Classification of queueing models for a workstation with interruptions : a review

Wu, Kan

2013

Wu, K. (2014). Classification of queueing models for a workstation with interruptions: a review. International Journal of Production Research, 52(3), 902-917.

<https://hdl.handle.net/10356/104585>

<https://doi.org/10.1080/00207543.2013.843799>

© 2013 Taylor & Francis. This is the author created version of a work that has been peer reviewed and accepted for publication by International Journal of Production Research, Taylor & Francis. It incorporates referee's comments but changes resulting from the publishing process, such as copyediting, structural formatting, may not be reflected in this document. The published version is available at:
[<http://dx.doi.org/10.1080/00207543.2013.843799>].

Downloaded on 26 Apr 2023 15:28:16 SGT

PROOF COVER SHEET

Journal acronym: TPRS

Author(s): Kan Wu

Article title: Classification of queueing models for a workstation with interruptions: a review

Article no: 843799

Enclosures: 1) Query sheet
2) Article proofs

Dear Author,

1. Please check these proofs carefully. It is the responsibility of the corresponding author to check these and approve or amend them. A second proof is not normally provided. Taylor & Francis cannot be held responsible for uncorrected errors, even if introduced during the production process. Once your corrections have been added to the article, it will be considered ready for publication.

Please limit changes at this stage to the correction of errors. You should not make insignificant changes, improve prose style, add new material, or delete existing material at this stage. Making a large number of small, non-essential corrections can lead to errors being introduced. We therefore reserve the right not to make such corrections.

For detailed guidance on how to check your proofs, please see
<http://journalauthors.tandf.co.uk/production/checkingproofs.asp>

2. Please review the table of contributors below and confirm that the first and last names are structured correctly and that the authors are listed in the correct order of contribution. This check is to ensure that your name will appear correctly online and when the article is indexed.

Sequence	Prefix	Given name(s)	Surname	Suffix
1		Kan	Wu	

Queries are marked in the margins of the proofs.

AUTHOR QUERIES

General query: You have warranted that you have secured the necessary written permission from the appropriate copyright owner for the reproduction of any text, illustration, or other material in your article. (Please see <http://journal-authors.tandf.co.uk/preparation/permission.asp>.) Please check that any required acknowledgements have been included to reflect this.

AQ1	The reference 'Belz (1976)' is cited in the text but is not listed in the references list. Please either delete in-text citation or provide full reference details following journal style [http://www.tandf.co.uk/journals/authors/style/reference/tf_ChicagoAD.pdf].
AQ2	Please check the edits made in the sentence "The uptime 2250 min" and correct if necessary.
AQ3	Please provide the location details for reference 'Adan (2001)'.
AQ4	The reference 'Ashcroft (1950)' is listed in the references list but is not cited in the text. Please either cite the reference or remove it from the references list.
AQ5	The reference 'Asmussen (1992)' is listed in the references list but is not cited in the text. Please either cite the reference or remove it from the references list.
AQ6	The reference 'Buzacott and Shanthikumar (1993)' is listed in the references list but is not cited in the text. Please either cite the reference or remove it from the references list and also provide the city for publisher location.
AQ7	Please check the last page number for reference 'Federgruen and Green (1986)'.
AQ8	The reference 'Gross and Harris (1998)' is listed in the references list but is not cited in the text. Please either cite the reference or remove it from the references list.
AQ9	The reference 'Jaiswal (1968)' is listed in the references list but is not cited in the text. Please either cite the reference or remove it from the references list.
AQ10	Please provide the location details for reference 'Kingman (1965)'.
AQ11	The reference 'Kingman (1965)' is listed in the references list but is not cited in the text. Please either cite the reference or remove it from the references list.
AQ12	Please provide the volume number for reference 'Levy and Kleinrock (1986)'.
AQ13	Please provide the volume number for reference 'Neuts and Lucantoni (1979)'.
AQ14	Please provide the editors name and page range for reference 'SEMI E10 (2001)'.
AQ15	Please check the last page number for reference 'Takagi (1988)'.
AQ16	Please provide the publisher and location, page range, and editors name for reference 'Takagi (2000)'.
AQ17	The reference 'Wolff (1982)' is listed in the references list but is not cited in the text. Please either cite the reference or remove it from the references list.
AQ18	Please provide the location for reference 'Wu et al. (2008)'.
AQ19	Please check the equations have been correctly formatted throughout the article.

How to make corrections to your proofs using Adobe Acrobat

Taylor & Francis now offer you a choice of options to help you make corrections to your proofs. Your PDF proof file has been enabled so that you can edit the proof directly using Adobe Acrobat. This is the simplest and best way for you to ensure that your corrections will be incorporated. If you wish to do this, please follow these instructions:

1. Save the file to your hard disk.
2. Check which version of Adobe Acrobat you have on your computer. You can do this by clicking on the "Help" tab, and then "About."

If Adobe Reader is not installed, you can get the latest version free from <http://get.adobe.com/reader/>.

- If you have Adobe Reader 8 (or a later version), go to "Tools" / "Comments & Markup" / "Show Comments & Markup."
- If you have Acrobat Professional 7, go to "Tools" / "Commenting" / "Show Commenting Toolbar."

3. Click "Text Edits." You can then select any text and delete it, replace it, or insert new text as you need to. If you need to include new sections of text, it is also possible to add a comment to the proofs. To do this, use the Sticky Note tool in the task bar. Please also see our FAQs here: <http://journalauthors.tandf.co.uk/production/index.asp>.

4. Make sure that you save the file when you close the document before uploading it to CATS using the "Upload File" button on the online correction form. A full list of the comments and edits you have made can be viewed by clicking on the "Comments" tab in the bottom left-hand corner of the PDF.

If you prefer, you can make your corrections using the CATS online correction form.

Classification of queueing models for a workstation with interruptions: a review

Kan Wu*

School of Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore, Singapore

(Received 15 July 2012; accepted 20 August 2013)

Queueing theory is commonly used to evaluate the performance of production systems. Due to the complexity of practical production lines, choosing correct queueing models under the existence of interruptions can be intricate, but is critical to the evaluation of system performance. This paper gives a review of research in this area, proposes a systematic way to classify different types of interruptions in manufacturing systems and suggests proper queueing models for each category. The operational definition of service time is given and the decomposition properties for the run-based state-induced and time-based ~~pre-emptive~~ events are explained. The corresponding G/G/1 approximate models are proposed.

Keywords: manufacturing system; Queueing theory; classification

1. Introduction

In applying queueing theory to evaluate the performance of manufacturing systems, M/G/1 and G/G/1 models are widely used. For the M/G/1 queue, the Pollaczek–Khintchine (P–K) formula determines average cycle time (CT) in a single server with Poisson arrivals. Although each arrival process may not be Poisson, based on the Palm–Khintchine Theorem, the overall arrival process could be approximated by a Poisson process when a workstation is fed by a large number of independent and renewal processes. The study of the G/G/1 model can be traced back to Lindley (1952) and a complete solution of the equilibrium queue time (QT) distribution was obtained by Smith (1953). However, the solution is not easy to evaluate in general. By observing the upper bound derived by Kingman (1962), Heyman (1975) derived the heavy approximation for G/G/1 queue, the so-called Kingman's approximation. Kramer and Lagenbach-Belz (1978) proposed an improved G/G/1 model based on a heuristic extension of P–K formula and Kingman's approximation.

Although the M/G/1 and G/G/1 models point out a way to evaluate the performance of manufacturing systems, they do not suffice when systems are subject to interruptions, which are common in practical manufacturing systems. Manufacturing applications of queueing models, which incorporate interruptions, have been studied since White and Christie (1958). Gaver (1962), Keilson (1962), Avi-Itzhak and Naor (1963), and Thiruvengadam (1963) studied the M/G/1 models for queues subject to service interruptions. Federgruen and Green (1986) and Sengupta (1990) generalised the M/G/1 ~~pre-emptive~~-resume models with general interruption arrivals. Hopp and Spearman (2008) classified interruptions as ~~pre-emptive~~ and ~~nonpre-emptive~~, and incorporated those interruptions with the true process time to create a modified process time. Adan and Resing (2001) gave M/G/1 models for unreliable machines and machines with set-up times. Boxma, Mandjes, and Kella (2008) extended earlier results to the situation that allows the length of an off-period to depend on the length of the previous on-period.

Because ~~resource is~~ limited in practice, machine interference (or repairman) problems attempt to model the impact from limited ~~resources~~ in production lines. Interference refers to undesirable waiting caused by using limited ~~resources~~ (such as repairmen or operators) to serve multiple machines. Research on this topic has been reported since 1930s (Khintchine 1933). Stecké and Aronson (1985) and Haque and Armstrong (2007) have gave comprehensive reviews on this topic.

Queueing models launched in other fields could also be applied to manufacturing systems. Queueing systems with vacations were inspired by the computer and communication systems where servers require maintenance after each busy period. A generalised vacation queue covers all types of queueing systems where servers switch between production and nonproduction modes. Hence, interruptions could become special cases of vacation queues. Doshi (1986) gave a

*Email: kan626@gmail.com

Figure 1. Summary of time from SEMI E10 (2001).

survey on this topic and gave Laplace–Stieltjes transforms for G/G/1-based vacation models. An important property of vacation queues is the stochastic decomposition property, first demonstrated by Gaver (1962) and explained by Fuhrmann and Cooper (1985) and Doshi (1986).

A polling model is a system of multiple queues accessed by a single server based on a prescribe manner. An extensive study on polling model was carried out to model computer communication networks in 1970s and local-area networks in 1980s. Takagi (1988, 2000) summarised the applications of polling models. Abundant queueing models have been proposed cross-applications. To effectively apply those models to solve practical manufacturing problems, proper classification schemes should be proposed based on the fundamental property of queues in manufacturing systems rather than on the mathematical property.

Our classification is inspired by the interruptions observed in a semiconductor fab. In Figure 1, the total time of a workstation has been decomposed into smaller portions from the view point of productivity. Based on Figure 1, a workstation is subject to various types of interruptions, such as breakdowns, set-ups, routine maintenance, etc. As explained in Section 2, those interruptions can be time-based or run-based, and pre-emptive or nonpre-emptive. The resource to handle the interruptions can be limited or ample. Thus, a critical issue in performance evaluation is to choose the right model when machines are subject to various types of interruptions.

In practical production lines, resource for productivity improvement is limited in general. The correct evaluation of systems performance is important for managers to allocate the limited resource. A systematic classification is crucial in this aspect. For example, pre-emptive breakdowns, set-ups and preventive maintenance (PM) are commonly seen in production lines but have totally different impact on QT. As we will see in Section 4, a run-based breakdown has the most direct impact on QT and a state-induced set-up has limited impact in heavy traffic. Since a PM has no impact on service time variability, its impact on system QT is relatively small.

This paper is written for both practitioners and researchers. For practitioners, we clarify the definitions of availability and service times from the view point of queueing models and point out the queueing models appropriate for different interruption types under practical production settings. Practitioners can treat this paper as a guideline and choose the correct model in each identified situation. For researchers, we give the literature review for interruption queueing models, classify prior results and propose the G/G/1 approximate models for different types of interruptions.

This paper is organised as follows: We classify different event types from the perspective of queueing theory in Section 2 and clarify the definitions of service time in Section 3. The queueing models for a single machine under each specific interruption category are proposed in Section 4. Conclusion is given in Section 5.

2. Classifying events

The behaviour of manufacturing systems is the mutual effects among three basic entities: jobs, machines and resources, under the influence of interruption events. Resources are used to support the intended functions (e.g. job processing) when machines are in the production mode, or to support the recovery from interruptions when machines are in the

nonproduction mode. Typical resources are operators, engineers or tool kits. Depending on the situations, resource can be ample or limited. Our classifications start with the assumption of ample resource, and then are modified to include situations under resource contention.

2.1 Classification under ample resource

Queueing theory predicts system delay performance under the existence of randomness. When resource is ample, randomness comes from natural variability of job inter-arrival times and machine service times in the production mode, or from interruptions, which will switch machine states to the nonproduction mode. Natural variability refers to the randomness caused by the system performing its intended functions. The sources of randomness come from natural properties designed into or inherent to the system. Hopp and Spearman (2008) pointed out that there is typically more natural variability in a manual process than in an automated one. However, even in fully automated machine operations, the service time variations can be caused by robot motions, product mix or the composition of the material, etc.

Interruptions are events whose occurrence prevents the system from performing its intended function. They are inherent in manufacturing systems and have negative impacts on productivity. A majority of interruptions are caused by downtime events. Based on the specifications in semiconductor industry, SEMI E10 (2001, 1), downtime is 'the time when the equipment is not in a condition, or is not available, to perform its intended function.' Examples are break-downs, process or equipment experiments, PM and set-ups. There are two types of down states: unscheduled downtime and scheduled downtime. Since scheduled downtime implies controllability when it happens, it is nonpre-emptive in nature. Since unscheduled downtime is not controllable, it is pre-emptive in general. Furthermore, failure in SEMI E10 (2001, 1) is defined as 'any unscheduled downtime event that changes the equipment to a condition where it cannot perform its intended function.' Hence, we specifically define failures as pre-emptive interruptions. Failures will decrease the availability of machines, but scheduled downtime events will not from the view point of queueing models. Since the occurrence of scheduled downtime events is under control, it will have no impact on service time. For example, in the auto industry, the production line commonly has a two-hour break between any two shifts and the entire production line is shut down for necessary maintenance. Hence, the nonpre-emptive event has no impact on service time. This clarification for availability definition is crucial to the correct use of queueing models in practical manufacturing systems. More rigorous definition will be given in Section 3.

Buzacott and Hanifin (1978) have classified interruptions as run-based or time-based events. Run-based events refer to the interruptions induced by processing jobs, while time-based events are associated with time. That is, the occurrence of a time-based interruption is a function of time, and the occurrence of a run-based interruption is a function of job arrivals. Since job arrivals are also a function of time, a run-based event can be viewed as a special case of a time-based event. Run-based events occur during or right before (or after) processing, but time-based events could occur anytime, whether or not the machine is busy. For example, power outages caused by lightning are time-based. Set-ups due to change of recipes are run-based events. Typical examples are:

- (1) Run-based events: out-of-spec input, set-up.
- (2) Time-based events: power outage, PM, experiment.

Both run-based and time-based events can be further classified as pre-emptive or nonpre-emptive. A pre-emptive event can occur during processing, but a nonpre-emptive event can only occur before or after processing. Hence, for nonpre-emptive events, the occurrence of an event and its impact on machine status (i.e. interruptions) can be at different time epochs. For example, the thickness of accumulated deposit on the chamber exceeds the limit during processing, but PM will be conducted when the machine is idle. Since what we want to model is essentially the period when the event has impact on machine status, instead of modelling the deposit thickness, we can simply transform the deposition process into a usage-based problem (i.e. conduct PM after 2000 jobs of usage).

A run-based nonpre-emptive event can only occur before job processing starts or after job processing ends, since it cannot pre-empt processing yet is associated with the processing of jobs. Set-ups are typical examples of this type. Run-based nonpre-emptive events can be further classified as state-induced and product-induced events. State-induced events correspond to machine changing state from idle to busy. For example, a machine that goes into a sleep mode when it is idle requires a warm-up period when it returns to production mode. It experiences a state-induced interruption. Product-induced events correspond to switching machine settings for different products. For example, a machine may need some set-up time when switched from one recipe to another. State-induced set-ups vanish when a machine is fully utilised, but product-induced set-ups are driven by customer demands and cannot be completely avoided.

Figure 2. Classification of factors which affect QTs.

For the other event types, out-of-spec inputs are run-based ~~pre-emptive~~ events, power outages are time-based pre-emptive events, and equipment or process experiments are time-based ~~nonpre-emptive~~ events. Note that the occurrence of PM can be associated with process usage (i.e. total job processed since the last PM), time or both. For example, a car needs an oil change every 3000 miles or 6 months so PM can be run-based (3000 miles) or time-based (6 months) or both. If an event is both run- and time-based, we can classify it as a time-based event, since a run-based event is special case of time-based events, and could be approximated by a time-based event in the steady state. For example, when the mean arrival rate is 100 jobs per day and a special PM is needed whenever 2000 jobs are processed, the PM is needed every 20 days on average. The above classifications are summarised in Figure 2.

An interruption may or may not occur during another interruption. An out-of-spec input will not occur when a machine is down for another out-of-spec input, and a set-up will not occur during another set-up. It is also unlikely that a power outage occurs during a power outage. However, due to its long duration and multiple resources (e.g. PM and experiments, etc.), a time-based ~~nonpre-emptive~~ event may possibly occur during another time-based ~~nonpre-emptive~~ event. Hence, except for the time-based ~~nonpre-emptive~~ events, we assume no simultaneous interruption exists for all other event types.

2.2 Classification under resource contention

When ~~resource is~~ limited, machines suffer resource contention. Our interest focuses on the impact on job QT due to the undesirable machine QT, which occurs when ~~resource (r) is~~ less than machines (n) as shown in Figure 3.

The upper part in Figure 3 shows the queuing model of a resource contention system. Jobs arrive in job queue and then are processed by one of the n machines. When a machine needs service, it goes into the machine queue and then is served by one of the r resources. The lower part in Figure 3 shows the relation among entities, corresponding to the upper graph. Jobs need to be processed by machines. Machines can be in either production mode or nonproduction mode.

When it is in the production mode, it may process a job automatically by a robot, or manually by an operator. In the latter case, a job can only be processed with the attendance of an operator. If no operator is available, the machine has to wait in the machine queue. If a machine is in the nonproduction mode (i.e. suffering interruptions), depending on

Figure 3. Graphical explanation for resource contention problems.

Figure 4. Classification of resource contention problems.

the type of interruptions, it may need an engineer to fix the breakdown or an operator to perform a set-up. Furthermore, operators or engineers may suffer time-based nonpre-emptive interruptions (e.g. taking breaks).

If the recovery of a breakdown needs a precious tool kit, an extra resource queue loop should be added to the graph. This argument can be carried on. For example, one may claim operators and engineers share a water fountain. Exact analysis for real situations is almost impossible and may not be necessary.

Controllability of interruptions plays an important role in resource contention problems. The occurrences of some types of interruptions are controllable. Ideally, equipment experiments can be properly modelled as ample resource problems if we can choose not to start an experiment until resources, such as engineers and parts, are ready. Product-induced set-ups have the similar property, since we can run the same product until operators are ready to do a set-up. In the ideal situations, the classification for resource contention problems is summarised in Figure 4. The categories 2-1-1, 2-2-1 and 2-2-2 in Figure 4 are the resource contention counterparts of the categories 2-1-2.a, 2-2-1 and 2-1-1 in Figure 2, respectively.

2.3 Classification under ample resource in prior research

The idea to classify interruptions based on the fundamental properties has been addressed by Gaver (1962). He distinguished interruptions as active and independent, where each case is further classified as pre-emptive or postponable. However, he did not further classify the run-based nonpre-emptive events as state-induced or product-induced, but simply modelled them as Poisson arrivals during process times. Furthermore, for the time-based nonpre-emptive events, he assumed no event could occur during an interruption (e.g. the needs of PM could not occur during an experiment), and all interruptions had high priority, which were not always true in practice.

Keilson (1962) discussed three types of interruptions: postponable, pre-emptive-resume and pre-emptive-repeat disciplines under M/G/1 assumptions. Avi-Itzhak and Naor (1963) presented the extensions of M/G/1 models for five different types of interruptions. Compared with our classifications, their Model A is indeed the same as time-based pre-emptive events, and Model B is the same as run-based pre-emptive events. However, except for the above two models, the rest of their models focus on the applications in telecommunication instead of manufacturing systems. Thiruvengadam (1963) also generalised the M/G/1 models for three different types of interruptions. Model II models a time-based pre-emptive event and Model III models a run-based pre-emptive event.

Doshi (1986) gave a survey on queueing systems with vacations and classified the models for computer, communication and production systems. Three types of interruptions are covered by those vacation models: product-induced, state-induced and time-based pre-emptive events. They are named as cyclic server queues, set-up time models and pre-emptive vacations, respectively. Cyclic server queues are polling models with cyclic order assumptions. However, in general, queue-tending sequences are not cyclic in manufacturing systems.

Table 1. Comparison of different classifications for ample resource models.

Classification in Figure 2	Gaver	Keilson	Avi-Itzhak and Naor	Thiruvengadam	Doshi	Hopp and Spearman	Adan and Resing	Wu
2-1-1	●		●	●		●		●
2-1-2.a					●		●	●
2-1-2.b					○			●
2-2-1	●	●	●	●	●		●	●
2-2-2	○	○			○			●

Hopp and Spearman (2008) classified interruptions as ~~pre-emptive~~ and ~~nonpre-emptive~~, but failed to differentiate time-based and run-based interruptions, nor state-induced and product-induced set-ups. Adan and Resing (2001) gave M/G/1 models for two types of interruptions: unreliable machines and machines with set-up times. However, their unreliable machine models only considered time-based ~~pre-emptive~~ events, and their set-up models only considered the state-induced set-ups.

The prior classifications for ample resource problems have been compared with our classification in Table 1. When the model has been fully addressed before, it is noted as a fully filled circle. When the model has only been incompletely addressed (i.e. with less practical assumptions), it is noted as a partially filled circle.

2.4 Resource contention in prior research

The classification for resource contention problem is not noted to the best of our knowledge. Based on Figure 3, job QT is affected by the availability of machines while machine availability is affected by resources, and a resource can be affected by other resources. However, if we limited our attention to machine (not job) QTs caused by a specific resource, the machine interference problems have been addressed by abundant prior research, e.g. Stecké and Aronson (1985) and Haque and Armstrong (2007). The classification of Haque and Armstrong is based on the basic assumptions of queueing terms (such as inter-arrival time and service time distributions), which brings limited insight to the application of queueing models in manufacturing systems.

If ~~resource is~~ not always available when needed, it should be modelled to reflect the extra QT. Based on the discussion in Section 2.2, the resource contention problems can occur in either production mode or nonproduction mode.

In the production mode, we have scenario 1–1 in Figure 4. In this case, each job needs to be processed by one of the r operators on one of the n machine (where $r < n$). If the arrival process is Poisson and service time is generally distributed, the model reduces to an M/G/r time-based ~~nonpre-emptive~~ event model. The interruption can be nonpre-emptive, if we assume operators will not take breaks in the middle of job processing.

In the nonproduction mode, by assuming no interruption (but only natural variability) on repairman, Neuts and Lucantoni (1979) modelled 2–2–1 as a particular case of the M/M/N queue under an Markovian setting. The model for more general settings is not found. No model is available for the run-based scenarios of 2–1–1 and 2–2–2 to the best of our knowledge. The resource contention models are very difficult to model due to the following reasons: (1) when the interruption is run-based, the arrival process at the resource depends on the machine status (i.e. busy or idle); (2) in addition to machines, ~~resource~~ may also suffer interruptions (e.g. taking breaks or warm-up); and (3) since the M/G/N model is analytically intractable, more general settings for resource contention are unlikely to attain.

Although the analytical models are difficult to obtain, not all interruptions have to be modelled as resource contention problems, especially when their occurrence are controllable as we have explained in Section 2.2 (e.g. 2–1–2, b and 2–2–2 in Figure 2). Fortunately, a large portion of interruptions on shop floors belong to 2–2–2 according to SEMI E10 (2001).

Since exact analytical models for the resource contention problems do not exist, we have to resort to approximations. One possibility is simply approximating resource contention by its ample resource counterpart. In this situation, the ample resource model gives a lower bound to the resource contention problem. This approach should work well when the cost of ~~resource~~ is relatively low and the resource is indeed close to ample. Morrison and Martin (2007) treated a resource contention problem as a special case of the run-based ~~pre-emptive~~ event. Comparing their approximate model with the simulation results, the errors are reported from 1.4 to 21.3%.

While analysing resource contention problems with general settings seems formidable, analysing ample resource problems is relatively easier. By giving a comprehensive classification, we will explain how to use ample resource models correctly in practical production environment in the following sections.

3. Definition of service time

The fundamental parameters in queueing models are service time distribution, inter-arrival time distribution and number of servers. For example, in an M/M/1 model, the arrival rate and service rate are the only two parameters when we want to compute QT. The mean service time in an M/M/1 model is just the inverse of capacity, where capacity is the potential maximum throughput rate that a system could achieve in the long run. The concept of service time in queueing theory is indeed derived from capacity. While inter-arrival time and server count are clearly defined in practice, the definition of service time sometimes causes confusions in the presence of interruptions.

Under the existence of interruptions, the concept of service time has to be generalised. From the view point of capacity, generalised service time (GST) is defined as

$$\text{GST} = \text{Job departure time} - \text{The time epoch when the job first claims capacity of the machine} \quad (1)$$

where job departure time is the time epoch that a job releases machine capacity. A job claims capacity of a machine if (i) the job is present at the machine, (ii) the preceding job has released machine capacity, and (iii) the machine is ready to process this job, or is ready to be engaged in a product-induced event (2–1–2.b).

If a job arrives when the machine is down, it cannot claim capacity until the machine is ready for production, or to be engaged in a product-induced event. Furthermore, the set-up times caused by product-induced events are counted into GST, but the set-up times caused by state-induced events are not, since they have no impact on capacity.

By definition, GST is the summation of product-induced set-up time, service time and the downtimes of all ~~pre-emptive~~ events occurring during that service time,

$$G = S + \sum_{i=1}^{N(S)} D_i + T_p, \quad (2)$$

where G is GST, S is service time, $N(S)$ is the number of ~~pre-emptive~~ events during S , D_i is the i -th downtime and T_p is the duration of a run-based ~~nonpre-emptive~~ product-induced event experienced by a job. $N(S)$ is a counting process. G , S , D_i and T_p are mutually independent and are identically distributed (i.i.d.) random variables. Note that both time-based and run-based ~~pre-emptive~~ interruptions will be counted into GST if they occur during S . The state-induced set-up time (2–1–2) does not go into GST since it vanishes in heavy traffic and thus has no impact on capacity. Furthermore, GST is independent of utilisations (or arrival rates) under both run-based and time-based events. When no other failure occurs during a ~~pre-emptive~~ interruption, the mean of GST is

$$\begin{aligned} E(G) &= E\left(S + \sum_{i=1}^{N(S)} D_i\right) + E(T_p) = E_S \left[E_N \left[E \left[S + \sum_{i=1}^{N(S)} D_i \mid N(S) \right] \mid S \right] \right] + E(T_p) = E_S [S + m_r \cdot E[N(S) \mid S] + E(T_p)] \\ &= E(S)[1 + m_r/m_f] + E(T_p) = E(S)/A + E(T_p), \end{aligned} \quad (3)$$

$$A = \frac{m_f}{m_f + m_r} = \frac{1/\eta}{1/\eta + E(D)} = \frac{1}{1 + \eta E(D)}, \quad (4)$$

where A is availability, D is the downtime of the ~~pre-emptive~~ events during the service period, η is machine failure rate, m_f is mean time to failures (MTTF) after repair and m_r is mean time to repair (MTTR) or $E(D)$. Based on Equation (4), only failures (or ~~pre-emptive~~ events) decrease availability, but not other types of interruption do.

Equation (3) is different from effective process time (EPT) proposed by Hopp and Spearman (2008). EPT is ‘the total time seen by a job at a station’, which is Equation (1) with conditions (i) and (ii) only. Although the equation of EPT looks the same as Equation (3), the interpretation of variables is different. EPT considers all ~~pre-emptive~~ breakdowns no matter it occurs during service or idle periods, while the breakdowns in Equation (3) only consider the ~~pre-emptive~~ events during the service period. This subtle difference introduces systematic errors to EPT under time-based events and makes their approximation only applicable to run-based events. Since, under time-based events, their approach will cause dependency between arrivals and EPT. For example, if a job arrives during a time-based interruption, the waiting time will be counted into EPT (but not in GST). Higher arrival rates imply higher probability to see the downtime. Wu and Hui (2008) have detailed discussion on the issues caused by EPT. Their findings eventually lead to the concept of GST in this paper.

The definition of service time is different from process time. Process time is the period that a job stays with a machine, while mean service time is the inverse of machine capacity. This difference can lead to different values of service time and process time in practice. When applying queueing theory, we should stick to the definition of service time rigorously.

4. Models for a single server with interruptions and ample resource

For practical needs, we unified the prior single-server models into an easily accessible form under our classification scheme. Starting with M/G/1 models, we then propose the G/G/1 approximations. We assume service time, inter-arrival

8

K. Wu

time, uptime and downtime are i.i.d. and mutually independent of each other. Dispatching rules are first-come-first-serve if not specified.

4.1 Run-based ~~pre-emptive~~ interruptions with Poisson arrivals

For run-based ~~pre-emptive~~ events (2–1–1), a machine breaks down only when it is processing jobs. In the M/G/1 run-based ~~pre-emptive~~ event model, jobs arrive according to a Poisson process with rate λ . Service time and downtime are generally distributed. The service time devoted to a job before an interruption is not lost and service is resumed once the machine recovers from failure. The time to failure between two interruptions is exponentially distributed with mean $1/\eta$. Since a run-based breakdown can only occur during processing, the time to failure in a run-based model only considers the elapsed time during processing. For stability, we assume $\rho_G = \lambda E(G) < 1$. The mean QT and CT are

$$E(QT) = \frac{\rho_G E(R_G)}{(1 - \rho_G)} = \left(\frac{1 + c_G^2}{2} \right) \left(\frac{\rho_G}{1 - \rho_G} \right) E(G), \quad (5)$$

$$E(CT) = E(QT) + E(G), \quad (6)$$

where $\rho_G = \lambda E(G)$, $E(R_G) = E(G^2)/2E(G)$; $E(G) = E(S)/A$, $E(G^2) = E(S^2)[1 + \eta E(D)]^2 + E(S)\eta E(D^2)$, $c_G^2 = c_0^2 + (1 + c_r^2)A(1 - A)E(D)/E(S) = (E(G^2) - E(G)^2)/E(G)^2$ G is GST; R_G is the residual service time of G ; S is service time; D is downtime; and c_G , c_0 and c_r are the coefficient of variation (CV) of GST, service time and downtime, respectively. The first equality in Equation (5) is given by Gaver (1962). The rightmost term is given by Hopp and Spearman (2008). These two forms are indeed identical, because

$$E(R_G) = E(G^2)/2E(G) = (1 + c_G^2)E(G)/2. \quad (7)$$

4.2 Run-based ~~nonpre-emptive~~ interruptions with Poisson arrivals

Set-ups and PM are common resources of run-based ~~nonpre-emptive~~ interruptions in production systems. There are two types of set-up: state-induced or product-induced. Their models will be introduced in the following two sections respectively. The run-based ~~nonpre-emptive~~ events caused by PM could also be modelled as a product-induced set-up as explained in Section 4.2.2.

4.2.1 State-induced set-ups

State-induced set-ups are commonly seen on shop floors. For example, a machine is turned off when it is idle, and turned on again when a new job arrives. However, restarting encounters a delay. This delay can be a qualification process or simply a warm-up. The M/G/1 run-based ~~nonpre-emptive~~ queue with state-induced events was first studied by Welch (1964). Jobs arrive according to a Poisson process with rate λ . Service time (S) and set-up time (T_s) are generally distributed with mean $1/\mu$ and $1/\theta$, respectively. For stability, we assume $\rho = \lambda E(S) < 1$. The mean QT and CT are (Adan and Resing 2001),

$$E(QT) = \frac{\rho E(R_S)}{(1 - \rho)} + \frac{1/\lambda}{1/\lambda + E(T_s)} E(T_s) + \frac{E(T_s)}{1/\lambda + E(T_s)} E(R_T), \quad (8)$$

$$E(CT) = E(QT) + E(S), \quad (9)$$

where $E(R_S) = E(S^2)/2E(S)$, and $E(R_T) = E(T_s^2)/2E(T_s)$.

From Equation (8) and the stability condition, it can be seen that state-induced events only affect QT but not capacity. The mean QT of an M/G/1 state-induced event model is simply the mean QT of an M/G/1 model plus an extra gap caused by the state-induced set-up. That is

$$\text{Gap} = \frac{1/\lambda}{1/\lambda + E(T_s)} E(T_s) + \frac{E(T_s)}{1/\lambda + E(T_s)} E(R_T). \quad (10)$$

Theorem 1 (Decomposition property of state-induced set-ups)

Under Poisson arrivals, the gap between the mean QTs of an M/G/1 queue and the M/G/1 queue with state-induced set-ups is characterised by Equation (10).

This decomposition property has been formally proved by Levy and Kleinrock (1986). When λ approaches μ , QT of the M/G/1 model goes to infinity but the gap is bounded. State-induced set-ups have less impact on QT in heavy traffic.

4.2.2 *Product-induced set-ups*

Product-induced set-ups occur due to changes in production processes induced by switching products. Inspired by seemingly similar situations in computer systems, product-induced set-ups are sometimes modelled by polling models. Under the assumptions of cyclic order and Poisson arrivals, Takagi (1988, 2000) summarised the polling models under four dispatching rules: exhaustive, gated, limited and decrementing service. The cyclic order can also be replaced by a periodic service order table, or a polling table (Baker and Rubin 1987; Eisenberg 1972). However, both the assumptions of cyclic order or polling tables are not very general in practice.

In practical production lines, switching products may be attributed to one of the following conditions: (1) the daily move target has been achieved, (2) the other product has a more urgent due date, or (3) the downstream bottleneck is starving for the other product, conditioning on the queue length of the current processing product. For example, when the queue length of the current product is zero, the next product could be chosen based on their due dates. However, even if the queue length is not zero, a switch may still occur if the due date is urgent enough (e.g. a super hot lot in a fab).

Since a product-induced set-up is induced by multiple independent random factors, it is reasonable to assume the occurrence of a product-induced set-up is memoryless and follows geometric distributions, i.e. the workstation processes an average of N_p jobs between two consecutive set-ups, and the probability of doing a set-up after any job is $1/N_p$. In this situation,

$$\text{GST} = S + T_p, \quad (11)$$

$$E(G) = E(S + T_p) = E(S) + E(T_p) = E(S) + E(P)/N_p, \quad (12)$$

where P is the product-induced set-up times and T_p is the product-induced set-up time experienced by each job.

For stability, we assume $\rho_G = \lambda E(G) < 1$. The comparison of the stability conditions between the state-induced and product-induced models reveals that the set-up times of product-induced models have direct impact on capacity, while the set-up times of state-induced models do not. The mean QT and CT are

$$E(\text{QT}) = \frac{\rho_G E(R_G)}{(1 - \rho_G)} = \left(\frac{1 + c_G^2}{2} \right) \left(\frac{\rho_G}{1 - \rho_G} \right) E(G), \quad (13)$$

$$E(\text{CT}) = E(\text{QT}) + E(G). \quad (14)$$

where $E(R_G) = E(G^2)/2E(G)$, $E(G^2) = E(S^2) + 2E(S)E(T_p) + E(T_p^2)$, $c_G^2 = \left(\sigma_0^2 + \frac{\sigma_p^2}{N_p} + \frac{N_p - 1}{N_p^2} E(P)^2 \right) / E(G)^2$, or $c_G^2 = (E(G^2) - E(G)^2) / E(G)^2$, c_G is the CV of GST, and σ_p is the standard deviation of product-induced set-up times. In Equation (13), the right-hand side is given by Hopp and Spearman (2008). Based on Equation (7), these two forms are indeed identical.

The product-induced set-up model can be applied to more general situations than its name. For example, the spark plug of a car needs to be replaced every 60,000 miles. It is obviously a run-based ~~nonpre-emptive~~ interruption, since one would not change the spark plug when driving on the highway and it will not reach 60,000 miles unless one is driving. If the duration between replacements is exponentially distributed with mean 60,000 miles and the mileage you drive the car each time is fixed, we can determine the value of N_p and apply it to Equation (13). However, the memoryless property is not always a proper assumption. When facing generally distributed durations, an alternative is to approximate the performance of a run-based ~~nonpre-emptive~~ event by a time-based ~~nonpre-emptive~~ model (which will be given in Sections 4.4 and 4.5) as explained in Section 2.1.

4.3 Time-based ~~pre-emptive~~ interruptions with Poisson arrivals

For time-based ~~pre-emptive~~ events (2–2–1), a machine can break down anytime. In the M/G/1 time-based ~~pre-emptive~~ event model, jobs arrive according to a Poisson process with rate λ . Service time and downtime are generally distributed. The uptime between two breakdowns is exponentially distributed with mean $1/\eta$. Since a time-based breakdown can occur anytime, the time to failure in a time-based model considers the overall elapsed time. For stability, we assume $\rho_G = \lambda E(G) < 1$. The mean QT and CT are

$$E(QT) = \frac{\rho_G E(R_G)}{(1 - \rho_G)} + (1 - A_{NP})E(R_D), \quad (15)$$

$$E(CT) = E(QT) + E(G). \quad (16)$$

where $\rho_G = \lambda E(G)$, $E(R_G) = E(G^2)/2E(G)$, $E(G) = E(S)/A$, $E(G^2) = E(S^2)[1 + \eta E(D)]^2 + E(S)\eta E(D^2)$, $E(R_D) = E(D^2)/2E(D) = \left(\frac{1+c_D^2}{2}\right)E(D)$, $A_{NP} = \frac{1/(\lambda+\eta)}{1/(\lambda+\eta)+E(D)\eta/(\lambda+\eta)} = \frac{1}{1+E(D)\eta}$, S is service time, D is downtime and A_{NP} is availability of the machine during nonprocessing period. Note that the uptime in A considers the complete time horizon, which is different from the time to failure in the run-based ~~pre-emptive~~ event.

Equation (15) was first derived by Gaver (1962) with the variance of QT. Note that based on Equations (2) and (15), some of the time-based ~~pre-emptive~~ interruptions go into GST (when it occurs during service times) and some of them go into QT (when it occurs during idle periods). However, all time-based ~~pre-emptive~~ interruptions will go into GST and become run-based events in heavy traffic.

Comparing Equations (15) with (5), the gap between M/G/1 time-based and run-based ~~pre-emptive~~ event models is

$$\text{Gap} = (1 - A_{NP})E(R_D). \quad (17)$$

Theorem 2 (Decomposition property of time-based ~~pre-emptive~~ events)

Under Poisson arrivals, the gap between the mean QTs of (2–1–1) an M/G/1 queue with run-based ~~pre-emptive~~ interruptions and (2–2–1) the M/G/1 queue with time-based ~~pre-emptive~~ interruptions is characterised by Equation (17), and it is independent of job arrival rates.

Since the gap is independent of utilisation, QT in Equation (15) is dominated by the utilisation term (i.e. its run-based model) instead of the gap in heavy traffic. From Properties 1 and 2, we observe that stochastic decomposition is rooted in the fact that a more complex model approaches a simpler one in heavy traffic, i.e. in Theorem 1, the state-induced set-up vanishes, and in Theorem 2, the time-based event becomes a run-based one. The difference between these two models constitutes the decomposition property. Later, we will use this stochastic decomposition property to approximate the mean QT when the inter-arrival times are generally distributed.

4.4 Time-based ~~nonpre-emptive~~ interruptions with Poisson arrivals

Time-based ~~nonpre-emptive~~ events (2–2–2) are commonly seen in practice. Indeed, most of the events listed in Summary of Time in SEMI E10 (2001) are time-based ~~nonpre-emptive~~. Examples are process or equipment experiments, PM, tool modifications and change of consumables. If the impact on QT from different sources is modelled together, the various independent sources justify the use of Palm–Khintchine Theorem. Hence, we assume Poisson arrivals for the following model. This assumption will be relaxed in Section 4.5.

For this type of events, we usually have some level of control on the occurrence of each interruption. For example, we may postpone the execution of an equipment experiment until the machine is idle (without any WIP in queue), or until the machine completes its current processing job. They can be classified as, Case 1: postpone PM to the completion of all jobs in queue, and Case 2: postpone PM to the completion of the current job.

In Case 1, the interruptions have low priority and the WIP has high priority. In Case 2, the interruptions have high priority and the WIP has low priority. Both cases can be modelled by the ~~nonpre-emptive~~ priority queues. If the service times are generally distributed with FIFO discipline, the M/G/1 model for the ~~nonpre-emptive~~ priority queues with two priorities is as follows (Adan and Resing 2001),

$$E(QT_1) = (\rho_1 E(R_{S1}) + \rho_2 E(R_{S2})) / (1 - \rho_1), \quad (18)$$

$$E(QT_2) = \frac{\rho_1 E(R_{S1}) + \rho_2 E(R_{S2})}{(1 - \rho)(1 - \rho_1)}, \quad (19)$$

$$E(QT) = \frac{\lambda_1}{\lambda_1 + \lambda_2} E(QT_1) + \frac{\lambda_2}{\lambda_1 + \lambda_2} E(QT_2), \quad (20)$$

$$E(CT_i) = E(QT_i) + E(S_i), \quad i = 1, 2 \text{ and } E(CT) = E(QT) + \frac{\lambda_1}{\lambda_1 + \lambda_2} E(S_1) + \frac{\lambda_2}{\lambda_1 + \lambda_2} E(S_2),$$

where $\rho_1 = \lambda_1 E(S_1)$, $\rho_2 = \lambda_2 E(S_2)$, $\rho = \rho_1 + \rho_2$, and $E(R_{Si}) = E(S_i^2)/2E(S_i)$, $i = 1, 2$. Furthermore, one represents high priority jobs and two represents low priority jobs. λ_1 and λ_2 are the arrival rates of high and low priority jobs, respectively. For stability, we assume $\rho < 1$.

In practice, the control can be more complicated. We may postpone the occurrence of PM to some period when the machine is less busy, but may not postpone it too much to jeopardise the quality of products. The actual mean QT is therefore between Cases 1 and 2. The situation could be closer to Case 1 when forecast utilisation is low and closer to Case 2 when forecast utilisation is high. When utilisation is low, there are chances to perform PM without interrupting production runs. When utilisation is high, there is no motivation to wait and take the risk of sacrificing product quality. When the postponable duration is given, this model is a special type of limited service vacation models (Doshi 1986).

Note that in priority queues, an interruption could occur during another interruption. If an interruption cannot occur during interruptions and the WIP has low priority, the time-based ~~nonpre-emptive~~ model is the same as the time-based ~~pre-emptive~~ model under Poisson arrivals (Gaver 1962).

4.5 Extension to general arrival processes

Because the arrival process may not be Poisson in practical manufacturing systems, it is important to extend the arrival process of previous queueing models to general arrivals. Under general and independent arrival processes, QT of G/G/1 queues can be estimated by Equation (21), which is known as Kingman's heavy traffic approximation (Heyman 1975):

$$E(QT) \cong \left(\frac{c_a^2 + c_s^2}{2} \right) \left(\frac{\rho}{1 - \rho} \right) E(S), \quad (21)$$

where ρ is utilisation, and c_a and c_s are CV of inter-arrival time and service time, respectively. If we replace the service time in Equation (21) by GST, run-based ~~pre-emptive~~ (2-1-1) and run-based ~~nonpre-emptive~~ product-induced events (2-1-2.b) can be approximated as

$$E(QT) \cong \left(\frac{c_a^2 + c_G^2}{2} \right) \left(\frac{\rho_G}{1 - \rho_G} \right) E(G), \quad (22)$$

Table 2. G/G/1 approximate models for different interruption types.

Category	G/G/1 approximate model
2-1-1	$E(QT) \cong \left(\frac{c_a^2 + c_G^2}{2} \right) \left(\frac{\rho_G}{1 - \rho_G} \right) E(G)$
2-1-2.a	$E(QT) \cong \left(\frac{c_a^2 + c_s^2}{2} \right) \left(\frac{\rho}{1 - \rho} \right) E(S) + \frac{1/\lambda}{1/\lambda + E(T_s)} E(T_s) + \frac{E(T_s)}{1/\lambda + E(T_s)} E(R_T)$
2-1-2.b	$E(QT) \cong \left(\frac{c_a^2 + c_G^2}{2} \right) \left(\frac{\rho_G}{1 - \rho_G} \right) E(G)$
2-2-1	$E(QT) \cong \left(\frac{c_a^2 + c_G^2}{2} \right) \left(\frac{\rho_G}{1 - \rho_G} \right) E(G) + \left(\frac{c_{aD}^2 + c_D^2}{2} \right) \frac{E(D)^2 \eta}{1 + E(D) \eta}$
2-2-2	$E(QT_1) \cong \left(\frac{c_{a1}^2 + c_{s1}^2}{2} \right) \frac{\rho_1}{1 - \rho_1} E(S_1) + \left(\frac{c_{a2}^2 + c_{s2}^2}{2} \right) \frac{\rho_2}{1 - \rho_1} E(S_2)$ $E(QT_2) \cong \left(\frac{c_{a1}^2 + c_{s1}^2}{2} \right) \frac{\rho_1}{(1 - \rho)(1 - \rho_1)} E(S_1) + \left(\frac{c_{a2}^2 + c_{s2}^2}{2} \right) \frac{\rho_2}{(1 - \rho)(1 - \rho_1)} E(S_2)$

where $E(G)$ can be computed by Equation (3), since the derivation does not assume Poisson arrivals. Based on Theorems 1 and 2, Equation (7) and Equation (22), we can easily extend the other M/G/1 models to G/G/1 approximate models as shown in Table 2.

From 2–2–2, one can observe that QT is caused by not only congestion of jobs but also interruptions, i.e. ~~a job can have QT even if there is no job in queue~~. Another important observation is that the variability of low utilisation jobs has limited impact on total QTs. Hence, when the PM has low traffic intensity, its impact on product QT will be small, even if it has large variability. Furthermore, the variability of time-based ~~nonpre-emptive~~ events does not go into product service time variability. Hence, if one wants to reduce product QT, reducing product service time variability is more effective than reducing ~~nonpre-emptive~~ event variability.

4.6 Simulation validation

To validate the G/G/1 models, we conducted two simulation experiments. The first one shows the importance of choosing the right model under time-based ~~pre-emptive~~ interruptions. The second one validates the proposed models as shown in Table 2.

4.6.1 Time-based ~~pre-emptive~~ interruptions

In Factory Physics (Hopp and Spearman 2008), interruptions on the shop floors are classified as either ~~pre-emptive~~ or ~~nonpre-emptive~~ events. Failing to distinguish the difference between time-based and run-based events, they treated both power outage (time based) and running out of consumables (run based) as the same type of events (i.e. ~~pre-emptive~~ events). To demonstrate the importance of choosing correct models, three simulation experiments were conducted for a single machine system with time-based ~~pre-emptive~~ events. The first experiment has Poisson arrivals, while the second and third have gamma-distributed inter-arrival times. All experiments compare three types of QTs: simulated mean queue time (SQT), theoretical mean queue time (TQT) from the model of 2–2–1 and the approximated mean queue time (FQT) from the model proposed by Hopp and Spearman (2008).

In the first experiment, service time is gamma-distributed with mean 30 min and squared coefficient of variation (SCV) 0.2. We choose the service time variability small, since service time has to satisfy the tight product specification in practice. Downtime is gamma-distributed with mean 250 min and SCV 0.8. The uptime between two interruptions is exponentially distributed with mean 2250 min. (Therefore, availability is 90%). GST is 33.33 (=30/0.9) based on Equation (3). Arrivals follow Poisson distribution with 10 different mean arrival rates resulting in utilisations from 10 to 95% as shown in Table 3.

For each input rate, the mean and variance are computed based on 100 replications. Each replication is composed of 200,000 jobs while the first half of jobs is discarded. Table 3 compares the mean QTs from three different approaches. SQT and simulated mean service time (SST) are followed by the half width of 90% confidence intervals (CI), where the largest CI is less than 3% of its mean. TQT is computed based on Equation (15), and FQT is computed based on Equation (5) or the model in Factory Physics (Hopp and Spearman 2008). SST is obtained based on Equation (1). The percentage difference between TQT and FQT (compared with SQT) is reported in 'Diff %'. At all utilisations, TQT is within the CI as expected. The errors of FQT come from the gap between time-based and run-based events, because

Table 3. Comparison among different QTs with Poisson arrivals.

Unit: min Util. (%)	SQT	TQT	FQT	Diff %		Gap FQT–TQT	SST	GST
				TQT (%)	FQT (%)			
10	27.26±0.10	27.22	4.72	−0.2	−82.7	−22.50	33.34±0.02	33.33
20	33.07±0.15	33.13	10.63	0.2	−67.9	−22.50	33.34±0.02	33.33
30	40.68±0.22	40.71	18.21	0.1	−55.2	−22.50	33.32±0.02	33.33
40	50.78±0.28	50.83	28.33	0.1	−44.2	−22.50	33.33±0.02	33.33
50	64.89±0.42	65.00	42.50	0.2	−34.5	−22.50	33.34±0.02	33.33
60	85.97±0.67	86.25	63.75	0.3	−25.8	−22.50	33.31±0.02	33.33
70	121.56±0.95	121.67	99.17	0.1	−18.4	−22.50	33.34±0.02	33.33
80	194.13±2.07	192.50	170.00	−0.8	−12.4	−22.50	33.35±0.02	33.33
90	407.06±6.24	405.00	382.50	−0.5	−6.0	−22.50	33.33±0.02	33.33
95	847.27±23.35	830.00	807.50	−2.0	−4.7	−22.50	33.33±0.02	33.33

FQT is obtained by treating time-based events as run-based. Due to the heavy traffic property, FQT performs quite poorly except for very high utilisations. The error percentage caused by FQT can be as large as 12.4 and 80% utilisation, but decreases to 4.7 and 95% utilisation. The constant gap between FQT and TQT is consistent with Theorem 2.

The second and third experiments are basically the same as the first, except the inter-arrival times and uptimes are gamma distributed. The inter-arrival time SCVs are 0.6 and 2 for the second and third experiments, respectively. The uptimes between two interruptions are gamma-distributed with SCV 0.9 in the second and third experiments. The results are shown in Tables 4 and 5.

Approximate mean queue time (AQT) in Tables 4 and 5 is computed based on the equation of 2-2-1 in Table 2. Since the arrival process is not Poisson, the errors of AQT mainly come from Kingman's approximation. The errors of FQT come from: (1) Kingman's approximation, and (2) the gap. In Table 4, because these two errors of FQT possess different signs and coincidentally negate each other, the errors become smaller than those of AQT in heavy traffic. While the average absolute error of AQT is only 3.6% in Table 4 and 8.9% in Table 5, the errors from FQT are much larger in general. The above simulation results reinforce the importance of using the model appropriate to the types of interruptions.

4.6.2 Model validation

Among the models in Table 2, the model of 2-2-1 has been validated in the Section 4.6.1, and the models of 2-1-2.b and 2-1-1 are the same as the ones explained by Hopp and Spearman (2008). Hence, in the following experiments, we will validate the QT approximations for 2-1-2.a and 2-2-2, which are newly proposed in this paper.

We assume service time has small natural variability due to the motion of human (and/or robot). To validate the model of 2-1-2.a, we assume whenever the machine is idle, a state-induced set-up is needed before the machine can process jobs again. For 2-2-2, we assume the time-based ~~nonpre-emptive~~ events, such as experiments or ~~pre-emptive~~ maintenances, occur several times a month, but only when the machine is idle and no production job is in queue

Table 4. Comparison among different QTs when inter-arrival time SCV is 0.6.

Unit: min Util. (%)	SQT	AQT	FQT	Diff %		Gap FQT-AQT	SST	GST
				AQT (%)	FQT (%)			
10	25.16±0.10	25.20	3.98	0.2	-84.2	-21.22	33.24±0.02	33.33
20	29.39±0.13	30.20	8.96	2.8	-69.5	-21.24	33.27±0.02	33.33
30	35.35±0.18	36.60	15.36	3.5	-56.6	-21.24	33.27±0.02	33.33
40	43.18±0.26	45.10	23.89	4.4	-44.7	-21.21	33.26±0.02	33.33
50	54.70±0.36	57.10	35.83	4.4	-34.5	-21.27	33.27±0.02	33.33
60	71.83±0.54	75.00	53.75	4.4	-25.2	-21.25	33.27±0.02	33.33
70	100.87±0.79	104.90	83.61	4.0	-17.1	-21.29	33.29±0.02	33.33
80	156.78±1.49	164.60	143.33	5.0	-8.6	-21.27	33.29±0.02	33.33
90	330.90±4.88	343.80	322.50	3.9	-2.5	-21.30	33.33±0.02	33.33
95	679.41±18.74	702.10	680.83	3.3	0.2	-21.27	33.32±0.02	33.33

Table 5. Comparison among different QTs when inter-arrival time SCV is 2.

Unit: min Util. (%)	SQT	AQT	FQT	Diff %		Gap FQT-AQT	SST	GST
				AQT (%)	FQT (%)			
10	34.13±0.12	27.80	6.60	-18.5	-80.7	-21.20	33.24±0.02	33.33
20	43.40±0.16	36.00	14.80	-17.1	-65.9	-21.20	33.25±0.02	33.33
30	54.72±0.26	46.60	25.40	-14.8	-53.6	-21.20	33.27±0.02	33.33
40	69.37±0.32	60.70	39.40	-12.5	-43.2	-21.30	33.26±0.02	33.33
50	89.27±0.43	80.40	59.20	-9.9	-33.7	-21.20	33.31±0.02	33.33
60	118.48±0.69	110.00	88.80	-7.2	-25.1	-21.20	33.30±0.02	33.33
70	168.25±1.19	159.30	138.10	-5.3	-17.9	-21.20	33.31±0.02	33.33
80	264.97±2.42	257.90	236.70	-2.7	-10.7	-21.20	33.29±0.02	33.33
90	549.50±9.52	553.80	532.50	0.8	-3.1	-21.30	33.29±0.02	33.33
95	1140.46±40.48	1145.40	1124.20	0.4	-1.4	-21.20	33.32±0.02	33.33

Table 6. Parameters used in the experiments.

Unit: min	Service time ($1/\lambda_1, S_1$)	2-1-2.a Nonpre-emptive state-induced ($-, T_s$)	2-2-2 Time-based nonpre-emptive ($1/\lambda_2, S_2$)
IAT	—	—	7200
SCV	—	—	0.75
Mean or MTTR	30	10	180
SCV	0.2	0.3	0.5

Table 7. Error percentage of mean QT approximations.

Util. (%)	2-1-2.a			2-2-2		
	SQT	AQT	%	SQT	AQT	%
10	12.11 ± 0.01	11.89	−1.9	5.11 ± 0.02	4.50	−11.9
20	14.72 ± 0.01	14.28	−3.0	7.90 ± 0.03	7.23	−8.5
30	18.07 ± 0.01	17.40	−3.7	11.48 ± 0.05	10.71	−6.7
40	22.49 ± 0.03	21.59	−4.0	16.16 ± 0.05	15.30	−5.3
50	28.63 ± 0.04	27.50	−4.0	22.72 ± 0.08	21.64	−4.7
60	37.82 ± 0.06	36.42	−3.7	32.34 ± 0.13	30.97	−4.2
70	52.92 ± 0.13	51.34	−3.0	47.58 ± 0.18	46.04	−3.2
80	82.59 ± 0.30	81.26	−1.6	76.40 ± 0.30	74.50	−2.5
90	172.56 ± 1.11	171.19	−0.8	152.18 ± 1.07	148.50	−2.4
95	351.53 ± 5.40	351.16	−0.1	266.26 ± 2.65	259.50	−2.5

(i.e. the interruption has lower priority). The parameters for the above scenarios are summarised in Table 6, where IAT is mean inter-arrival times (for time-based nonpre-emptive events), MTTR is mean time to repair, and SCV is squared coefficient of variations. All random variables in Table 6 follow gamma distribution.

Job arrivals follow gamma distribution with 10 different arrival rates resulting in utilisations from 10 to 95% as shown in Table 7. The SCV of job arrival intervals is 0.5. For each input rate, the mean and variance are computed based on 100 replications. Each replication is composed of 200,000 jobs while the first half is discarded.

Table 7 compares the mean QTs from two different approaches: AQT is computed based on the models in Table 2. SQT is simulated mean queue time followed by the half width of 90% CI. In both 2-1-2.a and 2-2-2, the largest CI is less than 2% of its mean.

For the model of 2-1-2.a, the largest error is 4 and 50% utilisation. It performs very well in heavy traffic. For the model of 2-2-2, the largest error occurs in light traffic and the percentage errors are gradually reduced to less than 3% in heavy traffic. The simulation results show that the models of 2-1-2.a and 2-2-2 are reliable.

5. Conclusions

We give the literature review for interruption queueing models, classify prior results and propose a new classification of the queueing models for workstations with interruptions. Through the systematic classification, we point out the proper queueing models for each interruption type and the corresponding G/G/1 approximate models.

In addition to the proposed G/G/1 models, an important contribution of this paper is suggesting a systematic way to study manufacturing system performance, namely, (1) systematic classification of events, (2) defining service time from the view point of capacity, and (3) deriving queueing models accordingly. While the first and second steps are sometimes overlooked, they indeed play a key role in selecting the correct queueing model for practical manufacturing systems. The concept of service times has been commonly mixed up with the concept of process times. In Section 3, we explained that the determination of service time may not be trivial and it should be computed from the view point of capacity.

Since we assume all pre-emptive events are pre-emptive-resume, our results will be a lower bound when the interruptions are pre-emptive-repeat. However, this lower bound should be tight if the occurrence of the pre-emptive interruptions is rare (which is usually the case in practical production lines) and the service time is relatively short compared to the duration of the pre-emptive interruptions.

Cheng, Gupta, and Wang (2000) gave a comprehensive review on the static flow shop scheduling problems with set-up times. Our state-dependent set-up is one type of the sequence-independent job set-up times and the product-dependent set-up belongs to the sequence-independent family set-up times. While our set-up queueing models assume sequence-independent, sequence-dependent assumption is also seen in the static flow shop scheduling problems (Allahverdi et al. 2008; Jolai, Rabiee, and Asefi 2012; Mousavi, Zandieh, and Amiri 2011). The derivation of queueing models with sequence-dependent set-ups is left for future research.

In practical manufacturing systems, machine failure mode can be more complex than what we have described. For example, an interruption may degrade server performance over time but instead of shutting down the server, or the recovery time may depend on the previous uptime duration. Simultaneous interruptions may occur, although the chance can be small. Resource contention commonly exists. Dispatching rules can be set-up-oriented (Pickardt and Branke 2012). In some circumstances, PM is required if failure risk exceeds a predetermined threshold. This condition-based maintenance (Golmakani 2011) will also complicate the model. The trade-off is between model complexity and approximation accuracy. Newell (1979) stated. In fact, in most applications, one is lucky if one has a good estimate of the service rates (to within 5% say); the variance rates are often known only to within a factor of 2, seldom to within an accuracy of 20%. Even if the analytical model is exact, the results can still be approximations due to parameter uncertainty. It is impossible to take all details into account (such as resource contention on water fountains), but the key factors should be kept right whenever possible.

An important contribution of this paper is that it clarified the relations among different types of models, such as the decomposition properties, and proposed their G/G/1 models. Rather than investigating different models in detail, Wu, McGinnis, and Zwart (2011) derived an integrated model which considers all types of interruptions under ample resource. Without giving the G/G/1 models and comprehensive literature review, Wu, McGinnis, and Zwart (2008) compared the classification with the Summary of Time in SEMI E10 in detail.

For the product-induced set-up model, the memoryless property and single-service time distributions are assumed. More general models with generally distributed occurrence and different service time distributions are left for future research. In queueing network settings, the impact on intrinsic ratios (Wu and McGinnis 2013) from different types of interruptions should be further investigated. The G/G/N approximations with interruptions and the resource contention problems need more work. Further improvements on the classification and more precise queueing models to describe the behaviour of manufacturing systems are expected and left for future research.

References

- AQ3. Adan, I., and J. Resing. 2001. *Queueing Theory*. Eindhoven University of Technology.
- Allahverdi, A., C. T. Ng, T. C. E. Cheng, and M. Y. Kovalyov. 2008. "A Survey of Scheduling Problems with Setup Times or Costs." *European Journal of Operational Research* 187 (3): 985–1032.
- AQ4. Ashcroft, H. 1950. "The Productivity of Several Machines Under the Care of One Operator." *Journal of the Royal Statistical Society. Series B (Methodological)* 12 (1): 145–151.
- AQ5. Asmussen, S. 1992. "Queueing Simulation in Heavy Traffic." *Mathematics of Operations Research* 17 (1): 84–111.
- Avi-Itzhak, B., and P. Naor. 1963. "Some Queueing Problems with the Service Station Subject to Breakdown." *Operations Research* 11: 303–320.
- Baker, J., and I. Rubin. 1987. "Polling with a General-service Order Table." *IEEE Transactions on Communications* 35 (3): 283–288.
- Boxma, O., M. Mandjes, and O. Kella. 2008. "On a Queueing Model with Service Interruptions." *Probability in the Engineering and Informational Sciences* 22 (4): 537–555.
- Buzacott, J. A., and L. E. Hanifin. 1978. "Models of Automatic Transfer Lines with Inventory Banks – A Review and Comparison." *AIIE Transactions* 10 (2): 197–207.
- AQ6. Buzacott, J. A., and J. G. Shanthikumar. 1993. *Stochastic Models of Manufacturing Systems*. NJ: Prentice Hall.
- Cheng, T. C. E., J. N. D. Gupta, and G. Wang. 2000. "A Review of Flowshop Scheduling Research with Setup Times." *Production and Operations Management* 9 (3): 262–282.
- Doshi, B. 1986. "Queueing Systems with Vacations – A Survey." *Queueing Systems* 1 (1): 29–66.
- Eisenberg, M. 1972. "Queues with Periodic Service and Changeover Time." *Operations Research* 20 (2): 440–451.
- AQ7. Federgruen, A., and L. Green. 1986. "Queueing Systems with Service Interruptions." *INFORMS: Institute for Operations Research* 34: 752–769.
- Fuhrmann, S., and R. Cooper. 1985. "Stochastic Decompositions in the M/G/1 Queue with Generalized Vacations." *Operations Research* 33 (5): 1117–1129.
- Gaver, D. P. 1962. "A Waiting Line with Interrupted Service, including Priorities." *Journal of the Royal Statistical Society* 24: 73–90.
- Golmakani, H. R. 2011. "Condition-based Inspection Scheme for Condition-based Maintenance." *International Journal of Production Research* 50 (14): 3920–3935.

- ~~AQ8~~ Gross, D., and C. M. Harris. 1998. *Queueing Theory*. New York: Wiley.
- Haque, L., and M. Armstrong. 2007. "A Survey of the Machine Interference Problem." *European Journal of Operational Research* 179 (2): 469–482.
- 5 Heyman, D. P. 1975. "A Diffusion Model Approximation for the GI/G/1 Queue in Heavy Traffic." *Bell System Technical Journal* 54 (9): 1637–1646.
- Hopp, W. J., and M. L. Spearman. 2008. *Factory Physics*. Chicago, IL: IRWIN.
- ~~AQ9~~ Jaiswal, N. K. 1968. *Priority Queues*. New York: Academic Press.
- 10 Jolai, F., M. Rabiee, and H. Asefi. 2012. "A Novel Hybrid Meta-heuristic Algorithm for a No-wait Flexible Flow Shop Scheduling Problem with Sequence Dependent Setup Times." *International Journal of Production Research* 50 (24): 7447–7466.
- Keilson, J. 1962. "Queues Subject to Service Interruption." *Annals of Mathematical Statistics* 33: 1314–1322.
- Khinchine, A. Y. 1933. "On Mean Time of Stoppage of Machines." *Mathematicheski Sbornik* 40: 119–123.
- Kingman, J. F. C. 1962. "Some Inequalities for the Queue GI/G/1." *Biometrika* 49: 315–324.
- ~~AQ10~~ Kingman, J. F. C. 1965. "The Heavy Traffic Approximation in the Theory of Queues." In *Proceedings of the Symposium on*
~~AQ11~~ *Congestion Theory*, 137–159. University of North Carolina Press.
- ~~AQ12~~ Levy, H., and L. Kleinrock. 1986. "A Queue with Starter and a Queue with Vacations: Delay Analysis by Decomposition." *Operations Research*: 426–436.
- Lindley, D. V. 1952. "Theory of Queues with a Single Server." *Proceedings of the Cambridge Philosophical Society* 48: 277–289.
- Morrison, J., and D. Martin. 2007. "Practical Extensions to Cycle Time Approximations for the G/G/M-queue with Applications." *IEEE Transactions on Automation Science and Engineering* 4 (4): 523–532.
- 20 Mousavi, S. M., M. Zandieh, and M. Amiri. 2011. "Comparisons of Bi-objective Genetic Algorithms for Hybrid Flowshop Scheduling with Sequence-dependent Setup Times." *International Journal of Production Research* 50 (10): 2570–2591.
- ~~AQ13~~ Neuts, M., and D. Lucantoni. 1979. "A Markovian Queue with N Servers Subject to Breakdowns and Repair." *Management Science*: 849–861.
- 25 Newell, G. F. 1979. *Approximate Behavior of Tandem Queues*. New York: Springer.
- Pickardt, C. W., and J. Branke. 2012. "Setup-oriented Dispatching Rules – A Survey." *International Journal of Production Research* 50 (20): 5823–5842.
- ~~AQ14~~ SEMI E10. 2001. "Specification for Definition and Measurement of Equipment Reliability, Availability, and Maintainability." In *Book of SEMI Standards*. Mountain View, CA: SEMI.
- 30 Sengupta, B. 1990. "A Queue with Service Interruptions in an Alternating Random Environment." *Operations Research* 38 (2): 308–318.
- Smith, W. L. 1953. "Distribution of Queueing Times." *Proceedings of the Cambridge Philosophical Society* 49: 449–461.
- Stecke, K., and J. Aronson. 1985. "Review of Operator/Machine Interference Models." *International Journal of Production Research* 23 (1): 129–151.
- ~~AQ15~~ Takagi, H. 1988. "Queueing Analysis of Polling Models." *ACM Computing Surveys* 20 (1): 5–28.
- ~~AQ16~~ Takagi, H. 2000. "Analysis and Application of Polling Models." In *Performance Evaluation: Origins and Directions*, 423–442.
- Thiruvengadam, K. 1963. "Queueing with Breakdowns." *Operations Research* 11 (1): 62–71.
- Welch, P. 1964. "On a Generalized M/G/1 Queueing Process in Which the First Customer of Each Busy Period Receives Exceptional Service." *Operations Research* 12 (5): 736–752.
- 40 White, H., and L. Christie. 1958. "Queueing with Preemptive Priorities or with Breakdown." *Operations Research* 6: 79–95.
- ~~AQ17~~ Wolff, R. W. 1982. "Poisson Arrivals See Time Averages." *Operations Research* 30: 223–231.
- Wu, K., and K. Hui. 2008. "The Determination and Indetermination of Service Times in Manufacturing Systems." *IEEE Transactions on Semiconductor Manufacturing* 21 (1): 72–82.
- Wu, K., and L. F. McGinnis. 2013. "Interpolation Approximations for Queues in Series." *IIE Transactions* 45 (3): 273–290.
- ~~AQ18~~ Wu, K., L. F. McGinnis, and B. Zwart. 2008. "Queueing Models for Single Machine Manufacturing Systems with Interruptions." In *Winter Simulation Conference*: 2083–2092.
- ~~AQ19~~ Wu, K., L. F. McGinnis, and B. Zwart. 2011. "Queueing Models for a Single Machine Subject to Multiple Types of Interruptions." *IIE Transactions* 43 (10): 753–759.