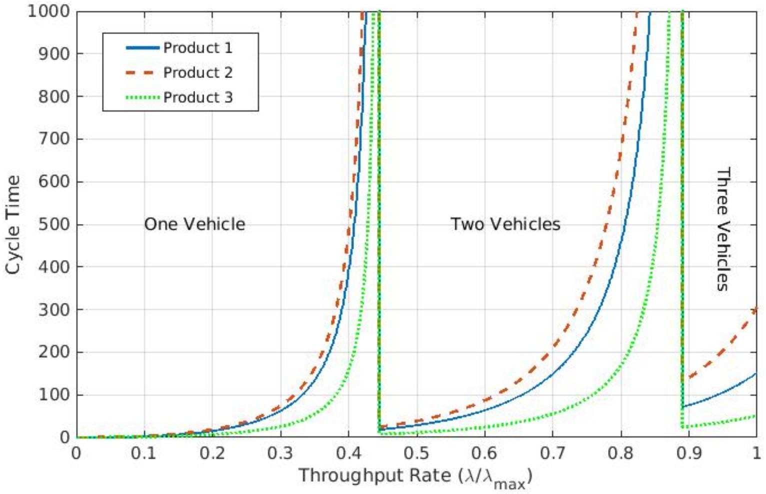

Figure 10. Impact of the number of vehicles on cycle times under different throughput rates

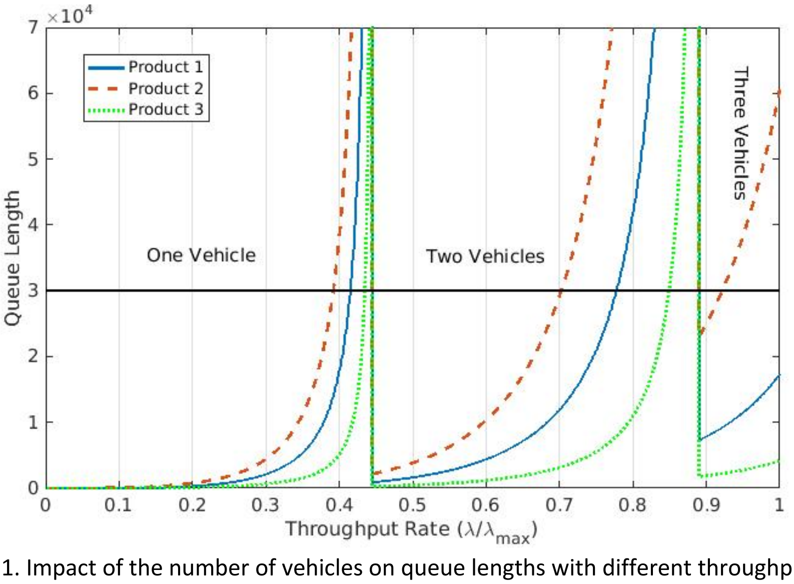

Figure 11. Impact of the number of vehicles on queue lengths with different throughput rates

Figure 11 illustrates the queue length of the products for different throughput rates as well as the impact of the number of vehicles on the queue lengths. The same interpretations than for Figure 10 can be derived for queue lengths. Figure 11 provides valuable information on capacity planning, and particularly the production storage. As an example from Figure 11, considering the maximum storage level equal to 30 000 unit of products, it can be observed that the production system can operate under three scenarios: Low throughput, medium throughput, and high throughput. These scenarios are created based on the number of vehicles as follows:

- *Low throughput scenario.* This scenario corresponds to an AMHS with one vehicle. When the production system operates with only one vehicle and wants to keep the storage below 30,000, the throughput rate of products 1 to 3 must be lower than 0.41, 0.39, and 0.43, respectively.
- *Medium throughput scenario.* This scenario corresponds to an AMHS with two vehicles. When the production system operates with only two vehicles and wants to keep the storage below 30,000, the throughput rate of products 1 to 3 must be lower than 0.77, 0.70, and 0.84, respectively.
- *High throughput scenario.* This scenario corresponds to an AMHS with three vehicles. When the production system operates with three vehicles and wants to keep the storage below 30,000, the throughput rate of product 2 must be lower than 0.92 and two other products can be processed with the maximum throughput rate. In the situation with the highest throughput rate for products 1 and 3, the queue length of these products are 17,417, and 4,168 units, respectively.

Figure 12 shows the cycle time of the products for different throughput rates as well as the impact of the number of vehicles on the cycle times when products are prioritized as follows: $P_2 > P_3 > P_1$, i.e. P_2 has the highest priority and P_1 has the lowest priority. Figure 10 shows that product 2 has the largest cycle times for all throughput rates and that product 1 has larger cycle times than product 3. However, when products are prioritized, their cycle times are significantly impacted. For instance, product 2 now has the lowest cycle times because it is the most prioritized. In Figure 12, cycle times are ordered based on the products' priority.

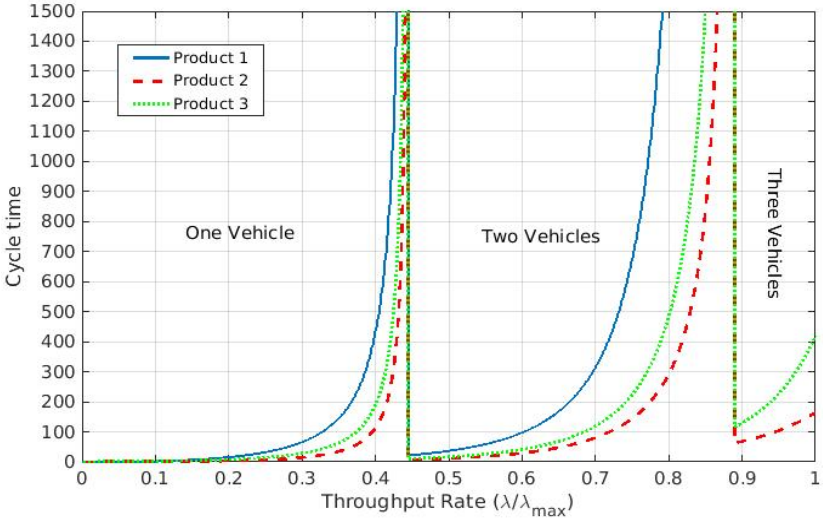

Figure 12. Impact of the number of vehicles on cycle times with different throughput rates when products have priorities

Another important result from Figure 12 relates to the saturation states of products for different number of vehicles. With two vehicles in the AMHS and no priority on products, saturation states are reached at throughput rates around 0.84, 0.82, and 0.87 for products 1, 2, and 3, respectively. However, when products are prioritized, saturation states are reached at throughput rates around 0.78, 0.87, and 0.85 for products 1, 2, and 3, respectively. The saturation state of product 1 has decreased from 0.84 to 0.78, i.e. the production system cannot reach large throughput rates for product 1 and thus assigns less capacity to this product. Consequently, product 1 must wait for products 2 and 3 with higher priority to be processed. This is the opposite for product 2 since its saturation state increases from 0.82 to 0.87, i.e., the production system assigns the maximum capacity to product 2. This capacity share significantly decreases the cycle time of product 2 which has the highest priority. In addition, for the largest throughput rates (i.e., $\lambda/\lambda_{max} > 0.87$) and when a third vehicle is added to the AMHS, the production system is no longer able to process product 1.

Figure 13 depicts the significant impact of prioritizing products on the cycle times. Products 1 and 3 are negatively impacted since their cycle times increase compared to the case with no priority between products. On the other hand, the cycle time of product 2, which has the highest priority, significantly decreases. Figure 13 also shows that the production system cannot handle the largest throughput rates of product 1 when products are prioritized.

4.3.3. Impact of products' priority on the cycle time

In another analysis, Table 5 reports the impact of imposing the priorities of products on the KPIs. Products are prioritized as follows: $P_2 > P_3 > P_1$, wherein P_2 has the highest priority and P_1 has the lowest priority. Table 5 shows that assigning the highest priority to P_2 leads to an improvement of its KPIs. On the contrary, the KPIs of P_1 and P_3 degrade since they have to wait whenever a new unit of P_2 arrives. Therefore, P_1 and P_3 are congested in the system and their KPIs degrade. NaN values refer to the saturation of the system for P_1 that has the lowest priority. This means that the queue length of P_1 goes to infinity. This happens in instances (i.e., I4, I8, I12, I16, and I20) with the highest throughput rate (i.e., 98%). Accordingly, prioritizing the products may make the production system unstable in case of high throughput rates.

Table 5. Impact of imposing the priorities of products

Instance	Cycle Time Increase (%)			Waiting for time Increase (%)			Queue Length Increase (%)		
	P_1	P_2	P_3	P_1	P_2	P_3	P_1	P_2	P_3
I1	17	-75	32	18	-77	34	17	-75	32
I2	41	-68	57	42	-69	61	41	-68	57
I3	181	-59	155	186	-60	164	181	-59	155
I4	**NaN**[*]	-47	640	**NaN**	-48	659	**NaN**	-47	640
I5	12	-77	21	12	-78	23	12	-77	21
I6	46	-67	65	47	-68	70	46	-67	65
I7	165	-59	137	170	-60	145	165	-59	137
I8	**NaN**	-47	555	**NaN**	-48	571	**NaN**	-47	555
I9	12	-77	22	13	-78	24	12	-77	22
I10	48	-67	69	50	-68	74	48	-67	69
I11	171	-59	143	175	-60	152	171	-59	143
I12	**NaN**	-47	602	**NaN**	-48	620	**NaN**	-47	602
I13	12	-77	23	13	-78	24	12	-77	23
I14	50	-67	71	51	-68	76	50	-67	71
I15	174	-59	146	179	-60	155	174	-59	146
I16	**NaN**	-47	621	**NaN**	-48	639	**NaN**	-47	621
I17	13	-77	23	13	-78	25	13	-77	23
I18	40	-68	55	41	-69	59	40	-68	55
I19	177	-59	149	181	-60	158	177	-59	149
I20	**NaN**	-47	631	**NaN**	-48	650	**NaN**	-47	631

[*] System saturation with no feasible solution

In another experiment, Figure 14 illustrates the impact of the number of vehicles on the products' cycle time without and with priorities on products. These results are based on instance I20 where the production system operates at 80% of its capacity (i.e., $\lambda/\lambda_{max} = 0.80$). The upper-left part of Figure 14 shows that increasing the number of vehicles to 4 helps to decrease the cycle time of all the products, and there is no significant change with more than 4 vehicles. This shows that the proposed QAG model supports manufacturers in determining the required number of vehicles for the AMHS not to impact the production system's performance. Other figures in Figure 14 show the impact of the number of vehicles on the cycle time of each product without and with priorities on products.

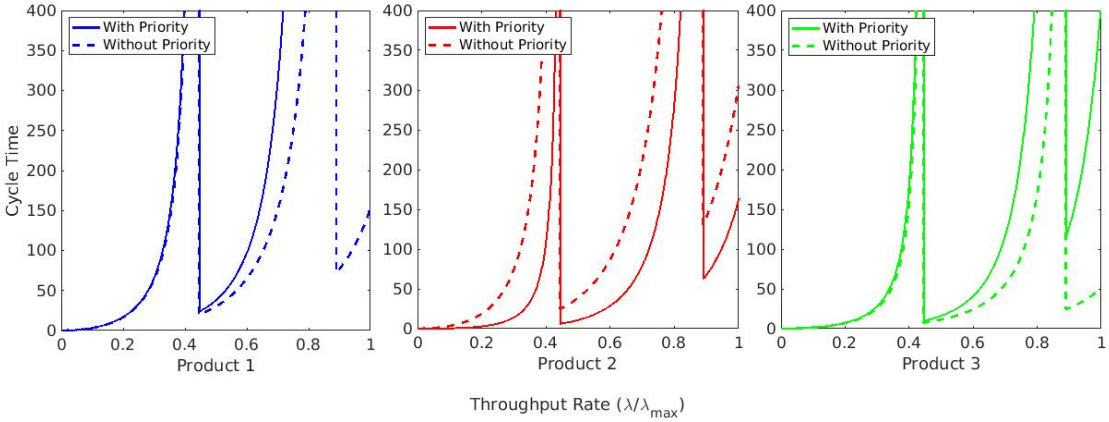

Figure 13. Impact of the priorities of products on cycle times

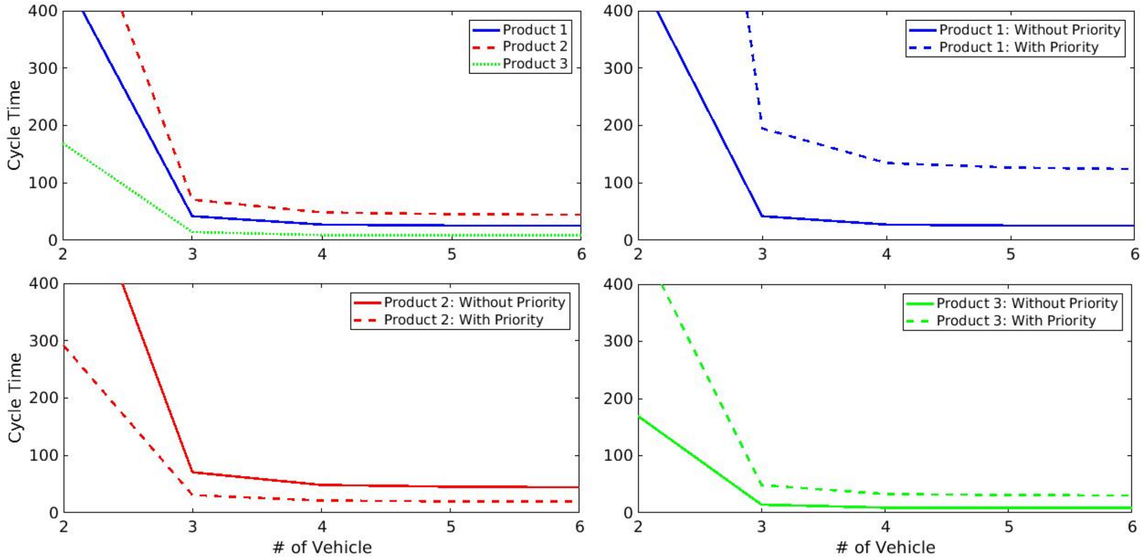

Figure 14. Number of vehicles vs. Product cycle time: Without and with priorities on products

4.4. Capacity planning of the AMHS

This section evaluates the performance of the AMHS to keep the troublesome workstation (*TW*) and the fastest workstation (*FW*) from being idle. Accordingly, two policies are considered: **Policy 1:** “Preventing *FW* from being idle” and **Policy 2:** “Preventing *TW* from being idle”. For such a capacity planning, instance I20 is considered as the base case where the production system operates at 85% of its capacity (i.e., λ/λ_{max} = 0.85). The service level (target probability) α is set to 95%, i.e. in 95% of the time, *TW* and *FW* are operating (i.e., they become idle less than 5% of the time) and at least one unit of product is waiting to be processed.

Table 6 reports the service levels for both *TW* and *FW* with the number of vehicles in the AMHS. In the case study of Figure 8, workstations 6 and 1 become *TW* and *FW*, respectively. Table 6 shows that, with Policy 1 and a service level of α = 95%, at least 5 vehicles are required in the AMHS. On the other hand, Policy 2 requires at least 3 vehicles in the AMHS. The difference between the numbers of vehicles between two policies is due to the fact that the *FW* releases the products quickly and more products are waiting to be transported to the next workstation(s) and consequently higher demand of vehicles. In opposite, the TW is lowest workstation in terms of service rate and consequently it is busy most of the time. Therefore, no hurry to feed the *TW*. Accordingly, the TW can be busy for 95% of the time with lower number of vehicles. Although Policy 1 requires more vehicles than Policy 2, the proposed QAG model provides analytical information on the capacity planning of the AMHS.

These number of vehicles, 5 vehicles and 3 vehicles, are the minimum number of vehicles to keep the FW and TW busy, respectively. Obviously, increasing the number of vehicles increases the service level up to a level that both FW and TW are always busy (i.e., service rate 100%). The 100% of service rate happens when operating 6 and 4 vehicles for the FW and TW, respectively. Increasing the number of vehicles from these values no more affects the service rate and the vehicles become idle for a percent of time.

Table 6. Capacity planning of the AMHS

# of Vehicles	Service Level	
	Fastest Workstation (*FW*)	Troublesome Workstation (*TW*)
1	0.38	0.65
2	0.57	0.87
3	0.80	**0.97**
4	0.92	1.00
5	**0.98**	1.00
6	1.00	1.00
7	1.00	1.00
8	1.00	1.00

5. Conclusion

Processing a wide range of products to answer dynamic customers' requirements has become one of the greatest challenges in almost all manufacturing systems. This makes production planning of such manufacturing systems more and more important and complex to improve the performance of factories. The main performance indicators to evaluate a production system are the cycle time and the throughput. Accordingly, an accurate estimation of the cycle time as a function of the throughput is critical. The material handling system often significantly impacts the performance of the overall system, in particular the cycle time. Today's manufacturing systems are very complex and include many processing and handling operations that make it impossible to evaluate the components one by one using simple analytical methods. Accordingly, a novel approach is required to accurately aggregate the production system's components into a simpler configuration, and then evaluate the aggregated system.

In this paper, we proposed a novel aggregation model based on a queueing network approach, so-called queue-based aggregation (QAG) model, to estimate the cycle time of a job-shop production system that consists of several processing workstations and in which products are transferred via an Automated Material Handling System (AMHS). In this model, the whole job-shop system is aggregated into a single-step workstation. The parameters of the single-step aggregated workstation are approximated based on the parameters of the original processing workstations and the AMHS. The proposed model aggregates both production and handling systems and provides an accurate and fast estimation of the overall cycle time. The performance and superiority of the proposed model are validated by comparing its results with those of a detailed discrete-event simulation model, but the utilization of the proposed QAG model with throughput rates higher than 80% is not recommended. However, it depends on the tolerance of the experts to the accuracy of the cycle time approximation.

Through a comprehensive sensitivity analysis, it is observed that, among various parameters, increasing the arrival rates as well as the failure rates of the processing workstation and the AMHS significantly increases the cycle time. On the other hand, increasing the speed of the vehicles in the AMHS, as well as the retrieving rates of the processing workstation and of the AMHS, helps to significantly decrease the cycle time. In another experiment, it is verified that the AMHS with a limited number of vehicles has a noticeable impact on the cycle time. This confirms that it is essential to consider the MHS when evaluating the performance of a production system. In addition, an experiment is conducted to analyze the capacity of the AMHS system by defining a service level for the fastest and the troublesome workstations.

The performance of the production system is also analyzed when the products are prioritized. In this case, a product must wait when a product with higher priority arrives and needs to be processed in the corresponding workstation. It is observed that the product's priority has a significant impact on the cycle time and the capacity of the production system to handle the products with lower priorities. In an extreme case with high throughput rates, the product with the lowest priority could no longer be processed.

We believe the following future research directions are relevant to investigate:

- In this paper, the storage capacity at the workstations is limited. An interesting future research direction could be to model the processing workstations with a finite buffer by using GI/G/*c*/K queuing models. In this case, a product arriving in the full buffer is ignored and leaves the system. What happens to the departing products should be defined.
- Another research direction consists in considering production systems with re-entrant/re-work operations. In this case, at each workstation, a given percentage of products should re-do some of their operations at already visited workstations. The proposed QAG model can be easily adapted for this case by accumulating the arrival rate of products at each workstation that is the sum of: 1) The arrival rate of

products that arrive for the first time to the workstation and 2) The arrival rate of products from downstream workstations that require re-working.

- Considering the lead time of jobs to account for the number of tardy jobs in the production system, and analyzing how the AMHS and the throughput rate impact this number, also constitutes an interesting research avenue.
- The proposed QAG model does not actually schedule the arriving jobs on the workstation, but only analyzes and evaluates the congestion at workstations. The dispatching rule in the QAG model (without priority) is FIFO. However, many production systems also require setup times and even sequence-dependent setup times. The setup times of different jobs can be integrated in the QAG model as an expected value depending on the arrival rate of each product to workstations. If we consider the setup time of product *p* on workstation *i* as st_{pi}, the expected setup time of the aggregated product at workstation *i*, $st_{\mathbb{P}i}$, would be $st_{\mathbb{P}i} = \sum_p \pi_p st_{pi}$. This value can be added to the processing time of the products at workstations. Studying this new property is also relevant.

References

[1] A. A. Taleizadeh, M. Karimi Mamaghan, and S. A. Torabi, “A possibilistic closed-loop supply chain: pricing, advertising and remanufacturing optimization,” *Neural Comput. Appl.*, Aug. 2018.

[2] A. Etienne *et al.*, “Cost engineering for variation management during the product and process development,” *Int. J. Interact. Des. Manuf. IJIDeM*, vol. 11, no. 2, pp. 289–300, May 2017.

[3] D. P. Martin, “Capacity and cycle time-throughput understanding system (CAC-TUS) an analysis tool to determine the components of capacity and cycle time in a semiconductor manufacturing line,” in *10th Annual IEEE/SEMI. Advanced Semiconductor Manufacturing Conference and Workshop. ASMC 99 Proceedings (Cat. No.99CH36295)*, 1999, pp. 127–131.

[4] J. R. Morrison and D. P. Martin, “Cycle Time Approximations for the G/G/m Queue Subject to Server Failures and Cycle Time Offsets with Applications,” in *The 17th Annual SEMI/IEEE ASMC 2006 Conference*, 2006, pp. 322–326.

[5] A. A. A. Kock, L. F. P. Etman, and J. E. Rooda, “Effective process times for multi-server flowlines with finite buffers,” *IIE Trans.*, vol. 40, no. 3, pp. 177–186, Jan. 2008.

[6] C. P. L. Veeger, L. F. P. Etman, E. Lefeber, I. J. B. F. Adan, J. van Herk, and J. E. Rooda, “Predicting Cycle Time Distributions for Integrated Processing Workstations: An Aggregate Modeling Approach,” *IEEE Trans. Semicond. Manuf.*, vol. 24, no. 2, pp. 223–236, May 2011.

[7] M. Manitz, “Analysis of assembly/disassembly queueing networks with blocking after service and general service times,” *Ann. Oper. Res.*, vol. 226, no. 1, pp. 417–441, Mar. 2015.

[8] P. Yan, S. Q. Liu, T. Sun, and K. Ma, “A dynamic scheduling approach for optimizing the material handling operations in a robotic cell,” *Comput. Oper. Res.*, vol. 99, pp. 166–177, Nov. 2018.

[9] A. Klausnitzer and R. Lasch, “Optimal facility layout and material handling network design,” *Comput. Oper. Res.*, Nov. 2018.

[10] D. Raman, S. V. Nagalingam, B. W. Gurd, and G. C. I. Lin, “Quantity of material handling equipment—A queuing theory based approach,” *Robot. Comput.-Integr. Manuf.*, vol. 25, no. 2, pp. 348–357, Apr. 2009.

[11] J. A. Tompkins, J. A. White, Y. A. Bozer, and J. M. A. Tanchoco, *Facilities Planning*. John Wiley & Sons, 2010.

[12] J. T. Lin, C.-H. Wu, and C.-W. Huang, “Dynamic vehicle allocation control for automated material handling system in semiconductor manufacturing,” *Comput. Oper. Res.*, vol. 40, no. 10, pp. 2329–2339, Oct. 2013.

[13] J. R. Morrison and D. P. Martin, “Performance evaluation of photolithography cluster tools,” *Spectr.*, vol. 29, no. 3, pp. 375–389, Jul. 2007.

[14] J. R. Morrison and D. P. Martin, “Practical Extensions to Cycle Time Approximations for the G/G/m-Queue With Applications,” *IEEE Trans. Autom. Sci. Eng.*, vol. 4, no. 4, pp. 523–532, Oct. 2007.

[15] B. Vahdani and M. Mohammadi, “A bi-objective interval-stochastic robust optimization model for designing closed loop supply chain network with multi-priority queuing system,” *Int. J. Prod. Econ.*, vol. 170, pp. 67–87, Dec. 2015.

[16] Y. Dallery, R. David, and X.- Xie, “Approximate analysis of transfer lines with unreliable machines and finite buffers,” *IEEE Trans. Autom. Control*, vol. 34, no. 9, pp. 943–953, Sep. 1989.

[17] Y. Dallery and Y. Frein, “On Decomposition Methods for Tandem Queueing Networks with Blocking,” *Oper. Res.*, vol. 41, no. 2, pp. 386–399, Apr. 1993.

[18] M. K. Govil and M. C. Fu, “Queueing theory in manufacturing: A survey,” *J. Manuf. Syst.*, vol. 18, no. 3, pp. 214–240, Jan. 1999.

[19] H. TEMPELMEIER and M. BÜRGER, “Performance evaluation of unbalanced flow lines with general distributed processing times, failures and imperfect production,” *IIE Trans.*, vol. 33, no. 4, pp. 293–302, Apr. 2001.

[20] N. Guerouahane, D. Aissani, N. Farhi, and L. Bouallouche-Medjkoune, “M/G/c/c state dependent queuing model for a road traffic system of two sections in tandem,” *Comput. Oper. Res.*, vol. 87, pp. 98–106, Nov. 2017.

[21] K. R. Gue and H. H. Kim, “An approximation model for sojourn time distributions in acyclic multi-server queueing networks,” *Comput. Oper. Res.*, vol. 63, pp. 46–55, Nov. 2015.

[22] Y. Rahimi, R. Tavakkoli-Moghaddam, M. Mohammadi, and M. Sadeghi, “Multi-objective hub network design under uncertainty considering congestion: An M/M/c/K queue system,” *Appl. Math. Model.*, vol. 40, no. 5, pp. 4179–4198, Mar. 2016.

[23] M. Mohammadi, R. Tavakkoli-Moghaddam, A. Siadat, and Y. Rahimi, “A game-based meta-heuristic for a fuzzy bi-objective reliable hub location problem,” *Eng. Appl. Artif. Intell.*, vol. 50, pp. 1–19, Apr. 2016.

[24] M. Mohammadi, P. Jula, and R. Tavakkoli-Moghaddam, “Design of a reliable multi-modal multi-commodity model for hazardous materials transportation under uncertainty,” *Eur. J. Oper. Res.*, vol. 257, no. 3, pp. 792–809, Mar. 2017.

[25] M. Mohammadi, S. Dehbari, and B. Vahdani, “Design of a bi-objective reliable healthcare network with finite capacity queue under service covering uncertainty,” *Transp. Res. Part E Logist. Transp. Rev.*, vol. 72, pp. 15–41, Dec. 2014.

[26] M. Mohammadi, F. Jolai, and H. Rostami, “An M/M/c queue model for hub covering location problem,” *Math. Comput. Model.*, vol. 54, no. 11, pp. 2623–2638, Dec. 2011.

[27] M. Mohammadi, S. A. Torabi, and R. Tavakkoli-Moghaddam, “Sustainable hub location under mixed uncertainty,” *Transp. Res. Part E Logist. Transp. Rev.*, vol. 62, pp. 89–115, Feb. 2014.

[28] A. Arisha and P. Young, “Intelligent simulation-based lot scheduling of photolithography toolsets in a wafer fabrication facility,” in *Proceedings of the 2004 Winter Simulation Conference, 2004.*, 2004, vol. 2, pp. 1935–1942 vol.2.

[29] N. Nayani and M. Mollaghasemi, “Validation and verification of the simulation model of a photolithography process in semiconductor manufacturing,” in *1998 Winter Simulation Conference. Proceedings (Cat. No.98CH36274)*, 1998, vol. 2, pp. 1017–1022 vol.2.

[30] N. G. Pierce and M. J. Drevna, “Development of Generic Simulation Models to Evaluate Wafer Fabrication Cluster Tools,” in *Proceedings of the 24th Conference on Winter Simulation*, New York, NY, USA, 1992, pp. 874–878.

[31] M. Calle, P. L. González-R, J. M. Leon, H. Pierreval, and D. Canca, “Integrated management of inventory and production systems based on floating decoupling point and real-time information: A simulation based analysis,” *Int. J. Prod. Econ.*, vol. 181, pp. 48–57, Nov. 2016.

[32] M. Thürer, M. Stevenson, C. Silva, and T. Qu, “Drum-buffer-rope and workload control in High-variety flow and job shops with bottlenecks: An assessment by simulation,” *Int. J. Prod. Econ.*, vol. 188, pp. 116–127, Jun. 2017.

[33] D. Nazzal, “A closed queueing network approach to analyzing multi-vehicle material handling systems,” *IIE Trans.*, vol. 43, no. 10, pp. 721–738, Oct. 2011.

[34] F. Yang, B. E. Ankenman, and B. L. Nelson, “Estimating Cycle Time Percentile Curves for Manufacturing Systems via Simulation,” *Inf. J. Comput.*, vol. 20, no. 4, pp. 628–643, Jul. 2008.

[35] E. J. Chen, “Metamodels for Estimating Quantiles of Systems with One Controllable Parameter,” *SIMULATION*, vol. 85, no. 5, pp. 307–317, May 2009.

[36] R. T. Johnson, J. W. Fowler, and G. T. Mackulak, “A Discrete Event Simulation Model Simplification Technique,” in *Proceedings of the 37th Conference on Winter Simulation*, Orlando, Florida, 2005, pp. 2172–2176.

[37] O. Rose, “Why do simple wafer fab models fail in certain scenarios?,” in *2000 Winter Simulation Conference Proceedings (Cat. No.00CH37165)*, 2000, vol. 2, pp. 1481–1490 vol.2.

[38] J. H. Jacobs, L. F. P. Etman, J. E. Rooda, and E. J. J. V. Campen, “Quantifying operational time variability: the missing parameter for cycle time reduction,” in *2001 IEEE/SEMI Advanced Semiconductor Manufacturing Conference (IEEE Cat. No.01CH37160)*, 2001, pp. 1–10.

[39] R. J. Brooks and A. M. Tobias, “Simplification in the simulation of manufacturing systems,” *Int. J. Prod. Res.*, vol. 38, no. 5, pp. 1009–1027, Mar. 2000.

[40] W. J. Hopp and M. L. Spearman, *Factory Physics: Foundations of Manufacturing Management*. Irwin Professional Publishing, 1996.

[41] W. J. Hopp and M. L. Spearman, *Factory Physics: Third Edition*. Waveland Press, 2011.

[42] J. H. Jacobs, L. F. P. Etman, E. J. J. van Campen, and J. E. Rooda, “Characterization of operational time variability using effective process times,” *IEEE Trans. Semicond. Manuf.*, vol. 16, no. 3, pp. 511–520, Aug. 2003.

[43] J. R. Morrison, “Equipment models for fab level proudction simulation: Practical features and computational tractability,” in *Proceedings of the 2009 Winter Simulation Conference (WSC)*, 2009, pp. 1581–1591.

[44] H. S. Lee, A. Bouhchouch, Y. Dallery, and Y. Frein, “Performance evaluation of open queueing networks with arbitrary configuration and finite buffers,” *Ann. Oper. Res.*, vol. 79, no. 0, pp. 181–206, Jan. 1998.

[45] K. Satyam and A. Krishnamurthy, “Performance evaluation of a multi-product system under CONWIP control,” *IIE Trans.*, vol. 40, no. 3, pp. 252–264, Jan. 2008.

[46] K. Wu and L. McGinnis, “Interpolation approximations for queues in series,” *IIE Trans.*, vol. 45, no. 3, pp. 273–290, Mar. 2013.

[47] K. Satyam, A. Krishnamurthy, and M. Kamath, “Solving general multi-class closed queuing networks using parametric decomposition,” *Comput. Oper. Res.*, vol. 40, no. 7, pp. 1777–1789, Jul. 2013.

[48] R. Schelasin, “Using static capacity modeling and queuing theory equations to predict factory cycle time performance in semiconductor manufacturing,” in *Proceedings of the 2011 Winter Simulation Conference (WSC)*, 2011, pp. 2040–2049.

[49] H. P. Hillion and J.- Proth, “Performance evaluation of job-shop systems using timed event-graphs,” *IEEE Trans. Autom. Control*, vol. 34, no. 1, pp. 3–9, Jan. 1989.

[50] R. Suri, J. L. Sanders, and M. Kamath, “Chapter 5 Performance evaluation of production networks,” in *Handbooks in Operations Research and Management Science*, vol. 4, Elsevier, 1993, pp. 199–286.

[51] C. R. N. da Silva and R. Morabito, “Performance evaluation and capacity planning in a metallurgical job-shop system using open queueing network models,” *Int. J. Prod. Res.*, vol. 47, no. 23, pp. 6589–6609, Dec. 2009.

[52] M. R. Abdi and A. W. Labib, "Performance evaluation of reconfigurable manufacturing systems via holonic architecture and the analytic network process," *Int. J. Prod. Res.*, vol. 49, no. 5, pp. 1319–1335, Mar. 2011.

[53] M. Mohammadi, S. Dauzère-Pérès, and C. Yugma, "Performance evaluation of single and multi-class production systems using an approximating queuing network," *Int. J. Prod. Res.*, vol. 0, no. 0, pp. 1–27, Jul. 2018.

[54] P. Bedell and J. M. Smith, "Topological arrangements of M/G/c/K, M/G/c/c queues in transportation and material handling systems," *Comput. Oper. Res.*, vol. 39, no. 11, pp. 2800–2819, Nov. 2012.

[55] F. Benson and G. Gregory, "Closed Queueing Systems: A Generalization of the Machine Interference Model," *J. R. Stat. Soc. Ser. B Methodol.*, vol. 23, no. 2, pp. 385–393, 1961.

[56] M. Posner and B. Bernholtz, "Two stage closed queueing systems with time lags," *CORS J. J. Can. Oper. Res. Soc.*, vol. 5, no. 2, 1967.

[57] S. Benjaafar, "Modeling and Analysis of Congestion in the Design of Facility Layouts," *Manag. Sci.*, vol. 48, no. 5, pp. 679–704, May 2002.

[58] E. Koenigsberg, "Twenty Five Years of Cyclic Queues and Closed Queue Networks: A Review," *J. Oper. Res. Soc.*, vol. 33, no. 7, pp. 605–619, Jul. 1982.

[59] D. Nazzal and L. F. McGinnis, "Analytical approach to estimating AMHS performance in 300 mm fabs," *Int. J. Prod. Res.*, vol. 45, no. 3, pp. 571–590, Feb. 2007.

[60] Y.-M. Tu, C.-W. L. P. D, and A. H. I. Lee, "AMHS capacity determination model for wafer fabrication based on production performance optimization," *Int. J. Prod. Res.*, vol. 51, no. 18, pp. 5520–5535, Sep. 2013.

[61] M. E. Johnson, "Modelling empty vehicle traffic in AGVS design," *Int. J. Prod. Res.*, vol. 39, no. 12, pp. 2615–2633, Jan. 2001.

[62] R. Azizmohammadi, M. Amiri, R. Tavakkoli-Moghaddam, and M. Mohammadi, "Solving a Redundancy Allocation Problem by a Hybrid Multi-objective Imperialist Competitive Algorithm," *Int. J. Eng. - Trans. C Asp.*, vol. 26, no. 9, pp. 1031–1042, May 2013.

[63] M. Mohammadi, P. Jula, and R. Tavakkoli-Moghaddam, "Reliable single-allocation hub location problem with disruptions," *Transp. Res. Part E Logist. Transp. Rev.*, vol. 123, pp. 90–120, Mar. 2019.

[64] C. P. L. Veeger, L. F. P. Etman, J. van Herk, and J. E. Rooda, "Generating Cycle Time-Throughput Curves Using Effective Process Time Based Aggregate Modeling," *IEEE Trans. Semicond. Manuf.*, vol. 23, no. 4, pp. 517–526, Nov. 2010.

[65] W. Whitt, "The Queueing Network Analyzer," *Bell Syst. Tech. J.*, vol. 62, no. 9, pp. 2779–2815, Nov. 1983.

[66] G. R. Bitran and D. Tirupati, "Multiproduct Queueing Networks with Deterministic Routing: Decomposition Approach and the Notion of Interference," *Manag. Sci.*, vol. 34, no. 1, pp. 75–100, Jan. 1988.

[67] M. Zhalechian, S. A. Torabi, and M. Mohammadi, "Hub-and-spoke network design under operational and disruption risks," *Transp. Res. Part E Logist. Transp. Rev.*, vol. 109, pp. 20–43, Jan. 2018.

[68] D. Delp *, J. Si, Y. Hwang, B. Pei, and J. Fowler, "Availability-adjusted X-factor," *Int. J. Prod. Res.*, vol. 43, no. 18, pp. 3933–3953, Sep. 2005.

[69] D. Delp, J. Si, and J. W. Fowler, "The development of the complete X-factor contribution measurement for improving cycle time and cycle time variability," *IEEE Trans. Semicond. Manuf.*, vol. 19, no. 3, pp. 352–362, Aug. 2006.

[70] W. Whitt, "APPROXIMATIONS FOR THE GI/G/m QUEUE," *Prod. Oper. Manag.*, vol. 2, no. 2, pp. 114–161, Jun. 1993.

[71] J. Köllerström, "Heavy traffic theory for queues with several servers. I," *J. Appl. Probab.*, vol. 11, no. 3, pp. 544–552, Sep. 1974.

[72] A. H. I. Lee, T.-H. Huang, and H.-Y. Kang, "A priority mix planning model for semiconductor fabrication under an uncertain information environment," *J. Inf. Optim. Sci.*, vol. 29, no. 2, pp. 377–400, Mar. 2008.

[73] G. Bolch, S. Greiner, H. de Meer, and K. S. Trivedi, *Queueing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications*. John Wiley & Sons, 2006.