International Journal of Production Research

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tprs20

Application of a novel approach of production system modelling, analysis and improvement for small and medium-sized manufacturers: a case study

Yuting Sun & Liang Zhang

To cite this article: Yuting Sun & Liang Zhang (2022): Application of a novel approach of production system modelling, analysis and improvement for small and medium-sized manufacturers: a case study, International Journal of Production Research, DOI: 10.1080/00207543.2022.2079015

To link to this article: https://doi.org/10.1080/00207543.2022.2079015

Published online: 01 Jun 2022.

Submit your article to this journal

Article views: 153

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=tprs20

INTERNATIONAL JOURNAL OF PRODUCTION RESEARCH
https://doi.org/10.1080/00207543.2022.2079015

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

Check for updates

Application of a novel approach of production system modelling, analysis and improvement for small and medium-sized manufacturers: a case study

Yuting Sun and Liang Zhang

Department of Electrical and Computer Engineering, University of Connecticut, Storrs, CT, USA

ABSTRACT

With the great opportunities created by the new advances in Industry 4.0, many manufacturers are testing and investing in new equipment and infrastructure to deploy these technologies. However, there are a huge number of small and medium-sized manufacturers (SMMs) that are lagging behind due to the lack of in-house R&D capabilities and workforce shortage and/or financial constraints to afford such investment. Additionally, application of theoretical production research in SMMs often confront challenges such as low data availability and data quality, etc. In this paper, we describe a case study at a local medium-sized manufacturer of electromechanical devices for industrial, consumer, and medical applications, who was struggling to meet ever-growing market demand, and apply a novel approach of production system modelling to overcome the challenge of unavailability of the operation up- and downtime data. Specifically, the parametric model of the production system is identified using several system performance metrics derived based on the parts flow data of the in-process buffer. With the mathematical model constructed, the system bottleneck is analysed and a number of improvement scenarios are explored that can potentially enhance the system throughput. Finally, model sensitivity is analysed by calculating the deviation of the model-predicted performance metrics to those produced by a reference nominal model. This analysis demonstrates that the model constructed using our proposed approach is robust even when the system parameters vary from the baseline ones.

ARTICLE HISTORY
Received 1 December 2021
Accepted 11 May 2022

KEYWORDS
Production system; exponential machine; parameter identification; modelling; smart manufacturing

1. Introduction

According to the data gathered by SCORE, 98.6% of American manufacturing companies are small businesses, and 75.3% of those businesses have fewer than 20 employees due 2019 (Weston 2019). While the rapid development of Industry 4.0 technologies are creating vast opportunities, small and medium-sized manufacturers (SMMs) are facing various challenges in adapting to this transformation. One of the areas of concern is the adoption of advanced data analytics in improving the operating efficiency of their production systems. Indeed, although the popular machine learning/AI techniques have made great strides in tasks such as pattern recognition, natural language processing, computer vision, etc., its reliance on pre-labelled training/testing data has greatly impeded its way into system-level production analysis, coordination and control, especially for SMMs. On the other hand, theoretical studies in production systems have accumulated a great amount of results and most of the methods developed are based on mathematical models of manufacturing processes (see, for instance, Li and Meerkov 2009; Gershwin 1994; Yao 1994; Papadopoulos, Heavy, and Browne 1993). In manufacturing practice, however, building a high-fidelity model of a real manufacturing process is not trivial, which usually requires solid training, extensive experience, sharp intuition, and a large amount of time (Li and Meerkov 2009; Sun et al. 2020).

Generally speaking, the first step of production system modelling is to transform the overall system layout to a standard topological/structural model (Cox 1990; Li and Meerkov 2009). This step is typically straightforward. The next step is to identify the parameters of the machines and buffers in the model using factory floor production data. A commonly used approach is to collect the operating status data (i.e. up- and downtimes) for each operation (either manually or using automatic data collection mechanisms such as PLCs). In certain situations where a work cell contains multiple operations, algorithms must be devised to calculate the aggregated parameters of the work cell as a whole in order to feed the theoretical system model used for analysis. In practice,

CONTACT Yuting Sun yuting.2.sun@uconn.edu

© 2022 Informa UK Limited, trading as Taylor & Francis Group

this conventional approach (collecting operation status data) typically faces the following challenges or limitations, especially to SMMs:

- *Data unavailability*: In some cases, operating status data are not readily available, which is very common in SMMs. One of the main contributing factors is the presence of a high number of manual operations involved. Many SMMs do not have automatic data collection modules or equipment that connect with machines to record the machine status in real-time. Collecting up- and downtime data for such systems usually imposes additional burden on the workforce (e.g. via manual time study), which the manufacturers usually cannot afford.
- *Data complexity*: In manufacturing facilities, where IT infrastructure is available for automated collection of machine operating status data, it is very common that the operating data of different equipment are programmed using different operation/failure codes. For example, a failure mode for Equipment A may not be applicable to other equipment. As a result, one must learn and decode all operating/failure modes of various types of equipment in order to extract the machine up- and downtime data correctly and feed them into a production system model.
- *Data quality*: For either automatically or manually collected data, errors and invalid entries are often inevitable due to various sources of noise or disruption in an industrial environment. As a result, the raw data usually cannot be directly used to feed the mathematical model. A critical step is perform data cleaning to improve the quality of the data and ensure data integrity. This process (e.g. cross-checking data validity) may be very complex and time-consuming, especially for operation up- and downtime data.

These challenges, along with the lack of in-house expertise, usually make it difficult for SMMs to effectively apply the conventional approach to carry out modelling, analysis, and control for the production systems in their practice. To overcome these challenges, a new approach is proposed in our prior work (Sun et al. 2020; Sun and Zhang 2020), based on which one can identify the parameters of a production system based on standard performance metrics derived from parts flow data. To our best knowledge, this is the first work of reversely calculating machine parameters based on the analytical expressions of performance metrics in production systems research. The advantages of this new approach include:

- *Standard performance metrics data*: This approach uses data with commonly accepted definition, such as throughput (average number of parts produced per time unit), work-in-process (average number of parts in a buffer). This greatly reduces the ambiguity that may be contained in operating status data and makes it easier for the approach to be generalised to different manufacturing facilities/industries (without having to learn the complex operation/failure modes of equipment in a new facility).
- *Convenience for automatic data collection*: It should be noted that the performance metrics data used in this approach are measured based on part-counting. This can be accomplished by deploying sensors, such as weight sensors, photoelectric sensors, cameras (e.g. on a smart phone), into the manufacturing process.

With these advantages above, we apply this new modelling approach to a case study in a local SMM, which designs and manufactures electromechanical devices for industrial, consumer, and medical applications. Based on the mathematical model identified by the proposed approach, improvement of system operations to enhance the throughput is discussed and the model validity/sensitivity is demonstrated by numerical experiments.

The rest of the paper is organised as follows: Section 2 reviews typical approaches for mathematical modelling of production systems as well as industrial case studies reported in the literature. Section 3 overviews the background of the case study, the system studied, and the problems addressed in this paper and lay out the challenges and the approach to be used. Mathematical modelling, including structural and parametric modelling, is carried out in Section 4, while the identification of the system parameters is studied in Section 5. Section 7 analyses the baseline performance of the system and investigates several improvement plans that can increase the system throughput. The sensitivity of the model is discussed in Section 8. Finally, the conclusions and future work are summarised in Section 9.

2. Literature review

2.1. Mathematical models for production systems analysis

In manufacturing systems research, a production system is typically modelled as a stochastic process, where the operation of the machines are characterised by randomly distributed uptimes, downtimes, and/or cycle times, and the buffers are defined by their storing capacity (see Papadopoulos, Heavy, and Browne 1993; Gershwin 1994; Yao 1994; Li and Meerkov 2009). Note that these models are universally applicable to production systems in both SMMs and large manufacturers with the

difference typically being the detailed model assumptions used.

The commonly used mathematical models for characterising such random behaviour of production operations include the Bernoulli reliability model, geometric reliability model, and exponential reliability model. Under the Bernoulli reliability model, the machine status (up or down) is modelled as Bernoulli random variables, while the geometric reliability model formulates the up- and downtime of a machine as geometric random variables. Production system models with Bernoulli and/or geometric reliability machines are characterised by discrete-time Markov chains. Similarly, the exponential reliability model formulates the up- and downtime of a machine as exponential random variables and production system models with exponential reliability machines are characterised by continuous-time Markov chains.

Using Markovian analysis and an iterative aggregation-based analytical approach, various theoretical problems have been studied for production system models with Bernoulli, geometric, and exponential machines. Representative results include performance metrics calculation (see Jia et al. 2015; Ju, Li, and Deng 2016; Feng et al. 2018; Jia and Zhang 2019; Bai et al. 2021), lean buffering design (see Chiang, Hu, and Meerkov 2008), bottleneck identification and continuous improvement (see Biller et al. 2008, 2009; Xie and Li 2012; Li 2013; Tu and Zhang 2022), multi-job production (see Zhao and Li 2013; Zhao, Li, and Huang 2014; Zhao and Li 2015; Alavian, Denno, and Meerkov 2017), preventive maintenance (see Ambani, Meerkov, and Zhang 2010; X. Liu, Wang, and Peng 2015; Y. Liu et al. 2021), production control (see Zhang and Yue 2011; Chen et al. 2012; Biller, Meerkov, and Yan 2013; Jia et al. 2016; Wang and Ju 2021), etc. These production system models have also been successfully applied to numerous industrial case studies (see below).

2.2. Conventional approach for parameter identification when modelling a production system

In theoretical studies of production systems, the machine reliability parameters (i.e. machine average up- and downtime, efficiency) are usually assumed to be known and/or randomly generated in statistical experiments (for instance, Y. Liu, Li, and Chiang 2010; Yan and Zhao 2013; Meerkov and Yan 2014). In industrial case studies, these parameters (and other parameters of the theoretical model adopted) are usually identified from factory floor measurements of the equipment operating status directly.

For example, in a case study at an automotive paint shop reported by Arinez et al. (2009), the production system is modelled as a serial production line with rework. The Bernoulli reliability model is adopted and the machine model parameters are identified using one-month of equipment operating status data measured by the plant PLC system. Moreover, customised formulas are devised to calculate the Bernoulli model parameters based on the operation up- and downtime data. Similarly, in the case study of a multi-product machining line at motorcycle manufacturing plant in Park and Li (2019), the authors simplify a complex production process through aggregation to transform the overall system into a two-stage production line and built a Bernoulli line model for a series of model-based analysis. To identify the efficiencies of the two aggregated machines, the up-/downtime data are collected from 15 machines and customised procedures are developed to obtain the parameters of the aggregated model. For another example, Zandieh, Joreir-Ahmadi, and Fadaei-Rafsanjani (2017) studies the buffer and preventive maintenance period allocation problem in a single-type water heater production line. In this case study, the system is modelled as a non-homogeneous, unreliable production line. The model parameters, including processing time, repair time, maintenance time, time between failures, and time between preventive maintenance, are extracted from preprocessed real production data. This preprocessing step also involves aggregating the data of different operating status recorded on the equipment. Moreover, the workstation failure rates are estimated by the managers' experience. In addition, Liberopoulos and Tsarouhas (2005) investigate the improvement of an automated pizza production line and present a statistical analysis of a set of field failure data. In this study, the average downtime of each machine is computed based on hand-written records of failures spanning a period of four years. In Nwika, Umoh, and Amaewhule (2017), to evaluate the reliability of individual section machines in a glass manufacturing facility, the mean time to repair and the mean time between failures are identified using five years' of data. In a case study at a Toyota manufacturing plant, Li (2013) builds an exponential assembly system model and uses three-month data to calculate the average up/downtime in order to implement the continuous improvement. Similar studies using this operation up/down status data-based production systems modelling approach (referred to as the *conventional approach*) can be found in Li and Meerkov (2009), Jia et al. (2015), Du, Xu, and Li (2016), and Tu et al. (2020), etc.

It should be noted that, although the conventional modelling approach has been successfully applied in a number of studies at large and small/medium-sized manufacturing plants, its limitations (see discussions in Section 1) are also commonly acknowledged by

researchers in this area. Specifically, the availability and quality of equipment operating status data in production systems can be hard to ensure in many practical scenarios (e.g. in some small and medium-sized manufacturers), which usually leads to significant efforts spent on collecting, cleaning, and processing the data. In addition, designing customised algorithms to calculate the model parameters of aggregated/combined operations often requires knowledge and expertise in production systems modelling and analysis.

2.3. New approach for parameter identification when modelling production systems

To overcome these challenges, a new modelling approach is proposed in Sun et al. (2020) to reversely compute the parameters of a production system model based on measured performance metrics of the system, i.e. through inverse modelling (Reddy and Andersen 2002; Anish and Shankar 2020). It should be noted that, matching system performance metrics is a commonly-used approach to identify parameters of a system model in many different fields of engineering. For instance, in Lima, Jacobina, and de Souza Filho (1997), in order to determine the values of the steady-state equivalent circuit parameters of a three-phase squirrel-cage induction machine, the authors formulate and solve a nonlinear optimisation problem to minimise the least-squared error between the theoretical values of stator current, input power, and electromagnetic torque calculated by the analytical expression with circuit parameters and the experimental data value collected from a machine test. Additionally, Reddy and Andersen (2002) investigates different inverse modelling methods with application to off-line model parameter estimation of a field-operated chiller. In the area of production systems engineering, two different methods have been studied to inversely estimate machine parameters in multiple-machine Bernoulli production line models: analytical expression-based method (Sun et al. 2020) and statistical/machine learning-based method (Sun and Zhang 2020). In both studies, standard system performance metrics (throughput, work-in-process, etc.) are used as the input to identify the machine parameters. In a follow-up work, Sun and Zhang (2021) developed an analytical expression-based method to identify parameters in a two-machine synchronous exponential line and Tu et al. (2020) implement a neural network method for parameter identification in synchronous exponential lines with multiple machines. The neural network approach, however, usually involves long training processes and may be difficult to generalise to other cases (e.g. asynchronous operations, systems with parallel construction, etc.). While our prior work (Sun et al. 2020; Sun and Zhang 2020, 2021; Tu et al. 2020) has laid a theoretical foundation for the new production system modelling approach, its efficacy has not been tested in a practical industrial environment, especially the sensitivity of the identified model when the system parameters vary from the baseline values. Therefore, the goal of this paper is to extend the analytical inverse modelling approach to an asynchronous exponential production system model and test the applicability of this method in a practical industrial case study. Additionally, based on the identified model, productivity improvement is investigated and model sensitivity is discussed under improvement scenarios.

2.4. Research originality and contributions

The main contributions of the paper are as follows. From the theoretical aspects, this paper

- develops an efficient and robust search algorithm for machine parameter estimation in two-machine asynchronous exponential models;
- tests the accuracy of the new parameter identification algorithm through numerical/statistical experiments;
- justifies the efficacy of this production system modelling approach via an industrial case study; and
- verifies the applicability of the identified model through a model sensitivity analysis.

For the industrial perspective, this paper

- describes the implementation process of the new modelling approach in a practical system;
- describes in detail the baseline analysis and the design of potential improvement projects; and
- explains the selection and implementation of the improvement project.

This work not only provides a stepping stone for furthering the theoretical research in innovating the production system modelling approach but also offers a detailed reference for researchers and practitioners to implement this approach in real manufacturing practice.

3. System description and problems addressed

3.1. Project background

This study is carried out at a local medium-sized manufacturing company, which designs and manufactures electromechanical devices for industrial, consumer, and medical applications. The company produces over 100 end products in the factory from over 15 product

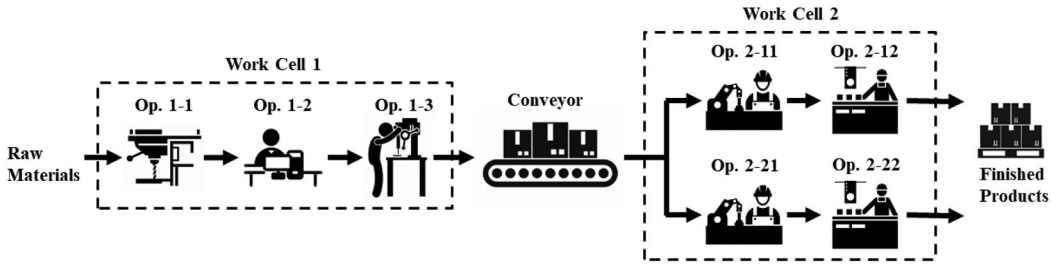


Figure 1. System layout.

families. The production system studied in this paper is dedicated to one of its staple products, which was expected to see a significant increase in volume. This volume increase mainly comes from the company's decision to move the production from overseas back to the U.S. to comply to the *Made-in-USA* certification required by majority of its customers. In addition, due to highly stable demand and reliably forecast of this product, the manufacturer manages the production of this product in a make-to-stock regime based on sales forecast from historical records.

Before the case study was carried out, this production system constantly failed to meet production target. Since the manufacturer also produces several other products simultaneously in the factory, it didn't have extra workforce to be allocated to this particular production system or to produce this particular product in other parts of the factory. Thus, to ensure on-time delivery of the product orders, the manufacturer had to use overtime and extra shifts in this system. On the other hand, the manufacturer did not have any in-house expertise and resources to solve this issue. As a result, the manufacturer applied for assistance through the Quiet Corner Innovation Cluster (QCIC) at the authors' institution funded by the U.S. Economic Development Administration. The objective of the QCIC is to leverage university resources to create an innovation network to drive economic development in the New London, Tolland, and Windham Counties in Connecticut, USA, by sponsoring collaborative projects with small and medium-sized manufacturing enterprises in the region. The study reported in this paper was one of the projects conducted through the QCIC with focus on improving the manufacturer's production efficiency.

3.2. System layout

The layout of the production system studied is shown in Figure 1. The system has two work cells. The enclosures of the device are retrieved from the warehouse to be made available at the input of Work Cell 1. The operator at Op. 1–1 uses a drill press to make several holes on the enclosure. At Op. 1–2, an operator affix labels on the exterior and interior of the enclosures. Then, at Op. 1–3, an operator visually inspects the quality of the previous two steps and make sure locations of the holes as well as the placement and orientation of the labels comply to the product design. The intermediate products are then places on a conveyor to be transported to Work Cell 2. Work Cell 2 is comprised of two parallel lines, which split the work load and carry out identical processing work. Specifically, the operators at Op. 2–11 and Op. 2–21 install the printed circuit board (PCB), electrical wires, switch, and several small mechanical parts into the enclosure and fix them in place by applying screws into the holes drilled in Op. 1–1. Finally, Op. 2–12 and Op. 2–22 visually inspect the assembled devices (e.g. gaps, alignment), test the electrical/mechanical functions, fix any issue discovered, and package the finished products with instruction manual cards into cardboard boxes.

3.3. Problems addressed

As one can see, all operations in this production system involve human labour and it is often difficult for the management to identify or trace production issues/interruptions since there is no production data being recorded outside of the total number of units produced in a shift before this case study was carried out.

To better understand the manufacturing process, a series of meetings were held in the manufacturing facility with the engineers, operators, supervisors, and managers, to learn about the demand/sales/inventory information of the product, its components and production bill-of-materials, detailed operations of each individual manufacturing step, the tricks and challenges in completing each individual step, training and management of the operators, typical problems/failures/stoppages in the production process, actions that have been taken or plans that will be taken to fix these issues, concerns about production disruption caused by this study, etc.

As an outcome from the initial meetings between the manufacturer and the academic team, it was determined that the goals of this case study would include studying the baseline performance of the production

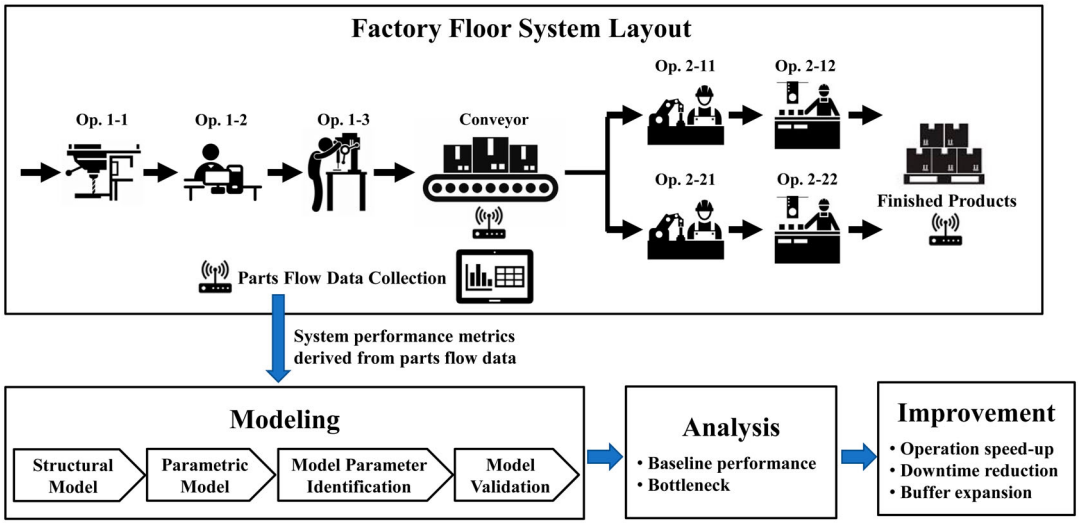


Figure 2. Diagram of the problems addressed in this research.

system, developing improvement plan, and helping the manufacturer implement the plan to enhance the system performance in order to meet potential demand growth in future years.

Thus, in this case study, the following problems are addressed by the academic team (see Figure 2 for an illustration):

- *Modeling*: Collect data and construct a mathematical model for the production system.
- *Analysis*: Use the mathematical model to quantitatively understand and analyse the baseline performance of the system operation.
- *Improvement*: Based on the system model and the baseline performance, investigate and develop potential improvement plans to enhance system performance.

3.4. Challenges and proposed approach

Clearly, to ensure the accuracy of the quantitative analysis of the production system, the mathematical model must have high fidelity. In the conventional approach for mathematical modelling of production systems, occurrences and duration up- and downtimes of all operations are recorded in order to obtain the reliability model and parameters of the operations (see Li and Meerkov 2009). This approach, unfortunately, is not feasible in this case study, because all operations in the system under consideration are manual. First of all, defining up and down states of manual operations is usually challenging and strongly depends on the nature of the processing work conducted at each operation. Secondly, recording the up and down events of manual operations typically requires additional labour to conduct a full on-site time study of the operations involved over an extended period of time. Neither the manufacturer nor the academic team carrying out this case study can afford this much effort. Lastly, measuring up- and downtimes of manual operations may require some close-up observation of the operators. It was learned from the meetings with the operators that this *intrusive* approach may cause stress and negatively impact their productivity and performance.

To overcome these challenges, the new modelling approach is adopted. In particular, parts flow data, i.e. the number of parts at different stages of the system, will be recorded (instead of operation up- and downtimes) to measure system performance metrics (e.g. throughput, work-in-process). Then, we inversely identify the parameters of the individual operations by matching the given performance metrics data. This approach is first proposed for production systems modelling in Sun et al. (2020) which studies the Bernoulli production system models. Using the basic framework of inverse modelling approach, but different from Sun et al. (2020), we formulate this model parameter identification problem as an optimisation problem to minimise the sum of squared errors of estimated performance metrics:

$$\min_{\mathbf{x}} \sum_{k} (F_k(\mathbf{x}) - \theta_k^*)^2, \quad \text{s.t. } \mathbf{x} \in \mathbf{X}, \tag{1}$$

where $\mathbf{x}$ is the model parameters vector, $F_k(\mathbf{x})$ is the estimated performance metric k under parameter $\mathbf{x}$, θ_k^* is the observed/measured value of the performance metric, and $\mathbf{X}$ is the feasible set of $\mathbf{x}$. Detailed implementation of this approach to the system at hand, further analysis, and productivity improvement are described in the following sections.

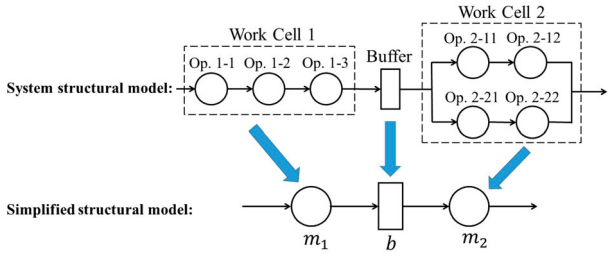


Figure 3. Structural modelling of the medical device production system.

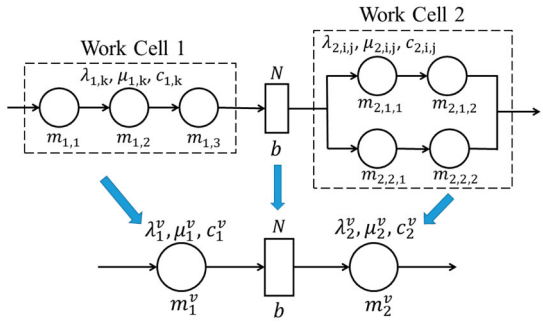


Figure 4. Structural and parametric modelling of the production system studied.

4. Mathematical models

4.1. Structural/parametric modelling and model assumptions

Based on the physical layout of the system (Figure 1), it is not difficult to construct the structural model of the system shown in the block diagram of Figure 3, where circles represent the operations/workstations and the rectangle in the middle represents the in-process buffer connecting the two work cells. Note that no buffers are present within the work cells as the operators in the same work cells are expected to work synchronously inside a work cell. We can further aggregate the operations in each work cell to reach the simplified structural model shown in Figure 3.

To obtain the parametric model of the system, we model the operations as *unreliable machines*, because the operations may experience cycle overruns and delays due to various factors. In addition, since no prior knowledge is available about the up- and downtime of the operations, it is assumed that they can be modelled using the *exponential reliability model*. Under this model, it is assumed that the up- and downtime of a machine/operation are exponential random variables with parameters λ (1/min) and μ (1/min), respectively. The parameters λ and μ are referred as the *breakdown rate* and *repair rate* of the machine, respectively. Given parameters λ and μ, the machine's efficiency e, can be calculated as

$$e = \frac{\mu}{\lambda + \mu} = \frac{T_{up}}{T_{up} + T_{down}}, \tag{2}$$

where $T_{up} = 1/\lambda$ and $T_{down} = 1/\mu$ are the average up- and downtimes, respectively. In addition, the processing speed of a machine is denoted as c (parts/min), while the cycle time is denoted as τ (min). Thus, for the complete structural model, an operation can be characterised by a vector: $(\lambda_{1,k}, \mu_{1,k}, c_{1,k})$, $k = 1, 2, 3$, for the ones in Work Cell 1, and $(\lambda_{2,i,j}, \mu_{2,i,j}, c_{2,i,j})$, $i, j = 1, 2$, for the ones in Work Cell 2. Similarly for the simplified structural model, the aggregated/virtual machines are characterised by $(\lambda_i^v, \mu_i^v, c_i^v)$, $i = 1, 2$. Finally, the in-process buffer is characterised by its storing capacity N, i.e. the maximum number of parts that the buffer can hold (see Figure 4).

Furthermore, based on the production operations of the *actual* system, the following assumptions are made for operations of the *mathematical models* of the production system:

- Full model (seven-machine-one-buffer model):
  (1) (1)For consecutive machines with no intermediate buffers in-between, part processing takes place only when all machines involved are up.
  (2) (2)For parallel lines, parts are evenly allocated to each line involved.
  (3) (3)Machines $m_{2,i,1}$ in Work Cell 2 are said to be *starved* is they are up, the buffer is empty, and the machines in Work Cell 1 are processing jobs with rate slower than $c_{2,i,1}$ (or $(c_{2,1,1} + c_{2,2,1})$ if both are up).
  (4) (4)Machines $m_{1,3}$ in Work Cell 1 are said to be *blocked* is it is up, the buffer is full, and the machines in Work Cell 2 are processing jobs with rate slower than $c_{1,3}$.
  (5) (5)Machine $m_{1,1}$ in Work Cell 1 is not starved for raw materials and machines $m_{2,i,2}$ in Work Cell 2 are not blocked by finished goods inventory.
  (6) (6)If a machine is up and neither blocked nor starved, then it processes parts with rate up to the minimum capacity among the consecutive machines connected with it.
- Simplified model (two-machine-one-buffer model):
  (1) (1)Machine m_2^v is said to be *starved* if it is up, the buffer is empty, and machine m_1^v is either down or processing jobs with rate slower than c_2^v. Assume that machine m_1^v is never starved for raw material.
  (2) (2)Machine m_1^v is said to be *blocked* if it is up, the buffer is full, and machine m_2^v is either down is processing jobs with rate slower than c_2^v. Assume that machine m_2^v is never blocked.

(3) (3)If machine m_i^v is up and neither blocked nor starved, then it processes parts with rate c_i^v (parts/min).

In summary, the production system model considered is a single-product (with unlimited arrival) serial production line with unreliable operations (with randomly distributed up- and downtimes), zero in-process buffering within a cell, and a finite-capacity buffer connecting the cells. Under these model assumptions, the throughput and work-in-process of the production system model are defined as:

- *Throughput*, *TP*: the average number of parts produced by machines $m_{2,1,2}$ and $m_{2,2,2}$ combined in the complete structural model or by machine m_2^v in the simplified structural model per unit of time (e.g. minute, hour) during steady state;
- *Work-in-process*, *WIP*: the average number of parts contained in buffer b during steady state.

In addition, we define two other performance indices to be used in this work:

- *Probability that buffer b is empty*, P_0,
- *Probability that buffer b is full*, P_N.

Note that *TP* and *WIP* are considered as the most important metrics in manufacturing research and practice and are commonly monitored on the factory floor. In the case of P_0 and P_N, although they are not commonly measured on the factory floor, they can still be estimated using the fractions of time that the buffer is empty and full, respectively, during a certain observation period. For the system under consideration, due to the difficulties of monitoring the operations up- and downtime data described in Subsection 3.4, we measure the parts flow and occupancy of the buffer (i.e. the entrances/exits of parts to/from the buffer and the number of parts in the buffer) instead. There are three advantages of measuring the buffer data. First, the measurement (counting the number of parts) is straightforward and involves no ambiguity. Second, all four performance metrics can be derived from the buffer data. Third, the data collection does not require much labour and does not put extra stress on the operators.

In the subsequent analysis, we mainly use the two-machine simplified model since this model has a smaller machine parameter set that can possibly be identified using just *TP*, *WIP*, P_0, and P_N, and these performance metrics can be calculated using analytical formulas (Li and Meerkov 2009). The seven-machine complete model are only used in Subsection 7.4 when we discuss the effects of reducing individual operation downtimes under additional assumptions about the relationships among machine parameters. The sensitivity of the models with respect to system parameter change and to those addition assumptions are discussed in Section 8.

5. Model parameter identification

5.1. Data collected

It should be noted that, due to confidentiality, data masking was applied to the original data collected from the manufacturing floor before being presented in this paper. However, the ones presented still retain similar qualitative features from the actual data. The storing capacity of the conveyor buffer is measured based on the total length of the conveyor space and dimensions of the product. It was determined that the capacity of buffer b is $N = 15$. The reliability parameters (i.e. breakdown rate λ and repair rate μ) of the operations were not collected due to the challenges discussed in Subsection 3.4. Instead, we only measured the (average) cycle time of each operation under typical operating conditions via a time study. The results are as follows:

$$\boldsymbol{\tau_1} = [0.89 \quad 0.80 \quad 0.84] \text{ (min)},$$
$$\boldsymbol{\tau_2} = \begin{bmatrix} 2.06 & 2.02 \\ 1.99 & 2.11 \end{bmatrix} \text{ (min)}. \tag{3}$$

Thus, the processing speeds of $m_{1,k}$'s and $m_{2,i,j}$'s in the complete 7-machine structural model can be obtained as follows:

$$\boldsymbol{c_1} = [1.1200, 1.2524, 1.1915] \text{ (parts/min)},$$
$$\boldsymbol{c_2} = \begin{bmatrix} 0.4854 & 0.4944 \\ 0.5032 & 0.4739 \end{bmatrix} \text{ (parts/min)}. \tag{4}$$

Using the parts flow and occupancy of the buffer measured during a 4-week span (typically 2–3 times a week and about 4 h each time), the following performance metrics were obtained:

$$TP^* = 0.6724 \text{ (parts/min)}, \quad WIP^* = 9.4745 \text{ (parts)},$$
$$P_0^* = 0.0788, \quad P_N^* = 0.3297. \tag{5}$$

5.2. Method to parameter identification for the simplified model

Using these data above, we first identify the parameters of the simplified two-machine-one-buffer model. To accomplish this, based on the machine connection relationship, the equivalent/aggregated processing speeds of

m_1^v and m_2^v can be calculated as:

$$\begin{aligned} c_1^v &= \min\{c_{1,1}, c_{1,2}, c_{1,3}\} = 1.12 \text{ (parts/min)}, \\ c_2^v &= \min\{c_{2,1,1}, c_{2,1,2}\} + \min\{c_{2,2,1}, c_{2,2,2}\} \\ &= 0.9594 \text{ (parts/min)}. \end{aligned} \tag{6}$$

Given c_1^v, c_2^v, and N, system performance metrics TP, WIP, P_0, and P_N are functions of λ_1^v, μ_1^v, λ_2^v, μ_2^v. The formulas of these functions are derived in Li and Meerkov (2009) and are omitted here due to space limitations. In addition, since λ_i^v can be determined based on μ_i^v and e_i^v using

$$\lambda_i^v = \mu_i^v(1 - e_i^v)/e_i^v, \tag{7}$$

instead of identifying $(\lambda_1^v, \mu_1^v, \lambda_2^v, \mu_2^v)$ directly, we define and identify the machine reliability parameter in the form of $\mathbf{x} = (\mu_1^v, \mu_2^v, e_1^v, e_2^v)$. Note that these two formulations are equivalent but the latter gives more convenience in determining the feasible region of the machine parameters. Next, we define the vector-valued function $\mathbf{F}(\mathbf{x})$ as:

$$\begin{aligned} \mathbf{F}(\mathbf{x}) &= \begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \\ f_3(\mathbf{x}) \\ f_4(\mathbf{x}) \end{bmatrix} \\ &= \begin{bmatrix} TP(\mu_1^v, \mu_2^v, e_1^v, e_2^v) - TP^* \\ WIP(\mu_1^v, \mu_2^v, e_1^v, e_2^v)/N - WIP^*/N \\ P_0(\mu_1^v, \mu_2^v, e_1^v, e_2^v) - P_0^* \\ P_N(\mu_1^v, \mu_2^v, e_1^v, e_2^v) - P_N^* \end{bmatrix}. \end{aligned} \tag{8}$$

where TP^*, WIP^*, P_0^*, and P_N^* are the observed system performance metrics given in (5). Based on the above, we formulate the following constrained optimisation problem:

Find machine parameter $\mathbf{x} = (\mu_1^v, \mu_2^v, e_1^v, e_2^v)$ that minimises the 2-norm of error function $\mathbf{F}$ over a certain box-constraint set $\mathbf{X}$, i.e.

$$\begin{aligned} \min_{\mathbf{x}} \quad & f(\mathbf{x}) = \|\mathbf{F}(\mathbf{x})\|^2, \\ \text{s.t.} \quad & \mathbf{x} \in \mathbf{X}, \end{aligned} \tag{9}$$

where $\mathbf{X} = \{\mathbf{x} \in \mathbb{R}^4 \mid \mu_l \le \mu_i^v \le \mu_u, e_l \le e_i^v \le e_u, i = 1, 2\}$ is a box-set containing all possible values of μ_i^v's and e_i^v's and μ_l, μ_u, e_l, and e_u are the lower- and upper-bounds for the parameters.

Since the objective function $f(\mathbf{x})$ of (9) can be computed analytically, it is possible to apply gradient-based numerical optimisation algorithms to solve the optimisation problem considered in this paper. A commonly used gradient-based method to solve box-constrained optimisation problems is the *projected gradient method* (Bertsekas 2016). However, as a first-order method, it usually suffers from slow convergence. To overcome this issue and incorporate higher-order information into the projected search, the projected quasi-Newton method (Schmidt, Kim, and Sra 2012) is introduced by calculating the projection step using an approximate Hessian matrix of the objective function. Projected quasi-Newton method has been proved to be globally convergent and achieves superlinear convergence rate under certain conditions in Bertsekas (2016), Schmidt, Kim, and Sra (2012) and Kim, Sra, and Dhillon (2010). However, if the objective function is not convex, a local minimum may be obtained by the gradient-based method. To overcome this issue, the multi-start strategy of global search can be used, which is capable of exploring more than a single basin of attraction of the objective function (Peri and Tinti 2012).

In this case study, to solve the parameter identification problem (9) and to ensure global optimum, we thus develop the following multi-start modified projected quasi-Newton method (MMPQN). The steps of this algorithm are as follows:

Step 1: Tighten the feasible region. Given the observed throughput TP^*, we can easily obtain $e_i \ge TP^*/c_{min}$, where $c_{min} = \min\{c_1^v, c_2^v\}$. Therefore, the lower bound of e_i can be rewritten as $e_l' = \max\{TP^*/c_{min}, e_l\}$ and the corresponding feasible region becomes

$$\mathbf{X}' = \{\mathbf{x} \in \mathbb{R}^4 \mid \mu_l \le \mu_i \le \mu_u, e_l' \le e_i \le e_u, i = 1, 2\} \tag{10}$$

Step 2: Multi-start policy. Partition the feasible region $\mathbf{X}'$ using multi-dimensional grid and create initial point set P. There are four decision variables in optimisation problem (9) and each variable has a box-constraint. For the i-th decision variable (dimension), we uniformly partition its constraint interval into d_i segments. This results in $D = d_1 \times d_2 \times d_3 \times d_4$ sub-regions. In each sub-region, we randomly sample a point $x_n^{(0)} = (\mu_1^{(0)}, \mu_2^{(0)}, e_1^{(0)}, e_2^{(0)})$ as an initial point. This leads to the initial point set P consisting of D initial points.

Step 3: From each initial point selected, we compute a candidate feasible solution using a modified projected quasi-Newton method, in parallel with other initial points together.

(i) (i)*Initialization of gradient scaling matrix.* Let $k = 0$. With the nth initial point $\mathbf{x}_n^{(0)} = (\mu_1^{(0)}, \mu_2^{(0)}, e_1^{(0)}, e_2^{(0)})$ in P, calculate the initial gradient scaling matrix as $\mathbf{S}^{(0)} = \delta \|\nabla f(\mathbf{x}_n^{(0)})\|^{-1}\mathbf{I}$,

where $\mathbf{I}$ is the identity matrix and $\delta \in (0,1)$ is a parameter of the algorithm (Nocedal and Wright 2006).

(ii) (ii)*Armijo rule along the projection arc.* Using $[\cdot]^+$ to denote the projection on the constraint set $\{\mathbf{x} \in \mathbb{R}^n \mid a_i \le x_i \le b_i,\ i = 1, \ldots, n\}$, the ith coordinate of the projection of vector $\mathbf{x}$ is given by Bertsekas (2016):

$$[x_i]^+ = \begin{cases} a_i & \text{if } x_i \le a_i, \\ b_i & \text{if } x_i \ge b_i, \\ x_i & \text{otherwise.} \end{cases} \tag{11}$$

Then, the projection arc is defined as

$$\mathbf{x}_n^{(k)}(\alpha) = [\mathbf{x}_n^{(k)} - \alpha \mathbf{S}^{(k)} \nabla f(\mathbf{x}_n^{(k)})]^+, \tag{12}$$

where $\alpha \in (0,1)$ and $\mathbf{S}^{(k)}$ is the gradient scaling matrix (Schmidt, Kim, and Sra 2012). By Armijo rule (Bertsekas 2016), under algorithm parameters $\beta \in (0,1)$ and $\sigma \in (0,1)$, we find the smallest non-negative integer q such that

$$f(\mathbf{x}_n^{(k)}) - f(\mathbf{x}_n^{(k)}(\beta^q)) \ge \sigma \nabla f(\mathbf{x}_n^{(k)})^T [\mathbf{x}_n^{(k)} - \mathbf{x}_n^{(k)}(\beta^q)]. \tag{13}$$

Let $\alpha = \beta^q$. The feasible projection arc vector is given by

$$\bar{\mathbf{x}}_n^{(k)} = [\mathbf{x}_n^{(k)} - \alpha \mathbf{S}^{(k)} \nabla f(\mathbf{x}_n^{(k)})]^+ \tag{14}$$

and the feasible direction is $d = \bar{\mathbf{x}}_n^{(k)} - \mathbf{x}_n^{(k)}$.

(iii) (iii)*Update* $\mathbf{x}_n^{(k+1)}$. With diminished step size $s = \gamma^q$, where $\gamma \in (0,1)$ and $q = \lceil k/10 \rceil$, we calculate $\mathbf{x}_n^{(k+1)}$ as

$$\mathbf{x}_n^{(k+1)} = \mathbf{x}_n^{(k)} + s \cdot d. \tag{15}$$

(iv) (iv)*Update* $\mathbf{S}^{(\mathbf{k+1})}$. To calculate the new gradient scaling matrix to be used in the feasible projection arc vector (14) in the next iteration, we first divide the variables into two groups: *free* and *restricted*. The latter refers to the subset of variables that are close to their bounds:

$$\mathfrak{R}^{(k)} = \{i \mid \bar{x}_i^{(k)} < a_i + \epsilon,\ \partial_i f(\bar{\mathbf{x}}^{(k)}) > 0\} \cup \{i \mid \bar{x}_i^{(k)} > b_i - \epsilon,\ \partial_i f(\bar{\mathbf{x}}^{(k)}) < 0\}, \tag{16}$$

where ϵ is very small positive number. All other variables constitute the set of free variables, denoted as $\mathfrak{F}^{(k)}$. Without loss of generality, assume that $\mathfrak{F}^{(k)} = \{1, \ldots, K_F\}$ and $\mathfrak{R}^{(k)} = \{K_F + 1, \ldots, n\}$, where K_F is the number of free variables in the current iteration and K is the total number of variables ($K = 4$ for the problem addressed in this paper). Then, the new gradient scaling matrix is calculated as

$$\mathbf{S}^{(k+1)} = \begin{bmatrix} \bar{\mathbf{S}}^{(k+1)} & 0 \\ 0 & \mathbf{I} \end{bmatrix}, \tag{17}$$

where $\bar{\mathbf{S}}^{(k+1)}$ is given by the principal submatrix of the approximated inverse of the Hessian matrix as induced by the free variables calculated via the *Broyden-Fletcher-Goldfarb-Shanno* (BFGS) update (Nocedal and Wright 2006). In other words, the Newton-type Hessian matrix-based gradient scaling only applies to the free variables, while the descent of the restricted variables is along the gradient direction.

(v) (v)*Convergence criteria.* Let $k = k+1$. If $\|\nabla f(\mathbf{x}_n^{(k)})\| \le \epsilon_g$, terminate algorithm and output the nth candidate solution $\widehat{\mathbf{x}}_n = \mathbf{x}_n^{(k)}$; otherwise, return Step (ii).

Step 4: Select optimal solution. With D different initial points from P, we obtain D candidate feasible solutions $\{\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \ldots, \hat{\mathbf{x}}_D\}$. The optimal solution $\hat{\mathbf{x}}$ is selected as the one that leads to the lowest value of objective function.

Note that the analytical formula of $\nabla f(\mathbf{x})$ is all but impossible to derive. Therefore, we use the central difference formula (Nocedal and Wright 2006; Bertsekas 2016) as an approximation:

$$\frac{\partial f(\mathbf{x})}{\partial x_i} \approx \frac{f(x_1, \ldots, x_i + \Delta, \ldots, x_n) - f(x_1, \ldots, x_i - \Delta, \ldots, x_n)}{2\Delta}, \tag{18}$$

where Δ is a very small positive number. In the following numerical experiments, we set $\Delta = 10^{-12}$. In summary, the procedure of the MMPQN algorithm is also described in the pseudo-code as below.

5.3. Identified model parameters

The proposed algorithm MMPQN applies second-order gradient-based method with the reduction of searching space and utilises multiple start points, which are well distributed in the feasible space. These features make the algorithm highly computationally efficient and capable of obtaining high-quality solution. Using MMPQN, we solve the parameter identification problem (9) under observed performance metrics (5) and obtain the parameters of the simplified two-machine-one-buffer model as

Algorithm Multi-start Modified Projected Quasi-Newton Method (MMPQN)

Determine the tightened feasible region **X'** based on (10).
Based on the multi-start policy, select D different initial points in **X'** to construct the initial point set $P = \{\mathbf{x}_1^{(0)}, \ldots, \mathbf{x}_D^{(0)}\}$.
for $n = 1, \ldots, D$ **do**
  Initialization with $\mathbf{x}_n^{(0)}$: Set $k = 0$ and $\mathbf{S}^{(0)} = \delta \|\nabla f(\mathbf{x}_n^{(0)})\|^{-1}\mathbf{I}$, where $\delta \in (0, 1)$.
  while $\|\nabla f(\mathbf{x}_n^{(k)})\| \geq \epsilon_g$ **do**
    1. Find appropriate value for $\alpha^{(k)}$ using Armijo rule;
    2. Compute the projection arc: $\bar{\mathbf{x}}_n^{(k)} = [\mathbf{x}_n^{(k)} - \alpha^{(k)}\mathbf{S}^{(k)}\nabla f(\mathbf{x}_n^{(k)})]^+$;
    3. Compute the feasible direction: $d = \bar{\mathbf{x}}_n^{(k)} - \mathbf{x}_n^{(k)}$;
    4. Update $\mathbf{x}_n^{(k+1)} = \mathbf{x}_n^{(k)} + \gamma^q \cdot d$, where $\gamma \in (0, 1)$ and $q = \lceil k/10 \rceil$;
    5. Compute the approximated inverse of the Hessian matrix $\bar{\mathbf{S}}^{(k+1)}$ via *BFGS* update method;
    6. Determine restricted variables set $\mathfrak{R}^{(k)}$ by (16);
    7. Set $\mathbf{S}^{(k+1)} = \mathbf{I}$, and then, let $\mathbf{S}_{i,j}^{(k+1)} = \bar{\mathbf{S}}_{i,j}^{(k+1)}$, $\forall i, j \notin \mathfrak{R}^{(k)}$;
    8. $k = k + 1$;
  end while
  $\widehat{\mathbf{x}}_n = \mathbf{x}_n^{(k)}$.
end for
Return $\widehat{\mathbf{x}} = \arg\min_{\widehat{\mathbf{x}}_n} f(\widehat{\mathbf{x}}_n)$.

follows:

$$T_{up,1}^{v} = 24.94, \quad T_{down,1}^{v} = 9.40, \quad e_1^{v} = 0.7263,$$
$$T_{up,2}^{v} = 17.14, \quad T_{down,2}^{v} = 5.21, \quad e_2^{v} = 0.7669. \tag{19}$$

6. Model validation

The conventional production system modelling approach usually validates the identified model by comparing the model-predicted system performance metrics with the ones measured on the factory floor (Li and Meerkov 2009). Following this idea, we calculate the system performance metrics based on the two-machine-one-buffer model and the identified system parameters (19) and evaluate the errors compared with the observed ones (5) using

$$\epsilon_{TP} = \frac{|\widehat{TP} - TP^*|}{TP^*} \cdot 100\%, \quad \epsilon_{P_0} = |\widehat{P}_0 - P_0^*|,$$
$$\epsilon_{WIP} = \frac{|\widehat{WIP} - WIP^*|}{N} \cdot 100\%, \quad \epsilon_{P_N} = |\widehat{P}_N - P_N^*|, \tag{20}$$

where $\widehat{\cdot}$ denotes the performance metrics obtained from the two-machine line model. As a result, it was obtained that

$$\epsilon_{TP} = 3.48 \times 10^{-8}\%, \quad \epsilon_{WIP} = 8.63 \times 10^{-8}\%,$$
$$\epsilon_{P_0} = 3.19 \times 10^{-10}, \quad \epsilon_{P_N} = 1.43 \times 10^{-10}. \tag{21}$$

In other words, the parameters identified can provide a perfect match to observation data from the factory floor (in the sense of system performance metrics).

In addition, it is desirable to determine how close the identified machine parameters (19) match the actual operation of the two work cells. Unfortunately, a direct comparison is not feasible due to no measurement data available from the operations. As an alternative, an indirect approach is considered. Specifically, instead of focussing on the identified machine parameters (19) for this particular system, we turn to evaluate the performance of the MMPQN algorithm in identifying the machine parameters in general, including the parameter estimation accuracy, convergence, computation efficiency. To accomplish this, we generates 10,000 *two-machine* exponential lines with machine parameters randomly and equiprobably selected from

$$\mu_i \in (1/15, 1/5), \quad e_i \in (0.7, 0.95),$$
$$c_i \in (0.75, 1.25), \quad i = 1, 2. \tag{22}$$

The buffer capacity of each line is selected as $N_i = K \cdot \max(1/\mu_1, 1/\mu_2)$, where K is randomly generated from $(1, 3)$. These parameter ranges are selected to reflect typical manufacturing cases where the exponential line model is appropriate: the downtime about 5–15 times of the machine cycle time and the buffer can accommodate roughly 1–3 downtimes on average. For the 10,000 lines generated above, we calculate their performance metrics based on the analytical expressions derived in Li and Meerkov (2009), which are used as the input to identify the machine parameters with MMPQN algorithm. All computations reported in this paper were implemented in MATLAB R2021a on an HP ENVY TE01-2275xt workstation with 11th Gen Intel(R) Core(TM) i7-11700 CPU 2.50 GHz processor and 16.0 GB of RAM. For 10,000 lines studied in this experiment, the MMPQN algorithm converges in finite steps and reaches the near-zero optimisation objective function value (average $f_{\min} = 2.94 \times 10^{-12}$) with very short computation time (0.1032 s on average).

Table 1. Average estimation errors of performance metrics.

ϵ_{TP} (%)	ϵ_{WIP}(%)	ϵ_{P_0}	ϵ_{P_N}
4.38×10^{-6}	3.46×10^{-6}	2.93×10^{-8}	3.29×10^{-8}

For each line generated above, the estimation errors of the machine parameters identified by MMPQN, compared with the true machine parameters, $T^*_{up,i}$, $T^*_{down,i}$, and e^*_i are calculated based on

$$\begin{aligned}
\epsilon_{T_{up,i}} &= \frac{|\hat{T}_{up,i} - T^*_{up,i}|}{T^*_{up,i}} \cdot 100\%, \\
\epsilon_{T_{down,i}} &= \frac{|\hat{T}_{down,i} - T^*_{down,i}|}{T^*_{down,i}} \cdot 100\%, \\
\epsilon_{e_i} &= \frac{|\hat{e}_i - e^*_i|}{e^*_i} \cdot 100\%, \quad i = 1, 2.
\end{aligned} \tag{23}$$

where $\hat{\cdot}$ denotes the estimated machine parameters obtained by MMPQN algorithm. Moreover, the errors of the performance metrics calculated using the identified machine parameters for the 10,000 lines are evaluated based on (20). The average of these estimation errors are summarised in Table 1. As one can see, the MMPQN method can consistently obtain machine parameters that match the input (observed) performance metrics almost perfectly. For individual machine parameters, the estimated average up- and downtimes typically have about 0.0001% error compared with the true machine parameters, while the estimated machine efficiencies are only about 8×10^{-5}% different from the true parameters (see Table 2).

Therefore, based on the low performance metrics estimation error obtained in (21) for the production system at hand and the high accuracy of parameter estimation of the MMPQN method, in general, demonstrated in Table 2, we claim that the two-machine line model and the machine parameters (19) for the production system studied in this paper are validated.

7. System baseline analysis and improvement

In this section, based on the mathematical model constructed in Section 4 and identified model parameters in Section 5, we analyse the baseline performance of the system and investigate options for performance improvement to meet the manufacturer's goal of increasing the throughput by 10–15%. This is achieved by studying and combining the effects of increasing processing speed, increasing buffer capacity and reducing machine/operation downtime.

7.1. System bottleneck

The *bottleneck* (BN) of a production system is defined as the machine m_i that leads to the maximal increase of TP when the processing speed of one machine is increased. Mathematically, Work Cell i, $i \in \{1, 2\}$, is the BN, if

$$\frac{\partial TP}{\partial c_i} > \frac{\partial TP}{\partial c_j}, \quad \forall\, j \neq i. \tag{24}$$

Using the two-machine-one-buffer model, the parameters identified in (19) and the bottleneck identification method introduced in Li and Meerkov (2009), we obtain that the BN of the system is Work Cell 2. Similarly, we define the BN in Work Cell 2 as the machine that leads to the maximal increase of Work Cell 2's overall processing speed when the processing speed of one machine is increased. Based on (4) and (6), we can easily find out that Op. 2–22 is the BN of this work cell, and, thus, the BN of this production system.

This conclusion is consistent with the observation on the factory floor and discussion with the production personnel as well. Indeed, Op. 2–12 and Op. 2–22 involve some dexterous manipulations of small objects while holding the main product in place. The operator at Op. 2–22 during this study was a new employee that just finished training on the job not long ago and, thus, was less proficient compared to her peer in Op. 2–12. Similarly, although Op. 2–11 is not the BN, it is another constraining operation besides Op. 2–22, because it is also staffed by a new employee. The plant intentionally pair an inexperienced operator (Op. 2–11 and Op. 2–22) with an experienced one (Op. 2–12 and Op. 2–21) in each of the parallel lines in Work Cell 2 to enable teaching and assisting on the spot.

7.2. Effects of increasing machine processing speed

Given that Work Cell 2 is the BN of the system, we discuss the following four improvement scenarios of increasing the processing speeds of its operations (Table 3):

- *Improvement Scenario I*:
  - *Action*: Increase the processing speed of Op. 2–22 to $c_{2,2,2} = 0.4944$.

Table 2. Average estimation errors of machine parameters.

$\epsilon_{T_{up,1}}$ (%)	$\epsilon_{T_{up,2}}$ (%)	$\epsilon_{T_{down,1}}$ (%)	$\epsilon_{T_{down,2}}$ (%)	ϵ_{e_1} (%)	ϵ_{e_2} (%)
9.79×10^{-4}	3.42×10^{-4}	9.98×10^{-4}	3.67×10^{-4}	7.19×10^{-5}	8.50×10^{-5}

Table 3. Improvement scenarios of increasing machine processing speeds in Work Cell 2.

	$c_{2,i,j}$	c_2^V	*TP* (parts/hour)	*TP* improvement %
Baseline parameters	$\begin{bmatrix} 0.4854 & 0.4944 \\ 0.5032 & 0.4739 \end{bmatrix}$	0.9594	40.34	–
Improvement Scenario I	$\begin{bmatrix} 0.4854 & 0.4944 \\ 0.5032 & \mathbf{0.4944} \end{bmatrix}$	0.9798	40.96	1.53%
Improvement Scenario II	$\begin{bmatrix} \mathbf{0.4944} & 0.4944 \\ 0.5032 & 0.4739 \end{bmatrix}$	0.9683	40.62	0.68%
Improvement Scenario III	$\begin{bmatrix} \mathbf{0.4944} & 0.4944 \\ 0.5032 & \mathbf{0.4944} \end{bmatrix}$	0.9888	41.22	2.17%
Improvement Scenario IV	$\begin{bmatrix} \mathbf{0.5267} & \mathbf{0.5267} \\ \mathbf{0.5267} & \mathbf{0.5267} \end{bmatrix}$	1.0534	42.95	6.45%

- *Rationale*: Since Op. 2–22 is the BN of the system, it is natural to prioritise its improvement over other operations in the system. This can be accomplished, for instance, by offering better training and retaining more experienced workers by the manufacturer.

- *Improvement Scenario II*:
  - *Action*: Increase the processing speed of Op. 2–11 to $c_{2,1,1} = 0.4944$.
  - *Rationale*: Although Op. 2–11 is the not the BN, it is also an operation that strongly limits the productivity of the system. In addition, the processing accomplished at Op. 2–11 and Op. 2–21 are less complex than that at Op. 2–12 and Op. 2–22, which may lead to an easier implementation for improvement in practice.
- *Improvement Scenario III*:
  - *Action:* Increase the processing speeds of both Op. 2–11 and Op. 2–22 to $c_{2,1,1} = c_{2,2,2} = 0.4944$.
  - *Rationale*: This scenario is the combination of the above two, where the operations staffed by inexperienced operators (Op. 2–11 and Op. 2–22) are to be improved to gain similar proficiency as an experienced one (Op. 2–12).
- *Improvement Scenario IV* :
  - *Action*: Increase the processing speeds of all operations in Work Cell 2 to $c_{2,i,j} = 0.5267$, so that the BN of this system shifted to Work Cell 1. We refer this scenario as maximum speed-up on system BN.
  - *Rationale*: As mentioned above, Work Cell 2 involves some quite dexterous manipulations of small objects using one hand, while holding the main product in place to align the pre-drilled holes using the other hand. To make things more challenging, the holding was performed while suppressing a relatively strong spring inside the product enclosure. This improvement scenario is created to explore the potential of designing and introducing a fixture that can mitigate the complexity of the operations and thus improve productivity.

The results of the improved throughput under each scenario are also summarised in Table 3. As one can see, these improvement scenarios can increase the throughput by 1.5–6.5%, which cannot reach our improvement target. Thus, in the following, we further investigate the effects of increasing buffer capacity and reducing operation downtime on system throughput.

7.3. Effects of increasing buffer capacity

In this subsection, we study the effects of increasing buffer capacity on top of the four improvement scenarios discussed above. Specifically, with the identified model parameters (19), we increase the buffer capacity from $N = 15$ to $N = 30$ and calculate the throughput of the system under the baseline processing speed as well as under the four scenarios of improved processing speeds. The effects of buffer capacity increase are evaluated based on the *additional* improvement in *TP* introduced by increasing the buffer capacity (on top of the improvement reported in Table 3). The results are illustrated in Figure 5.

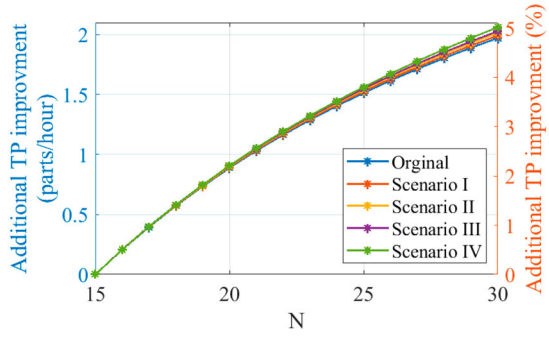


Figure 5. Improvement in *TP* by adjusting *N* under each scenario.

As shown in this figure, increasing buffer capacity has practically identical effects on throughput in all four improvement scenarios as well in the baseline parameter case and can make another 4.5% improvement on system throughput when the buffer capacity is doubled. While the result suggests that combining the maximum processing speed improvement on BN (i.e. Scenario IV) and doubled buffer capacity can increase the throughput by about 11.5%, this is achieved by stretching both avenues to the limit with no headroom to spare, which may not be ideal in practical implementation. Thus, potential improvement by reducing machine/operation downtime is further discussed in next subsection.

7.4. Effects of reducing machine downtime

In addition to machine/operation processing speed and buffer capacity, reducing downtime of the operations is also a commonly used method to improve system throughput in manufacturing practice. In this case study, since the up- and downtime data of each individual operation are not available, a direct analysis of effects of operation downtime reduction is not possible. Thus, an approximation procedure is developed to estimate the resulting model parameters of the work cells when the downtime of an operation is reduced. Based on the estimated model parameters, the effects of reducing operation downtime on system throughput are then calculated.

7.4.1. Algorithm to parameter estimation for the complete model

In order to estimate the new model parameters of a work cell when the downtime of a certain operation in this work cell is reduced, the first step is to identify the parameters of the seven individual machines. Specifically, according to the knowledge gained from the interviews with the operators, engineers, and managers on the factory floor, the workload within each work cell is usually balanced by design such that the stand-alone throughput of each operation in the same work cell is similar. Moreover, for the operations that are connected consecutively in a group, i.e. (Op. 1–1, Op. 1–2, Op. 1–3), (Op. 2–11, Op. 2–12), and (Op. 2–21, Op. 2–22), their 'stoppages' are also similar. Therefore, based on this information, we assume the following for the seven-machine-one-buffer model:

- The operations in the same work cell have identical stand-alone throughput, i.e.

$$\text{Work Cell 1}: c_{1,1}e_{1,1} = c_{1,2}e_{1,2} = c_{1,3}e_{1,3}, \tag{25}$$

$$\text{Work Cell 2}: \begin{aligned} &c_{2,1,1}e_{2,1,1} = c_{2,1,2}e_{2,1,2},\\ &c_{2,2,1}e_{2,2,1} = c_{2,2,2}e_{2,2,2},\\ &\min(c_{2,1,1}, c_{2,1,2}) \cdot e_{2,1,1}e_{2,1,2}\\ &\quad = \min(c_{2,1,1}, c_{2,1,2}) \cdot e_{2,1,1}e_{2,1,2}. \end{aligned} \tag{26}$$

- The consecutive operations in a group have identical downtime (breakdown rate), i.e.

$$\text{Work Cell 1}: T_{down,1,1} = T_{down,1,2} = T_{down,1,3}, \tag{27}$$

$$\begin{aligned} &\text{Work Cell 2}: T_{down,2,1,1} = T_{down,2,1,2},\\ &T_{down,2,2,1} = T_{down,2,2,2}, \end{aligned} \tag{28}$$

or

$$\begin{aligned} &\mu_{1,1} = \mu_{1,2} = \mu_{1,3},\\ &\mu_{2,1,1} = \mu_{2,1,2}, \quad \mu_{2,2,1} = \mu_{2,2,2}, \end{aligned} \tag{29}$$

In addition, Yan and Zhao (2013) derive formulas to approximate the aggregated parameters of a group of parallel machines or consecutive machines. Specifically, for a group of W parallel machines, the aggregated parameters of the group as a whole can be calculated as

$$e^{par} = \frac{\sum_{i=1}^{W} c_i e_i}{\sum_{i=1}^{W} c_i}, \quad T_{down}^{par} = \frac{1 - e^{par}}{\sum_{i=1}^{W} \frac{1}{T_{down,i} + T_{up,i}}}. \tag{30}$$

For a group of Z consecutive machines, the aggregated parameters of the group can be calculated as

$$e^{con} = \prod_{j=1}^{Z} e_j, \quad T_{down}^{con} = \frac{1 - e^{con}}{\sum_{j=1}^{Z} \frac{1}{T_{down,j} + T_{up,j}}}. \tag{31}$$

Thus, based on (30), (31) and the assumptions above, we can reversely calculate the parameters of each individual operations in the seven-machine model as follows:

- Work Cell 1: It follows from (31) that the parameters of the individual machines should satisfy:

$$e_1^v = e_{1,1}e_{1,2}e_{1,3}, \quad T_{down,1}^v = \frac{1 - e_1^v}{\sum_{i=1}^{3} \frac{1 - e_{1,i}}{T_{down,1,i}}}. \tag{32}$$

Solving these equations with assumptions (25) and (27), we can obtain:

$$\begin{aligned} e_{1,i} &= \frac{(e_1^v \prod_{j=1}^{3} c_{1,j})^{\frac{1}{3}}}{c_{1,i}},\\ T_{down,1,i} &= T_{down,1}^v \frac{\sum_{j=1}^{3}(1 - e_{1,j})}{1 - e_1^v}. \end{aligned} \tag{33}$$

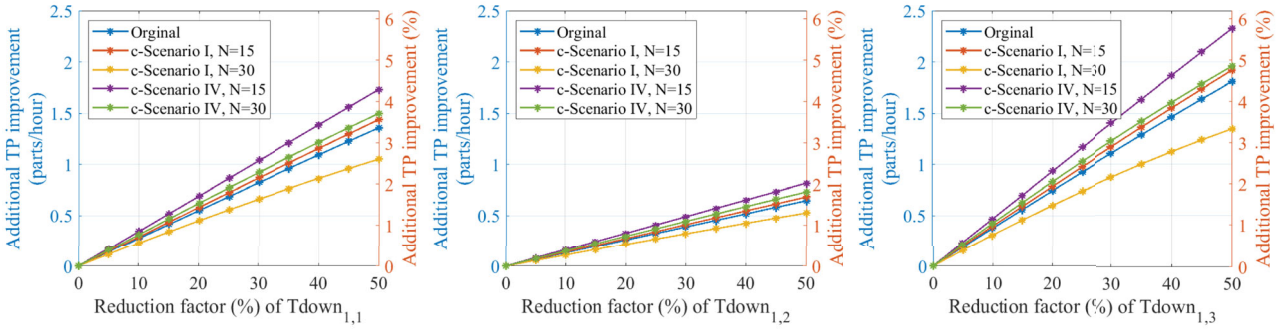


Figure 6. Improvement in *TP* by adjusting $T_{down,1,k}$ under each scenario.

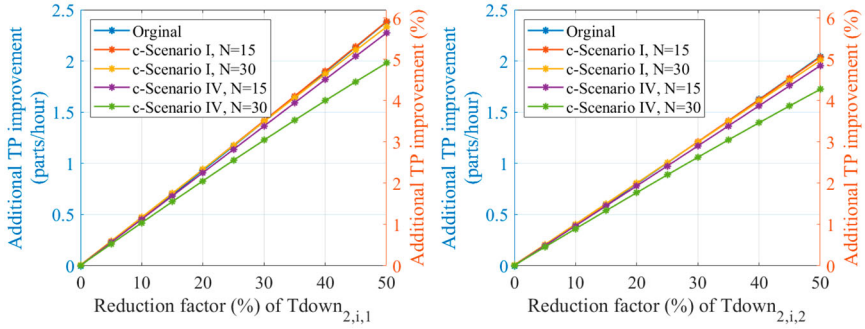


Figure 7. Improvement in *TP* by adjusting $T_{down,2,i,j}$ under each scenario.

Using the two-machine model parameters, these equations lead to

$$\begin{aligned} &e_{1,1} = 0.8931, \quad e_{1,2} = 0.9489, \quad e_{1,3} = 0.8571, \\ &T_{down,1,1} = T_{down,1,2} = T_{down,1,3} = 10.33. \end{aligned} \tag{34}$$

- Work Cell 2: Similarly, the parameters of the operations should satisfy:

$$\begin{aligned} e_2^v &= \frac{\min(c_{2,1,1}, c_{2,1,2}) \cdot e_{2,1,1}e_{2,1,2} + \min(c_{2,1,1}, c_{2,1,2}) \cdot e_{2,2,1}e_{2,2,2}}{\min(c_{2,1,1}, c_{2,1,2}) + \min(c_{2,1,1}, c_{2,1,2})}, \\ T_{down,2}^v &= \frac{1 - e_2^v}{\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{1-e_{2,i,j}}{T_{down,2,i,j}}}. \end{aligned} \tag{35}$$

Solving these equations with assumptions (26) and (28) under the two-machine model parameters results in

$$\begin{aligned} &e_{2,1,1} = 0.8786, \quad e_{2,1,2} = 0.8626, \quad e_{2,2,1} = 0.8550, \\ &e_{2,2,2} = 0.9079, \\ &T_{down,2,1,1} = T_{down,2,1,2} = 11.13, \\ &T_{down,2,2,1} = T_{down,2,2,2} = 11.04. \end{aligned} \tag{36}$$

7.4.2. Improvement scenarios

With the identified parameters of the seven-machine model (34) and (36) and the equations to calculate aggregated machine parameters in the two-machine model (32) and (35), the effects of reducing downtime on system throughput can be analysed. Specifically, the analysis is conducted under the baseline parameters and processing speed modification scenarios Improvement Scenario I and Improvement Scenario IV in Subsection 7.2, in combination with the original buffer capacity $N = 15$ and doubled capacity $N = 30$. For each case, we reduce the downtime of an operation by a factor of $a\%$, $a \in [0, 50]$, with a referred as the *reduction factor*, and calculate the additional improvement in throughput resulted from the downtime reduction. Note that, due to the parallel structure in Work Cell 2, we reduce the downtime of the operations carrying out the same processing work simultaneously (i.e. $T_{down,2,1,1}$ and $T_{down,2,2,1}$ together and $T_{down,2,1,2}$ and $T_{down,2,2,2}$ together) to mimic the elimination/alleviation of certain common causes for failures/stoppages. The modified individual machine downtime ($(1 - a\%)T_{down,1,k}$ or $(1 - a\%)T_{down,2,i,j}$) is first used to update its own efficiency ($e_{1,k}$ or $e_{2,i,j}$). Then, the modified operation downtime and efficiency are plugged back into (32) or (35) to obtain the aggregated machine parameters (e_i^v and $T_{down,i}^v$) of the two-machine-one-buffer model. Finally, the improved throughput is calculated based on the closed-form expression for *TP*

given in Li and Meerkov (2009) for two-machine asynchronous exponential lines. The resulting additional improvement in *TP* from downtime reduction are shown in Figures 6 and 7.

As one can see, for the operations in Work Cell 1, the throughput is the most sensitive to the downtime reduction at Op. 1–3. While within Work Cell 2, the throughput is more sensitive to the downtime reduction at Op. 2–11/2–21. Between Op. 1–3 and Op. 2–11/2–21, the latter appears to be more impactful. Furthermore, reducing their downtime in half may lead to an addition 5–6% improvement in *TP*. On the factory floor, reducing downtime at Op. 2–11/2–21 can be achieved by identifying and alleviating the dominant stoppage/failure causes/modes and will be considered in the recommendation/implementation stage next.

7.5. Recommendations and implementation

To facilitate selecting the improvement actions, a total of 10 improvement plans (IPs) are designed. For operation speed-up, we consider Scenario I (speed-up on BN of Work Cell 2) and Scenario IV (maximum speed-up on system BN) and select three levels of downtime reduction factor (10%, 25% and 40%) for $T_{down,2,i,1}$. Then, buffer capacities are selected to further boost the throughput under various combinations of operation speed-up and downtime reduction. This leads to IPs 1–9, grouped into three levels of *TP* improvement (7.5%, 10%, and 12.5%), and IP 10, which is the potentially maximal improvement with each improvement action stretched to the maximum. The modifications of system parameters required for each IP are given in Table 4. These IPs are intended to provide the production management a mix of different options to improve their operations. The *TP* improvement for each IP as well as the breakdown of the improvement by actions (operations speed-up, downtime reduction, and buffer expansion) are given in Figure 8. The figure can further help them visualise the results of different improvement plans and make decision based on their desired target, preferences, and available resources.

With the analysis and follow-up discussion with the production supervisor and company officials, IP 6 was eventually selected by the team. To implement this plan, we designed and delivered a 3D-printed fixture that can hold the part steady in place. This relieves both hands of the operators to performance the tasks more efficiently, leading to shorter processing times in all operations in Work Cell 2. In addition, using this fixture also effectively eliminates the most common stoppages in these operations, thus, reducing the downtime of the operations. Finally, since the product has a relatively small footprint in dimension, the current buffer is capable of accommodating 3–4 more parts on top of the nominal capacity of $N = 15$. After implementation of these actions, the system throughput was improved to 0.7477 parts/min or 44.86 part/hour, about 11% improvement over the baseline and similar to what is predicted by the analysis.

Table 4. Improvement plans of each *TP* improvement target.

TP Impr. Target	Impr. Plan	Operation Speed-up	Reduction of $T_{down,2,i,1}$	N
7.50%	Impr. Plan 1	Scenario I	10%	30
	Impr. Plan 2	Scenario I	25%	23
	Impr. Plan 3	Scenario I	40%	18
10.00%	Impr. Plan 4	Scenario I	40%	26
	Impr. Plan 5	Scenario IV	10%	21
	Impr. Plan 6	Scenario IV	25%	17
12.50%	Impr. Plan 7	Scenario IV	10%	30
	Impr. Plan 8	Scenario IV	25%	24
	Impr. Plan 9	Scenario IV	40%	19
	Impr. Plan 10	Scenario IV	50%	30

8. Model sensitivity analysis

Although the model parameters identified in Subsections 5.2 (for the two-machine-one-buffer model) and Subsection 7.4.1 (for the seven-machine-one-buffer model) can perfectly fit the observed system performance metrics. It does not directly imply that these model parameters are indeed the *true* parameters of the production system. In particular, the *identical downtime and identical stand-alone throughput* assumptions used to estimate the seven-machine model parameters in Subsection 7.4.1 may not hold exactly in practice and the parameters calculated under these assumptions may be different from the actual up- and downtime characteristics of the operations. Such issues may lead to deviation of the performance metrics from the true values when we design improvement actions discussed in Section 7, which requires modification of the model parameters. Thus, in this section, we investigate the sensitivity of the models identified through the proposed modelling approach with respective to these assumptions.

To carry this out, we first search for a set of machine parameters that leads to the same performance metrics of (5) under the seven-machine model. A solution is given below:

$$\begin{aligned}
\boldsymbol{\lambda_1} &= [0.0091, 0.0119, 0.0170], \\
\boldsymbol{\lambda_2} &= \begin{bmatrix} 0.0047 & 0.0075 \\ 0.0095 & 0.0053 \end{bmatrix}; \\
\boldsymbol{\mu_1} &= [0.1379, 0.1222, 0.1092], \\
\boldsymbol{\mu_2} &= \begin{bmatrix} 0.0487 & 0.0449 \\ 0.0460 & 0.0508 \end{bmatrix}.
\end{aligned} \tag{37}$$

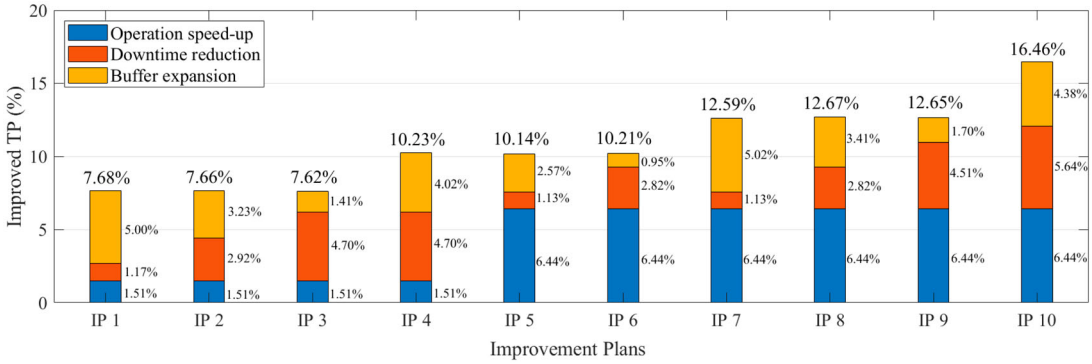


Figure 8. Improvement in *TP* (%) under each improvement plans.

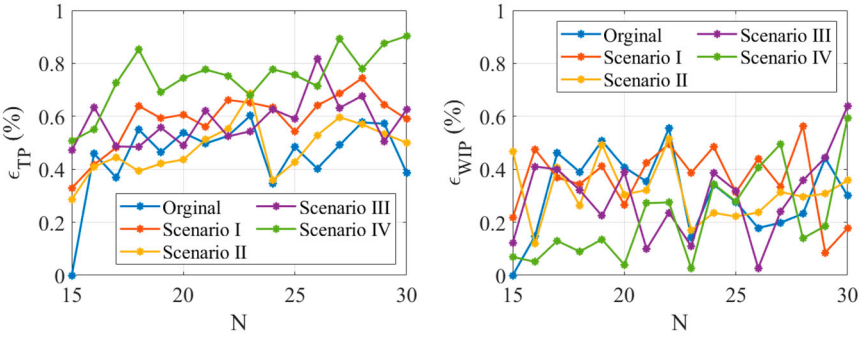


Figure 9. Errors of *TP* and *WIP* under modified buffer capacity and machine processing speed.

Equivalently, based on (7), the machine reliability parameters can also be expressed as

$$\begin{aligned} \boldsymbol{T_{up,1}} &= [109.69, 83.73, 58.69], \\ \boldsymbol{T_{up,2}} &= \begin{bmatrix} 212.97 & 133.40 \\ 105.30 & 187.48 \end{bmatrix}; \\ \boldsymbol{T_{down,1}} &= [7.25, 8.18, 9.16], \\ \boldsymbol{T_{down,2}} &= \begin{bmatrix} 20.55 & 22.26 \\ 21.72 & 19.68 \end{bmatrix}; \\ \boldsymbol{e_1} &= [0.9380, 0.9110, 0.8650], \\ \boldsymbol{e_2} &= \begin{bmatrix} 0.9120 & 0.8570 \\ 0.8290 & 0.9050 \end{bmatrix}. \end{aligned} \tag{38}$$

As one can see, these parameters are quite different from the ones calculated in Subsection 7.4.1.

8.1. Sensitivity analysis of the two-machine model

The approach of parameter identification proposed in Section 5 guarantees that the two-machine line model parameters obtained fit the observed system performance metrics (5) (almost) perfectly. However, it provides no guarantee that this still holds when the machine processing speeds are modified (as discussed in Subsection 7.2) and/or when the buffer capacity is changed (as discussed in Subsection 7.3). Therefore, we calculate the performance metrics of the system based on the hypothesised 'true' parameters (37) or (38) using simulations under the four operation speed-up scenarios in Subsection 7.2, for buffer capacity up to $N = 30$. The results are compared with the performance metrics analytically calculated from the two-machine line model. The errors of throughput and work-in-process are computed based on (20) and plotted in Figure 9.

As one can see, the errors of both performance metrics are low for all cases studied: within 1% for *TP* and 0.7% for *WIP*. This implies that the inference made with the two-machine line model remains valid for the improvement scenarios under the hypothesised system parameters above.

8.2. Sensitivity analysis of the seven-machine model

In the seven-machine model, it is assume that the operations in the same work cell have identical stand-alone throughput (see (25) and (26)) and the operations in the same group have identical downtime (27) and (28)). The hypothesised 'true' system parameters (37) or (38) apparently do not meet these assumptions. Based on these parameters (37) or (38), we calculate again the throughput and work-in-process of the system using simulations for downtime reduction cases at each operation. The

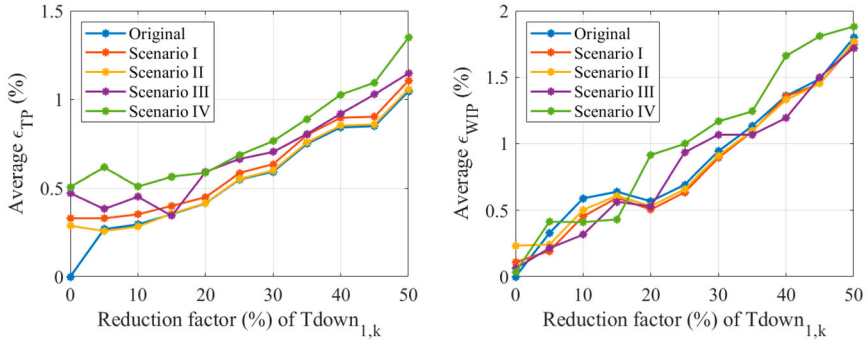


Figure 10. Errors of *TP* and *WIP* by reducing $T_{down,1,k}$ under each scenario ($N = 15$).

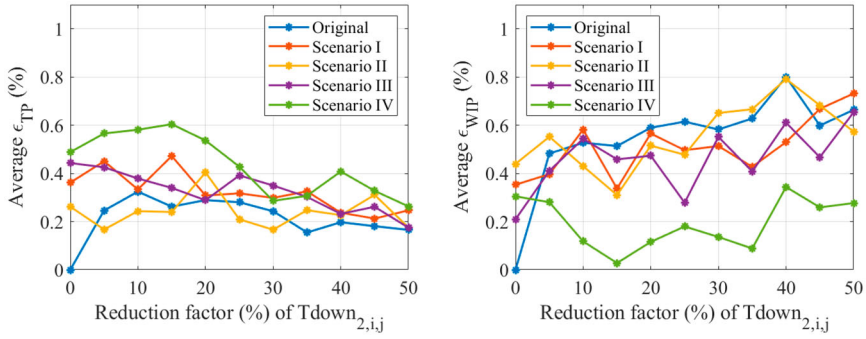


Figure 11. Errors of *TP* and *WIP* by reducing $T_{down,2,i,j}$ under each scenario ($N = 15$).

average errors, when compared to those calculated in Subsection 7.4, are summarised in Figures 10 and 11.

As one can see, the seven-machine model can still provide a highly accurate estimation of both throughput and work-in-process (less 1% errors) when modifying the average downtime of operations in Work Cell 2 by as much as 50%. In the case of operation downtime reduction in Work Cell 1, the errors increase as the T_{down} moves away from the baseline value, because of the larger deviation of hypothesised 'true' parameters of the operations in Work Cell 1 from assumptions (25) and (27). Despite the increasing trend, the system performance estimates still remain in a relatively reasonable range (less than 1.5% error for *TP* and less than 2% error for *WIP*).

8.3. Summary

In this section, we analyse the sensitivity of the model identified using our proposed approach. As demonstrated by the data, this approach can predict the performance metrics accurately not only under the nominal parameters but also in a reasonable range of parameter change even when some of the assumptions about the relationship among individual machine parameters are violated. This property of the modelling approach allows construction of a robust mathematical model for the production system without precise knowledge of T_{up} and T_{down} of individual machines. This is critical in designing continuous improvement projects in manufacturing practice, where the prediction accuracy of the model is the key to the assessment of various what-if scenarios.

Finally, it should be noted again that, implementing the conventional modelling approach (measuring up- and downtimes of the operations) and the algorithms in Li and Meerkov (2009) and Yan and Zhao (2013) requires the up- and downtime data of all seven machines in the two work cells, which may face a number of challenges in practice as discussed in Sections 1 and 3. On the other hand, monitoring and measuring the parts flow data in the buffer is sufficient for the proposed approach to collect the input data, *TP*, *WIP*, P_0, and P_N. This can be more efficient and convenient in practice.

9. Conclusions

In this paper, considering the challenges of collecting and processing the machine status data in small and medium-sized manufacturing plants, we develop a novel approach to modelling a production system, in which parts flow data of in-process buffer are collected and used for identifying the model parameters. This new approach is applied in a case study at a local small manufacturer of medical devices to understand its baseline performance and to design continuous improvement actions to enhance its productivity. To accomplish this, we first

construct the structural and parametric model of the system. In particular, two models are considered: a complete seven-machine model and a simplified two-machine production line model. Then, we formulate the parameter identification problem for the two-machine line model as a constrained optimisation problem that aims at searching for the optimal machine parameters to match the given performance metrics data (*TP*, *WIP*, P_0, and P_N). To solve this optimisation problem, multi-start modified projected quasi-Newton method is developed, which shows great optimisation performance and computational efficiency via numerical experiments. Based on the identified model parameters, we investigate potential *TP* improvement via operation speed-up, buffer expansion, and downtime reduction. These analyses lead to several improvement plans that were handed over to the production personnel. One improvement plan was selected for implementation, resulting in increased throughput as predicted by the analysis. Finally, the efficacy, validity, and robustness of the models are investigated through various numerical experiments of model sensitivity analysis.

The implications of this work are two-fold:

- From the academic perspective, the paper develops a method for machine reliability parameter identification in a production system parts flow model and tests its performance through numerical/statistical experiments and an industrial case study. This work justifies the efficacy of the approach of performance metrics-based manufacturing system modelling and lays the foundation for extending the theoretical methods to more complex system models.
- From the industrial perspective, this work describes a model-based analysis and improvement design approach for a practical manufacturing system. Multiple scenarios were explored based on the model identified. Based on the analysis, recommendations were formulated and offered to the manufacturer, who eventually adopted the recommendations and saw the productivity performance improvement expected from the model. This modelling-analysis-improvement design process can be generalised to other manufacturing systems, where collecting parts flow data in in-process buffers is relatively easy and convenient.

In future work, this new modelling approach will be extended to more complex production system models, such as multi-stage production line models and assembly system models, and generalised to systems with machines following other reliability models (e.g. geometric, Gamma, etc.). Furthermore, we will continue our effort in promoting and applying this approach in real manufacturing systems.

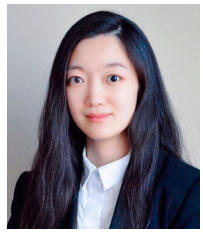
Disclosure statement

Dr. Liang Zhang have financial interests and/or other relationships with Smart Production Systems LLC, Ann Arbor, MI, USA.

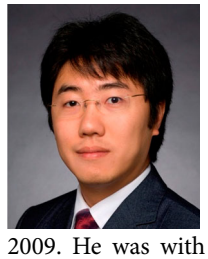
Funding

This work is supported in part by the U.S. National Science Foundation (NSF), [grant number FM-2134367].

Notes on contributors



Yuting Sun received the B.S. and M.S. degrees in Statistics, and the M.S. degree in Industrial Engineering from University of Minnesota Twin Cities, Minneapolis, MN, USA, in 2014 and 2018, respectively. She is currently pursuing the Ph.D. degree with the Smart Production Systems Lab, Department of Electrical and Computer Engineering, University of Connecticut, Storrs, CT, USA. Her research interests include modelling, analysis, improvement and visualisation of smart manufacturing systems, and the application of machine learning, numerical computation and engineering optimisation methods.



Liang Zhang received the B.E. and M.E. degrees from the Center for Intelligent and Networked Systems, Department of Automation, Tsinghua University, Beijing, China, in 2002 and 2004, respectively, and the Ph.D. degree in Electrical Engineering-Systems from the University of Michigan, Ann Arbor, MI, USA, in 2009. He was with the Department of Industrial and Manufacturing Engineering, University of Wisconsin–Milwaukee from 2009 to 2013. He is currently an Associate Professor with the Department of Electrical and Computer Engineering, University of Connecticut, Storrs, CT, USA. His research interests include modelling, analysis, improvement, design, control, and smart operations of manufacturing systems. He is a co-founder of Smart Production Systems LLC.

Data availability statement

The authors confirm that the data supporting the findings of this study are available within the article [and/or] its supplementary materials.

References

Alavian, P., P. Denno, and S. M. Meerkov. 2017. "Multi-Job Production Systems: Definition, Problems, and Product-Mix Performance Portrait of Serial Lines." *International Journal of Production Research* 55 (24): 7276–7301.

Ambani, S., S. M. Meerkov, and L. Zhang. 2010. "Feasibility and Optimization of Preventive Maintenance in Exponential Machines and Serial Lines." *IIE Transactions* 42 (10): 766–777.

Anish, R., and K. Shankar. 2020. "Identification of Nonlinear Structural Parameters Using Combined Power Flow and Acceleration Matching Approaches." In *Advances in Mechanical Engineering*, 1139–1149. Singapore: Springer.

Arinez, J., S. Biller, S. M. Meerkov, and L. Zhang. 2009. "Quality/Quantity Improvement in an Automotive Paint Shop: A Case Study." *IEEE Transactions on Automation Science and Engineering* 7 (4): 755–761.

Bai, Y., J. Tu, M. Yang, L. Zhang, and P. Denno. 2021. "A New Aggregation Algorithm for Performance Metric Calculation in Serial Production Lines with Exponential Machines: Design, Accuracy and Robustness." *International Journal of Production Research* 59 (13): 4072–4089.

Bertsekas, D. P. 2016. *Nonlinear Programming*. Belmont, MA: Athena Scientific.

Biller, S., J. Li, S. P. Marin, S. M. Meerkov, and L. Zhang. 2009. "Bottlenecks in Bernoulli Serial Lines with Rework." *IEEE Transactions on Automation Science and Engineering* 7 (2): 208–217.

Biller, S., S. P. Marin, S. M. Meerkov, and L. Zhang. 2008. "Closed Bernoulli Production Lines: Analysis, Continuous Improvement, and Leanness." *IEEE Transactions on Automation Science and Engineering* 6 (1): 168–180.

Biller, S., S. M. Meerkov, and C.-B. Yan. 2013. "Raw Material Release Rates to Ensure Desired Production Lead Time in Bernoulli Serial Lines." *International Journal of Production Research* 51 (14): 4349–4364.

Chen, G., L. Zhang, J. Arinez, and S. Biller. 2012. "Energy-Efficient Production Systems Through Schedule-Based Operations." *IEEE Transactions on Automation Science and Engineering* 10 (1): 27–37.

Chiang, S.-Y., A. Hu, and S. M. Meerkov. 2008. "Lean Buffering in Serial Production Lines with Nonidentical Exponential Machines." *IEEE Transactions on Automation Science and Engineering* 5 (2): 298–306.

Cox, M. G. 1990. "Mathematical Modelling in Manufacturing Metrology." In *Proceedings of the Twenty-Eighth International*, 533–539. New York, NY: Springer.

Du, S., R. Xu, and L. Li. 2016. "Modeling and Analysis of Multiproduct Multistage Manufacturing System for Quality Improvement." *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 48 (5): 801–820.

Feng, Y., X. Zhong, J. Li, and W. Fan. 2018. "Analysis of Closed-Loop Production Lines with Bernoulli Reliability Machines: Theory and Application." *IISE Transactions* 50 (3): 143–160.

Gershwin, S. B. 1994. *Manufacturing Systems Engineering*. Englewood Cliff, NJ: Prentice Hall.

Jia, Z., and L. Zhang. 2019. "Serial Production Lines with Geometric Machines and Finite Production Runs: Performance Analysis and System-theoretic Properties." *International Journal of Production Research* 57 (8): 2247–2262.

Jia, Z., L. Zhang, J. Arinez, and G. Xiao. 2015. "Finite Production Run-Based Serial Lines with Bernoulli Machines: Performance Analysis, Bottleneck, and Case Study." *IEEE Transactions on Automation Science and Engineering* 13 (1): 134–148.

Jia, Z., L. Zhang, J. Arinez, and G. Xiao. 2016. "Performance Analysis for Serial Production Lines with Bernoulli Machines and Real-time Wip-Based Machine Switch-on/off Control." *International Journal of Production Research* 54 (21): 6285–6301.

Ju, F., J. Li, and W. Deng. 2016. "Selective Assembly System with Unreliable Bernoulli Machines and Finite Buffers." *IEEE Transactions on Automation Science and Engineering* 14 (1): 171–184.

Kim, D., S. Sra, and I. S. Dhillon. 2010. "Tackling Box-Constrained Optimization Via a New Projected Quasi-Newton Approach." *SIAM Journal on Scientific Computing* 32 (6): 3548–3563.

Li, J. 2013. "Continuous Improvement at Toyota Manufacturing Plant: Applications of Production Systems Engineering Methods." *International Journal of Production Research* 51 (23–24): 7235–7249.

Li, J., and S. M. Meerkov. 2009. *Production Systems Engineering*. New York, NY: Springer Science Business Media.

Liberopoulos, G., and P. Tsarouhas. 2005. "Reliability Analysis of an Automated Pizza Production Line." *Journal of Food Engineering* 69 (1): 79–96.

Lima, A. N., C. B. Jacobina, and E. B. de Souza Filho. 1997. "Nonlinear Parameter Estimation of Steady-State Induction Machine Models." *IEEE Transactions on Industrial Electronics* 44 (3): 390–397.

Liu, Y., J. Li, and S.-Y. Chiang. 2010. "Performance Approximation of Re-Entrant Lines with Unreliable Exponential Machines and Finite Buffers." *The International Journal of Advanced Manufacturing Technology* 49 (9–12): 1151–1159.

Liu, X., W. Wang, and R. Peng. 2015. "An Integrated Production, Inventory and Preventive Maintenance Model for a Multi-product Production System." *Reliability Engineering & System Safety* 137: 76–86.

Liu, Y., Q. Zhang, Z. Ouyang, and H.-Z. Huang. 2021. "Integrated Production Planning and Preventive Maintenance Scheduling for Synchronized Parallel Machines." *Reliability Engineering & System Safety* 215: 107869.

Meerkov, S. M., and C.-B. Yan. 2014. "Production Lead Time in Serial Lines: Evaluation, Analysis, and Control." *IEEE Transactions on Automation Science and Engineering* 13 (2): 663–675.

Nocedal, J., and S. Wright. 2006. *Numerical Optimization*. New York, NY: Springer Science & Business Media.

Nwika, C. A., G. Umoh, and F. B. Amaewhule. 2017. "A Method of Reliability Estimation of Individual Section Machine (IS) of Glass Manufacturing Facility; a Systematic Approach with Exponential Model, Availability and Mean Time Between Failures." *International Journal of Emerging Technologies in Engineering Research* 5: 126–131.

Papadopoulos, H. T., C. Heavy, and J. Browne. 1993. *Queueing Theory in Manufacturing Systems Analysis and Design*. London, UK: Chapman & Hill.

Park, K., and J. Li. 2019. "Improving Productivity of a Multi-Product Machining Line at a Motorcycle Manufacturing Plant." *International Journal of Production Research* 57 (2): 470–487.

Peri, D., and F. Tinti. 2012. "A Multistart Gradient-Based Algorithm with Surrogate Model for Global Optimization." *Communications in Applied and Industrial Mathematics* 3 (1): 1–22.

Reddy, T. A., and K. K. Andersen. 2002. "An Evaluation of Classical Steady-State Off-Line Linear Parameter Estimation

Methods Applied to Chiller Performance Data." *Hvac&R Research* 8 (1): 101–124.

Schmidt, M., D. Kim, and S. Sra. 2012. "Projected Newton-Type Methods in Machine Learning." In *Optimization for Machine Learning*. Vol. 1. Cambridge, MA: The MIT Press.

Sun, Y., and L. Zhang. 2020. "Parameter Identification for Multiple-Machine Bernoulli Lines Using Statistical Learning Methods." In *2020 IEEE 16th International Conference on Automation Science and Engineering (CASE)*. IEEE.

Sun, Y., and L. Zhang. 2021. "Parameter Identification for Synchronous Two-Machine Exponential Production Line Model." In *2021 IEEE 17th International Conference on Automation Science and Engineering (CASE)*, 1884–1889. IEEE.

Sun, Y., T. Zhu, L. Zhang, and P. Denno. 2020. "Parameter Identification for Bernoulli Serial Production Line Model." *IEEE Transactions on Automation Science and Engineering* 18 (4): 2115–2127.

Tu, J., Y. Bai, M. Yang, L. Zhang, and P. Denno. 2020. "Real-time Bottleneck in Serial Production Lines with Bernoulli Machines: Theory and Case Study." *IEEE Transactions on Automation Science and Engineering* 18 (4): 1822–1834.

Tu, J., and L. Zhang. 2022. "Performance Analysis and Optimisation of Bernoulli Serial Production Lines with Dynamic Real-time Bottleneck Identification and Mitigation." *International Journal of Production Research* 1–17. doi:10.1080/00207543.2021.2019343.

Tu, J., T. Zhu, Y. Bai, and L. Zhang. 2020. "Estimation of Machine Parameters in Exponential Serial Lines using Feedforward Neural Networks." In *2020 IEEE 16th International Conference on Automation Science and Engineering (CASE)*. IEEE.

Wang, F., and F. Ju. 2021. "Decomposition-Based Real-Time Control of Multi-Stage Transfer Lines with Residence Time Constraints." *IISE Transactions* 53 (9): 943–959.

Weston, B. 2019. "How Small Manufacturing Businesses Drive the U.S. Economy." https://www.score.org/blog/how-small-manufacturing-businesses-drive-us-economy.

Xie, X., and J. Li. 2012. "Modeling, Analysis and Continuous Improvement of Food Production Systems: A Case Study at a Meat Shaving and Packaging Line." *Journal of Food Engineering* 113 (2): 344–350.

Yan, C.-B., and Q. Zhao. 2013. "A Unified Effective Method for Aggregating Multi-Machine Stages in Production Systems." *IEEE Transactions on Automatic Control* 58 (7): 1674–1687.

Yao, D. D. 1994. *Stochastic Modeling and Analysis of Manufacturing Systems*. New York, NY: Springer-Verlag.

Zandieh, M., M. Joreir-Ahmadi, and A. Fadaei-Rafsanjani. 2017. "Buffer Allocation Problem and Preventive Maintenance Planning in Non-Homogenous Unreliable Production Lines." *The International Journal of Advanced Manufacturing Technology* 91 (5-8): 2581–2593.

Zhang, L., and X. Yue. 2011. "Operations Sequencing in Flexible Production Lines with Bernoulli Machines." *IEEE Transactions on Automation Science and Engineering* 8 (3): 645–653.

Zhao, C., and J. Li. 2013. "Analysis of Multiproduct Manufacturing Systems with Homogeneous Exponential Machines." *IEEE Transactions on Automation Science and Engineering* 11 (3): 828–838.

Zhao, C., and J. Li. 2015. "Analysis and Improvement of Multiproduct Bernoulli Serial Lines: Theory and Application." *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 45 (9): 1218–1230.

Zhao, C., J. Li, and N. Huang. 2014. "Performance Evaluation of Multi-Product Manufacturing Systems with Asynchronous Exponential Machines." In *2014 IEEE International Conference on Automation Science and Engineering (CASE)*, 692–697. IEEE.