MODELLING AND ANALYSIS OF PULL PRODUCTION SYSTEMS

A THESIS
SUBMITTED TO THE DEPARTMENT OF INDUSTRIAL ENGINEERING
AND THE INSTITUTE OF ENGINEERING AND SCIENCE
OF BİLKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

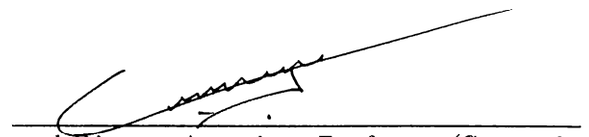
By
Nureddin Kırkavak
July 1995

MODELLING AND ANALYSIS OF PULL PRODUCTION SYSTEMS

A THESIS
SUBMITTED TO THE DEPARTMENT OF INDUSTRIAL ENGINEERING
AND THE INSTITUTE OF ENGINEERING AND SCIENCE
OF BİLKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

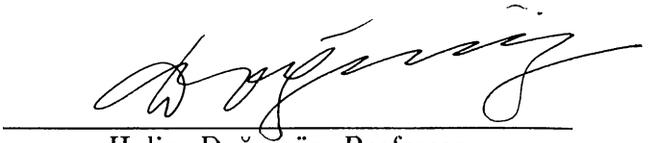
By
Nureddin Kırkavak
July 1995

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.



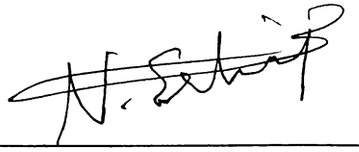
Cemal Dinçer, Associate Professor (Supervisor)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.



Halim Doğrusöz, Professor

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.



Nesim Erkip, Professor

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

Osman Oğuz, Associate Professor

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

Erdal Erel, Associate Professor

Approved for the Institute of Engineering and Science:

Prof. Dr. Mehmet Baray,
Director of Institute of Engineering and Science

Abstract

MODELLING AND ANALYSIS
OF
PULL PRODUCTION SYSTEMS

Nureddin Kırkavak
Ph.D. in Industrial Engineering
Supervisor: Cemal Dinçer, Associate Professor
July 1995

A variety of production systems appearing in the literature are reviewed in order to develop a classification scheme for production systems. A number of pull production systems appearing in the classification are found to be equivalent to a tandem queue so that accurate tandem queue decomposition methods can be used to find the performance of such systems. The primary concern of this dissertation is to model and analyze non-tandem queue equivalent periodic pull production systems.

In this research, an exact performance evaluation model is developed for a single-item periodic pull production system. The processing and demand interarrival times are assumed to be Markovian. For large systems, which are difficult to evaluate exactly because of large state spaces involved, an approximate decomposition method is proposed. A typical approximate decomposition procedure takes individual stages or pairs of stages in isolation to analyze the system and

then it aggregates the results to obtain an approximate performance for the whole system. An experiment is designed in order to investigate the general behavior of the decomposition. The results are worth attention.

A second aspect of this study is to investigate an allocation methodology to achieve the maximum throughput rate with providing two sets of allocation parameters regarding the number of kanbans and the workload at each stage of the system. Together with some structural properties, the experimental results provide some insight into the behavior of pull production systems and also provide a basis for the proposed allocation methodology.

Finally, we conclude our findings together with some directions for future research.

Keywords: Production/Inventory Systems, Performance Evaluation, Markov Processes, Approximate Decomposition, Throughput Maximization, Workload–Kanban Allocation.

Özet

ÇEKME TİPİ ÜRETİM SİSTEMLERİNİN MODELLENMESİ VE ANALİZİ

Nureddin Kırkavak

Endüstri Mühendisliği Doktora

Tez Yöneticisi: Doç. Dr. Cemal Dinçer

Temmuz 1995

Üretim sistemlerine yönelik bir sınıflandırma sistemi geliştirmek amacıyla literatürde yer alan çok değişik tipte üretim sistemleri incelendi. Ele alınan üretim sistemleri içinde yer alan Çekme Tipi Üretim Sistemlerinin büyük bir çoğunluğu seri akışlı kuyruk modellerine eşdeğer bulundu. Bu nedenle, bu tip eşdeğer sistemlerin performans değerlendirmesinde, oldukça iyi sonuç veren, seri akışlı kuyruk modelleri için geliştirilmiş, çözüm tekniklerinden yararlanılabilir. Bu araştırma çalışmasının en temel amacı eşdeğer olmayan Çekme Tipi Periyodik Üretim Sistemlerinin modellenmesi ve analizidir.

Bu çalışmada, Tek Ürünlü Çekme Tipi Periyodik Üretim Sistemleri için bir performans değerlendirme modeli geliştirildi. Üretim sistemi içindeki parça işleme ve talebin varış ara zamanları Markof özelliklidir. Durum uzayının büyüklüğünden dolayı tam olarak çözümlenemeyecek kadar büyük sistemler için yaklaşık sonuç veren bir ayrıştırma yöntemi geliştirildi. Bu tip yaklaşık sonuç veren ayrıştırma yöntemleri, üretim aşamalarını birer birer ele alarak sistemi parçalara ayırırlar. Daha sonra da, elde edilen sonuçları bir araya getirerek tüm sistemin performansını bulurlar. Önerilen yaklaşık çözüm yönteminin genel doğruluk seviyesini belirlemek amacıyla, sonuçları oldukça olumlu bir nümerik deney gerçekleştirildi.

Çalışmanın ikinci bölümünde, üretim hızının, üretim aşamalarına dağıtılacak iş yükü ve ara–stok kapasitelerini belirleyen parametrelere en uygun değerlerin bulunması suretiyle, maksimizasyonuna yönelik bir dağıtım metodolojisi üzerinde çalışıldı. Yapılan yoğun deneysel çalışmalar sonucunda, bulunan bir takım yapısal özelliklere ek olarak, Çekme Tipi Üretim Sistemlerinin genel işleyişi ile ilgili bilgi elde edilerek, bir dağıtım metodolojisi önerildi.

Çalışmanın sonunda ise, yapılan tüm işler özetlenerek, ileriki araştırma çalışmalarına yönelik çeşitli noktaların üzerinde duruldu.

Anahtar kelimeler: Üretim/Envanter Sistemleri, Performans Değerlendirme, Markof Süreçleri, Yaklaşık Ayrıştırma, Üretim Hızı Maksimizasyonu, İş Yükü ve Ara–Stok Dağıtımı.

To my wife and my son,

Acknowledgement

I would like to express my sincere gratitude to the numerous people who have supported me in various ways during the process of this dissertation. I am indebted to Assoc. Prof. Cemal Dinçer. Although, his time-consuming administrative activities as the Assistant Dean of the Faculty of Engineering prevented him from being daily involved, his supervisory support has been invaluable. Above all, I gained the experience of conducting independent research and I thank him for his contribution to that.

I am grateful to Prof. Halim Doğrusöz, Prof. Nesim Erkip, Prof. Akif Eyler, Assoc. Prof. Ömer Benli, Assoc. Prof. Osman Oğuz, Assoc. Prof. Erdal Erel, Asst. Prof. Selçuk Karabatı, and Asst. Prof. Cemal Akyel for their valuable comments. In particular, I believe the remarks of Prof. Halim Doğrusöz and Prof. Nesim Erkip, will also be useful in improving the papers to be published from this dissertation. I also wish to thank to Assoc. Prof. Ülkü Gürler also for her help in the statistical analysis part of this work, especially for supplying me with the most valuable recent SAS Manuals.

I wish to express my appreciation to all BCC personnel for their help. I would like to offer my sincere thanks to my comrade Dr. Levent Kandiller for his extensive help, understanding, morale support, and encouragement. I wish to extend my appreciation to my friends Dr. Hakan Polatoğlu, Abdullah Daşcı, and Yavuz Günalay for their help and morale support during this study.

I would like to extend my deepest gratitude and thanks to my parents for their stimulating attention and morale support.

Last but not least, I owe special thanks to my wife, İlknur Kırkavak. This study could have taken much longer without her understanding, patience, continuous morale support, and encouragement. It is to her that this study is affectionally dedicated, without whom it would not be possible.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The traditional approach in production system design has been to assume a deterministic world in which the impact of variability on performance can be resolved by providing adequate surplus capacity. It is assumed that production managers would take the necessary steps in order to eliminate the sources of variability, using the approaches of simplification, standardization and control. However, in reality a few manufacturing industries use sufficiently stable processes for this approach to work well. New processes and new products continually appear and coping with the resulting uncertainties becomes the most important concern of production managers.

Since the seventies discrete event simulation has been extensively used in modelling the production systems to assess the impact of variability and to explore various ways of coping with the change and uncertainty. Apart from the problems of validating, large and complex simulation models often result in limited insight into the factors determining the behavior of the system. So in recent years there have been considerable development in using queuing theory to model production systems. These queuing models can often be used for performance evaluation and comparison of designs of production systems both in comparing alternative

configurations and in selecting best parameter values.

1.1 Production Systems

Production systems consist of materials, work areas and storage areas. Materials flow from one storage area to a work area and after they are placed in another storage area. There is a storage area and work area combination through which materials enter and another storage area and work area combination through which they leave the production system. The times, that materials spend in work areas are random. This randomness may be due to random processing times or random failure and repair events. Storage areas can hold only a finite amount of materials. The work areas are usually called *machines*. In general, it is assumed that the machines are never allowed to be idle while they have materials to work on and there is space at storage areas in which to put the materials they have worked on. Storage areas are often called *buffers*. The materials in general consist of discrete parts.

The production systems, in which each part travels the same sequence of machines and buffers, are called *production lines*. In the language of queuing theory, a production line can be represented as a *finite buffer tandem queuing system*. In that case, machines are called servers, storage areas are called queues and discrete parts are called customers or jobs (See Figure 1.1).

There have been many researchers in this field, and almost as many different sets of notation. In a finite buffer tandem queuing system, like the one in Figure 1.1, servers are numbered from 1 to N, where N is the number of servers in the system. There are N queue–server couples in which queue Q_i feeds the server S_i. Job arrivals to the system are placed at the first queue, Q_1. The arrived jobs are processed sequentially at each server up to the last server, S_N, after which they leave the system. All the jobs have to be processed on all the servers. A great

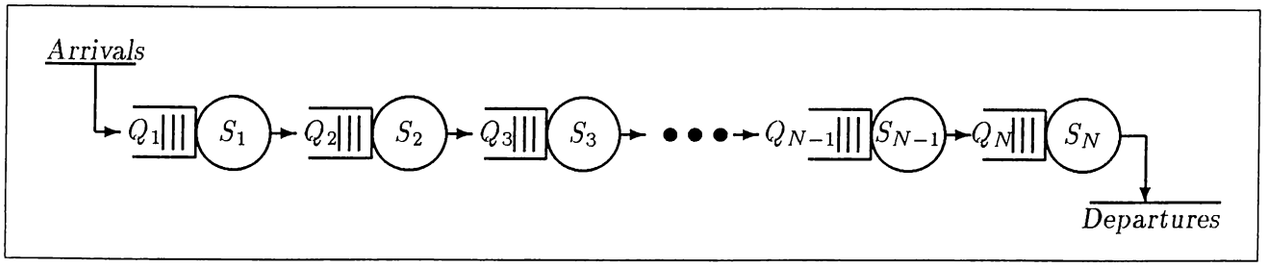


Figure 1.1: Arrangement of servers (S_j: $j = 1, 2, ..., N$) with finite capacity queues (Q_j: $j = 1, 2, ..., N$) in a tandem queueing system.

deal of additional notation is defined throughout the text and a list is given at the end of the text.

In the majority of studies on production systems reported in the literature, the goal has been primarily to calculate the maximum rate of flow of material through the system. The maximum flow rate of materials is often called *production rate* or *throughput*. In the production research literature, other performance measures, especially the average amount of materials in the buffers, are also important.

Like all types of mathematical models, the models of production systems are compromises between reality and tractability. But, the use of results that are based on simplifications of reality is essential in the design and implementation of large and complex production systems.

1.2 Outline of the Thesis

There are many different kinds of production systems, and many different kinds of models in the literature. The purpose of this study is first to overview a variety of production systems appearing in the literature to develop a classification scheme for production systems. Therefore, in the next chapter the distinguishing literature on production systems is briefly reviewed and a unifying classification

scheme for production systems with respect to design and operating characteristics is proposed.

In the rest of the study, we consider modelling and analysis of a non–tandem–queue (NTQ) equivalent periodic pull production system. It is a periodically controlled serial production system in which a single–item is processed at each stage with an exponential processing time to satisfy the Poisson finished product demand. The exact performance evaluation model of this system, using discrete–time Markov processes, is given in Chapter 3. Note that, these systems are difficult to evaluate exactly because of large state spaces involved, an approximate solution method is also proposed. In addition, the results of a numerical experiment is reported in order to investigate the accuracy level of the approximation. A resource allocation problem, related with allocation of both the workload and kanbans in pull production systems, is defined in Chapter 4. Together with some structural properties of such systems, the experimental results that form a basis for the proposed allocation methodology concludes the chapter. In the last chapter of the text, the major contributions of this dissertation research and some further research directions are discussed.

Chapter 2

Performance Evaluation of Production Systems: A Review

In the last decade, there have been numerous attempts for modelling production systems as queueing systems for the purpose of understanding their behavior. So far, the models in the literature usually involved single–product systems with single or multiple stages for tractability purposes. Cases with multiple products, although closer to reality, proved to be quite difficult to tackle analytically. A production system is usually viewed as an arrangement of production stages in a particular configuration, where each stage consists of a single workstation or several identical workstations in parallel. These workstations may consist of workers, machines and work-in-process materials.

Performance evaluation in general is concerned with finding out how well the system is functioning provided that certain policies and parameters are set. Typical performance measures for the evaluation of production systems are throughput, average inventory levels, utilizations, customer service levels and average flow times among others. In obtaining these measures, when analytical techniques

become insufficient often numerical techniques, such as simulation or approximations could be used.

An important part of production research literature appeared in the area of production lines. During the last thirty years, performance evaluation models have been developed for many different types of production lines using exact and approximate approaches.

2.1 Basic Terminology

To avoid ambiguity throughout the text, we specify below the usage of some key terms. Our usage of these terms conforms closely to that in the production literature.

Raw Material: A raw material is a distinct commodity that is supplied to the system, but not processed in the system yet.

Raw Material Supply: It is the process through which the raw materials are supplied to the system.

Operation: An operation is an elemental task which requires resources such as materials, machines, tools, fixtures and labor.

Component: It is used to identify a part, subassembly or assembly on which some operations are performed in the system.

Product: Any commodity produced for sale. Associated with each product, there is a set of operations and a precedence relationship that may constrain the sequence in which those operations can be executed in the system. A component on which all required set of operations are performed is called a product (finished component).

Finished Product Demand: It is the need for a particular product. The demand could come from any number of sources, i.e. customer order, branch warehouse, spare part, forecast or the next production stage.

Item: Item is an inclusive term to denote any distinct product produced or purchased by the system, that is, an end product, assembly, subassembly, component or raw material.

Machine: A machine is an appliance or mechanical device by which some operations are performed on materials.

Fixtures and Tools: In some production systems, the materials are required to be fitted on some fixtures before they are released into the system. Because, without those fixtures the operations could not be accomplished on materials. Also, a machine is required to be equipped with some special tools in order to execute a set of operations.

Buffer: Buffer is a storage area for some physical commodities to be placed in.

Workstation: It is a specific production facility, consisting of labor, machine and buffer, which can be considered as one unit for purposes of planning and scheduling.

Stage: A stage is a set of workstations grouped together to operate more efficiently, either because of some shared resources or because of the dependency relations of some operations.

Tandem Line: A serial arrangement of stages is called a tandem line in which all items are processed at all stages of the system with a unique sequence of flow.

Flowshop: Flowshop is a tandem line in which the flow of materials is unidirectional but, there are alternative stages for some set of operations to be performed on.

Jobshop: It is an arbitrary arrangement of stages in which each item receives processing in a variety of orders, from one stage to another.

2.2 Characteristics of Production Systems

In this section, we informally discuss some of the features, attributes and properties of production systems in order to develop a framework for classification.

Non–perishability: In the literature we survey, the material in buffers is assumed to be non–perishable. That is, it does not decay or loose value, no matter how long it waits in the buffers.

Failures and Repairs: Failures and repairs are related to the machines in the system. When a failure occurs at some machine in the system, it may not process any material until it is repaired and operational again. A variety of assumptions about the conditions under which a failure may occur and after a repair how the operation is continued, the time until the next failure, the time to repair a failure, and so forth, are considered in the literature [7, 33, 39, 50, 54, 55, 114]. Note that, a machine which cannot fail is called a *reliable machine* or otherwise, an *unreliable machine.*

Parallel Workstations: Production stages are built with workstations in parallel for two reasons: either to achieve a greater production rate or to achieve a greater reliability [5, 33, 52, 53, 88].

Up, Down and Operation Times: The productive time between two failures of a machine is called the up–time and consequently the non–productive time spend for the repair of the machine is called the down–time [39, 40]. On the other hand, an *operation time* is the time required to execute a single operation. In this respect, it is assumed that an up–time period is composed of operation times of parts processed in the duration between two consecutive failures.

Synchronous / Asynchronous Production: In most real systems, the machines are not constrained to start and stop their operations at the same instant. Even the stages have fixed and equal cycle time (the time required to accomplish the operations at each stage), uncertain failure and repair times can lead to *asynchronous* operations.

Blocking, Starvation and Decoupling: The presence of buffers between the stages allows them to start and stop independently, as long as the intermediate buffers are neither empty nor full. With the asynchronous flow of material throughout the system, some buffers might become empty or some buffers might become full. Consequently, the production is delayed at some stages because of starvation and blocking. A stage is *starved* when there is no material to be processed in the buffer and it is *blocked* when there is no space in the buffer to put the material it has processed. In this respect, the function of a buffer is to decouple production stages [12, 16, 24, 31, 39, 76].

Discrete / Continuous Production: The production process through which a system produces discrete units of a product is called discrete, otherwise it is continuous. In a *discrete* production environment, individual items are treated, and each requires a non–zero, finite amount of processing time at each stage. On the other hand, systems that treat *continuous* material, share some characteristics, such as the stages can fail and finite buffers can become empty or full synchronously, hence the disturbances are propagated and as a consequence the system waste significant amount of production resources [33, 52, 53]. Automobile industries and oil refineries are good examples for discrete and continuous production environments, respectively.

Saturated / Non–saturated Systems: Materials arrive at and leave a production system in a variety of different ways. In reality, it is always possible that some raw materials are absent or some finished product buffers are full in the system. Because, for some reasons, the shipments of those raw materials and finished products are failed. However, in the literature, it is almost always assumed that the

first stage is never starved and the last is never blocked [7, 12, 33, 39, 49, 50, 69]. Such systems are called *saturated systems*. This assumption is appropriate for addressing the most important performance issue which is the throughput of a production system without considering the environmental (external) uncertainties. But, in a production environment uncertainties in raw material supply and finished product demand are essential. To represent a system with uncertain raw material supply as a saturated system, an additional dummy production stage is attached prior to the first stage of the system. A similar approach could also be applicable in formulating a production system with uncertain finished product demand as a saturated system. That means, an unsaturated system with some external uncertainties can equivalently be represented by a saturated system.

Equivalence of Two Queueing Systems: For two particular queuing systems to be *equivalent* to each other, they must have the same joint queue length distribution [16, 23, 73, 76]. This is simply because most of the key performance measures are obtained using the joint queue length distribution.

Scrap / Rework of Material: *Scrapping* refers to the rejection of bad components which are out of specifications with no possibility to recover the material. When such a bad component is rejected, it leaves the system [56, 115]. But, in some cases, bad components could be returned to production process by some *rework* on that material [115]. This rework could be either a set of last operations to be re–executed or it could be totally a new set of operations.

Assembly / Dis–assembly Operations: In a production line with manufacturing operations, each stage feeds a single buffer and each buffer feeds a single stage. However, in an *assembly* operation, several components from two or more buffers are assembled together to produce a single component. On the other hand, in a *dis–assembly* operation, a single component is separated into several components. A production system in which some stages perform assembly or dis–assembly type of operations is called an *assembly system* [12, 13, 36, 40, 52, 53, 91, 107].

Split / Merge Configurations: *Split* is the configuration in which several workstations are supplied through a single buffer and *merge* is the configuration in which several workstations supply a single buffer [5, 33, 40, 91, 107].

Set–up Times and Batch Sizes: *Set–up time* is the time needed to prepare a set of machines by attaching the proper tooling in order to execute a set of production processes. This preparation is required to be executed when the set of operations is changed. In a single product system, once the set–ups are done the production continues, so that the set–up times could be ignored. But, in a multi product system, every time a machine takes a different product to process with a different set of operations, a set–up is required. In order to minimize, the lost production during these set–up times, the products are processed in batches [1, 6, 21, 22, 60, 116]. The *production batch size* is the amount of a particular item that is produced once a set–up is done. On the other hand, in order to minimize the cost of material handling in the system, the materials are transferred from one stage to another in quantities of *transfer batch sizes.*

Production Scheduling: There are various levels of *scheduling* within a production system. In general, determining when and in which sequence to produce is referred to as production scheduling. To achieve this a desired start or completion time is established for each operation in order to satisfy the finished product demand on time while minimizing the operating costs.

Stockout / Backorder / Lost Sales: *Stockout* is the lack of materials or components which are needed to be on hand in stocks. An immediate demand against a finished product whose inventory is insufficient to satisfy the demand, could be either *backordered* or *lost* [3, 6, 11, 24, 60, 70, 93, 107, 116].

Reversibility: The production system obtained from the original system by reversing the direction of material flow is called a reversed system. The property that the production rate of the reversed system is the same as that of the original system is called *reversibility* [31, 67, 72, 103, 111].

Duality: The *duality* is related with the idea of equivalence between flow of material in one direction and the empty containers in the opposite direction [43]. Note that, in some systems, the behavior of parts in the reversed system is the same as the behavior of empty containers in the original system. Also, a starvation in the reversed system corresponds to blocking in the original system, and vice–versa. As a result, the steady–state joint distribution of parts in the reversed system is exactly the same as the steady–state joint distribution of empty carriers in the original system. This equivalence implies that the original system is *reversible*.

Bowl Phenomenon: It refers to the increase in production rate obtained by unbalancing a production system such that the operation time increases progressively on either side of the central stage(s) or alternatively the buffers in the middle get more and the buffers at the ends get less storage space [29, 37, 49, 51, 67, 74, 81, 84, 86, 104].

Operating Policy: An operating policy is a set of rules and procedures through which the operation of the system is controlled. In most of the studies, it is assumed that machines are not allowed to be idle if they can be operated. That is, whenever a production stage is neither blocked nor starved, it is executing some operations. In this respect, the description of how the system starts and stops production is closely related with the operating policy of the system. There is a variety of policies, such as *push*, *pull* or *conwip* [26, 30, 35, 45, 83, 95, 105].

Push System: The system that authorizes the production in advance of physical demand called a push system. In a *push* system, the demand for finished product and the demand for materials in–process at each stage are forecasted. Then, a release date for each material is computed considering the expected flow time (lead time) up to the final stage. Based on this plan, the materials are released into the system from the first stage and then, these in–process materials are pushed through the stages up to the final stage.

Pull System: Pull systems trigger a production order when the inventory is physically removed from the buffer stock. That is, the amount and time of material flow in the system are determined by the rate and time of the actual consumption in buffers. In *pull* production systems, materials are pulled from one stage to another to meet the finished product demand at the last stage on time [11, 17, 26, 28, 34, 61, 62, 70, 90, 102].

Kanban System: It is an information system for the management of materials in a production system. It acts as the nervous system of a pull production system whose functions are to direct the materials through the stages and to pass information as to what and how much to produce through the use of *kanbans* (cards) [62, 101].

Conwip System: It is a *hybrid* push–pull based production system in which the level of work–in–process materials is kept constant [36, 55, 96]. Under conwip operating policy, only the first stage is operated as a pull system; then the work–in–process materials are pushed between stages without any buffer space limitation up to the final stage, as it is in a push system. Note that, a conwip production system operates as a closed queueing network model in which the same number of jobs are circulating around the system.

2.3 A Classification of Production Systems

There have been many alternative forms of production system described in the literature. Unfortunately, the diversity of these descriptions has made it difficult to organize and synthesize these research studies. To overcome this problem, a large number of articles related with production systems has been considered and a classification scheme is developed. Most of the differentiating attributes contributing to the classification of production systems are discussed in the previous sections. They are similar to the studies reported in the literature (See

Aneke and Carrie [9], Berkley [16, 18], Bitran and Dasu [20], Buxey, Slack and Wild [25], Buzacott and Shanthikumar [27], Dallery and Gershwin [32], Kalkunte, Sarin and Wilhelm [59], Stidham And Weber [100], and Sarker [85] for alternative classification and review of the research studies on production systems).

2.3.1 Deterministic vs Stochastic Systems

The production systems could be classified into two categories according to the way they are formulated:

- *Deterministic*, or
- *Stochastic.*

The nature of system parameters is very important in developing a model in order to evaluate the performance of a production system. If all the parameters of the system are assumed to be *deterministic*, then the model to be developed for a *deterministic system* could be *generative.* Note that, a generative model is capable of finding the best values for various system parameters in order to optimize a given set of performance measures of the system. Price, Gravel and Nsakanda [82] reviewed a variety of optimization models for kanban–based production systems covering tandem production lines, bottleneck workstations, assembly and jobshop production.

Bitran and Chang [19] developed a mathematical programming model of a flowshop structured deterministic production system in order to optimize the operating costs. The system they utilized in their study is a Kanban–type pull production system. One of the major problem in such systems is to determine the number of kanbans (buffer capacity) required to achieve a predetermined level of system performance (See Bard and Golany [14]). In another study, Li and Co [65] proposed a dynamic programming model for the formulation of the

same problem for multi–stage multi–period deterministic production systems. Determination of lot sizes in a deterministic production environment is another problem issue. Philipoom, Rees, Taylor and Huang [79] and Luss and Rosenwein [66] utilized similar integer programming approaches in order to minimize inventory holding costs subject to capacity availability and the required mix of items.

In a production environment, the following items are usually assumed to be the major source of randomness:

- raw material supply,
- production process,
  - human interventions,
  - defective production,
  - failures and repairs,
- finished product demand.

If at least one of the parameters of the system inherits *randomness*, then the model to be developed becomes *evaluative*. That is, an evaluative model can only compute the performance of the system given the pre–determined values of the system parameters. In this respect, almost all of the queueing models of production systems are assumed to be evaluative.

The nature of the system parameters is assumed to be either *stationary* or *non–stationary* with respect to time scale. In terms of modelling and analysis, dealing with stationary–deterministic parameters is the easiest and dealing with non–stationary stochastic parameters is the most difficult one.

Production systems with stationary stochastic demand and production processes received a great attention in the production research literature. In most of the

analytical models (e.g. Altıok and Stidham [7], Brandwajn and Jow [24], Hillier and Boling [47], Karmarkar and Kekre [60], Mitra and Mitrani [70], Siha [90], Springer [97], and Wang and Wang [107]), the Poisson demand arrivals and exponential processing times are simplifying assumptions that preserve the Markovian property, even if it leads to pessimistic results on performance of the system.

Since, the coefficient of variation is more important parameter than the shape of the processing time distribution in affecting the performance of the production system (as it is reported in Hillier and So [49]), the phase–type distributions could be used for approximating more general distributions (See Altıok [2]). The attempt by Yao and Buzacott [113] to transform a queueing network with general processing times into an approximately equivalent exponential network is another approximation approach through exponentialization. As a consequence, Altıok [3, 4], Altıok and Stidham [7], Berkley [17], De Koster [33], and Hillier and So [50] utilized phase–type distributions for processing times in their models in order to represent more general processing time distributions.

On the other hand, Gershwin [39, 40] utilized non–exponential distributions in his model. The processing times were deterministic and workstations are subject to failures. He assumed that, time to failure and repair time were geometrically distributed.

2.3.2 Manufacturing vs Assembly/Dis–assembly Systems

The operations involved in the production process could have some degree of complications;

- *Manufacturing–type operation*,
- *Assembly–type operation*,
- *Disassembly–type operation*.

A manufacturing type of operation is involved with a production process in which one unit of material is withdrawn from the buffer for processing and after completion one unit of material is sent to the buffer. In most of the analytic studies in the literature, manufacturing type of operations are utilized since it is the simplest type of operation to be formulated mathematically. A production system in which all operations are of this type is called a *manufacturing system.*

In an assembly type of operation, the availability of all assembly parts is necessary at the feeding buffers in order to start the involved assembly process. This way, an assembly operation produces one unit of assembled component by withdrawing the required number of assembly parts from buffers. On the other hand, an operation that produces more than one unit of material from one unit of input material is called a dis–assembly type of operation. An *assembly system* is a production system in which some assembly and/or dis–assembly type of operations are performed (See Baker, Powell and Pyke [12, 13], Gershwin [40], Hodgson and Wang [52, 53], Smith and Daskalaki [91], and Wang and Wang [107]).

2.3.3 Single–Stage vs Multi–Stage Systems

A production system is usually assumed to be composed of several stages in tandem. In each stage, a set of production operations is to be executed through the use of some machines, fixtures and tools in order to produce the finished product. The network configuration of a production system regarding the stages could be either:

- *Single–stage*, or
- *Multi–stage.*

Although, developing a single–stage model of a production system is almost unrealistic, in the literature researchers have continually developed such models

mainly for two reasons.

First, a single–stage model is easier to formulate and solve, because it has less number of parameters than a multi–stage model. For multi–stage production systems with complicated characteristics that are analytically intractable to formulate and solve, the analysis of a single–stage model may provide helpful insights (See Altıok [3], Altıok and Shiue [6], Bitran and Tirupati [22], and Zipkin [116]).

Second, a single–stage model is mostly utilized in an approximate decomposition technique which is applicable for multi–stage production systems. In a typical decomposition approach, all production stages are analyzed separately, then the results are aggregated with resolving the inter–relation between the stages in order to obtain the overall performance of the whole system. For most of the production systems studied in the literature, an exact decomposition could not be possible. In case of an approximate decomposition, the trade–off between the precision of the results and the complexity of the computations becomes important. There is a considerable amount of literature on the performance evaluation of multi–stage tandem production lines through the use of approximate decomposition techniques, e.g. Altıok [4], Berkley [17], Brandwajn and Jow [24], Gershwin [39], Hillier and Boling [47], Hong, Glassey and Seong [54], and Springer [97].

Note that, a multi–stage system in which alternative routes for materials are allowed with utilizing split/merge configurations is called a flowshop. Allowing this type of configuration in production systems increases the scheduling flexibility especially in case of machine failures. Altıok and Perros [5], De Koster [33], Gershwin [40], and Smith and Daskalaki [91] developed efficient approximate decomposition techniques for production systems in flowshop configuration.

Although, the computational complexity remains feasible even for large–scale systems through the use of such approximate techniques, the well–known queueing models of serial production systems could be exactly evaluated up to three stages in tandem (See Altıok and Stidham [7], Badinelli [11], Deleersnyder *et al.* [34],

and Muth and Alkaff [75]).

On the other hand, the jobshop configuration is the most difficult case for modelling and analysis. Dealing with every possible route between stages complicates the formulation of the system. In a jobshop system, several different products are assumed to be processed using the same facilities. This causes another complexity in modelling and formulation of the jobshop production system with including the related scheduling issues. So, there is no analytically tractable multi–item jobshop model reported in the literature beyond a few studies formulating the system under very restrictive assumptions (See Akyildiz and Huang [1], Bitran and Tirupati [21], and Zipkin [116]).

2.3.4 Single–Item vs Multi–Item Systems

The production systems could be classified into two categories according to the number of products produced in the system;

- *Single–item*, or
- *Multi–item.*

In a single–item system, there is only one finished product to be processed in the system. The set of operations to be executed at each stage is unique. For that reason, the set–ups required at each stage in order to start production are done once and for all. Modelling a single–stage system with a single demand arrival and production processes is easier relative to a multi–item system in which there are set–ups and production batches to be scheduled.

With respect to modelling and analysis, the assumption of producing single item with random operation times is not so unrealistic. Because, the random behavior of operation times could be accepted as an alternative representation of the

variation of operation times from one product to the other in a multi–item environment. But, a multi–item model with assuming zero set–up times between different operations and one unit batch sizes might be unrealistic to some extend. In spite of this, a tremendous number of single–item performance evaluation models with random processing times is developed in the literature (See Altıok [4], Altıok and Stidham [7], Badinelli [11], Brandwajn and Jow [24], De Koster [33], Deleersnyder *et al.* [34], Hillier and Boling [47], Hillier and So [50], Mitra and Mitrani [69, 70], Muth and Alkaff [75], Springer [97], Wang and Wang [107]).

Producing more than one product introduces some resource sharing issues at workstations which process more than one product. Then, the terms set–up and batch size (in order to minimize the lost productive times due to set–ups) come into the scene and complicate the formulation and analysis of the system. In recent years there have been considerable developments in modelling multi–item production systems analytically. Akyildiz and Huang [1], Altıok and Shiue [6], Bitran and Tirupati [21, 22], Karmarkar and Kekre [60], and Zipkin [116] developed multi–item queueing models for production systems either for single–stage or for more general network configurations under some restrictive assumptions.

Note that, if two different production processes producing two different items have no interactions on production resources during the whole production process, then the system could be evaluated using two single–item models separately.

2.3.5 Reliable vs Unreliable Systems

In a production system:

- the production facilities could be either;
  - *Reliable* (machines cannot fail), or
  - *Unreliable* (machines can fail),

- the production operations could be either;
  - *Reliable* (no defective parts produced), or
  - *Unreliable* (with scrap or rework).

The reliability issue primarily refers to the production facilities in the system. Since, tracking of these failures and repairs in the system complicates the formulation and analysis, a lot of studies reported in the literature deals with reliable production systems in which machines cannot fail (See Baker, Powell and Pyke [12], Berkley [17], Brandwajn and Jow [24], Buzacott, Price and Shanthikumar [28], Hillier and Boling [47], Mitra and Mitrani [69, 70], Muth [71], Siha [90], Springer [97], and Wang and Wang [107] for reliable systems).

Some of the studies in the literature are focused on time to failure and repair time distributions. The most common and analytically more tractable assumption is, these distributions are exponential, as in the studies of Altıok and Stidham [7], De Koster [33], Hong, Glassey and Seong [54], and Hopp and Spearman [55]. In the literature, there are few analytical models with non–exponential failures and repairs. The analytical model studied by Gershwin [39, 40] is an example for non–exponential (geometrically distributed) failures and repairs.

The operation times could be alternatively defined in terms of *operation completion times* in order to incorporate both the variability in operation times and unreliability due to failures and repairs at machines and defectives in the production process (See Altıok [3, 4], Altıok and Stidham [7], and Hillier and So [49, 50]). Altıok and Stidham [7] further utilized a two–stage phase–type distribution which is an exact representation of the distribution of operation completion time of parts in a system of exponential servers subject to exponential failures and repairs.

In another view, reliability may refer to production operations. An operation which yields scrap or requires rework of material could be defined as an unreliable operation (See Jafari and Shanthikumar [56] and Yu and Bricker [115]). Almost

all of the studies in the literature deal with reliable systems in which non–defective parts are produced.

2.3.6 Push vs Pull Systems

The production systems could be classified into following categories according to the control strategies utilized;

- *Push*,
- *Pull*, or
- *Hybrid*.

The modelling and analysis of production systems within the framework of *push* control strategy received a great attention in the production literature (See Buzacott [26], Deleersnyder *et al.* [35], Pyke and Cohen [83], and Spearman and Zazanais [95]). In push systems, the independent demand for finished products and the dependent demand for materials in-process at each stage are forecasted. Then, a release date for each material in-process is computed considering the expected flow time (lead time) up to the final stage. Materials Requirements Planning (MRP) makes it possible to construct a time-phased requirements plan for this system. Based on this plan, the materials are released into the system from the first stage and then, these in-process materials are pushed through the stages up to the final stage. So, any workstation operating in a push system could not stay idle if the input queue is not empty. These systems are controlled through the work-in-process (WIP) inventories in the system. Thus an incorrect forecast or drastic changes in demand, in most cases, are overcome by the in-process inventories including the safety stocks which can result in unnecessarily high carrying costs. This is because of the difficulties faced during the renewal of the production plan for each process and for each part in the system.

After 1970s, the *Just-In-Time* (JIT) philosophy has been introduced into the production literature and it has produced an alternative production control system (Kanban System) as offspring. The basic tenets of the JIT philosophy are the elimination of waste (in terms of materials, manpower, productive time, energy etc.), participation of employee in decision-making to improve productivity, participation of supplier for reduced lead-times and total quality control. To a certain extent, JIT has come to refer to all that is good in production. Golhar and Stamm [42] offer a comprehensive review of the JIT literature and provide a framework for classifying the related JIT literature.

The first successful example of development and implementation of JIT concept as a material management system has been reported by Sugimori *et al.* [101] at Toyota whose production system is actually operated by means of kanbans. The kanban material management system is well described by Sugimori *et al.* [101] and Kimura and Terada [62]. It acts as the nerve of the JIT production system whose functions are to direct in-process materials just-in-time to the workstations and to pass information as to what and how much to produce.

When the JIT philosophy is applied to a material management system, it is called a *pull* system, which means that the amount and time of material flow are determined by the rate and time of the actual consumption. In pull production systems, the kanban system pulls in-process materials from one workstation to another to meet the demand at each workstation at the right time. There are many alternative forms of pull production control in practice; Badinelli [11], Berkley [17], Buzacott [26], Buzacott, Price and Shanthikumar [28], Deleersnyder *et al.* [34], Golhar and Sarker [41], Karmarkar and Kekre [60], Mitra and Mitrani [69, 70], Sarker and Parija [87], Siha [90], and Tayur [102, 103]. However, the common thread that distinguishes the pull system from conventional push method of production control is the existence of finite buffers for in-process materials and the triggering process for workstations to start and stop producing depends on the inventory level of the succeeding buffer stock.

Some implementations of pull production systems utilize two-card while others use only one card or some of them use computerized systems (no card at all). See Berkley [17], Karmarkar and Kekre [60], Kimura and Terada [62], and Sugimori *et al.* [101] for two–card; Deleersnyder *et al.* [34], Karmarkar and Kekre [60], Mitra and Mitrani [69, 70], and So and Pinault [93] for single–card; and Kim [61] for computerized kanban systems. In a computer-controlled pull production system, all of the transactions within the system can be collected and recorded automatically and instantaneously, so that, the continuous monitoring of the whole system could be possible. However, in a two-card kanban system, the production kanban cards serve as work orders to replace the empty containers of finished items withdrawn from the output buffer stock of the workstation and the withdrawal kanban cards act as material requisitions to the input buffer stock of the workstation. Finally, in single–card systems in which the workstations are physically located close together so that the material handling function between the workstations could be ignored and only the production kanbans are utilized.

Generally, each one of the push and pull type control strategies is thought to have both advantages and disadvantages. In this respect, there is a great potential in developing a system that possesses the benefits of both pull and push systems and can be used in a wide variety of production environments. There are several hybrid control strategies reported in the literature, i.e. Deleersnyder *et al.* [35], Duenyas and Hopp [36], Hodgson and Wang [52, 53], Hopp and Spearman [55], and Spearman, Woodruff and Hopp [96]. Among those the most well–known is the *conwip* (constant work–in–process) control strategy which is first introduced by Spearman, Woodruff and Hopp [96].

2.3.7 Periodic vs Continuous Review Systems

In the context of the classical inventory theory, the production systems are classified according to the management and control of work–in–process inventories as;

- *Continuous review*, or
- *Periodic review.*

In order to support decision making in inventory management and production control, a production system is to be reviewed and the status of the system should be monitored. According to the characteristics of the production system and the conditions of the environment, this review process is done either periodical or continuous basis.

In a periodic review system, the status of materials flow and the production at all stages are reviewed at regular intervals. The material withdrawals and all other production activities start immediately after the review as decided with respect to the status of the system. The time required for the review and decision making process is generally assumed to be negligible. The periodic review models of production systems mostly developed for the analysis of pull systems (See Berkley [17], Deleersnyder *et al.* [34], and Kim [61]).

The continuous review production systems have been investigated by many researchers (See Altıok [3], Altıok and Shiue [6], Badinelli [11], and So and Pinault [93]). Note that, in almost all of the tandem queueing models of production systems it is assumed that the system is reviewed continuously.

Srinivasan and Lee [98] studied a production system in which the time interval between two successive reviews is a random variable following an arbitrary distribution. Under a cost structure which includes set–up, holding and backorder costs, they obtained the optimal policy by minimizing the expression for the expected cost per unit time.

2.3.8 Instantaneous vs Periodic/Batch Order Systems

In a production system, the following types of items are assumed to be ordered;

- *Raw materials and parts to be* **purchased** *from outside vendors,*
- *Work–in–process materials to be* **handled** *between stages,*
- *Work–in–process materials to be* **processed** *at each stage,*
- *Finished products to be* **shipped** *to the customers.*

In most of the analytical models reported in the production literature the raw material supply is assumed to be infinite. Because of this, the orders related with raw materials and parts to be purchased from outside vendors are considered to be external to the system.

The ordering policy of a production system in order to replenish the orders for production, material handling and finished product shipment is another feature in the context of classification:

- *Instantaneous ordering,*
- *Periodic ordering* (fixed period, T time units; variable quantity of items ordered),
- *Batch ordering* (fixed quantity, Q units of items ordered; variable period length).

In this respect, almost all of the tandem queueing models of production systems are instantaneous order systems (See Altıok and Stidham [7], Brandwajn and Jow [24], Gershwin [39], Hillier and Boling [47], Mitra and Mitrani [69, 70], and Springer [97]).

There are relatively few analytical studies in the literature that investigates periodic or batch ordering policies since, dealing with periodic or batch orders in a production system is more difficult to formulate and analyze (See Berkley [17], Bitran and Tirupati [22], Karmarkar and Kekre [60], and Kim [61]).

As the physical distance between workstations increases, the instantaneous transfer of materials from one stage to the other becomes impractical. For these situations, the solution is to perform the material handling operations periodically. When material handling operations are carried out periodically, each workstation must have both an input and an output buffer stock. Within the period, material requirements of a workstation are satisfied from its input buffer stock and the processed items are placed in its output buffer stock. At the end of the period, either all of the processed items collected at the output buffer of a workstation or some of the processed items depending on the size of a transfer batch are transferred to the input buffer of the next workstation in the production route. Further in order to minimize material handling costs, the handling of materials could be made in batches.

2.3.9 Conclusion

So far, a number of major attributes of production systems are examined and most of the distinguishing analytical studies in production literature are reviewed in order to build the framework for a classification scheme.

Illustratively, Hillier and Boling [47] have reported one of the pioneering studies on finite queues in series with exponential service times. Brandwajn and Jow [24], and Springer [97] studied the same system to improve both the accuracy and the complexity of the computations. Next, Altıok and Stidham [7], and Hong, Glassey and Seong [54] extended this work with including exponential failures and repairs for the servers. On top of this unreliability, split and merge configurations are allowed in a model for continuous production environments in De Koster [33]. On the other hand, Gershwin [39] studied a tandem queueing system in which service times are deterministic but servers subject to geometric failures and repairs. Further, in order to generalize the service times, Altıok [4], and Hillier and So [50] utilized phase–type distributions in their models.

This way, a production system could be well described by identifying the attributes discussed in the previous sections, hence one can compare the results of the related research studies.

There has been a significant accumulation in the literature on tandem queuing models of production lines within the last thirty years. Various design and operating aspects of these systems have been studied. Some of the emerging design problems in the literature are **workload and buffer capacity allocation** for which no simple solutions exist in non-trivial cases. The *exact* analysis mostly focused on the special structure of the underlying Markov chains and solves the associated Chapman-Kolmogorov balance equations for the steady state probabilities [7, 11, 34, 47, 75].

As the state space of the system under study increases, the use of exact methods becomes computationally infeasible because of the magnitude of computational effort and the computer space requirements. The only remaining viable approach for the analysis of large-scale systems appears to be the use of *approximation* techniques. In an approximate analysis the system is decomposed into smaller (one or two-node) subsystems which are analyzed in isolation and then relates them to each other in an iterative manner to obtain the performance measures of the whole system [4, 17, 24, 33, 39, 47, 54, 97, 107].

2.4 Pull Production Systems: A review

In practice, there are many alternative forms of pull production systems which differ in some design and operating characteristics [18]. Among others the well known pull systems are the kanban–controlled pull production systems. Some of the kanban implementations utilize two–cards while others use only one card for the flow of information through the stages of the system related with the production and material requisitions.

The simplest form of kanban production control system has a fixed order quantity of one unit and is called a *base stock* system. There exists a single inventory buffer between each workstation. The maximum inventory level permitted in this intermediate buffer is called the base stock level. Each time the downstream workstation (the one being closer to final demand) requires work–in-process material, it withdraws one unit from the intermediate buffer. Production of one unit is then triggered at the upstream workstation since the inventory level falls below the base stock level. Production stops (workstation is blocked) when the inventory level of the buffer reaches the base stock level. Note that, the downstream workstation pulls the required amount of materials which are processed at the upstream workstation.

The base stock system defined above operates exactly the same as a *tandem queue* with communication–system blocking. This type of blocking means an upstream workstation is not allowed to start processing a material until space is available in the intermediate buffer. Recall that, there is another type of blocking mechanism that is called production–system blocking. It occurs when, at the moment of process completion in the upstream workstation the intermediate buffer is full. See Onvural and Perros [76] on the equivalencies of blocking mechanisms in queueing networks.

Many of the kanban systems described in the production literature are equivalent to a tandem queue [16, 18]. Berkley [16] showed when and how tandem queueing models can be used to obtain the performance measures of kanban–controlled pull production lines. He showed that a two-card kanban–controlled production line is equivalent to a tandem queue, with communication–system blocking, if and only if the two-card kanban–controlled production line is operated with a fixed order quantity of one or with a fixed order cycle time of zero (see Figure 2.1). He also gave some numerical examples to demonstrate this equivalence relation of two-card systems and tandem queues. The Markov chain model, the states and the corresponding steady–state balance equations of the tandem queues were generated following the procedure by Hillier and Boling [47].

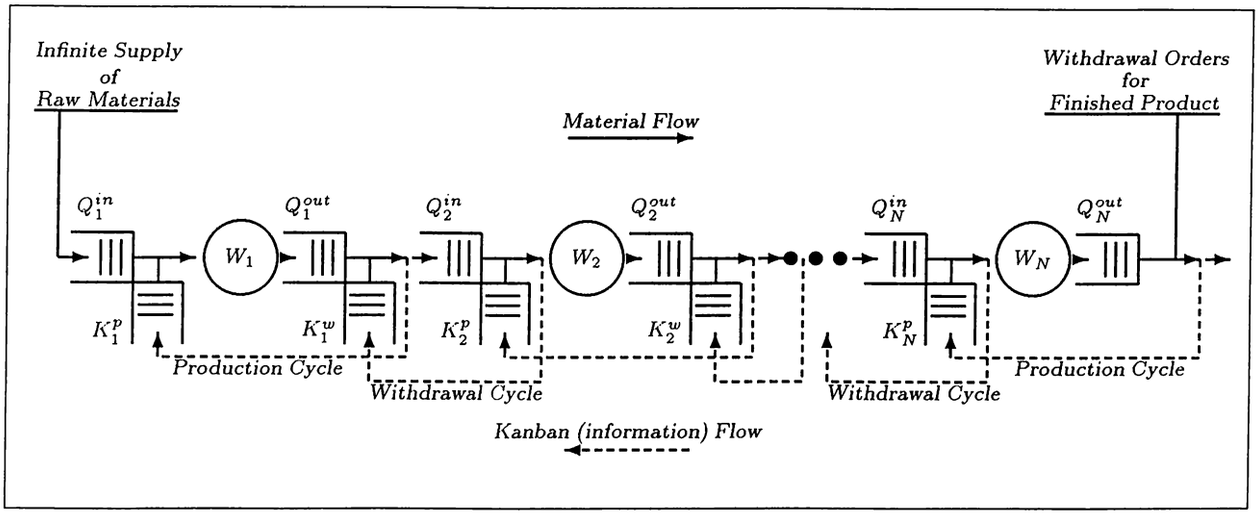


Figure 2.1: Tandem arrangement of workstations (W_j: $j = 1, 2, ..., N$) in a two-card kanban-controlled pull production line. Each workstation has both an input material queue and an output material queue, Q_j^{in} and Q_j^{out}, respectively. In the context of Kanban System, K_j^p refers to the number of production kanbans and K_j^w refers to the number of withdrawal kanbans at stage j.

Using the classification scheme, a pull production system;

- having stationary stochastic demand arrival and production processes,
- producing single-item with manufacturing-type operations,
- with a configuration of multi-stages in tandem and where each stage composed of a single machine,
- with instantaneous transfer of work-in-process materials between stages,
- monitoring the status of the system with a continuous review policy, and
- releasing instantaneous production orders at times when items are withdrawn from the related buffers,

is said to be equivalent to a tandem queue as stated by Berkley [16]. He has introduced a classification as either tandem queue (TQ) equivalent or not in his

study. For TQ equivalent pull production systems, any decomposition approximation existing in the literature developed for tandem queues are well applicable. There are very accurate tandem queue approximation procedures in which the relative error on performance measures is less than one percent. Buzacott, Price and Shanthikumar [28], Mitra and Mitrani [69, 70], Siha [90], So and Pinault [93], and Wang and Wang [107, 108] developed such TQ equivalent models of pull production systems. Note that, tandem queues can also be used to obtain upper bounds for the production rates and the work–in–process inventory levels of NTQ equivalent pull production systems [16].

Mitra and Mitrani [69] described an evaluative model for a single–card kanban system equivalent to a tandem queue (see Figure 2.2). The finished products were assumed to be immediately withdrawn from the system. In another study of Mitra and Mitrani [70], that is the second in a publication series on a particular scheme of coordination between production cells, an exogenous demand process was introduced so that, the first study turned out to be a special case corresponding to heavy demand arrivals. Analyzing the sample path descriptions of both cases, they also showed that systems under consideration become equivalent to a tandem queue when the input material queues are eliminated.

So and Pinault [93] alternatively decomposed the production line into individual M/M/1 queues with bulk service in estimating the amount of buffer stocks needed at each station in order to meet a predetermined level of performance (average percentage of backlogged demand). The analysis of the individual workstations were then combined using a heuristic procedure to approximate the performance of the entire system. They also reported that since workstations were assumed to have an infinite supply of raw materials, the proposed method is only valid when the number of kanbans is sufficiently large to prevent workstations from starvation. Buzacott, Price and Shanthikumar [28] considered a kanban–controlled serial pull production line. They proposed simple approaches for approximating the performance of the system and also provided some insights into behavior of the system.

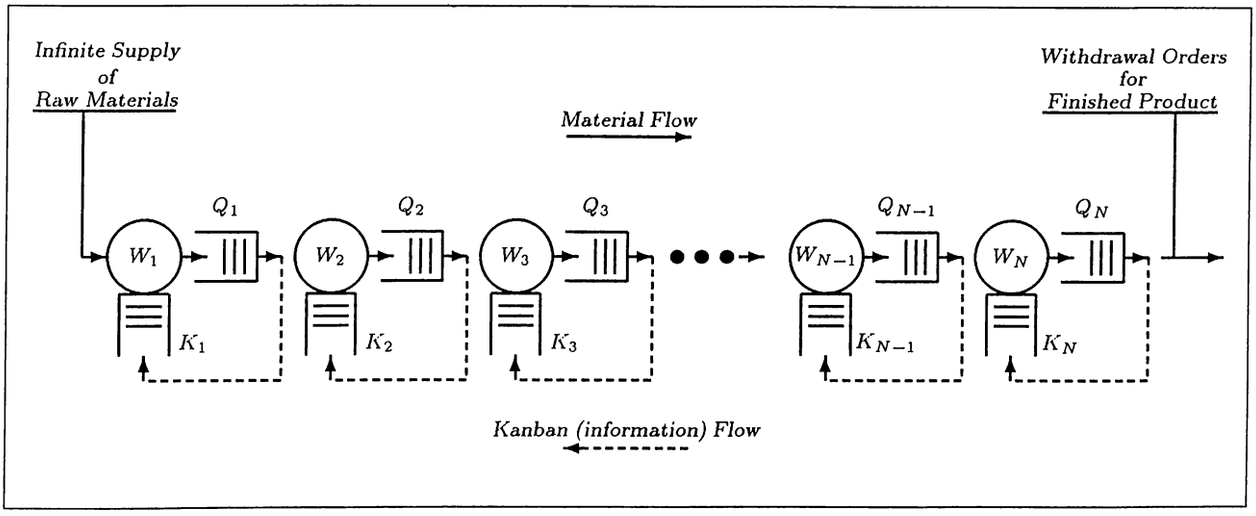


Figure 2.2: Single-card kanban-controlled pull production line is a tandem arrangement of workstations in which there is only an output material queue for each workstation.

Wang and Wang [107, 108] developed a Markov model for determining the number of kanbans required in a serial JIT production system, in which assembly and dis-assembly type of operations were allowed. By evaluating Markov chains for alternative number of production kanbans, they found a solution that minimizes total inventory holding and shortage costs. Since the order points were assumed to be one, the system could be operated with only one withdrawal kanban between each pair of workstations. They also proposed a decomposition approximation in which the workstations were assumed independent. Jordan [58] modeled workstation interdependence by formulating a two-stage line as a continuous-time Markov chain to determine the profit maximizing number of kanbans.

Siha [90] developed a continuous time Markov model for pull production systems in order to analyze some allocation patterns of kanban capacity and mean production time over the workstations of the system. The results were contradictory with some of the findings in the literature where the bowl pattern is suggested. However, some design guidelines are reported that could be useful in applications.

Recently, Berkley [17] introduced a decomposition approximation using imbedded Markov chains for kanban–controlled pull production lines with periodic material handling and Erlang processing times. In these systems, the number of withdrawal kanbans required is a direct function of material handling frequency or the withdrawal cycle time. In his study, several examples were given to show how the approximation could be used to find the required number of kanbans, the required withdrawal cycle time or both. Tayur [102, 103] developed some structural and theoretical results that characterize the dynamics of kanban–controlled serial lines and provided insight into their behavior and helped greatly in order to reduce the effort required in a simulation study. Based on these results, he developed a heuristic for the allocation of kanbans to a balanced line.

It was generally thought that kanban–controlled pull production systems are not applicable to multi–item jobshop environments. However, Gravel and Price [44] showed how the kanban method of control can be adapted to a multi–item jobshop environment. They illustrated the adaptation with examples drawn from a pilot study. Before the implementation stage, they extensively tested the system through the use of a simulation model. As a result, actual performance of the system indicated that both WIP inventories and cycle times are reduced. The reduction of production lead time allows inventories to be reduced without incurring high stockout costs. For a pull production system operating in a varying demand environment lead time reduction is crucial. Karmarkar and Kekre [60] studied the effect of batch sizing policy on the production lead times and hence on the inventory levels and on the performance of the system. They utilized approximate Markovian models in formulating a two–stage multi–item Kanban system.

Buzacott [26] developed a linked queuing network model to describe the behavior of a kanban–controlled production system. He pointed out that kanban–controlled systems can be shown to be particular cases of a more general inventory level triggered approach to production control in multi–stage systems.

Altıok [4] considered a single–stage pull system within the context of single facility production/inventory system with an (R,r) continuous inventory policy. This particular policy, indicates that the workstations start producing as soon as the stock on-hand drops to r and it continues producing until the stock on-hand reaches R. On the other hand, Badinelli [11] presented a descriptive model for steady–state performance of a serial inventory system in which each facility follows a continuous–review pull policy under stochastic demand. In the model of this serial inventory system, the popular kanban control system was represented by a conventional (Q,R) policy, in which each downstream facility orders a fixed amount, Q, from the upstream facility whenever the inventory position at the intermediate buffer reaches a reorder point, R.

In a JIT production environment, a supplier is expected to frequently deliver goods in small lot–sizes. Many suppliers are responding to this by producing goods in big lots and carrying excess finished good inventories. There is a few analytical studies directed to examine the economic impact of such supply strategies. Golhar and Sarker [41] developed a general cost model considering both supplier (of raw material) and buyer (of finished products) sides. They determined an optimal ordering policy for procurement of raw materials, and a production batch size to minimize the total cost. Then, in a further study Sarker and Parija [87] developed a mathematical model to find optimal batch size for a JIT production system operating under a fixed–quantity, periodic delivery policy. The system they considered procures raw materials from suppliers, processes them and finally it delivers the finished products demanded by outside buyers at fixed interval points in time.

A variety of approaches reviewed in this section are analytical studies dealing with mathematical models of performance evaluation in stochastic pull production systems. In most of the studies, uncertainties such as the variability in processing and demand inter–arrival times are assumed to be exponential. Most of the researchers proposed an approximate decomposition procedure for large–scale systems.

2.5 Potential Research Area

In the recent years, with parallel to the developments in manufacturing and computer technology, classical production facilities are being replaced by advanced systems and the companies have entered into a new age of global competitiveness. Because of the scarcity of world's natural resources, it becomes necessary to look for ways of improving productivity and reducing costs through a system of waste elimination. One such system is the JIT production system in which the waste is greatly reduced by adapting to changes. Thus, having all processes produce the necessary parts at the necessary time and having on hand only the minimum stock needed to hold the processes together. The pull production system is a way of implementing the JIT principles, with the finished product 'pulled' from the system at the actual demand rate. Since production systems generally suffer from demand, production and supply fluctuations, a stochastic model might facilitate the design and operation of such systems.

There has been a number of attempts in the literature to develop analytical models for the performance evaluation of stochastic pull production systems. Most of the existing studies address TQ equivalent systems. In the light of the proposed classification scheme, there are numerous NTQ equivalent pull production systems to be considered in a research study:

- periodic review systems with:
  - exponential/non-exponential distributions,
  - periodic/batch transfer of in-process materials,
  - batch ordering.
- continuous review systems with:
  - non-exponential distributions,
  - batch transfer of WIP,
  - batch ordering.

- multi–item multi–stage systems with:
  - non–zero setup times,
  - priority scheduling.

The major decisions for pull production systems are concerned with the allocation of workload (operations) to workstations, the determination of the number of kanbans between workstations and the production/transfer batch sizes.

As a result, modelling and analysis of pull production systems would attract more attention from researchers in a number of directions, especially with approximate evaluation methods handling more general inventory level triggered multi–stage multi–item pull production systems.

Chapter 3

Model Development: Periodic Pull Production Systems [1]

In the context of operational design, the periodic review and periodic material handling issues are the widely encountered characteristics in practice for pull production systems [61]. In such periodic pull production systems, the transfer of WIP inventories between stages and the release of collected kanbans as production orders to workstations are initiated at the beginning of the periods. In this study we investigate the steady–state behavior of a NTQ equivalent periodic pull production system. To this end it is formulated as a *discrete–time Markov process*. Note that, a discrete–time model can satisfactorily approximate the continuous model by sufficiently squeezing the time periods.

In this chapter, we developed the exact and approximate analytical models for performance evaluation of periodic pull production systems. The system we considered is described in the following section. Then in the next section, the formulation of the production system under consideration together with some key performance measures is presented. The third section is for the development of

[1] This chapter draws heavily on the (forthcoming) paper of Kırkavak and Dinçer [63].

an approximate model that decomposes the system into individual single–stage systems and aggregates the single–stage results to obtain the performance of the whole system. In the fourth section, the results of a numerical experiment on the accuracy level of the approximation is summarized. Finally, we conclude the chapter with a discussion on the applicability of both exact and approximate performance evaluation models.

3.1 Description of the System

This basic production system consists of N stages in tandem (see Figure 3.1). At each stage there is only one workstation processing a single–item, so that the term "stages" and "workstations" could be used interchangeably. W_j $(1 \leq j \leq N)$ represents workstations. At any workstation W_j, there are two stocks Q_j^{in} and Q_j^{out} respectively for storing incoming and outgoing WIP inventory items at workstation W_j. W_1 is responsible for the first operation of the item, converting raw material RM (or alternatively denoted by component C_0 stored in stock Q_1^{in}) into component C_1 (stored in stock Q_1^{out} till the end of the period then instantaneously transferred to stock Q_2^{in}). W_j $(2 \leq j \leq N-1)$ converts component C_{j-1} (from stock Q_j^{in}) into component C_j (stored in Q_j^{out} till the end of the period then instantaneously transferred to stock Q_{j+1}^{in}). Finally, W_N performs the final operation of the item, converting component C_{N-1} (from stock Q_N^{in}) into finished product FP (could be alternatively denoted by C_N and stored in Q_N^{out} till the end of the period then instantaneously transferred to Q_{FP} or alternatively Q_{N+1}^{in}).

The maximum number of items allowed in stocks Q_j^{out} and Q_{j+1}^{in} is K_j which is the maximum capacity of buffer space allocated for component C_j at workstation W_j. Note that, I_j^{in} $(0 \leq I_j^{in} \leq K_{j-1})$ and I_j^{out} $(0 \leq I_j^{out} \leq K_j)$ denote the level of WIP inventories at stocks Q_j^{in} and Q_j^{out} $(1 \leq j \leq N)$, respectively. Consider the total number of component C_j items between workstations W_j and W_{j+1}, then the inequality for the level of WIP inventories at stocks Q_j^{out} and

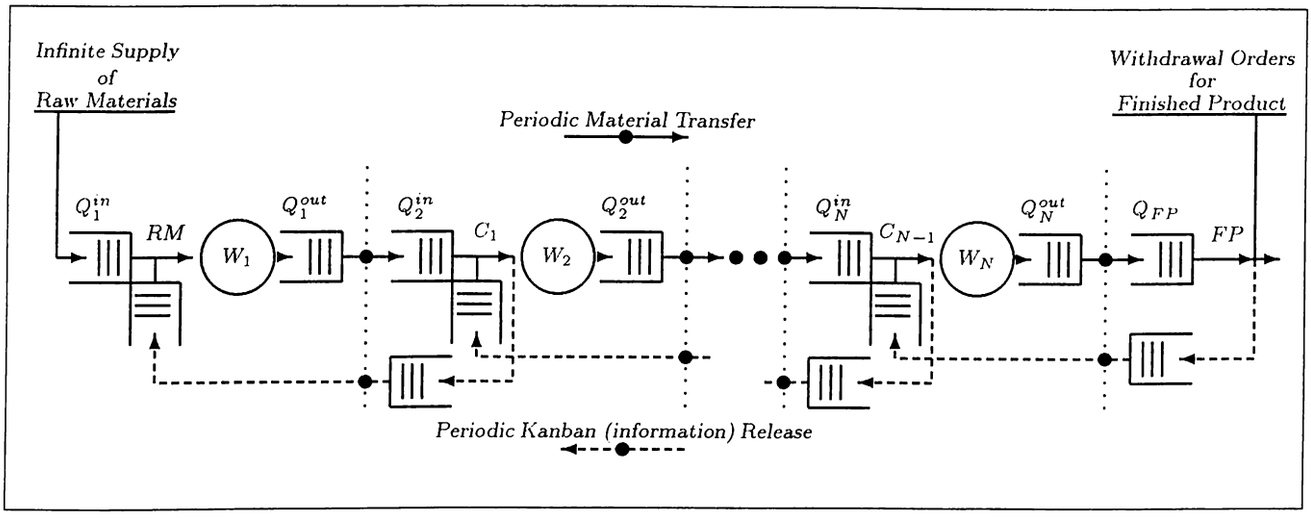


Figure 3.1: Kanban-controlled periodic pull production line.

Q_{j+1}^{in}; $I_j^{out} + I_{j+1}^{in} \leq K_j$ holds for all stages. However, at the finished product stock Q_{FP} (or alternatively Q_{N+1}^{in}) backordering is allowed up to a maximum allowable amount of B_{FP}. The inventory level at finished product stock is I_{FP} (or alternatively I_{N+1}^{in}, $-B_{FP} \leq I_{N+1}^{in} \leq K_N$).

For simplification, the rate of supply of RM is assumed to be infinite. Since a kanban-controlled pull production system typically operates with small lot sizes, it is assumed that one kanban corresponds to one item of inventory in this formulation. The analysis can be easily extended to cover the systems operating with lot sizes greater than one at a cost of dimensionality problem in evaluating transition matrices.

In these periodic pull systems, the production is only initiated just for the replenishment of items removed from the buffer stocks during the material handling and inventory review period of T time units (transfer/review cycle time). That is workstation W_j produces components C_j in order to maintain the inventory level of stock Q_{j+1}^{in} at K_j. Without loss of generality, the production system is assumed to have the same transfer/review cycle times among all stages.

At the end of period k, first the components collected at outgoing stocks ($I_j^{out}(k)$ units of component C_j) are transferred to incoming stocks Q_{j+1}^{in} in the context of material handling function. Then, in the context of production/inventory control function, the total number of kanbans released as production orders to start production of components C_j at workstation W_j for the period $k+1$ becomes $K_j - I_{j+1}^{in}(k+1)$. Note that, the time convention used in this study is *beginning of period* in evaluating any state parameter of the system. But, $I_j^{out}(k)$ denotes the inventory level at stock Q_j^{out} at the end of the period k, since all output buffers are empty at the beginning of any period.

The two sources of uncertainty considered in the production system are the demand and processing time variability. The demand for the finished product FP arrives with exponentially distributed inter–arrival times to the buffer stock Q_{FP}. The mean inter–arrival time of the demand is $(1/\lambda)$ time units. Although backordering is allowed, an arriving finished product demand finding an amount of B_{FP} backordered FP items (that means, I_{N+1}^{in} or alternatively I_{FP} is equal to $-B_{FP}$) is lost. The processing times are assumed to be exponentially distributed. The mean processing time at workstation W_j is $(1/\mu_j)$ time units. For simplification, the workstations are assumed to be reliable. As a result, there are $N+1$ stochastic processes involved in the formulation of the system.

3.2 Exact Performance Evaluation Model

3.2.1 The Formulation of the system

Satisfied/backordered finished product demand during a period. Considering the Poisson demand arrival process for finished product FP, $\{N_D(t), t \geq 0\}$, and the satisfied/backordered demand during period k, $D_s(k)$ $(0 \leq D_s(k) \leq I_{FP}(k) + B_{FP}$, because of backordering), the probability distribution is:

$$P\left[D_s(k) = d_s^0 \mid I_{FP}(k),\, B_{FP}\right] = \begin{cases} \frac{(\lambda T)^{d_s^0}}{d_s^0!} e^{-\lambda T} & 0 \le d_s^0 < I_{FP}(k) + B_{FP} \\ 1 - \sum_{l=0}^{d_s^0 - 1} \frac{(\lambda T)^l}{l!} e^{-\lambda T} & d_s^0 = I_{FP}(k) + B_{FP} \end{cases}$$

Production during a period. Considering the production/inventory control system, the production orders to be released for period k are determined at the beginning of the period k. After the periodic transfer of WIP inventory at the end of the period $k-1$, at the last stage the backordered items are delivered and their kanban cards are taken out of the system. Note that, for each backordered finished product demand an additional production kanban is inserted at the final stage in order to produce one more unit in the next period. A production order which is a total number of production kanbans collected till the end of the period $k-1$ (including the additional kanbans for backordered items at the final stage) and are still waiting for production is released at workstation W_j for producing component C_j in period k. This sum of all undelivered production orders at workstation W_j at the beginning of period k becomes $K_j - I_{j+1}^{in}(k)$. This targeted amount of production could be achieved if there is sufficient amount of component C_{j-1} at workstation W_j. That is, if $K_j - I_{j+1}^{in}(k) \le I_j^{in}(k) + W_j^{on}(k)$ where $W_j^{on}(k)$ is equal to one if workstation W_j is busy processing component C_{j-1} at the beginning of period k, to zero if the workstation W_j is idle at the beginning of period k. The target production, $O_j(k)$, is then adjusted according to the availability of component C_{j-1} at the beginning of period k as:

$$O_j(k) = \min\{K_j - I_{j+1}^{in}(k), I_j^{in}(k) + W_j^{on}(k)\}, \quad 1 \le j \le N.$$

On the other hand, the actual amount of production during period k at workstation W_j is referred as $P_j(k)$ $(0 \le P_j(k) \le O_j(k))$. Considering the exponential

production process of component C_j at workstation W_j, the probability distribution of producing $P_j(k)$ units of component C_j during period k is:

$$P\left[P_j(k) = p_j^0 \mid O_j(k)\right] = \begin{cases} \frac{(\mu_j T)^{p_j^0}}{p_j^0!} e^{-\mu_j T} & 0 \leq p_j^0 < O_j(k) \\ 1 - \sum_{l=0}^{p_j^0 - 1} \frac{(\mu_j T)^l}{l!} e^{-\mu_j T} & p_j^0 = O_j(k) \end{cases}$$

States of the system. The state of workstation W_j at the beginning of period k can be described by a pair of system parameters, $(I_j^{in}(k), W_j^{on}(k))$, where $0 \leq I_j^{in}(k) \leq K_{j-1}$, $W_j^{on}(k) \in \{0, 1\}$ and moreover, $I_j^{in}(k) + W_j^{on}(k) \leq K_{j-1}$. Then, the state of the whole system at the beginning of period k can be satisfactorily described by $2 * N$ parameters:

- WIP inventory levels, $I_j^{in}(k)$, $2 \leq j \leq N$,
  (since Q_1^{in} is assumed to be infinite $I_1^{in}(k)$ is deleted),
- state of the machines, $W_j^{on}(k)$, $1 \leq j \leq N$,
- FP inventory level, $I_{FP}(k)$,
  ($-B_{FP} \leq I_{FP}(k) \leq K_N$, since backordering is allowed at Q_{FP}).

In our discrete-time Markov process model, the state of the periodic pull production system at the beginning of period k is simply denoted by:

$$\vec{S}(k) = [W_1^{on}(k), I_2^{in}(k), W_2^{on}(k), I_3^{in}(k), W_3^{on}(k), ..., I_N^{in}(k), W_N^{on}(k), I_{FP}(k)]$$

It should be noted that the size of the state space, $\mathcal{E}$, is not simply the multiplication of the numbers of possible states for workstations and buffer stocks,

because some combinations of workstation and buffer stock states cannot occur simultaneously.

$$\mathcal{E} = \{\ \vec{\mathcal{S}}(k)\colon\ W_j^{on}(k) \in \{0,1\},\ 1 \le j \le N,$$
$$0 \le I_j^{in}(k) \le K_{j-1},\ 2 \le j \le N,\ -B_{FP} \le I_{FP}(k) \le K_N;$$
$$W_j^{on}(k) + I_j^{in}(k) \le K_{j-1},\ 2 \le j \le N\ \}$$

$$|\mathcal{E}| = 2 * \left[\prod_{j=2}^{N}(2 * K_{j-1} + 1)\right] * (B_{FP} + 1 + K_N)$$

The size of the state space given above could also be reduced by a factor of two in practice, because workstation W_1 could never be starved since the raw material supply is infinite. The one–step transition equations, determining the system state $\vec{\mathcal{S}}(k)$ are as follows:

- **Workstation status:**

$$W_1^{on}(k) = \begin{cases} 1 & if\ \ I_2^{in}(k-1) < K_1 \\ 0 & if\ \ I_2^{in}(k-1) = K_1 \end{cases}$$

$$W_j^{on}(k) = \begin{cases} 1 & if\ \ W_j^{on}(k-1) = 1\ \ and\ \ P_j(k-1) = 0 \\ & \qquad or \\ & W_j^{on}(k-1) = 0\ \ and\ \ O_j(k-1) > 0\ \ and\ \ P_j(k-1) = 0 \\ & \qquad or \\ & 0 \le P_j(k-1) < O_j(k-1) \\ \\ 0 & if\ \ W_j^{on}(k-1) = 0\ \ and\ \ O_j(k-1) = 0 \\ & \qquad or \\ & P_j(k-1) = O_j(k-1) \end{cases}$$

$$2 \le j \le N.$$

- **Inventory status:**

$$I_j^{in}(k) = I_j^{in}(k-1) + W_j^{on}(k-1) + P_{j-1}(k-1) - (P_j(k-1) + W_j^{on}(k))$$

$$2 \leq j \leq N,$$

$$I_{FP}(k) = I_{FP}(k-1) + P_N(k-1) - D_s(k-1).$$

The system state transition. All alternative transitions from $\vec{\mathcal{S}}(k-1)$ to $\vec{\mathcal{S}}(k)$ can be found by enumerating all possible values of $N+1$ stochastic processes. The entries of the resulting one-step transition probability matrix M, of size $|\mathcal{E}|\mathrm{x}|\mathcal{E}|$, are therefore given as follows:

$$m[\vec{\mathcal{S}}(k-1), \vec{\mathcal{S}}(k)] = \sum_{\vec{P}(k-1)\in\mathcal{R}} \xi(\vec{P}(k-1)) P\left[D_s(k-1) = d_s^0 \mid I_{FP}(k-1), B_{FP}\right] \prod_{j=1}^{N} P\left[P_j(k-1) = p_j^0 \mid O_j(k-1)\right]$$

where

$$\begin{aligned} \mathcal{R} \;=\; \{\; \vec{P}(k-1) = [P_1(k-1), ..., P_N(k-1), D_s(k-1)] \;:\; & \\ 0 \leq P_j(k-1) \leq O_j(k-1), \;\; 1 \leq j \leq N, & \\ 0 \leq D_s(k-1) \leq I_{FP}(k-1) + B_{FP} \;\}. & \end{aligned}$$

$$\xi(\vec{P}(k-1)) = \begin{cases} 1 & \text{if } \vec{P}(k-1) \text{ cause a transition from } \vec{\mathcal{S}}(k-1) \text{ to } \vec{\mathcal{S}}(k) \\ 0 & \text{otherwise} \end{cases}$$

The long–term behavior of the system. In this formulation, the limiting distribution of the states of the system $\vec{\pi}$, of size $|\mathcal{E}|$, could be found (if it exists) by solving the stationary equations of the Markov chain under consideration with the boundary condition imposed:

$$\vec{\pi}\, M = \vec{\pi} \qquad \text{and} \qquad \vec{\pi}\, \vec{e}^{\,T} = 1$$

where $\vec{e}$ is a row vector with all elements equal to one, $\vec{\pi}$ is the unique solution of the above transition and the boundary equations. A discussion on variety of methods to compute the stationary probabilities of large Markov chains could be find in [15, 80].

3.2.2 Key Performance Measures

In this section, a brief discussion about the information that can be extracted from the model will be given.

Mean throughput rate. Considering the long–term behavior of the system, the throughput rates of the workstations are equal to each other because of the conservation of material flow in the system. The mean throughput rate of workstation W_j is denoted by $\mathbf{MTR}_j$ and defined as the expected number of component C_j items produced per unit time.

- The mean throughput rate of the system is:

$$\mathbf{MTR} = \mathbf{MTR}_N = \mathbf{MTR}_{N-1} = \bullet\bullet\bullet = \mathbf{MTR}_2 = \mathbf{MTR}_1$$

where

$$\mathbf{MTR}_j = \begin{cases} \displaystyle\sum_{w_1^0=0}^{1} \sum_{i_2^0=0}^{K_1} \sum_{p_1^0=0}^{O_1} \left(\frac{p_1^0}{T}\right) \; P[W_1^{on} = w_1^0, I_2^{in} = i_2^0] \; P[P_1 = p_1^0 \mid O_1] & j = 1 \\ \displaystyle\sum_{i_j^0=0}^{K_{j-1}} \sum_{w_j^0=0}^{1} \sum_{i_{j+1}^0=0}^{K_j} \sum_{p_j^0=0}^{O_j} \left(\frac{p_j^0}{T}\right) \; P[I_j^{in} = i_j^0, W_j^{on} = w_j^0, I_{j+1}^{in} = i_{j+1}^0] \; P[P_j = p_j^0 \mid O_j] & 2 \le j \le N-1 \\ \displaystyle\sum_{i_N^0=0}^{K_{N-1}} \sum_{w_N^0=0}^{1} \sum_{i_{FP}^0=-B_{FP}}^{K_N} \sum_{p_N^0=0}^{O_N} \left(\frac{p_N^0}{T}\right) \; P[I_N^{in} = i_N^0, W_N^{on} = w_N^0, I_{FP} = i_{FP}^0] \; P[P_N = p_N^0 \mid O_N] & j = N \end{cases}$$

Mean utilization of workstations. Although the long–term mean throughput rates of the workstations are equal, the utilization of workstations $\mathbf{MU}_j$ could be different because the production rates of workstations may be different.

- The mean utilization of workstation W_j is:

$$\mathbf{MU}_j = \frac{\mathbf{MTR}_j}{\mu_j} = \frac{\mathbf{MTR}}{\mu_j}$$

Mean inventory levels. According to the above formulation of the system, there are N buffer stocks under consideration, Q_j^{in}, $2 \leq j \leq N+1$. These measures of performance could be analyzed in several ways:

- The mean inventory level at Q_j^{in} at the beginning of period is:

$$\mathbf{MBI}_j = \sum_{i_j^0=0}^{K_{j-1}} i_j^0 P[I_j^{in} = i_j^0]$$

- The mean inventory level at Q_j^{in} at the end of the period before the transfer of WIP inventory from Q_{j-1}^{out} to Q_j^{in} is:

$$\mathbf{MEI}_j = \begin{cases} \displaystyle\sum_{i_j^0=0}^{K_{j-1}} \sum_{w_j^0=0}^{1} \sum_{i_{j+1}^0=0}^{K_j} \sum_{p_j^0=0}^{O_j} (i_j^0 - p_j^0) \;\; P[I_j^{in} = i_j^0, W_j^{on} = w_j^0, I_{j+1}^{in} = i_{j+1}^0] \; P[P_j = p_j^0 \mid O_j] & 2 \leq j \leq N \\ \displaystyle\sum_{i_{FP}^0=-B_{FP}}^{K_N} \sum_{d_s^0=0}^{i_{FP}^0+B_{FP}} \max\{0, i_{FP}^0 - d_s^0\} \;\; P[I_{FP} = i_{FP}^0] \; P[D_s = d_s^0 \mid I_{FP}, B_{FP}] & j = N+1 \end{cases}$$

- The mean inventory level at Q_j^{in} during the period:

$$\mathbf{MI}_j = \begin{cases} \displaystyle\sum_{i_j^0=0}^{K_{j-1}} \sum_{w_j^0=0}^{1} \sum_{i_{j+1}^0=0}^{K_j} P[I_j^{in} = i_j^0, W_j^{on} = w_j^0, I_{j+1}^{in} = i_{j+1}^0] \left[(i_j^0 - O_j) + \sum_{p_j^0=1}^{O_j} (O_j + 1 - p_j^0) \frac{\mathrm{MTTP}_j(p_j^0) - \mathrm{MTTP}_j(p_j^0 - 1)}{T} \right] & 2 \le j \le N \\ \displaystyle\sum_{i_{FP}^0=0}^{K_N} P[I_{FP} = i_{FP}^0] \left[\sum_{d_s^0=1}^{i_{FP}^0} (i_{FP}^0 + 1 - d_s^0) \frac{\mathrm{MTTD}_s(d_s^0) - \mathrm{MTTD}_s(d_s^0 - 1)}{T} \right] & j = N+1 \end{cases}$$

where

$$\mathrm{MTTP}_j(p_j^0) = \begin{cases} 0 & p_j^0 = 0 \\ \displaystyle\int_0^T t \frac{\mu_j^{(p_j^0)} t^{(p_j^0 - 1)}}{(p_j^0 - 1)!} e^{-\mu_j t} dt + \int_T^\infty T \frac{\mu_j^{(p_j^0)} t^{(p_j^0 - 1)}}{(p_j^0 - 1)!} e^{-\mu_j t} dt & 1 \le p_j^0 \end{cases} \qquad 2 \le j \le N$$

$$\mathrm{MTTD}_s(d_s^0) = \begin{cases} 0 & d_s^0 = 0 \\ \displaystyle\int_0^T t \frac{\lambda^{(d_s^0)} t^{(d_s^0 - 1)}}{(d_s^0 - 1)!} e^{-\lambda t} dt + \int_T^\infty T \frac{\lambda^{(d_s^0)} t^{(d_s^0 - 1)}}{(d_s^0 - 1)!} e^{-\lambda t} dt & 1 \le d_s^0 \end{cases}$$

In the context of cost evaluation model, one of the above three values of inventories could be utilized in computing the mean inventory carrying cost of the system. But in our further analysis in this research, the time–mean value of inventories will be used.

Mean backorder level. According to formulation in which backordering is allowed at finished product level, Q_{FP} is under consideration. This measure of performance could again be analyzed in several ways:

- The mean backorder level at Q_{FP} at the beginning of period is:

$$\mathbf{MBB} = \sum_{i^0_{FP}=-B_{FP}}^{0} -i^0_{FP} P[I_{FP} = i^0_{FP}]$$

- The mean backorder level at Q_{FP} at the end of the period before the transfer of WIP inventory from Q_N^{out} to Q_{FP} is:

$$\mathbf{MEB} = \sum_{i^0_{FP}=-B_{FP}}^{K_N} \sum_{d^0_s=0}^{i^0_{FP}+B_{FP}} -\min\{0, i^0_{FP} - d^0_s\} \;\; P[I_{FP} = i^0_{FP}] \; P[D_s = d^0_s \mid I_{FP}, B_{FP}]$$

- The mean backorder level at Q_{FP} during the period:

$$\mathbf{MB} = \sum_{i^0_{FP}=-B_{FP}}^{K_N} P[I_{FP} = i^0_{FP}] \left[\sum_{d^0_s=1}^{i^0_{FP}+B_{FP}} -\min\{0, i^0_{FP} + 1 - d^0_s\} \; \frac{\mathrm{MTTD}_s(d^0_s) - \mathrm{MTTD}_s(d^0_s - 1)}{T} \right]$$

In the context of cost evaluation model, one of the above three values of backorders could be utilized in computing the mean backordering cost of the system. But in our further analysis in this research, the time–mean value of inventories will be used.

Mean service level. The current periodic formulation considers a system in which the aim of the production is to satisfy the stochastic finished product

demand on time. But, because of the uncertainty in both the production and demand arrival processes, backordering and lost sales are also allowed in the system. A demand for finished product arriving at times when Q_{FP} is empty is backordered up to a maximum level of B_{FP}. Finally, the arrivals finding an amount of B_{FP} finished products backordered, are lost. Because of this, there is a number of performance measures available in determining the mean service level of the system.

- The mean demand rate satisfied on time:

$$\mathbf{MDR}_s = \lambda \sum_{i^0_{FP}=-B_{FP}}^{K_N} \sum_{d^0_s=0}^{i^0_{FP}+B_{FP}} \delta\{i^0_{FP} - d^0_s \geq 0\} \;\; P[I_{FP} = i^0_{FP}] \; P[D_s = d^0_s]$$

- The mean demand rate backordered:

$$\mathbf{MDR}_{b/o} = \lambda \sum_{i^0_{FP}=-B_{FP}}^{K_N} \sum_{d^0_s=0}^{i^0_{FP}+B_{FP}} \delta\{-B_{FP} \leq i^0_{FP} - d^0_s < 0\} \;\; P[I_{FP} = i^0_{FP}] \; P[D_s = d^0_s]$$

- The mean demand rate lost:

$$\mathbf{MDR}_{lost} = \lambda \sum_{i^0_{FP}=-B_{FP}}^{K_N} \sum_{d^0_s=-\infty}^{B_{FP}-1} \delta\{i^0_{FP} - d^0_s \leq -B_{FP}\} \;\; P[I_{FP} = i^0_{FP}] \; P[D_s = d^0_s]$$

$$= \lambda - \mathbf{MDR}_s - \mathbf{MDR}_{b/o}$$

where $$\delta\{E\} = \begin{cases} 1 & \text{if } E \text{ is true} \\ 0 & \text{otherwise} \end{cases}$$

and $$E = \{i^0_{FP} - d^0_s \leq -B_{FP}\}.$$

Mean backorder time. An arrived finished product demand is backordered, if the finished product inventory level at stock Q_{FP} is less than zero. Backordering is continued until a maximum allowable level B_{FP} is reached. The mean waiting time for these backordered finished product demand items is another performance measure to be considered in this formulation.

- The mean backorder time at stock Q_{FP} is:

$$\mathbf{MBT} = \frac{\mathbf{MB}}{\mathbf{MDR}_{b/o}}$$

Probabilities related with the periodic control of the system. The production system formulated in this study is a periodic review – instantaneous order system. In this respect, it is essential to define several probabilities directly related with the length of the control period T.

- The probability of achieving production objective O_j at workstation W_j at the end of the period is:

$$\mathbf{PAPO}_j = \begin{cases} \sum_{w_1^0=0}^{1} \sum_{i_2^0=0}^{K_1} P[W_1^{on} = w_1^0, I_2^{in} = i_2^0]\ P[P_1 = O_1 \mid O_1] & j = 1 \\ \sum_{i_j^0=0}^{K_{j-1}} \sum_{w_j^0=0}^{1} \sum_{i_{j+1}^0=0}^{K_j} P[I_j^{in} = i_j^0, W_j^{on} = w_j^0, I_{j+1}^{in} = i_{j+1}^0]\ P[P_j = O_j \mid O_j] & 2 \le j \le N-1 \\ \sum_{i_N^0=0}^{K_{N-1}} \sum_{w_N^0=0}^{1} \sum_{i_{FP}^0=-B_{FP}}^{K_N} P[I_N^{in} = i_N^0, W_N^{on} = w_N^0, I_{FP} = i_{FP}^0]\ P[P_N = O_N \mid O_N] & j = N \end{cases}$$

- The probability of no material handling between workstations W_j and W_{j+1} at the end of the period is:

$$\mathbf{PNOM/H}_j = \begin{cases} \sum_{w_1^0=0}^{1} \sum_{i_2^0=0}^{K_1} P[W_1^{on} = w_1^0, I_2^{in} = i_2^0]\ P[P_1 = 0 \mid O_1] & j = 1 \\ \sum_{i_j^0=0}^{K_{j-1}} \sum_{w_j^0=0}^{1} \sum_{i_{j+1}^0=0}^{K_j} P[I_j^{in} = i_j^0, W_j^{on} = w_j^0, I_{j+1}^{in} = i_{j+1}^0]\ P[P_j = 0 \mid O_j] & 2 \leq j \leq N-1 \\ \sum_{i_N^0=0}^{K_{N-1}} \sum_{w_N^0=0}^{1} \sum_{i_{FP}^0=-B_{FP}}^{K_N} P[I_N^{in} = i_N^0, W_N^{on} = w_N^0, I_{FP} = i_{FP}^0]\ P[P_N = 0 \mid O_N] & j = N \end{cases}$$

So far in this section, we describe a numerous performance measures for the exact evaluation of periodic pull production system considered in this study. The list of performance measures might be enriched by adding several others, but we assure that the ones considered in this section form a fundamental subset. One could also define a unique performance measure to represent the profitability of the system as a whole.

3.3 Approximate Performance Evaluation Model

The approximation method decomposes the production system into several individual subsystems: starting with the last stage, each of the stages is approximated by a single stage model with an appropriately revised material supply, production and demand arrival functions. This decomposition procedure is repeated several times in order to adequately approximate the selected performance measures of the production system as a whole.

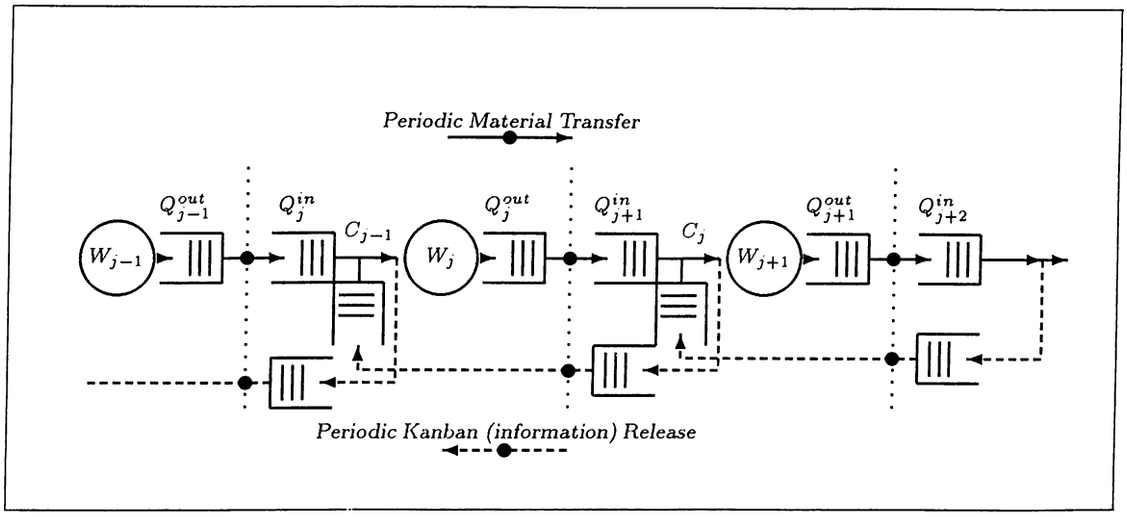


Figure 3.2: An isolated single–stage pull production subsystem $\mathcal{Z}_j$.

3.3.1 Isolated Single–stage Sub–system

Our goal is to approximate the whole production system given in Figure 3.1 by a sequence of isolated single–stage sub–systems, $\mathcal{Z}_j$ $1 \leq j \leq N$ (see Figure 3.2), where:

- the input material, component C_{j-1}, is supplied from input stock Q_j^{in},
- the production of component C_j, is initiated for the replenishment of items withdrawn from input stock Q_{j+1}^{in}.

The first and the last sub–systems are atypical, since in the first sub–system the raw material input is assumed to be infinite and in the last stage Poisson demand arrivals for finished product is external to the system.

States of the sub–system. The state of sub–system $\mathcal{Z}_j$ at the beginning of period k can be described by a pair of system parameters, $(W_j^{on}(k), I_{j+1}^{in}(k))$. In our formulation, the state of the isolated single–stage periodic pull production

sub–system at the beginning of period k is denoted by:

$$\vec{\mathcal{S}}_{\mathcal{Z}_j}(k) = [W_j^{on}(k), I_{j+1}^{in}(k)]$$

The one–step transition equations, determining the state of the sub–system remains the same as in the exact model.

State transition of the sub–system. All alternative transitions from $\vec{\mathcal{S}}_{\mathcal{Z}_j}(k-1)$ to $\vec{\mathcal{S}}_{\mathcal{Z}_j}(k)$ can be found by enumerating all possible realizations of related random variables; $I_j^{in}(k-1)$, $P_j(k-1)$, $W_{j+1}^{on}(k-1)$, $P_{j+1}(k-1)$ and $I_{j+2}^{in}(k-1)$. The entries of the resulting one–step transition probability matrix $M_{\mathcal{Z}_j}$ are given as follows:

$$m_{\mathcal{Z}_j}[\vec{\mathcal{S}}_{\mathcal{Z}_j}(k-1), \vec{\mathcal{S}}_{\mathcal{Z}_j}(k)] =$$

$$\begin{cases}
\displaystyle\sum_{w_2^0=0}^{1}\sum_{i_3^0=0}^{K_2}\sum_{p_1^0=0}^{O_1(k-1)}\sum_{p_2^0=0}^{O_2(k-1)} \psi\{\bullet\}\, P[W_2^{on}(k-1)=w_2^0, I_3^{in}(k-1)=i_3^0]* \\
\qquad P[P_1(k-1)=p_1^0 \mid O_1(k-1)]P[P_2(k-1)=p_2^0 \mid O_2(k-1)] & \text{j = 1} \\
\displaystyle\sum_{i_j^0=0}^{K_{j-1}}\sum_{w_{j+1}^0=0}^{1}\sum_{i_{j+2}^0=0}^{K_{j+1}}\sum_{p_j^0=0}^{O_j(k-1)}\sum_{p_{j+1}^0=0}^{O_{j+1}(k-1)} \psi\{\bullet\}\, P[I_j^{in}(k-1)=i_j^0, W_{j+1}^{on}(k-1)=w_{j+1}^0, I_{j+2}^{in}(k-1)=i_{j+2}^0]* \\
\qquad P[P_j(k-1)=p_j^0 \mid O_j(k-1)]P[P_{j+1}(k-1)=p_{j+1}^0 \mid O_{j+1}(k-1)] \\
\qquad 2 \le j \le N-1 \\
\displaystyle\sum_{i_N^0=0}^{K_{N-1}}\sum_{p_N^0=0}^{O_N(k-1)}\sum_{d_s^0=0}^{I_{FP}+B_{FP}} \psi\{\bullet\}\, P[I_N^{in}(k-1)=i_N^0]P[P_N(k-1)=p_N^0 \mid O_N(k-1)]* \\
\qquad P[D_s(k-1)=d_s^0 \mid I_{FP}(k-1), B_{FP}] & \text{j = N}
\end{cases}$$

where $\psi(\bullet) = \begin{cases} 1 & \text{if the realizations of the related random variables cause a transition from } \vec{\mathcal{S}}_{\mathcal{Z}_j}(k-1) \text{ to } \vec{\mathcal{S}}_{\mathcal{Z}_j}(k) \\ 0 & \text{otherwise.} \end{cases}$

The long–term behavior. In this formulation, the limiting distribution of the states of the jth sub–system $\vec{\pi}_{\mathcal{Z}_j}$ could be found (if it exists) by solving the stationary equations of the Markov chain under consideration with the following boundary condition imposed:

$$\vec{\pi}_{\mathcal{Z}_j} M_{\mathcal{Z}_j} = \vec{\pi}_{\mathcal{Z}_j} \qquad \text{and} \qquad \vec{\pi}_{\mathcal{Z}_j} \vec{e}^{\,T} = 1$$

where $\vec{e}$ is a row vector with all elements equal to one, $\vec{\pi}_{\mathcal{Z}_j}$ is the unique solution of the above transition and the boundary equations.

3.3.2 Decomposition Method

Our proposed decomposition method is based on the formulation of isolated single–stage sub–system given in the previous section. The aim is to represent the whole production system by a sequence of isolated single–stage periodic pull production sub–systems, where the streams of raw material and demand for component C_j to be produced at sub–system $\mathcal{Z}_j$ are provided by sub–systems $\mathcal{Z}_{j-1}$ and $\mathcal{Z}_{j+1}$, respectively (see Figure 3.3). The parameters of these isolated subsystems must be coordinated in such a way that the performance characteristics of the resulting sequence are as close as possible to those of the production system as a whole.

While decomposing the whole production system, we start with the last sub–system, $\mathcal{Z}_N$, and work backwards till we reach the first sub–system with considering infinite supply of raw material at all input buffer stocks, Q_j^{in}, in order to initialize the steady–state probabilities of states of all decomposed sub–systems. In this backward initialization pass, the starvation of all sub–systems is ignored and only blocking is considered. Then, two consecutive passes, backward and forward passes, are executed iteratively until obtaining a satisfactory level of approximation in evaluating the performance measures of the whole production system.

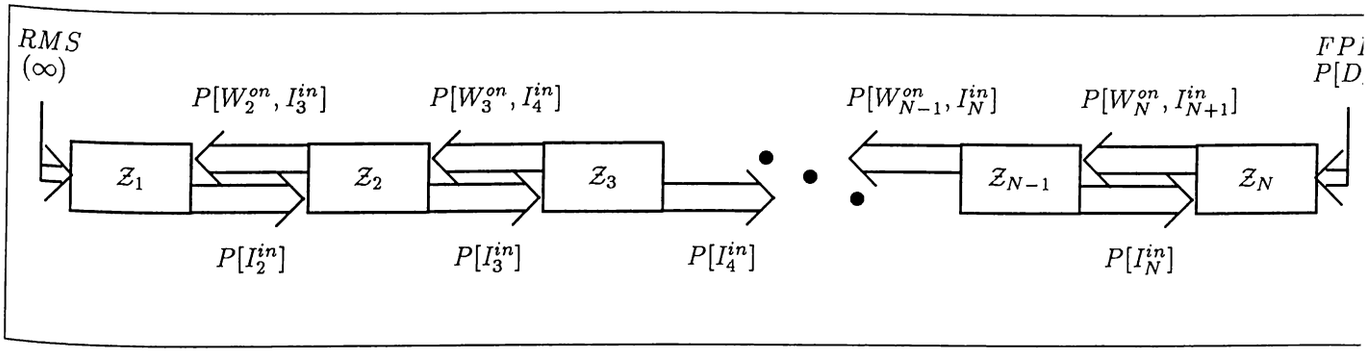


Figure 3.3: Model of the production system constituted from models of isolated single–stage sub–systems. $RMS(\infty)$ denotes the infinite supply of raw material and FPD $P[D_s]$ denotes the external finished product demand.

The level of approximation is determined by the deviation between throughput rates of the sub–systems at consecutive iterations. During these iterations, both starvation and blocking of sub–systems are considered. It is important to note that the marginal (with respect to the state of the whole system) probabilities used in the formulation of isolated sub–system still involve no approximation if they are computed exactly. More precisely, the steps summarizing the iterative decomposition approach are as follows:

Step 0. *Initialization.* Compute an initial approximation for the limiting distribution of the states of the system utilizing a backward pass, with assuming all input buffer stocks are full at the beginning of every period.

Set iteration index, $l \leftarrow 0$,
Set $P^{(l)}[I_j^{in} = K_{j-1}] = 1$, for $j = 2, .., N+1$,
Set sub–system (stage) index, $j \leftarrow N$,
Set the level of approximation ($\epsilon \leftarrow 10^{-8}$).
Backward loop. For $j := N$ downto 1
Compute one–step transition probability matrix, $M_{\mathcal{Z}_j}^{(l)}$;
Obtain the solution, $\vec{\pi}_{\mathcal{Z}_j}^{(l)}$.

Step 1. *Iterations.* At iteration l, solve each sub–system, $\mathcal{Z}_j$ for $j = 1, .., N$, twice; both in backward and forward passes.

Set $l \leftarrow l + 1$,
Backward loop. For $j := N$ downto 1
Compute one–step transition probability matrix, $M_{\mathcal{Z}_j}^{(l)}$,
Obtain the solution, $\vec{\pi}_{\mathcal{Z}_j}^{(l)}$,
Compute mean throughput rate, $\mathbf{MTR}_{\mathcal{Z}_j}^{(l_b)}$.
Forward loop. For $j := 2$ to N
Compute one–step transition probability matrix, $M_{\mathcal{Z}_j}^{(l)}$,
Obtain the solution, $\vec{\pi}_{\mathcal{Z}_j}^{(l)}$,
Compute mean throughput rate, $\mathbf{MTR}_{\mathcal{Z}_j}^{(l_f)}$.

Step 2. *Stopping Criteria.* If the maximum absolute deviation of mean throughput rates of the sub–systems between backward and forward passes is less than a given threshold value, stop. Otherwise continue iterations at *Step 1.*

if $\max_{2 \le j \le N} \mid \mathbf{MTR}_{\mathcal{Z}_j}^{(l_b)} - \mathbf{MTR}_{\mathcal{Z}_j}^{(l_f)} \mid < \epsilon$ then
Compute the performance measures of the system
and stop;

otherwise go to *Step 1.*

We do not have a proof of convergence for this method. However, in practice, in the many examples we have examined, the method has always converged within a reasonable number of iterations (low 10s), only moderately dependent on the number of stages. In an experiment in which 625 three–stage production systems are evaluated by executing an average of approximately 28 single–stage evaluations in order to obtain an acceptable level of approximation accuracy. As a result, the computational complexity of our approach grows relatively moderately (but more than linearly) with the number of stages in the system.

3.3.3 Key Performance Measures

In this section, a brief discussion about the information that can be extracted from the approximation model will be given.

- The mean throughput rate of sub-system $\mathcal{Z}_j$ is denoted by $\mathbf{MTR}_{\mathcal{Z}_j}$ and defined as the expected number of component C_j items produced per unit time. The approximated mean throughput rate of the whole system is:

$$\mathbf{AMTR} = \mathbf{MTR}_{\mathcal{Z}_N} \approx \mathbf{MTR}_{\mathcal{Z}_{N-1}} \approx \bullet\bullet\bullet \approx \mathbf{MTR}_{\mathcal{Z}_2} \approx \mathbf{MTR}_{\mathcal{Z}_1}$$

where

$$\mathbf{MTR}_{\mathcal{Z}_j} = \begin{cases} \sum_{w_1^0=0}^{1} \sum_{i_2^0=0}^{K_1} \sum_{p_1^0=0}^{O_1} (\frac{p_1^0}{T}) \; P[W_1^{on} = w_1^0, I_2^{in} = i_2^0] \; P[P_1 = p_1^0 \mid O_1] & j = 1 \\ \sum_{i_j^0=0}^{K_{j-1}} \sum_{w_j^0=0}^{1} \sum_{i_{j+1}^0=0}^{K_j} \sum_{p_j^0=0}^{O_j} (\frac{p_j^0}{T}) \; P[I_j^{in} = i_j^0, W_j^{on} = w_j^0, I_{j+1}^{in} = i_{j+1}^0] \; P[P_j = p_j^0 \mid O_j] & 2 \le j \le N-1 \\ \sum_{i_N^0=0}^{K_{N-1}} \sum_{w_N^0=0}^{1} \sum_{i_{FP}^0=-B_{FP}}^{K_N} \sum_{p_N^0=0}^{O_N} (\frac{p_N^0}{T}) \; P[I_N^{in} = i_N^0, W_N^{on} = w_N^0, I_{FP} = i_{FP}^0] \; P[P_N = p_N^0 \mid O_N] & j = N \end{cases}$$

- Although the long-term mean throughput rates of the sub-systems are assumed to be equal, the utilization of workstation at sub-systems $\mathbf{MU}_{\mathcal{Z}_j}$ could be different because the production rates of the sub-systems may be different. The approximated mean utilization of workstation W_j is:

$$\mathbf{AMU}_j = \mathbf{MU}_{\mathcal{Z}_j} = \frac{\mathbf{MTR}_{\mathcal{Z}_j}}{\mu_j} \approx \frac{\mathbf{AMTR}}{\mu_j}$$

- There are N buffer stocks to be considered, Q_j^{in}, $2 \leq j \leq N+1$, in the context of inventory control system. The approximated mean inventory level at Q_j^{in} during the period is:

$$\mathbf{AMI}_j = \begin{cases} \displaystyle\sum_{i_j^0=0}^{K_{j-1}} \sum_{w_j^0=0}^{1} \sum_{i_{j+1}^0=0}^{K_j} P[I_j^{in} = i_j^0, W_j^{on} = w_j^0, I_{j+1}^{in} = i_{j+1}^0] \left[(i_j^0 - O_j) + \sum_{p_j^0=1}^{O_j} (O_j + 1 - p_j^0) \frac{\mathrm{MTTP}_j(p_j^0) - \mathrm{MTTP}_j(p_j^0 - 1)}{T} \right] & 2 \leq j \leq N \\ \displaystyle\sum_{i_{FP}^0=0}^{K_N} P[I_{FP} = i_{FP}^0] \left[\sum_{d_s^0=1}^{i_{FP}^0} (i_{FP}^0 + 1 - d_s^0) \frac{\mathrm{MTTD}_s(d_s^0) - \mathrm{MTTD}_s(d_s^0 - 1)}{T} \right] & j = N+1 \end{cases}$$

- According to formulation in which backordering is allowed at finished product level, Q_{FP} is under consideration. The approximated mean backorder level at Q_{FP} during the period is:

$$\mathbf{AMB} = \sum_{i_{FP}^0=-B_{FP}}^{K_N} P[I_{FP} = i_{FP}^0] \left[\sum_{d_s^0=1}^{i_{FP}^0+B_{FP}} - \min\{0, i_{FP}^0 + 1 - d_s^0\} \; \frac{\mathrm{MTTD}_s(d_s^0) - \mathrm{MTTD}_s(d_s^0 - 1)}{T} \right]$$

- A demand for finished product arriving at times when Q_{FP} is empty is backordered up to a maximum level of B_{FP}. Then, the arrivals finding an amount of B_{FP} finished products backordered, are lost. Because of this, there is a number of performance measures available in determining the approximated mean service level of the system.

– The approximated mean demand rate satisfied on time:

$$\mathbf{AMDR}_s = \lambda \sum_{i^0_{FP}=-B_{FP}}^{K_N} \sum_{d^0_s=0}^{i^0_{FP}+B_{FP}} \delta\{i^0_{FP} - d^0_s \geq 0\} \;\; P[I_{FP} = i^0_{FP}] \; P[D_s = d^0_s]$$

– The approximated mean demand rate backordered:

$$\mathbf{AMDR}_{b/o} = \lambda \sum_{i^0_{FP}=-B_{FP}}^{K_N} \sum_{d^0_s=0}^{i^0_{FP}+B_{FP}} \delta\{-B_{FP} \leq i^0_{FP} - d^0_s < 0\} \;\; P[I_{FP} = i^0_{FP}] \; P[D_s = d^0_s]$$

– The approximated mean demand rate lost:

$$\mathbf{AMDR}_{lost} = \lambda \sum_{i^0_{FP}=-B_{FP}}^{K_N} \sum_{d^0_s=-\infty}^{B_{FP}-1} \delta\{i^0_{FP} - d^0_s \leq -B_{FP}\} \;\; P[I_{FP} = i^0_{FP}] \; P[D_s = d^0_s]$$

$$= \lambda - \mathbf{AMDR}_s - \mathbf{AMDR}_{b/o}$$

- An arrived finished product demand is backordered, if the finished product inventory level at stock Q_{FP} is less than zero. Backordering is continued until a maximum allowable level B_{FP} is reached. The approximated mean backorder time at stock Q_{FP} is:

$$\mathbf{AMBT} = \frac{\mathbf{AMB}}{\mathbf{AMDR}_{b/o}}$$

- The production system formulated in this study is a periodic review – instantaneous order system. Several performance probabilities are directly related with the control of the system with a control period length of T time units.

  – The approximated probability of achieving production objective O_j at workstation W_j at the end of the period is:

$$\mathbf{APAPO}_j = \begin{cases} \sum_{w_1^0=0}^{1} \sum_{i_2^0=0}^{K_1} P[W_1^{on} = w_1^0, I_2^{in} = i_2^0]\ P[P_1 = O_1 \mid O_1] & j = 1 \\ \sum_{i_j^0=0}^{K_{j-1}} \sum_{w_j^0=0}^{1} \sum_{i_{j+1}^0=0}^{K_j} P[I_j^{in} = i_j^0, W_j^{on} = w_j^0, I_{j+1}^{in} = i_{j+1}^0]\ P[P_j = O_j \mid O_j] & 2 \le j \le N-1 \\ \sum_{i_N^0=0}^{K_{N-1}} \sum_{w_N^0=0}^{1} \sum_{i_{FP}^0=-B_{FP}}^{K_N} P[I_N^{in} = i_N^0, W_N^{on} = w_N^0, I_{FP} = i_{FP}^0]\ P[P_N = O_N \mid O_N] & j = N \end{cases}$$

  – The approximated probability of no material handling between workstations W_j and W_{j+1} at the end of the period is:

$$\mathbf{APNOM/H}_j = \begin{cases} \sum_{w_1^0=0}^{1} \sum_{i_2^0=0}^{K_1} P[W_1^{on} = w_1^0, I_2^{in} = i_2^0]\ P[P_1 = 0 \mid O_1] & j = 1 \\ \sum_{i_j^0=0}^{K_{j-1}} \sum_{w_j^0=0}^{1} \sum_{i_{j+1}^0=0}^{K_j} P[I_j^{in} = i_j^0, W_j^{on} = w_j^0, I_{j+1}^{in} = i_{j+1}^0]\ P[P_j = 0 \mid O_j] & 2 \le j \le N-1 \\ \sum_{i_N^0=0}^{K_{N-1}} \sum_{w_N^0=0}^{1} \sum_{i_{FP}^0=-B_{FP}}^{K_N} P[I_N^{in} = i_N^0, W_N^{on} = w_N^0, I_{FP} = i_{FP}^0]\ P[P_N = 0 \mid O_N] & j = N \end{cases}$$

So far in this section, we describe the approximate computation of performance measures which are introduced in the exact evaluation model of the periodic pull production system.

3.3.4 Approximation

There are some long–term probabilities utilized both in the computation of one–step transition matrix and approximated performance measures which are successively estimated by the iterative decomposition method. The following approximations in estimating these probabilities are required:

1. The joint probability, used in the computation of one–step transition matrix of single–stage sub–systems ($\mathcal{Z}_j$, $2 \leq j \leq N-1$), is approximated as the product of two probabilities, i.e., I_j^{in} is assumed to be independent of W_{j+1}^{on} and I_{j+2}^{in}.

$$P[I_j^{in} = i_j^0, W_{j+1}^{on} = w_{j+1}^0, I_{j+2}^{in} = i_{j+2}^0] \approx P[I_j^{in} = i_j^0]\ P[W_{j+1}^{on} = w_{j+1}^0, I_{j+2}^{in} = i_{j+2}^0]$$

   The probabilities $P[I_j^{in} = i_j^0]$ and $P[W_{j+1}^{on} = w_{j+1}^0, I_{j+2}^{in} = i_{j+2}^0]$ are obtained from the solutions of sub–systems, $\mathcal{Z}_{j-1}$ and $\mathcal{Z}_{j+1}$, respectively (Figure 3.3).

2. The joint probability, used in the computation of performance measures after obtaining solutions to single–stage Markov chain models, is approximated as the product of two probabilities, i.e., I_j^{in} is assumed to be independent of W_j^{on} and I_{j+1}^{in}.

$$P[I_j^{in} = i_j^0, W_j^{on} = w_j^0, I_{j+1}^{in} = i_{j+1}^0] \approx P[I_j^{in} = i_j^0]\ P[W_j^{on} = w_j^0, I_{j+1}^{in} = i_{j+1}^0]$$

   The probabilities $P[I_j^{in} = i_j^0]$ and $P[W_j^{on} = w_j^0, I_{j+1}^{in} = i_{j+1}^0]$ are obtained from the solution of sub–systems, $\mathcal{Z}_{j-1}$ and $\mathcal{Z}_j$, respectively (Figure 3.3).

3.4 Numerical Experimentation

The exact performance evaluation model of the basic periodic pull production system is developed using Pascal programming language and some special data structures. In solving one–step transition matrices, sparse matrix solver which is coded in C programming language is utilized in order to increase the computational efficiency of the exact solution technique. It could be efficiently evaluated only up to three stages in tandem because of the dimensionality problem inherited in the exact model of the system through the use of discrete–time Markov processes. See Table A.1 in the Appendix for the dimensional properties of transition matrices of various production systems. The use of special data structures in the code of exact model becomes effective when there are three or more stages in the system. But, it is not computationally efficient to solve the exact model of such large systems. In this regard, the approximate decomposition method seems promising. The approximate decomposition procedure described in the previous section is also implemented using Pascal programming language.

An experiment is designed in order to investigate the general behavior and the accuracy level of the single–stage approximate decomposition procedure. A three stage periodic pull production system is selected, because it is the largest system that the solution of the exact model is computationally efficient. In the context of this experiment, 625 different three stage systems were solved both using the exact and the approximate solution techniques. The range of system parameters are as follows:

- *Mean arrival rate of FP demand*, $\lambda = 0.25, 0.50, 1.00, 2.00, 4.00,$
- *Number of kanbans at each stage*, $K = 2, 3, 4, 5, 6,$
- *Mean production rate at each stage*, $\mu = \frac{\lambda}{\rho}$, where ρ is the demand load (traffic intensity of the queuing system), $\rho = 0.50, 0.60, 0.70, 0.80, 0.90,$
- *Length of the transfer/review period*, $T = 0.25, 0.50, 1.00, 2.00, 4.00.$

The above pull systems consider single product with a Poisson demand which arrives at the third (last) stage of the system with a mean rate of λ. The demand arrivals during the times the finished product buffer Q_{FP} is empty are lost (backordering is not allowed). At each stage of the system, the processing times are exponential with the same mean $1/\mu$ and the number of kanbans allocated are equal to K. The status of the system is reviewed periodically with a period length of T. The production and material withdrawal orders are released at the beginning of periods. It is assumed that the raw material supply for the first stage is infinite and the material handling times between stages are zero.

The mean throughput rate is selected as a primary measure of performance for this experiment. All comparisons are based on this measure. Numerical experience suggests that when the mean throughput rates of the workstations converge to a unique solution during the iteration process, it agrees closely with the exact model. The percent absolute errors between the exact and the approximate mean throughput rates is computed as follows:

$$\% \text{ Absolute Error} = \left| \frac{\mathbf{AMTR - MTR}}{\mathbf{MTR}} \right| * 100$$

See Table 3.1 for the frequency distribution of percent absolute errors obtained from the results of the experiment. The cumulative relative frequency of having at most 3% average absolute error is about 0.7. Only 10% of the approximate evaluations result in an absolute error which is greater than 5%. In this experiment, the overall average and the maximum of percent absolute errors between **MTR** and **AMTR** is about 2.5 and 12.5, respectively. The effect of system parameters on the accuracy level of approximate decomposition technique is summarized by the use of some sub–averages given in Table 3.2. Note that, all of detailed results related with this experiment for three–stage systems could be found in Tables A.2, A.3, A.4, A.5 and A.6 in the Appendix.

Table 3.1: The summary report on percent absolute errors between the exact and the approximate mean throughput rates of the evaluated systems is given below. The frequency and the cumulative frequency distributions are given in order to clarify the variation of the percent absolute errors.

Frequency Distribution of Percent Absolute Errors				
Class Intervals	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
[0,1)	244	0.39040	244	0.39040
[1,2)	108	0.17280	352	0.56320
[2,3)	83	0.13280	435	0.69600
[3,4)	67	0.10720	502	0.80320
[4,5)	51	0.08160	553	0.88480
[5,6)	12	0.01920	565	0.90400
[6,7)	18	0.02880	583	0.93280
[7,8)	9	0.01440	592	0.94720
[8,9)	6	0.00960	598	0.95680
[9,10)	3	0.00480	601	0.96160
[10,11)	9	0.01440	610	0.97600
[11,12)	0	0.00000	610	0.97600
[12,13)	15	0.02400	625	1.00000

It is observed that the effect of the number of kanbans at each stage is very important in the accuracy of the approximation technique. When there are odd number of kanbans at each stage, the average of percent absolute errors seems to be greater than the case of even number of kanbans at each stage. But, for the case of increasing number of kanbans the average of percent absolute errors, although fluctuating within an acceptable range, is decreasing in the limit. This is as expected, because with increasing the number of kanbans at each stage the dependence between production stages is decreased and this makes it possible to estimate those unknown joint probabilities better in the approximation. As a result, the accuracy level is improved with increasing number of kanbans.

Table 3.2: Average percent absolute errors between the exact and the approximate solutions of the systems evaluated in the experiment with respect to the variation in system parameters.

Averages of Percent Absolute Errors					
Arrival Rate	$\lambda = 0.25$	$\lambda = 0.50$	$\lambda = 1.00$	$\lambda = 2.00$	$\lambda = 4.00$
	2.27178	2.13558	2.30268	2.57028	2.98307
Kanban Number	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$
	2.50787	4.66253	1.61783	2.36287	1.11229
Demand Load	$\rho = 0.50$	$\rho = 0.60$	$\rho = 0.70$	$\rho = 0.80$	$\rho = 0.90$
	1.62224	1.84171	2.25427	2.88278	3.66239
Period Length	$T = 0.25$	$T = 0.50$	$T = 1.00$	$T = 2.00$	$T = 4.00$
	2.30264	2.12503	2.29008	2.56435	2.98128

On the other hand, the demand arrival rate, the demand load (traffic intensity) and the transfer/review period length have similar effect on the approximation of the production system. While keeping the number of kanbans at some level, an increase in these parameters result in an increase in the utilization of the system and consequently the dependence between production stages is also increased. As a result, this increase in dependencies cause more errors in estimating those unknown joint probabilities.

Generally speaking, it is accepted that the error level of an approximation technique should not exceed 3%. In this regard, the proposed approximate decomposition technique with an absolute error of 2.45% on the average could be used for the evaluation of NTQ equivalent periodic pull production systems.

Chapter 4

Operating Characteristics: The Allocation Problem

The design of tandem production systems has been well studied in the production research literature with the primary focus being on how to improve their efficiency. Considering the large costs associated with these systems, a slight improvement in efficiency can lead to very significant savings over the life of the production system. Division of work among the workstations and allocation of buffer storage capacity between workstations are two critical design factors that have attracted the attention of many researchers and system designers. For a survey of the research in this area, see Sarker [85].

4.1 Review of Previous Results

One significant aspect of production line design is the so-called **line balancing problem**, i.e. allocating the total work content as evenly as possible to workstations and maximizing the utilization through minimizing idle times as well. The solution of line balancing problem specifies a system configuration capable of

producing a specified amount of finished product with minimum resource requirements. The operation times can be either deterministic or stochastic. However, line balancing techniques are based on the assumption of deterministic operation times. In practice, a perfect balance of workload may be impossible even with deterministic operation times, since, in most cases, equal allocation of total work content to workstations may be prevented by precedence and technological constraints, and continuous indivisibility of operations. In production systems with stochastic operation times, the balance of workload is attained through allocating the total work content evenly to the workstations based on the means of operation times. However, the balance of stochastic operation times may be impossible due to different variability of operation times at different workstations.

It is intuitively plausible that the variation in the operation times would decrease the mean production (throughput) rate of the system. This can happen in two ways; due to blocking and/or starvation. When there is considerable variability in the operation times at some respective workstations, a perfectly balanced production line may not be optimal. Previous work on **optimal allocation of workload** to production lines has found that, under certain assumptions, the mean throughput rate of a finite buffer production line is maximized by deliberately unbalancing the workload of the line in an appropriate way. In particular, the optimal allocation of work follows a "bowl phenomenon" whereby the center workstations are given preferential treatment (less workload) over the other workstations towards the beginning and the ending workstations; see Hillier and Boling [46, 48]. The analogous result of Stecke and Morin [99] is that the mean throughput rate of an infinite buffer production line is maximized by balancing the workload assigned to workstations. In other words, as buffer capacities increases, the degree of unbalance in the optimal workload decreases, until in the limit, a balanced allocation is optimal.

Hillier and Boling [48] reported that the improvement in mean throughput rate due to unbalancing grows up to 1.37% for a six workstation serial production line. On the other hand, Magazine and Silver [67] developed an approximation that

suggests the improvement from unbalancing is no larger than 1.65% for exponential operation times, regardless of the number of workstations in the system. One of the main insight emanating from these studies is that balanced systems give acceptable performance and further improvements in mean throughput rate can be made by unbalancing. However, the gains obtained from unbalancing are relatively small — on the order of 1%. The works of El-Rayah [37] and So [92] indicated that the bowl phenomenon is robust. That is, as long as the balance of workload is changed in the direction indicated by the bowl phenomenon, the mean throughput rate function is almost flat near the maximum. On the other hand, if the production line is unbalanced in a substantially different direction, the mean throughput rate decreases quite rapidly.

Muth and Alkaff [74] examined three stage serial production systems in a more general analytical setting in order to give the mean throughput rate as a function of several system parameters, subject to certain constraints. Rao [84] considered the generalization where the coefficient of variation of operation times are different for different workstations. The results found by Rao [84] indicated that unbalancing a serial production system can lead to substantial improvements in mean production rate when the variability of the stages differ from one to another. Optimum unbalancing could possibly be achieved by carrying out alternately the following two steps:

1. Workload from interior stages should be transferred to the exterior ones, (*Bowl Phenomenon*),
2. Workload from more variable stages should be transferred to less variable ones, (*Variability Imbalance*).

Step 1 is more important when the differences in the coefficient of variation of the stages are generally less than 0.5 while Step 2 predominates when they exceed 0.5. Then, Wolisz [109] showed that the idea of assigning less workload to more variable workstations is false for coefficient of variation greater than one.

For lines longer than three stages and for non–exponential distributions, analytic approaches are quite limited, and some studies have used simulation to study the workload allocation problem under more general conditions. Payne, Slack and Wild [77] simulated production lines with different patterns of processing time variances and observed that a great deterioration in the performance occurs either when processing time variances are increased, or when buffer capacities are highly restricted.

In a similar problem, Yamazaki, Sakasegawa and Shanthikumar [111] investigated the optimal ordering of workstations that maximizes the mean throughput rate of the system. Based on some theoretical and extensive empirical results, they proposed two rules for ordering workstations. The first rule recommends arranging the two worst workstations (apart from each other as far as possible) as the first and the last workstations. A worst workstation refers to the one either with the slowest production rate or with the most variable operation time. The second rule arranges the remaining workstations according to the bowl phenomenon.

All of the above studies have assumed that the production system has a serial structure. Baker, Powell and Pyke [13] have investigated the behavior of assembly systems in which two or more parts are produced at component lines and put together at an assembly workstation at the end. Their basic finding is that the assembly workstation in a balanced system is intrinsically a bottleneck. Villeda, Dudek and Smith [106] studied an assembly system in which three serial lines (each one composed of three workstations) merge at one assembly workstation which is operating as a pull system. They considered normal processing times with several coefficients of variation. They reported that mean throughput rate is maximized by assigning decreasing amounts of work closer to assembly workstation at which the mean processing time was fixed.

The effect of bowl phenomenon has been extensively studied in conventional type push production systems, however, studies exploring its effects and validity on

pull production systems are rare. The simulation studies made so far show conflicting results. In the simulation experiments performed by Meral [68], the bowl phenomenon is not confirmed for idealized just–in–time production systems. She found that balancing strategies are always superior to the unbalancing strategies based on bowl phenomenon. On the contrary, Villeda, Dudek and Smith [106] analyzed a just–in–time production system by investigating several unbalancing methods and they claimed that the only method giving a consistent improvement in the mean throughput rate of the system was the "high–medium–low" (decreasing) allocation. They also reported that the mean throughput rate with unbalanced workstations were always superior to the perfectly balanced configurations. On the other hand, Sarker and Harris [86] claimed that they observed the effect of bowl phenomenon on a just–in–time production system.

What ever the case, looking from a labor relations point of view, there may be difficulties in assigning significantly different workloads to different workstations. This raises the question as to whether there might be other ways of achieving this improvement in mean throughput rate by giving preferential treatment to the critical workstations without significantly unbalancing the workloads. One way of doing this is to provide such critical workstations with more buffer storage capacity than the other workstations. As surveyed by Sarker [85] various researchers have considered the general question of **optimal allocation of buffer storage capacity** in a variety of contexts. In the analogy to workload allocation problem there is a critical difference that the buffer allocation decision variables are discrete (integer) variables whereas the workload allocation decision variables are formulated as continuous variables in the previous studies.

Most of the research on buffer allocation has focused on analytical models of small systems simplified with restrictive assumptions [46, 74]. For larger systems, analytical approximations or simulation models have been utilized [12, 31]. Conway *et al.* [31] examined serial production systems via simulation. They found that buffers between workstations increase the production capacity of the system but the returns are reduced sharply with increasing inventory holding costs. They

also noted that the positioning as well as the capacity of the buffers are important. El–Rayah [38] utilized a computer simulation model to investigate the effect of unequal allocation of buffer capacity on the efficiency with an experiment limited to small production lines. He observed that the lines in which the center workstations are assigned larger buffer storage capacity than the ending workstations (inverted bowl phenomenon) are better (with respect to mean throughput rate) than the other unbalanced configurations. But, according to their experiment the inverted bowl configuration yielded more or less a similar mean throughput rate to that of a balanced line depending upon the total buffer storage capacity.

Hillier and So [49] studied the effect of the variability of processing times on the optimal allocation of buffer storage capacity between workstations. They concluded that either the center workstations or the workstations with high variability should be given more buffer storage capacity. Consequently, an inverted bowl phenomenon prevails regarding the optimal allocation of buffer storage capacity. In another study, Hillier and So [50] utilized an exact analytical model to conduct a detailed study of how the length of machine up and down times and interstage buffer storage capacity can effect the mean throughput rate of production lines with more than three stages. They developed a simple heuristic to estimate the amount of buffer storage capacity required to compensate for the decrease in mean throughput rate due to machine breakdowns. Sheskin [89] offered some guidelines for the allocation of buffer storage capacity in serial production lines subject to random failure and repair. If all machines have the same reliability, he maximized the mean throughput rate by allocating the buffers as nearly as possible equal in size. In case when the machines have different reliabilities, he proposed allocating more buffer storage capacity to less reliable machine. This intuitive result is also supported by Soyster, Schmidt and Rohrer [94].

Jafari and Shanthikumar [57] proposed a heuristic solution to determine the optimal allocation of a given total buffer storage capacity among workstations of

a serial production line. Their approximate solution is based on a dynamic programming model with an approximate procedure to compute the mean throughput rate of the line.

Smith and Daskalaki [91] have developed a design methodology for buffer storage capacity allocation within assembly lines to approximately solve the optimal buffer allocation problem by maximizing mean throughput rate while minimizing holding and buffer storage costs. Baker, Powell and Pyke [12] have examined the effect of buffers on the efficiency of systems in which two serial lines merge at an assembly workstation. They have concluded that small buffers are sufficient to regain most of the lost production capacity and buffer space should be allocated equally among the workstations.

So far, we review the researchers that proposed rules for allocating buffers to maximize the mean throughput rate in serial production lines operating with push control strategy. In contrast, Andijani and Clark [8] investigated the optimal allocation of buffers (kanbans) in a pull system by considering both the mean throughput rate and the WIP inventories in the maximized objective function. Recently, Askin, Mitwasi and Goldberg [10] utilized a continuous time, steady-state Markov model in determining the optimal number of kanbans to use for each part type at each workstation in a just-in-time production system. Their objective was to minimize the sum of inventory holding and backorder costs. Results indicated a need for increased safety stocks, for systems where many part types are produced in the same workstation.

Tayur [102, 103] developed some theoretical results — *reversibility* and *dominance* — that characterize the dynamics of kanban-controlled manufacturing systems. His study also provided some insights into the behavior of those systems and greatly reduced the simulation efforts required in an investigation.

The characterization of the optimal allocation of scarce resources in a production system requires further investigation with alternate models and techniques

through which the results may fit real–life better [51]. One direction is to try non–exponential processing times with different variations or another direction is *to broaden the allocation problem by combining the decisions on buffer storage capacity allocation with workload allocation.*

4.2 General Behavior of Periodic Pull Systems

The production system considered here in order to investigate the impacts of system parameters on the mean throughput rate is a single–item multi–stage stochastic periodic pull production system. The system is given in Figure 3.1 and all descriptive and modelling details are given in the previous chapter.

There are N production stages in series. Each production stage in the system is represented by a workstation with a processing rate of μ_j, an input buffer stock of capacity K_{j-1} and an output buffer stock of capacity K_j. Production kanbans authorize the production of components at a workstation acting as open work orders to be filled within the transfer/review period of T time units. The movement of materials at the end of the periods from the output buffer stock of a workstation to the input buffer stock of the succeeding workstation is controlled by withdrawal kanbans. In effect, the kanbans "pull" the loaded containers through the system just in time to meet the demand at each production stage and finally the customer orders. The external demand for finished product is assumed to be Poisson with a rate of λ.

The significance of this part is to provide an understanding into how these systems work, in particular to the effects of some significant system parameters on the mean throughput rate. These results may also provide some heuristic support for stochastic optimization of large–scale systems.

In a serial periodic pull production system with an infinite supply of raw material to the first stage and subject to stochastic demand for

finished product at the last stage:

RESULT 1: *Increasing the number of identical stages in series, with keeping all other system parameters the same, decreases the mean throughput rate of the system.* See Figure A.1 in the Appendix.

> Suppose there is only one stage in the system (original system). Then, an identical stage is added in series (modified system). Considering the second stage of the modified system, if all the other system parameters are the same, except that the first stage of the modified system has an infinite production rate, then both systems have the same mean throughput rate; otherwise original system has a greater mean throughput rate than the modified system. Using the same arguments, it is straightforward to show that the mean throughput rate decreases with increasing the number of identical stages in series. □

RESULT 2: *Increasing the demand arrival rate of finished product, with keeping all other system parameters the same, increases the mean throughput rate of the system.* See Figure A.2 in the Appendix.

> Suppose there is a single–stage system (original system) producing items in order to meet the demand arrivals with a rate of λ'. Then, the demand arrival rate is increased to λ'' for the modified system. Since $\lambda' \leq \lambda''$, the mean number of demand arrivals during a period and consequently the mean of the targeted value of production during a period is also increased in the modified system relative to the original system. This concludes that the mean throughput rate is increased with increasing the rate of demand arrivals. The extension of this result for multi–stage systems is straightforward. □

RESULT 3: *Increasing the length of the transfer/review period, with keeping*

all other system parameters the same, decreases the mean throughput rate of the system. See Figure A.3 in the Appendix.

Suppose, there is a multi–stage system (continuous system) producing items with a continuous review – instantaneous order policy in which the transfer/review period length, T', tends to zero. Whenever a demand arrives to the system, a production kanban is immediately released at the last stage in order to trigger the production process and the triggering process propagates up to the first stage instantaneously. Also, whenever a part is processed at any workstation it is immediately released to the succeeding stage for the remaining operations to be done.

On the other hand, let a non–zero transfer/review period length of T'' for the periodic system. Since, in periodic systems both the review and the decisions are made on periodical basis, the collected production kanbans and the parts processed at workstations should wait until the end of the period. This makes the periodic system more stationary and less reactive to demand variations than the continuous system.

Also concentrating on the last stage of the system, with increasing the transfer/review period length and keeping the buffer stock capacity of finished product inventory at the same level, the mean number of arrivals during a period and as a consequence the mean number of demand lost during a period are increased. This concludes that the mean throughput rate is decreased with increasing the transfer/review period length. □

RESULT 4: *Increasing the total work content to be allocated to the stages of the system, with keeping all other system parameters the same, decreases the mean throughput rate of the system.* See Figure A.5 in the Appendix.

Suppose there is a single–stage system (original system) with the total work content of $1/\mu'$. Then, the total work content is increased to $1/\mu''$ for the modified system. Since $1/\mu' \leq 1/\mu''$ and $\mu' \geq \mu''$, the production rate of the modified system is less than the original system and consequently the mean throughput rate of the modified system is decreased. Considering a two stage system, let the total work content is increased from $(1/\mu_1 + 1/\mu_2)$ to $(1/\mu_1 + 1/\mu_2 + \text{some additional work})$. If the additional work is assigned to the first stage then the production rate of the first stage and consequently the component supply rate to the second stage slows down. This increases the starvation probability and decreases the production rate of the second stage also. Finally, with the additional work assigned to the first stage the mean throughput rate of the whole system is decreased. Otherwise, if the additional work is assigned to the second stage then the production rate of the second stage and consequently the mean throughput rate of the whole system slows down. As a result, the assignment of additional work to any stage will slow down the mean throughput rate. Using the same arguments, it is straightforward to show this for systems having three or more stages in tandem. This concludes that the mean throughput rate is decreased with increasing total work content to be assigned to the production stages. □

RESULT 5: *Increasing the total number of kanbans to be allocated to the stages of the system, with keeping all other system parameters the same, increases the mean throughput rate of the system.* See Figure A.6 in the Appendix.

Suppose there is a single–stage system (original system) with the total number of kanbans, K'. Then, the total number of kanbans is increased to K'' for the modified system. Since $K' \leq K''$, the number of demand lost during a period in the modified system is less than the original system and consequently the mean throughput rate

of the modified system is increased. Considering a two stage system, let the total number of kanbans is increased from $(K_1 + K_2)$ to $(K_1 + K_2 + 1)$. If the additional kanban is assigned to the first stage then the production rate of the first stage and consequently the component supply rate to the second stage increases. This decreases the starvation probability and increases the production rate of the second stage. Finally, with the additional kanban assigned to the first stage the mean throughput rate of the whole system is increased. Otherwise, if the additional kanban is assigned to the second stage then the mean number of demand lost during a period decreases and as a consequence the mean throughput rate of the whole system increases. As a result, the assignment of additional kanban to any stage will increase the mean throughput rate. Using the same arguments, it is straightforward to show this for systems having three or more stages in tandem. This concludes that the mean throughput rate is increased with increasing total number of kanbans to be allocated to the production stages. □

RESULT 6: *Increasing the maximum level of allowed backorders, with keeping all other system parameters the same, increases the mean throughput rate of the system.* See Figure A.4 in the Appendix.

Suppose there is a single–stage system (original system) producing items in order to meet the demand arrivals with allowing a maximum of B'_{FP} items backordered. Then, the maximum level of allowed backorders is increased to B''_{FP} items for the modified system. Since $B'_{FP} \leq B''_{FP}$, some of the demand lost in the original system, will not be lost in the modified system and will be satisfied after some delay. This in turn, increases the mean throughput rate of the modified system relative to the original system because allowing more backorders acts as increasing the finished product buffer storage capacity. The

extension of this result for multi–stage systems is straightforward. □

After a brief discussion about the impact of system parameters on the mean throughput rate of the system, it appears that we must progress to the integration of all system parameters simultaneously in the setting of a scarce resource allocation problem. That is, given a set of parameters, the problem is to determine the best choice of these parameters in order to optimize the performance of the system.

4.3 Statement of the Problem

Other than the integration of two allocation problems, the basic model utilized here is essentially the same as the previous studies in the literature. The system consists of N production stages corresponding to N workstations in series. Suppose that the set of all production operations required to transform a raw material into a finished product (which is also called the total work content) requires a total of TWC time units. That is, the sum of processing times at all stages, $\sum_{j=1}^{N} 1/\mu_j$, is TWC. On the other hand, the total number of kanbans available for buffer storage in the system (excluding the input buffer stock of the first stage), $\sum_{j=1}^{N} K_j$, is TNK which corresponds to the maximum number of in–process materials and finished product allowed in the system at any instant.

The primary measure of performance of the system is assumed to be the mean throughput rate $\mathbf{MTR}(\vec{\mathcal{W}}, \vec{\mathcal{K}})$, where $\vec{\mathcal{W}} = (1/\mu_1, 1/\mu_2, ..., 1/\mu_N)$ represents the allocation of workload to workstations and $\vec{\mathcal{K}} = (K_1, K_2, ..., K_N)$ represents the allocation of kanbans between workstations.

The basic problem is to find the allocation vectors $\vec{\mathcal{W}}$ and $\vec{\mathcal{K}}$ which maximizes $\mathbf{MTR}(\vec{\mathcal{W}}, \vec{\mathcal{K}})$ subject to workload and kanban constraints. In the below formulation of the problem, the parameters N, TWC and TNK are fixed constants,

whereas the μ_j are continuous and the K_j are integer decision variables:

$$maximize \quad \mathbf{MTR}(\vec{\mathcal{W}}, \vec{\mathcal{K}})$$

subject to

$$\sum_{j=1}^{N} 1/\mu_j = TWC$$

$$\sum_{j=1}^{N} K_j = TNK$$

$$1/\mu_j > 0, \quad K_j > 0 \quad and \quad K_j \ integer \ for \ j = 1, 2, ..., N.$$

The above optimization model can be viewed as a linearly constrained mixed integer non–linear programming problem, where the non–linear function $\mathbf{MTR}(\vec{\mathcal{W}}, \vec{\mathcal{K}})$ cannot be expressed explicitly. Even if the processing and demand inter–arrival times are assumed to be exponential, the limitation imposed by the number of kanbans will cause the output process not to be Poisson. For this reason closed form solutions for the stationary probabilities of the system are not available and numerical methods should be used.

The evaluation of $\mathbf{MTR}(\vec{\mathcal{W}}, \vec{\mathcal{K}})$ for any given $\vec{\mathcal{W}}$ and $\vec{\mathcal{K}}$ involves formulating the underlying queuing process as a finite state, discrete time Markov chain, and then using an appropriate numerical procedure (such as the Gauss–Seidel method) to solve the resultant system of linear equations to obtain the stationary distribution of the system. Unfortunately, the number of states in the state space of the involved Markov chain, and so the number of equations to be solved, grows very rapidly with N, K_j and B_{FP}. Recall that, the size of the state space:

$$|\mathcal{E}| = 2 * \left[\prod_{j=2}^{N}(2 * K_{j-1} + 1)\right] * (B_{FP} + 1 + K_N)$$

heavily depends on the number of stages in the system, maximum buffer storage capacities and the maximum level of backorders allowed. For many of the cases considered in this study, this number is in the thousands. This rapid growth imposes definite limits on the size of the problem that will be computationally tractable.

For the allocation of workload and kanban, there are several empirically observed properties which are first reported by Hillier and Boling [46] in serial production lines. As summarized below, subsequent studies in the literature have supported the validity of these properties as well.

- *Reversibility*: The mean throughput rate of the system is the same if the allocations are reversed, that is:

  $$\mathbf{MTR}(\vec{\mathcal{W}}, \vec{\mathcal{K}}) = \mathbf{MTR}(\vec{\mathcal{W}}', \vec{\mathcal{K}}')$$

  for any arbitrary allocation of workload $\vec{\mathcal{W}} = (1/\mu_1, 1/\mu_2, ..., 1/\mu_N)$, its mirror image is $\vec{\mathcal{W}}' = (1/\mu_N, 1/\mu_{N-1}, ..., 1/\mu_1)$ and for any arbitrary allocation of kanban (buffer storage capacity) $\vec{\mathcal{K}} = (K_1, K_2, ..., K_N)$, its mirror image is $\vec{\mathcal{K}}' = (K_N, K_{N-1}, ..., K_1)$.

- *Symmetry*: The optimal allocation of both workload and kanban (buffer storage capacity) which maximizes the mean throughput rate is symmetric, that is:

  $$1/\mu_j = 1/\mu_{N+1-j} \quad and \quad K_j = K_{N+1-j} \quad for \quad j = 1, 2, ..., N.$$

- *Monotonicity (or Bowl Phenomenon)*: The workstations receive a decreasing amount of workload or an increasing amount of buffer storage capacity as they get closer to the center of the production line, that is:

– *in terms of workload allocation:*

$$1/\mu_{j-1} > 1/\mu_j \quad for \quad 2 \leq j \leq \lceil \tfrac{N}{2} \rceil,$$

$$1/\mu_j < 1/\mu_{j+1} \quad for \quad \lfloor \tfrac{N}{2} \rfloor < j \leq N-1 \quad or$$

– *in terms of kanban allocation:*

$$K_{j-1} < K_j \quad for \quad 2 \leq j \leq \lceil \tfrac{N}{2} \rceil,$$

$$K_j > K_{j+1} \quad for \quad \lfloor \tfrac{N}{2} \rfloor < j \leq N-1 \ .$$

None of these properties has been proven yet. However, note that the reversibility property immediately implies that if the optimal solution is unique then it must satisfy the symmetry property.

It is empirically shown that the number of serious candidates to be an optimal allocation is generally small. The number of feasible allocations that need to be evaluated can be reduced greatly by using two key theoretical results, reversibility and the concavity of the mean throughput rate function with respect to allocation of both workload and buffer storage capacity [102, 103, 110, 112].

4.4 Experimental Study

These structural results together with the performance of balanced systems (more or less similar to unbalanced systems within 1 or 2 percent of the optimal) imply that an optimal allocation could be found in some neighborhood of a balanced allocation. Therefore, rather than using an optimum seeking search procedure, an *enumeration approach* is to be used in this study. An unbalancing measure which shows the degree of imbalance in an arbitrary allocation is to be defined as follows:

- For the allocation of workload:

$$DI_w = \frac{\max_{1 \le j \le N}(1/\mu_j) - \min_{1 \le j \le N}(1/\mu_j)}{t^0}$$

where TWC is assumed to be equal to $N * 10 * t^0$ ($10 * t^0$ is the average processing time for each stage) and t^0 is the elemental operation time.

- For the allocation of kanban:

$$DI_k = \max_{1 \le j \le N}(K_j) - \min_{1 \le j \le N}(K_j)$$

An experiment is designed in order to investigate the optimal allocation of both workload and kanban in multi-stage single-item pull production systems in which the Poisson demand arrives at the last stage with a mean rate of λ. The demand arrivals during the times the finished product buffer is empty are lost (back-ordering is not allowed, $B_{FP} = 0$). At each stage of the system, the processing times are exponential with the mean $1/\mu_j$ where $\sum_{j=1}^{N} 1/\mu_j = TWC$ and the number of kanbans allocated is K_j where $\sum_{j=1}^{N} K_j = TNK$. The status of the system is reviewed periodically with a period length of T. The production and material withdrawal orders are released at the beginning of periods. It is also assumed that the raw material supply for the first stage is infinite and the material handling times between stages are zero.

In the context of this experiment, 48 two-stage systems, 36 three-stage systems and 20 four-stage systems are evaluated. The framework of the experiment is as follows:

- *CASE I:* Two–stage systems (see Table A.7 in the Appendix),
  - Mean demand arrival rate is fixed,
    $\lambda = 1.0$,
  - Total work content is set equal to three different levels,
    $TWC = 1.0, 1.50, 2.0$,
    corresponding to three different levels for the demand load,
    $\rho = 0.50, 0.75, 1.0$,
  - Total number of kanbans is varied from 2 to 9,
    $TNK = 2, 3, 4, 5, 6, 7, 8, 9$,
  - Length of the transfer/review period is set to two different values,
    $T = 0.0001, 1.0$, where $T = 0.0001$ approximates the continuous review instantaneous order pull system,
  - The maximum allowable value for the degree of imbalance is less than or equal to 5, that is $DI_w \leq 5$ and $DI_k \leq 5$.
- *CASE II:* Three–stage systems (see Table A.14 in the Appendix),
  - Mean demand arrival rate is fixed,
    $\lambda = 1.0$,
  - Total work content is set equal to three different levels,
    $TWC = 1.50, 2.25, 3.0$,
    corresponding to three different levels for the demand load,
    $\rho = 0.50, 0.75, 1.0$,
  - Total number of kanbans is varied within two disjoint sets,
    $TNK = 3, 4, 5$ *and* $12, 13, 14$,
  - Length of the transfer/review period is set to two different values,
    $T = 0.0001, 1.0$, where $T = 0.0001$ approximates the continuous review instantaneous order pull system,
  - The maximum allowable value for the degree of imbalance is less than or equal to 5, that is $DI_w \leq 5$ and $DI_k \leq 5$.

- *CASE III:* Four–stage systems (see Table A.21 in the Appendix),
  - Mean demand arrival rate is fixed,
    $\lambda = 1.0$,
  - Total work content is set equal to two different levels,
    $TWC = 2.0, 4.0$,
    corresponding to two different levels for the demand load,
    $\rho = 0.50, 1.0$,
  - Total number of kanbans is varied from 4 to 8,
    $TNK = 4, 5, 6, 7, 8$,
  - Length of the transfer/review period is set to two different values,
    $T = 0.0001, 1.0$, where $T = 0.0001$ approximates the continuous review instantaneous order pull system,
  - The maximum allowable value for the degree of imbalance is less than or equal to 4, that is $DI_w \leq 4$ and $DI_k \leq 4$.

In order to obtain the general behavior of the systems in some neighborhood of balanced allocations, 960 two–stage, 18786 three–stage and 26040 four–stage **MTR** functions are evaluated by solving the involved one–step transition matrices obtained from discrete–time Markov chain models of these systems.

4.5 Empirical Results

We will present our findings on the optimal allocation of workload and kanban by focusing on two–, three– and four–stage pull production lines, respectively. In the context of the designed experiment 104 different systems are evaluated in 500 (on the average) different configurations. Because of the huge amount of raw I/O data (input: 462,234 data items and output: 995,334 data items), we will briefly discuss some of the findings on various summarizing tables given in the

Appendix. The analysis of the output data is composed of three phases; empirical observations, factorial regression models and optimal allocations.

4.5.1 Empirically Observed Properties

Throughout the experiments, according to optimal allocation results the properties — *reversibility*, *symmetry* and *monotonicity* (or Bowl Phenomenon) — are not verified. **The periodic pull production system** modeled and analyzed in this thesis is **not reversible**. The stages closer to the finished product demand require more resources (more production rate and/or more buffer storage capacity) relative to the stages closer to raw material supply. This is because of our infinite assumption of raw material supply to the first stage.

Then, the empirical results show that the optimal allocation is **not symmetric**. The optimal allocation in general follows a pattern of decreasing workload and increasing kanban allocation towards the end of the production line. See Tables A.12, A.13, A.19, A.20 and A.27 in the Appendix. As a result, the bowl–phenomenon is not observed in these periodic pull production lines. Although we have evaluated all possible allocations within the limitations on DI_w and DI_k, giving preferential treatment to center workstations does not yield better mean throughput rates than we found by giving preferential treatment to the ending stages which are closer to finished product demand.

In the correlation analysis of the **MTR** and its independent factors (input parameters defining the whole system) this result is also verified. Mean throughput rate of the system is negatively correlated with TWC and positively correlated with TNK as it is intuitively clear. See Tables A.8, A.9, A.15, A.16, A.22 and A.23 in the Appendix. As it is observed from the tables, the correlation coefficients of both the amount of workload and the number of kanbans allocated to stages is monotone increasing towards the end of the production line. Thus, the preferential treatment should be focused on the last stages whose allocation variables

are the most significantly correlated to **MTR**. See Table A.9 for the correlation coefficients of K_1 and K_2 as -0.0062 and 0.6157, respectively. Although, TNK is positively correlated with MTR, small negative correlation of K_1 is simply because of $K_1 + K_2 = TNK$. This means that increasing the number of kanbans in the first stage directly decreases the number of kanbans in the second (last) stage. Since, the production capacity lost due to decreasing the number of kanbans in the second stage is significantly greater than the production capacity gained due to increasing the number of kanbans in the first stage, the correlation coefficient of K_1 is turned out to be negative. A similar effect is also observed in Table A.23 for four stage systems.

On the other hand, concavity is the only property of mean throughput rate function observed empirically in all cases. It is very difficult to visualize the concavity of **MTR** function of systems with three or more stages on a three–dimensional graph. See as an example of the mean throughput rate function of a two–stage periodic pull system around the balanced allocation in Figure 4.1.

In periodic systems, with decreasing the transfer/review period length T, the mean throughput rate is increased. Thus, the mean throughput rate of a system controlled periodically is always lower than its continuous counterpart. See Table 4.1 for the average **MTR** of the systems evaluated within the experiment. But, on the other hand, the periodic systems carry less inventory than the continuous systems. There is a trade–off between throughput and the inventory depending on the transfer/review period length so that one cannot prefer continuous control, simply that the system could produce more relative to its periodic counterpart, without further analysis of the cost structure.

4.5.2 Factorial Regression Models

The amount of output data obtained throughout the experiment is very large so that one cannot simply analyze the whole data and point out some rules for the

Table 4.1: The average of **MTR** and **MI** in pull production systems evaluated within the experiment in order to show the effect of periodicity.

TWO-STAGE PULL SYSTEMS	CONTINUOUS approximated by $T = 0.0001$	PERIODIC with $T = 1.0$
Average **MTR**	0.7616201	0.5824245
Average $\mathbf{MI}_1$	2.2732480	1.7347790
Average $\mathbf{MI}_2$	1.8009655	1.0638895

THREE-STAGE PULL SYSTEMS	CONTINUOUS approximated by $T = 0.0001$	PERIODIC with $T = 1.0$
Average **MTR**	0.7954141	0.6550668
Average $\mathbf{MI}_1$	2.9106434	2.3480141
Average $\mathbf{MI}_2$	2.5148292	1.7338911
Average $\mathbf{MI}_3$	2.2493926	1.4087271

FOUR-STAGE PULL SYSTEMS	CONTINUOUS approximated by $T = 0.0001$	PERIODIC with $T = 1.0$
Average **MTR**	0.6174996	0.3612814
Average $\mathbf{MI}_1$	1.2935214	0.9473989
Average $\mathbf{MI}_2$	1.1551158	0.7760223
Average $\mathbf{MI}_3$	1.0443473	0.5891467
Average $\mathbf{MI}_4$	0.8920616	0.4006487

optimal allocation of workload and kanban in pull production systems. In order to summarize the output data some regression models are utilized.

In this regression analysis, there is a single dependent variable (or response) $\mathbf{MTR}(\vec{\mathcal{W}}, \vec{\mathcal{K}})$, that depends on $2 * N$ independent (or regressor) variables $\vec{\mathcal{W}}$ and $\vec{\mathcal{K}}$. The relationship between these variables is characterized by a mathematical model. The regression model is fit to the output data obtained from the designed

experiment. However, the true functional relationship between the response and the regressors is unknown.

Linear Factorial Regression Model:

$$\mathbf{MTR}^1_{reg}(\vec{W},\vec{K}) = a_0 + \sum_{i=1}^{N} a_i 1/\mu_i + \sum_{i=1}^{N} a_{N+i} K_i$$

Here, we like to determine the linear relationship between the single response variable and the regressor variables. The unknown parameters in the above linear factorial regression model are called regression coefficients and the method of least squares is used to estimate them. Some of the statistical measures showing how well the linear factorial regression model fits the data for two–stage, three–stage and four–stage pull systems is summarized in Tables A.10, A.17 and A.24, respectively, in the Appendix. The linear factorial regression model fits better to data of continuous pull systems than the data of periodic pull systems. One of the most important measures, R-square, showing the proportion of variability in the data explained or accounted for by the regression model is above 0.8 for continuous pull systems and 0.6 for periodic pull systems. Another measure, mean square error, showing the average error per data point of the regression model is around 0.01. These are quite satisfactory results for linear factorial regression model.

The least square solution of coefficient estimates are given in Tables A.11, A.18 and A.25 for two–stage, three–stage and four–stage systems, respectively, in the Appendix. The significance of these linear models is that the coefficient estimates point the stage where the preferential treatment (less workload and more kanban) should be focused.

- *Two–stage systems:* See Table A.11 in the Appendix, for the coefficient estimates of linear factorial regression model, $a_1 > a_2$ and $a_3 < a_4$. This means that, in order to increase mean throughput rate of the system allocate less workload and more kanban to the second stage than the first stage.

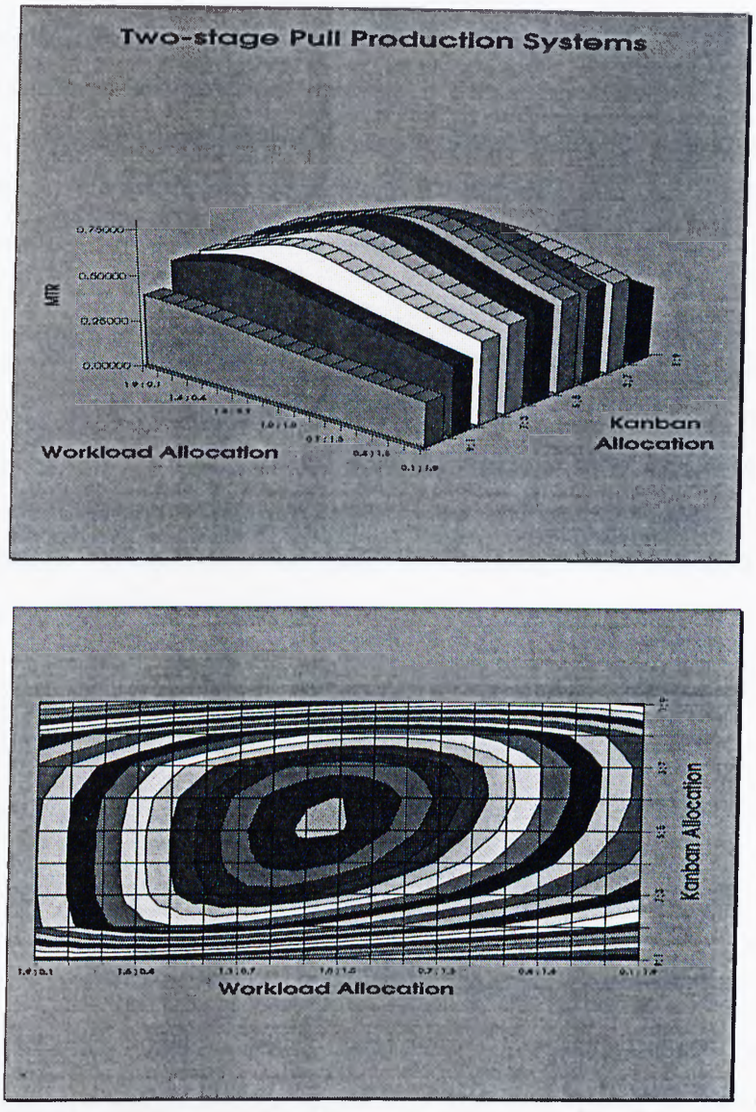


Figure 4.1: The mean throughput rate function in a two–stage periodic pull production system. The function is concave with respect to both allocation of workload and kanbans. In the contour plot, the maximum is at the quadrant in which the second stage gets less workload and more number of kanbans.

(Fixed parameters of the two–stage system: mean demand arrival rate $\lambda = 1.0$; transfer/review period length $T = 1.0$; total work content $TWC = 2.0$; total number of kanbans $TNK = 10$).

- *Three–stage systems:* See Table A.18 in the Appendix, for the coefficient estimates of linear factorial regression model, $a_1 > a_2 > a_3$ and $a_4 < a_5 < a_6$. This means that, in order to increase mean throughput rate of the system a decreasing workload and an increasing kanban allocation should be utilized. The most critical stage that requires preferential treatment is the last stage.

- *Four–stage systems:* See Table A.25 in the Appendix, for the coefficient estimates of linear factorial regression model, $a_1 > a_2 > a_3 > a_4$ and $a_5 < a_6 < a_7 < a_8$. This means that, in order to increase mean throughput rate of the system a decreasing workload and an increasing kanban allocation should be utilized. The most critical stage that requires preferential treatment is the last stage.

Quadratic Factorial Regression Model:

$$\mathbf{MTR}^2_{reg}(\vec{\mathcal{W}},\vec{\mathcal{K}}) = a_0 + \sum_{i=1}^{N} a_i 1/\mu_i + \sum_{i=1}^{N} a_{N+i} K_i + \sum_{i=1}^{N}\left[\sum_{j=i}^{N} a_{i,j} 1/\mu_i 1/\mu_j + \sum_{j=1}^{N} a_{i,N+j} 1/\mu_i K_j\right] + \sum_{i=1}^{N}\sum_{j=i}^{N} a_{N+i,N+j} K_i K_j$$

Response surface methodology is a collection of mathematical and statistical techniques that are useful for the modelling and analysis of problems in which a response, like mean throughput rate **MTR**, is influenced by several variables, like workload and kanban allocations $\vec{\mathcal{W}}$ and $\vec{\mathcal{K}}$, and the objective is to optimize the response. If the fitted surface is an adequate approximation of the response function, then analysis of the fitted surface will be approximately equivalent to analysis of the actual system. Since the form of the relationship between the response and the independent variables is unknown, a low–order (second order) polynomial is employed. The method of least squares is again used to estimate the regression coefficients. The quadratic factorial regression model better fits the data than the linear model in terms of all statistical measures considered.

R-Square is above 0.9 and 0.8 for continuous and periodic pull systems, respectively. Mean square error is reduced to 0.005. But, on the other hand, individual interpretation of regression coefficients with the inclusion of second order terms becomes meaningless. See Table 4.2 for the increase in number of terms to be utilized in a third order polynomial relative to linear and quadratic models.

Table 4.2: The number of terms utilized in factorial regression models developed for pull production systems evaluated within the experiment in order to summarize the huge amount of output data.

Size of Factorial Regression Models			
Factorial Regression Models	Number of Regression Terms 2–stage	3–stage	4–stage
$\mathbf{MTR}^1_{reg}(\vec{\mathcal{W}},\vec{\mathcal{K}})$	5	7	9
$\mathbf{MTR}^2_{reg}(\vec{\mathcal{W}},\vec{\mathcal{K}})$	15	28	45
$\mathbf{MTR}^3_{reg}(\vec{\mathcal{W}},\vec{\mathcal{K}})$	35	84	165
$\mathbf{MTR}^l_{reg}(\vec{\mathcal{W}},\vec{\mathcal{K}})$	$1+\sum_{j=1}^{l}\binom{2N-1+j}{j}$		

4.5.3 Optimal Allocations

Throughout this experiment an overall average of 1.35% improvement is obtained in the mean throughput rate over the balanced (as possible as) systems. Note that, in the design of experiment, there are several cases in which the total number of kanbans cannot be equally allocated to the stages in the system. In such cases, a composite measure of the degree of imbalance in both allocation of workload and kanban is defined as:

$$DI = \left| \left(\frac{TWC}{N^2} \sum_{i=1}^{N} \frac{1}{\mu_i} \right) - \left(\frac{TWC}{N} \prod_{i=1}^{N} \frac{1}{\mu_i} \right) \right| + \left| \left(\frac{TNK}{N^2} \sum_{i=1}^{N} K_i \right) - \left(\frac{TNK}{N} \prod_{i=1}^{N} K_i \right) \right|$$

This aids to find the most closely balanced configuration with maximized mean throughput rate. The level of the average improvement obtained is similar to the results reported in the literature. The results regarding the optimal allocation of both workload and kanban in pull production systems summarized in Tables A.12, A.13, A.19, A.20 and A.27, and a brief evaluation is as follows:

- *General Rule:* **Select kanbans to allocate first. Allocate kanbans in a monotone increasing pattern in which first stage gets less kanban than the last stage of the system. Allocate workload in a monotone decreasing pattern in which first stage gets more workload than the last stage of the system.**

- *Exceptions:* If TNK is low, then the effect of one unit of imbalance in the allocation of kanban is high. That is, giving one kanban to any stage results in high preferment to that stage, instead of taking some amount of this effect back, some extra workload could be transferred to that stage. As a result, in such cases an increasing pattern of workload may give the best performance.

- *Continuous vs Periodic:* The number of exceptions increases with the number of stages in the system and also with increasing the length of transfer/review period.

Note that, kanban allocation variables are discrete. On the other hand, although workload allocation variables were assumed continuous in the formulation, they are made discrete as multiples of elemental task time t_0 in the context of the experiment. This also causes some exceptions in the optimal allocation of workload.

4.6 Allocation Methodology

The allocation methodology we propose utilizes an evaluative modelling approach. The evaluation of mean throughput rate, $\mathbf{MTR}(\vec{\mathcal{W}}, \vec{\mathcal{K}})$, for any given $\vec{\mathcal{W}}$ and $\vec{\mathcal{K}}$ involves formulating the system as a finite state, discrete time Markov process and then using an appropriate technique to solve the resultant system of linear equations to obtain the stationary distribution of the system. The objective of the allocation methodology is to achieve the maximum mean throughput rate of the system with providing the best set of parameters regarding the allocation of total work content and the total number of kanbans among workstations. In this respect, the process through which the best set of allocation decisions generated is semi-generative.

Our proposed allocation methodology can be outlined as follows:

1. Allocate the number of kanbans to workstations as equal as possible,
2. Allocate the amount of workload to workstations as equal as possible,
3. If the resulting configuration is a pure balanced allocation,
   then all stages are identical to each other.
   In such a system the last stage which produces the finished product becomes the bottleneck because the other stages on top of their buffer stocks utilize the intermediate buffers of stages up to last stage as extra stocks.
   So, the system should be configured in such a way that all stages should be bottleneck (critical) at the same instant.
4. Either the resulting system has to possess imbalances because of indivisibility of the operations and precedence relations or not, depending on the total number of kanbans to be allocated, giving more preferential treatment to the last stage might improve the **MTR**. That is:

(a) If TNK is low,

i. allocate the kanbans as equal as possible,
if balanced allocation is not possible then allocate more kanban to the last stage(s),

ii. select a pattern (decreasing, balanced or increasing) for the allocation of workload depending on the effect of imbalance in the allocation of kanban.

(b) Otherwise, if TNK is sufficient,

i. select a monotone increasing pattern for kanban allocation with special emphasis given to the last stage.

ii. select a monotone decreasing pattern for workload allocation in which the first stage gets more workload than the last stage,

Note that, decreasing the workload and increasing the number of kanbans in a system have similar effect on mean throughput rate. In this respect they are treated as substitute of each other.

Chapter 5

Conclusion & Further Research Directions

This chapter provides a brief summary of the contributions of this dissertation research and addresses a wide range of directions for future research.

5.1 Contributions

A variety of production systems appearing in the literature is overviewed and a classification scheme is developed. Most of the approaches considered in the review are analytical studies dealing with mathematical performance evaluation models of stochastic production systems. Uncertainties such as the variability in processing and demand inter–arrival times are generally assumed to be exponential and the researchers proposed approximate decomposition procedures for large–scale systems.

There has been a number of attempts in developing analytical models for the performance evaluation of kanban–controlled stochastic pull production systems.

Majority of the existing models address tandem queue equivalent systems. In the light of the proposed classification scheme, there are a lot of non–tandem–queue equivalent pull production systems to be considered:

- Periodic review systems with:
  - exponential/non–exponential distributions,
  - periodic/batch transfer of in–process materials,
  - batch ordering.
- Continuous review systems with:
  - non–exponential distributions,
  - batch transfer of in–process materials,
  - batch ordering.
- Multi–item systems with:
  - non–zero setup times,
  - priority scheduling.

While increasing the complexity of the systems to be modeled by introducing some of the above characteristics such as order and transfer batches, set–up times and priorities, the feasibility of the exact evaluation of systems having more than one–stage decreases. That is, a multi–stage model with some of the above characteristics becomes both analytically and computationally intractable. For those systems, the only feasible approach is to develop approximate evaluation techniques with an acceptable level of accuracy.

A periodic review – instantaneous order / periodic transfer system is selected as the base system to start a research on modelling and analysis of non–tandem–queue equivalent pull production systems. This base system is formulated as a

discrete time Markov process. Because of the dimensionality problem inherited in the exact solution technique, the base system could exactly be evaluated up to five stages in tandem and the solution of the model remains computationally feasible by the use of special data structures and a sparse matrix solver which could be obtained from NETLIB via internet.

An approximate decomposition approach is developed to handle larger systems which are analytically intractable. Approximation is demonstrated on our base system in which the arrival and the production processes are both Markovian. The proposed approximate decomposition approach generates results which are quite close to the exact solution in an experiment designed for three–stage systems. The average percent absolute error which is the measure used in comparison is less than 2.5. Note that, the approximate decomposition approach has not been tested for larger systems. For this reason, we have limited experience about the accuracy level and the convergence properties of the approximation technique. However, in the many examples we have examined, the method has always converged within a reasonable number of iterations, only moderately on the number of stages. Since, there is no similar approximation method for periodic pull production systems, and exact solution of larger systems is formidable, simulation remains to be the only tool to be used for comparing the results of the approximate decomposition technique.

The extensions of the approximate decomposition technique to cover unreliable machines are straightforward. In terms of configuration of network, the approximation could be extended to cover periodic pull production systems in flow-shop configuration by formulating the split and merge sub–systems.

The design of tandem production systems has been well studied in the production research literature with the primary focus being on how to improve their efficiency. Considering the large costs associated with these systems, a slight improvement in efficiency can lead to very significant savings over the life of the production system. Division of work among the workstations and allocation of buffer storage

capacity between workstations are two critical design factors that have attracted the attention of many researchers.

Another contribution of this dissertation study is to provide an understanding into how these periodic pull systems work, in particular under the effects of some system parameters:

- the number of stages in the system (N),
- the external demand arrival rate (λ),
- the transfer/review period length (T),
- the total work content (TWC),
- the total number of kanbans (TNK),
- the maximum level of allowed backorders (B_{FP}),

on the mean throughput rate of the system. These theoretical results that characterize the dynamics of these pull systems can provide some heuristic support in the analysis of large–scale pull production systems.

Except for the integration of two allocation problems, the basic model utilized in this study is essentially the same as the previous studies in the literature. The basic problem is to find the allocation vectors $\vec{\mathcal{W}}$ and $\vec{\mathcal{K}}$ which maximize $\mathbf{MTR}(\vec{\mathcal{W}}, \vec{\mathcal{K}})$ subject to workload and kanban constraints.

An experiment is designed in order to investigate the optimal allocation of both workload and kanban in two–stage, three–stage and four–stage systems. The results do not support the properties — *reversibility*, *symmetry* and *monotonicity* — in periodic pull production systems. Similar to the results reported by Villeda, Dudek and Smith [106], a decreasing workload and an increasing kanban allocation strategy gives always a consistent improvement (1–to–10 percent relative to

balanced allocation) in the mean throughput rate. That is, the stages closer to demand are intrinsically bottleneck in a balanced system and requires preferential treatment (less workload and more buffer storage capacity) over the other stages.

5.2 Future Research Directions

At the end, there are several future research directions emanating from this dissertation research study as such:

- *further investigation of the base system:*
  In order to improve the overall accuracy level of the approximation, a further study could be the development and analysis of a **two–stage decomposition technique**. This type of approximation might lower the absolute errors on performance measures since one of the approximated probability utilized in the decomposition technique could be exactly evaluated. On the other hand, the computation requirements of a two–stage decomposition are increased both in terms of memory and cpu time.

  In another research, with the opportunity of parallel computing, the level of accuracy in approximate decomposition technique could be investigated for **larger systems having five or more stages in tandem**. If it could be possible to evaluate such larger systems efficiently, then with some further experimentation the findings related with the problem of optimal allocation of workload and kanban could be more generalized.

  The objective in the workload and kanban allocation problem could be generalized from throughput maximization to **cost minimization** or **profit maximization**. Then, an optimal seeking solution procedure should be implemented as another future research study. The definition of the cost model and the development of an optimal seeking solution procedure give the opportunity to investigate several **trade–offs**, for example between

backorders and finished product inventories. In another study, it could be possible to investigate conditions under which a periodic control policy is better than a continuous one or vice versa.

- *extensions to the current formulation of the base system:*
  By formulating **the split and the merge sub–systems**, the model could be extended to cover production systems in flow–shop configuration.

  Another future research could be based on the interaction of the variation coming from the stochastic processes and **several discrete distributions with different levels of variation** could be used in the formulation.

  With the inclusion of an **external raw material supply function** in a further study, the base system might become more realistic. Then, the effect of external raw material supply function on the performance of the system could be investigated.

- *further investigation of the extended systems:*
  In a series of related research study, the analysis on both approximation and optimal allocation problems could be extended to cover these systems. If the exact performance evaluation model of the extended systems become computationally intractable, then a **discrete–time simulation model** should be developed in order to carry on the analysis.

- *formulation and analysis of new systems:*
  With the insight gained in this study, developing both exact and approximate performance evaluation models for **multi–item multi–stage periodic pull production systems** could be an interesting future research. Note that, when there are more than one item in the system, because of some shared resources, set–up times and scheduling priorities the formulation becomes complicated. The use of vacation queues could be helpful in the development of the approximate model.

Modelling and analysis of pull production systems would attract more attention from researchers in a number of directions, particularly with approximate evaluation methods handling more general inventory level triggered multi–stage multi–item pull production systems.

List of Notation

Notation	Explanation	Remark
a_j and $a_{i,j}$	Factorial regression model coefficients	regression parameter
AMB	Approximated mean backorder level at Q_{FP}	performance measure
$\mathbf{AMI}_j$	Approximated mean inventory level at Q_j^{in}	performance measure
$\mathbf{AMU}_j$	Approximated mean utilization of W_j	performance measure
AMBT	Approximated mean backorder time at Q_{FP}	performance measure
AMTR	Approximated mean throughput rate of the system	performance measure
$\mathbf{AMDR}_s$	Approximated mean demand rate satisfied on time	performance measure
$\mathbf{AMDR}_{b/o}$	Approximated mean demand rate backordered	performance measure
$\mathbf{AMDR}_{lost}$	Approximated mean demand rate lost	performance measure
$\mathbf{APAPO}_j$	Approximated probability of achieving production objective at W_j	performance measure
$\mathbf{APNOM/H}_j$	Approximated probability of no material handling at W_j	performance measure
B_{FP}	Maximum allowable backorder level at Q_{FP}	system parameter
C_j	Component j	system description
d_s^0	Number of satisfied/backordered demand arrivals	realization
$D_s(k)$	Number of satisfied/backordered demand arrivals within period k	random variable
DI	Composite measure for the degree of imbalance	system description
DI_k	Degree of Imbalance in the allocation of kanbans	system description
DI_w	Degree of Imbalance in the allocation of workload	system description
$\vec{e}$	Row vector of all ones	vector
ϵ	Accuracy level of approximation	constant
$\mathcal{E}$	State space	matrix
FP	Finished product, or alternatively C_N	system description
i_j^0	Stock level at Q_j^{in}	realization
$I_{FP}(k)$	Stock level at Q_{FP} at the beginning of period k	random variable
$I_j^{in}(k)$	Stock level at Q_j^{in} at the beginning of period k	random variable
$I_j^{out}(k)$	Stock level at Q_j^{out} at the beginning of period k	random variable
j	Stage index	index
k	Period index	index
K	Average number of kanbans allocated per stage	system parameter
$\vec{\mathcal{K}}$	Kanban allocation vector	vector
K_j	Stock capacity of Q_j	
	or total number of kanbans allocated between W_j and W_{j+1}	system parameter
K_j^P	Number of production kanbans at W_j	system parameter
K_j^W	Number of withdrawal kanbans at W_j	system parameter

Notation	Explanation	Remark
l	Dummy index, e.g., summation or iteration index	index
λ	Demand arrival rate for FP	system parameter
$1/\lambda$	Mean demand inter–arrival time	system parameter
$m[\vec{\mathcal{S}}(k-1), \vec{\mathcal{S}}(k)]$	One step transition probability from $\vec{\mathcal{S}}(k-1)$ to $\vec{\mathcal{S}}(k)$	matrix element
$m_{\mathcal{Z}_j}[\vec{\mathcal{S}}_{\mathcal{Z}_j}(k-1), \vec{\mathcal{S}}_{\mathcal{Z}_j}(k)]$	One step transition probability from$\vec{\mathcal{S}}_{\mathcal{Z}_j}(k-1)$ to $\vec{\mathcal{S}}_{\mathcal{Z}_j}(k)$	matrix element
M	One step transition matrix	matrix
$M_{\mathcal{Z}_j}$	One step transition matrix of $\mathcal{Z}_j$	matrix
$M_{\mathcal{Z}_j}^{(l)}$	One step transition matrix of $\mathcal{Z}_j$ at iteration l	matrix
$\mathbf{MB}$	Mean backorder level at Q_{FP}	performance measure
$\mathbf{MI}_j$	Mean inventory level of Q_j^{in}	performance measure
$\mathbf{MU}_j$	Mean utilization of W_j	performance measure
$\mathbf{MBB}$	Mean period beginning backorder level at Q_{FP}	performance measure
$\mathbf{MBT}$	Mean backorder time at Q_{FP}	performance measure
$\mathbf{MEB}$	Mean period ending backorder level at Q_{FP}	performance measure
$\mathbf{MTR}$	Mean throughput rate of the system	performance measure
$\mathbf{MBI}_j$	Mean period beginning inventory level of Q_j^{in}	performance measure
$\mathbf{MEI}_j$	Mean period ending inventory level of Q_j^{in}	performance measure
$\mathbf{MTR}_j$	Mean throughput rate of W_j	performance measure
$\mathbf{MDR}_s$	Mean demand rate satisfied on time	performance measure
$\mathbf{MDR}_{b/o}$	Mean demand rate backordered	performance measure
$\mathbf{MDR}_{lost}$	Mean demand rate lost	performance measure
$\mathbf{MTR}_{\mathcal{Z}_j}$	Mean throughput rate of $\mathcal{Z}_j$	performance measure
$\mathbf{MTR}_{\mathcal{Z}_j}^{l_b}$	Mean throughput rate of $\mathcal{Z}_j$ in backward pass	performance measure
$\mathbf{MTR}_{\mathcal{Z}_j}^{l_f}$	Mean throughput rate of $\mathcal{Z}_j$ in forward pass	performance measure
$\mathbf{MTR}_{reg}^{l}$	lth–order factorial regression model of $\mathbf{MTR}$	regression model
$\mathrm{MTTP}_j(p_j^0)$	Mean time to process p_j^0 number of C_j at W_j	intermediate measure
$\mathrm{MTTD}_s(d_s^0)$	Mean time to arrival of d_s^0 number of FP demand	intermediate measure
μ	Average production rate per stage	system parameter
$1/\mu$	Average processing time per stage	system parameter
μ_j	Mean production rate at W_j	system parameter
$1/\mu_j$	Mean processing time at W_j	system parameter
N	Number of stages (workstations) in the system	system parameter
$N_D(t), t \geq 0$	Stochastic demand arrival process	stochastic process
$N_{P_j}(t), t \geq 0$	Stochastic production process at W_j	stochastic process
$O_j(k)$	Production objective of W_j at the beginning of period k	random variable
p_j^0	Number of C_j processed at W_j	realization
$\vec{P}$	Collection of $N+1$ random variables	vector
$P[\bullet]$	Probability function	function
$P_j(k)$	Number of C_j processed at W_j within period k	random variable
$\mathbf{PAPO}_j$	Probability of achieving production objective at W_j	performance measure
$\mathbf{PNOM/H}_j$	Probability of no material handling at W_j	performance measure

Notation	Explanation	Remark
$\vec{\pi}$	Limiting probabilities of states of the system	vector
$\vec{\pi}_{\mathcal{Z}_j}^{(l)}$	Limiting probabilities of states of the $\mathcal{Z}_j$	vector
Q	Order quantity	system parameter
Q_j	Input queue of S_j	system description
Q_{FP}	Finished product stock or alternatively Q_{N+1}^{in}	system description
Q_j^{in}	Input stock of W_j	system description
Q_j^{out}	Output stock of W_j	system description
(Q,R)	Inventory control policy	system parameter
ρ	Demand load (traffic intensity)	system parameter
$\mathcal{R}$	Feasible set for realization of $N+1$ random variables	set
RM	Raw material, or alternatively C_0	system description
(R,r)	Inventory control policy	system parameter
S_j	Server j in a tandem line	system description
$\vec{\mathcal{S}}(k)$	State of the system at the beginning of period k	vector
$\vec{\mathcal{S}}_{\mathcal{Z}_j}(k)$	State of $\mathcal{Z}_j$ at the beginning of period k	vector
t	Time variable	continuous variable
t^0	Elemental operation time	system parameter
T	Transfer/review period length in time units	system parameter
TNK	Total Number of Kanbans to be allocated to workstations	system parameter
TWC	Total Work Content to be allocated to workstations	system parameter
w_j^0	Status of W_j	realization
$\vec{\mathcal{W}}$	Workload allocation vector	vector
W_j	Workstation j in the system	system description
$W_j^{on}(k)$	Status of W_j at the beginning of period k either processing a component or not	random variable
$\mathcal{Z}_j$	Sub-system j in the system	system description
$\lvert\bullet\rvert$	Absolute value	function
$\lvert\bullet\rvert$	Cardinality, the number of distinct elements in	function
$\lfloor\bullet\rfloor$	Greatest integer smaller than the argument	function
$\lceil\bullet\rceil$	Smallest integer greater than the argument	function
$\delta[\bullet]$	Indicator of a given condition	function
$\psi[\bullet]$	Indicator of a transition from one state to the other	function
$\xi[\bullet]$	Indicator of a transition from one state to the other	function

APPENDIX

Table A.1: The dimensional properties of transition matrices of various periodic pull production systems in which no backordering is allowed:

- the number of non-zero elements,
- the number of total elements, $|\mathcal{E}|$x$|\mathcal{E}|$,
- the sparsity and density, (%).

In single-stage systems, increasing the number of kanbans increases the density of the transition matrix. On the other hand, for multi-stage systems, the density of the transition matrix decreases with increasing either the number of kanbans or the number of stages.

The number of kanbans/stage	Single Stage		Two Stage		Three Stage	
	Non-zeros	Total	Non-zeros	Total	Non-zeros	Total
	Sparsity	Density	Sparsity	Density	Sparsity	Density
1	3	2x2	16	6x6	72	18x18
	25.00	75.00	55.56	44.44	77.78	22.22
2	7	3x3	86	15x15	769	75x75
	22.22	77.78	61.78	38.22	86.33	13.67
3	13	4x4	286	28x28	4,180	196x196
	18.75	81.25	63.52	36.48	89.12	10.88
4	21	5x5	727	45x45	15,875	405x405
	16.00	84.00	64.10	35.90	90.32	9.68
5	31	6x6	1,556	66x66	47,748	726x726
	13.89	86.11	64.28	35.72	90.94	9.06
6	43	7x7	2,956	91x91	121,819	1183x1183
	12.25	87.75	64.30	35.70	91.29	8.71

Table A.2: Mean throughput rates obtained from both the exact and the approximate solution techniques are given below. The system parameters are in the range: transfer/review period length $T = 0.25$, mean demand arrival rate $\lambda = (0.25, 0.50, 1.00, 2.00, 4.00)$, mean production rate $\mu = \frac{\lambda}{\rho}$, where the demand load $\rho = (0.50, 0.60, 0.70, 0.80, 0.90)$ and number of kanbans $K = (2, 3, 4, 5, 6)$.

$\rho = 0.50$										$T = 0.25$
Demand Arrival Rate	Number of Kanbans allocated at each stage									
	$K = 2$		$K = 3$		$K = 4$		$K = 5$		$K = 6$	
	MTR	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**
$\lambda = 0.25$	0.21070	0.20624	0.23098	0.22995	0.23948	0.24056	0.24512	0.24547	0.24770	0.24780
$\lambda = 0.50$	0.41460	0.40624	0.46072	0.45701	0.48090	0.47976	0.48966	0.49031	0.49511	0.49530
$\lambda = 1.00$	0.79731	0.78537	0.90845	0.90085	0.95662	0.95329	0.97876	0.97766	0.98884	0.98918
$\lambda = 2.00$	1.45217	1.45224	1.74637	1.73590	1.88056	1.87370	1.94250	1.93939	1.97179	1.97068
$\lambda = 4.00$	2.35895	2.43495	3.10761	3.12902	3.54494	3.54297	3.77462	3.76939	3.88931	3.88601

$\rho = 0.60$										$T = 0.25$
Demand Arrival Rate	Number of Kanbans allocated at each stage									
	$K = 2$		$K = 3$		$K = 4$		$K = 5$		$K = 6$	
	MTR	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**
$\lambda = 0.25$	0.20048	0.19334	0.22297	0.21938	0.23419	0.23292	0.23925	0.24031	0.24396	0.24444
$\lambda = 0.50$	0.39408	0.38099	0.44376	0.43575	0.46833	0.46418	0.48128	0.47969	0.48745	0.48835
$\lambda = 1.00$	0.75700	0.73758	0.87339	0.85823	0.92996	0.92103	0.95982	0.95525	0.97623	0.97435
$\lambda = 2.00$	1.37935	1.36973	1.67499	1.65332	1.82282	1.80605	1.90006	1.89005	1.94224	1.93703
$\lambda = 4.00$	2.25607	2.32175	2.98409	2.99264	3.42504	3.41020	3.67531	3.65846	3.81474	3.80279

$\rho = 0.70$										$T = 0.25$
Demand Arrival Rate	Number of Kanbans allocated at each stage									
	$K = 2$		$K = 3$		$K = 4$		$K = 5$		$K = 6$	
	MTR	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**
$\lambda = 0.25$	0.19047	0.18050	0.21391	0.20715	0.22665	0.22251	0.23382	0.23200	0.23936	0.23807
$\lambda = 0.50$	0.37409	0.35598	0.42513	0.41142	0.45249	0.44325	0.46850	0.46287	0.47908	0.47540
$\lambda = 1.00$	0.71791	0.69049	0.83554	0.81032	0.89713	0.87877	0.93299	0.92078	0.95504	0.94756
$\lambda = 2.00$	1.30877	1.28854	1.59890	1.56284	1.75376	1.72136	1.84233	1.81832	1.89606	1.87990
$\lambda = 4.00$	2.15464	2.20778	2.85273	2.84548	3.28673	3.25311	3.54880	3.51136	3.70819	3.67696

$\rho = 0.80$										$T = 0.25$
Demand Arrival Rate	Number of Kanbans allocated at each stage									
	$K = 2$		$K = 3$		$K = 4$		$K = 5$		$K = 6$	
	MTR	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**
$\lambda = 0.25$	0.18059	0.16816	0.20431	0.19406	0.21770	0.20991	0.22590	0.22049	0.23220	0.22791
$\lambda = 0.50$	0.35495	0.33106	0.40589	0.38551	0.43451	0.41809	0.45247	0.43977	0.46443	0.45495
$\lambda = 1.00$	0.68061	0.64536	0.79606	0.75983	0.85978	0.82880	0.89918	0.87439	0.92534	0.90619
$\lambda = 2.00$	1.24152	1.21055	1.52058	1.46882	1.67613	1.62405	1.77063	1.72545	1.83223	1.79557
$\lambda = 4.00$	2.05695	2.09640	2.71841	2.69359	3.13523	3.07888	3.39775	3.33248	3.56800	3.50607

$\rho = 0.90$										$T = 0.25$
Demand Arrival Rate	Number of Kanbans allocated at each stage									
	$K = 2$		$K = 3$		$K = 4$		$K = 5$		$K = 6$	
	MTR	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**
$\lambda = 0.25$	0.17153	0.15659	0.19456	0.18089	0.20834	0.19615	0.21704	0.20671	0.22294	0.21446
$\lambda = 0.50$	0.33684	0.30944	0.38618	0.35952	0.41463	0.39076	0.43329	0.41230	0.44614	0.42807
$\lambda = 1.00$	0.64541	0.60298	0.75641	0.70940	0.81932	0.77498	0.85954	0.81985	0.88749	0.85253
$\lambda = 2.00$	1.17811	1.13694	1.44235	1.37533	1.59299	1.52068	1.68767	1.61858	1.75210	1.68911
$\lambda = 4.00$	1.96413	1.98954	2.58490	2.54215	2.97626	2.89658	3.22821	3.13331	3.39780	3.30056

Table A.3: Mean throughput rates obtained from both the exact and the approximate solution techniques are given below. The system parameters are in the range: transfer/review period length $T = 0.50$, mean demand arrival rate $\lambda = (0.25, 0.50, 1.00, 2.00, 4.00)$, mean production rate $\mu = \frac{\lambda}{\rho}$, where the demand load $\rho = (0.50, 0.60, 0.70, 0.80, 0.90)$ and number of kanbans $K = (2, 3, 4, 5, 6)$.

$\rho = 0.50$										$T = 0.50$
Demand Arrival Rate	Number of Kanbans allocated at each stage									
	$K = 2$		$K = 3$		$K = 4$		$K = 5$		$K = 6$	
	MTR	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**
$\lambda = 0.25$	0.20710	0.20312	0.23036	0.22850	0.24045	0.23988	0.24483	0.24516	0.24756	0.24765
$\lambda = 0.50$	0.39856	0.39268	0.45423	0.45042	0.47786	0.47665	0.48821	0.48883	0.49442	0.49459
$\lambda = 1.00$	0.72604	0.72612	0.87319	0.86795	0.94011	0.93685	0.97096	0.96970	0.98490	0.98534
$\lambda = 2.00$	1.17948	1.21748	1.55381	1.56451	1.77247	1.77148	1.88731	1.88469	1.94466	1.94301
$\lambda = 4.00$	1.67347	1.72371	2.29519	2.40578	2.88705	2.94586	3.31387	3.34416	3.60493	3.61569

$\rho = 0.60$										$T = 0.50$
Demand Arrival Rate	Number of Kanbans allocated at each stage									
	$K = 2$		$K = 3$		$K = 4$		$K = 5$		$K = 6$	
	MTR	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**
$\lambda = 0.25$	0.19689	0.19050	0.22188	0.21788	0.23340	0.23209	0.24064	0.23984	0.24373	0.24417
$\lambda = 0.50$	0.37841	0.36879	0.43670	0.42912	0.46466	0.46051	0.47931	0.47763	0.48811	0.48717
$\lambda = 1.00$	0.68964	0.68487	0.83750	0.82666	0.91126	0.90302	0.94980	0.94502	0.97112	0.96851
$\lambda = 2.00$	1.12802	1.16087	1.49204	1.49632	1.71252	1.70510	1.83765	1.82923	1.90737	1.90140
$\lambda = 4.00$	1.62249	1.68039	2.23990	2.33743	2.80742	2.85913	3.22657	3.24972	3.51994	3.52487

$\rho = 0.70$										$T = 0.50$
Demand Arrival Rate	Number of Kanbans allocated at each stage									
	$K = 2$		$K = 3$		$K = 4$		$K = 5$		$K = 6$	
	MTR	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**
$\lambda = 0.25$	0.18692	0.17799	0.21256	0.20571	0.22625	0.22162	0.23425	0.23143	0.23901	0.23770
$\lambda = 0.50$	0.35888	0.34524	0.41769	0.40516	0.44857	0.43938	0.46649	0.46039	0.47752	0.47378
$\lambda = 1.00$	0.65436	0.64427	0.79941	0.78142	0.87688	0.86068	0.92117	0.90916	0.94803	0.93995
$\lambda = 2.00$	1.07731	1.10389	1.42637	1.42274	1.64337	1.62656	1.77436	1.75568	1.85803	1.83848
$\lambda = 4.00$	1.56507	1.62938	2.17393	2.25766	2.71357	2.75537	3.11800	3.13108	3.40804	3.40282

$\rho = 0.80$										$T = 0.50$
Demand Arrival Rate	Number of Kanbans allocated at each stage									
	$K = 2$		$K = 3$		$K = 4$		$K = 5$		$K = 6$	
	MTR	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**
$\lambda = 0.25$	0.17737	0.16598	0.20281	0.19275	0.21708	0.20904	0.22596	0.21988	0.23173	0.22748
$\lambda = 0.50$	0.34024	0.32268	0.39794	0.37992	0.42977	0.41440	0.44941	0.43720	0.46267	0.45309
$\lambda = 1.00$	0.62074	0.60528	0.76025	0.73441	0.83800	0.81202	0.88531	0.86273	0.91612	0.89778
$\lambda = 2.00$	1.02846	1.04820	1.35918	1.34680	1.56759	1.53944	1.69887	1.66624	1.78397	1.75304
$\lambda = 4.00$	1.50693	1.57473	2.10194	2.17178	2.61061	2.64041	2.99307	2.99344	3.27172	3.25220

$\rho = 0.90$										$T = 0.50$
Demand Arrival Rate	Number of Kanbans allocated at each stage									
	$K = 2$		$K = 3$		$K = 4$		$K = 5$		$K = 6$	
	MTR	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**
$\lambda = 0.25$	0.16833	0.15472	0.19296	0.17976	0.20712	0.19538	0.21650	0.20615	0.22284	0.21403
$\lambda = 0.50$	0.32265	0.30149	0.37816	0.35470	0.40959	0.38749	0.42966	0.40993	0.44367	0.42627
$\lambda = 1.00$	0.58906	0.56847	0.72113	0.68766	0.79646	0.76034	0.84383	0.80929	0.87600	0.84455
$\lambda = 2.00$	0.98206	0.99477	1.29243	1.27107	1.48811	1.44829	1.61408	1.56666	1.69888	1.65028
$\lambda = 4.00$	1.45102	1.51912	2.02729	2.08350	2.50260	2.51929	2.85725	2.84336	3.11634	3.08056

Table A.4: Mean throughput rates obtained from both the exact and the approximate solution techniques are given below. The system parameters are in the range: transfer/review period length $T = 1.00$, mean demand arrival rate $\lambda = (0.25, 0.50, 1.00, 2.00, 4.00)$, mean production rate $\mu = \frac{\lambda}{\rho}$, where the demand load $\rho = (0.50, 0.60, 0.70, 0.80, 0.90)$ and number of kanbans $K = (2, 3, 4, 5, 6)$.

$\rho = 0.50$										$T = 1.00$
Demand Arrival Rate	Number of Kanbans allocated at each stage									
	$K = 2$		$K = 3$		$K = 4$		$K = 5$		$K = 6$	
	MTR	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**
$\lambda = 0.25$	0.19928	0.19634	0.22695	0.22521	0.23893	0.23832	0.24410	0.24441	0.24721	0.24729
$\lambda = 0.50$	0.36302	0.36306	0.43652	0.43398	0.47006	0.46842	0.48548	0.48485	0.49245	0.49267
$\lambda = 1.00$	0.58972	0.60874	0.77690	0.78225	0.88617	0.88574	0.94356	0.94235	0.97213	0.97150
$\lambda = 2.00$	0.83674	0.86186	1.14760	1.20289	1.44352	1.47293	1.65691	1.67208	1.80241	1.80784
$\lambda = 4.00$	0.98075	0.98364	1.31368	1.45601	1.88349	1.90520	2.21964	2.32114	2.64377	2.69544

$\rho = 0.60$										$T = 1.00$
Demand Arrival Rate	Number of Kanbans allocated at each stage									
	$K = 2$		$K = 3$		$K = 4$		$K = 5$		$K = 6$	
	MTR	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**
$\lambda = 0.25$	0.18921	0.18439	0.21822	0.21456	0.23233	0.23026	0.23966	0.23881	0.24318	0.24359
$\lambda = 0.50$	0.34482	0.34243	0.41869	0.41333	0.45563	0.45151	0.47490	0.47251	0.48531	0.48426
$\lambda = 1.00$	0.56401	0.58044	0.74600	0.74816	0.85621	0.85255	0.91875	0.91462	0.95355	0.95070
$\lambda = 2.00$	0.81125	0.84019	1.11995	1.16871	1.40370	1.42957	1.61328	1.62486	1.75994	1.76244
$\lambda = 4.00$	0.97812	0.98137	1.31070	1.44966	1.86718	1.89219	2.20373	2.29903	2.60843	2.66342

$\rho = 0.70$										$T = 1.00$
Demand Arrival Rate	Number of Kanbans allocated at each stage									
	$K = 2$		$K = 3$		$K = 4$		$K = 5$		$K = 6$	
	MTR	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**
$\lambda = 0.25$	0.17944	0.17262	0.20873	0.20258	0.22413	0.21969	0.23300	0.23019	0.23876	0.23689
$\lambda = 0.50$	0.32714	0.32214	0.39965	0.39071	0.43836	0.43034	0.46046	0.45458	0.47401	0.46997
$\lambda = 1.00$	0.53864	0.55194	0.71317	0.71137	0.82164	0.81328	0.88718	0.87784	0.92702	0.91924
$\lambda = 2.00$	0.78253	0.81469	1.08696	1.12883	1.35678	1.37769	1.55900	1.56554	1.70399	1.70141
$\lambda = 4.00$	0.97251	0.97665	1.30465	1.43760	1.83810	1.86904	2.17533	2.26205	2.55507	2.61195

$\rho = 0.80$										$T = 1.00$
Demand Arrival Rate	Number of Kanbans allocated at each stage									
	$K = 2$		$K = 3$		$K = 4$		$K = 5$		$K = 6$	
	MTR	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**
$\lambda = 0.25$	0.17012	0.16134	0.19887	0.18996	0.21473	0.20720	0.22471	0.21860	0.23118	0.22655
$\lambda = 0.50$	0.31034	0.30264	0.38007	0.36720	0.41900	0.40601	0.44260	0.43136	0.45797	0.44889
$\lambda = 1.00$	0.51423	0.52410	0.67959	0.67340	0.78376	0.76972	0.84941	0.83312	0.89193	0.87652
$\lambda = 2.00$	0.75347	0.78736	1.05096	1.08589	1.30529	1.32021	1.49652	1.49672	1.63584	1.62610
$\lambda = 4.00$	0.96318	0.96894	1.29494	1.41947	1.79793	1.83620	2.13491	2.21185	2.48906	2.54359

$\rho = 0.90$										$T = 1.00$
Demand Arrival Rate	Number of Kanbans allocated at each stage									
	$K = 2$		$K = 3$		$K = 4$		$K = 5$		$K = 6$	
	MTR	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**
$\lambda = 0.25$	0.16126	0.15074	0.18903	0.17735	0.20472	0.19375	0.21471	0.20496	0.22174	0.21313
$\lambda = 0.50$	0.29451	0.28423	0.36051	0.34383	0.39818	0.38017	0.42188	0.40464	0.43793	0.42228
$\lambda = 1.00$	0.49102	0.49738	0.64619	0.63554	0.74405	0.72415	0.80701	0.78333	0.84941	0.82514
$\lambda = 2.00$	0.72551	0.75956	1.01365	1.04175	1.25130	1.25965	1.42861	1.42168	1.55815	1.54028
$\lambda = 4.00$	0.95001	0.95823	1.28149	1.39600	1.75096	1.79570	2.08494	2.15174	2.41452	2.46250

Table A.5: Mean throughput rates obtained from both the exact and the approximate solution techniques are given below. The system parameters are in the range: transfer/review period length $T = 2.00$, mean demand arrival rate $\lambda = (0.25, 0.50, 1.00, 2.00, 4.00)$, mean production rate $\mu = \frac{\lambda}{\rho}$, where the demand load $\rho = (0.50, 0.60, 0.70, 0.80, 0.90)$ and number of kanbans $K = (2, 3, 4, 5, 6)$.

$\rho = 0.50$										$T = 2.00$
Demand Arrival Rate	Number of Kanbans allocated at each stage									
	$K = 2$		$K = 3$		$K = 4$		$K = 5$		$K = 6$	
	MTR	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**
$\lambda = 0.25$	0.18147	0.18153	0.21826	0.21699	0.23503	0.23421	0.24274	0.24242	0.24623	0.24633
$\lambda = 0.50$	0.29486	0.30437	0.38843	0.39113	0.44308	0.44287	0.47178	0.47117	0.48606	0.48575
$\lambda = 1.00$	0.41837	0.43093	0.57379	0.60144	0.72176	0.73646	0.82845	0.83604	0.90120	0.90392
$\lambda = 2.00$	0.49037	0.49182	0.65684	0.72800	0.94175	0.95260	1.10981	1.16057	1.32188	1.34772
$\lambda = 4.00$	0.49983	0.49985	0.66647	0.74946	0.99832	0.99859	1.16510	1.24668	1.49142	1.49309

$\rho = 0.60$										$T = 2.00$
Demand Arrival Rate	Number of Kanbans allocated at each stage									
	$K = 2$		$K = 3$		$K = 4$		$K = 5$		$K = 6$	
	MTR	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**
$\lambda = 0.25$	0.17238	0.17122	0.20934	0.20666	0.22782	0.22576	0.23745	0.23626	0.24265	0.24213
$\lambda = 0.50$	0.28199	0.29022	0.37300	0.37408	0.42811	0.42628	0.45938	0.45731	0.47678	0.47535
$\lambda = 1.00$	0.40562	0.42010	0.55996	0.58436	0.70185	0.71478	0.80661	0.81243	0.87997	0.88122
$\lambda = 2.00$	0.48906	0.49069	0.65535	0.72483	0.93359	0.94609	1.10186	1.14952	1.30421	1.33171
$\lambda = 4.00$	0.49983	0.49985	0.66646	0.74945	0.99829	0.99856	1.16506	1.24659	1.49117	1.49288

$\rho = 0.70$										$T = 2.00$
Demand Arrival Rate	Number of Kanbans allocated at each stage									
	$K = 2$		$K = 3$		$K = 4$		$K = 5$		$K = 6$	
	MTR	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**
$\lambda = 0.25$	0.16357	0.16107	0.19982	0.19535	0.21918	0.21517	0.23023	0.22729	0.23689	0.23499
$\lambda = 0.50$	0.26932	0.27597	0.35656	0.35568	0.41082	0.40664	0.44355	0.43892	0.46344	0.45962
$\lambda = 1.00$	0.39127	0.40734	0.54348	0.56442	0.67838	0.68884	0.77948	0.78277	0.85200	0.85070
$\lambda = 2.00$	0.48626	0.48832	0.65233	0.71880	0.91905	0.93452	1.08766	1.13102	1.27754	1.30597
$\lambda = 4.00$	0.49982	0.49983	0.66644	0.74939	0.99811	0.99840	1.16487	1.24617	1.49002	1.49193

$\rho = 0.80$										$T = 2.00$
Demand Arrival Rate	Number of Kanbans allocated at each stage									
	$K = 2$		$K = 3$		$K = 4$		$K = 5$		$K = 6$	
	MTR	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**
$\lambda = 0.25$	0.15514	0.15132	0.19004	0.18360	0.20945	0.20301	0.22122	0.21568	0.22898	0.22445
$\lambda = 0.50$	0.25710	0.26205	0.33978	0.33670	0.39188	0.38486	0.42466	0.41656	0.44596	0.43826
$\lambda = 1.00$	0.37673	0.39368	0.52548	0.54295	0.65265	0.66010	0.74826	0.74836	0.81792	0.81305
$\lambda = 2.00$	0.48159	0.48447	0.64747	0.70973	0.89896	0.91810	1.06744	1.10593	1.24453	1.27180
$\lambda = 4.00$	0.49976	0.49979	0.66638	0.74921	0.99751	0.99790	1.16428	1.24492	1.48670	1.48925

$\rho = 0.90$										$T = 2.00$
Demand Arrival Rate	Number of Kanbans allocated at each stage									
	$K = 2$		$K = 3$		$K = 4$		$K = 5$		$K = 6$	
	MTR	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**
$\lambda = 0.25$	0.14723	0.14212	0.18026	0.17192	0.19904	0.19008	0.21090	0.20232	0.21889	0.21114
$\lambda = 0.50$	0.24551	0.24869	0.32310	0.31777	0.37201	0.36207	0.40350	0.39166	0.42466	0.41257
$\lambda = 1.00$	0.36275	0.37978	0.50682	0.52087	0.62564	0.62982	0.71431	0.71084	0.77908	0.77014
$\lambda = 2.00$	0.47500	0.47911	0.64074	0.69800	0.87548	0.89785	1.04247	1.07587	1.20726	1.23125
$\lambda = 4.00$	0.49963	0.49966	0.66620	0.74875	0.99610	0.99672	1.16290	1.24217	1.47968	1.48369

Table A.6: Mean throughput rates obtained from both the exact and the approximate solution techniques are given below. The system parameters are in the range: transfer/review period length $T = 4.00$, mean demand arrival rate $\lambda = (0.25, 0.50, 1.00, 2.00, 4.00)$, mean production rate $\mu = \frac{\lambda}{\rho}$, where the demand load $\rho = (0.50, 0.60, 0.70, 0.80, 0.90)$ and number of kanbans $K = (2, 3, 4, 5, 6)$.

$\rho = 0.50$										$T = 4.00$
Demand Arrival Rate	Number of Kanbans allocated at each stage									
	$K = 2$		$K = 3$		$K = 4$		$K = 5$		$K = 6$	
	MTR	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**
$\lambda = 0.25$	0.14742	0.15218	0.19421	0.19556	0.22154	0.22144	0.23589	0.23559	0.24303	0.24288
$\lambda = 0.50$	0.20918	0.21546	0.28689	0.30072	0.36086	0.36823	0.41420	0.41802	0.45060	0.45196
$\lambda = 1.00$	0.24519	0.24591	0.32841	0.36400	0.47087	0.47630	0.55491	0.58029	0.66095	0.67386
$\lambda = 2.00$	0.24992	0.24992	0.33323	0.37473	0.49916	0.49929	0.58254	0.62334	0.74571	0.74655
$\lambda = 4.00$	0.25000	0.25000	0.33333	0.37500	0.50000	0.50000	0.58333	0.62500	0.75000	0.75000

$\rho = 0.60$										$T = 4.00$
Demand Arrival Rate	Number of Kanbans allocated at each stage									
	$K = 2$		$K = 3$		$K = 4$		$K = 5$		$K = 6$	
	MTR	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**
$\lambda = 0.25$	0.14100	0.14511	0.18647	0.18704	0.21400	0.21314	0.22960	0.22865	0.23839	0.23767
$\lambda = 0.50$	0.20281	0.21005	0.27998	0.29218	0.35091	0.35739	0.40331	0.40621	0.43995	0.44061
$\lambda = 1.00$	0.24453	0.24534	0.32767	0.36241	0.46680	0.47305	0.55093	0.57476	0.65211	0.66585
$\lambda = 2.00$	0.24991	0.24992	0.33323	0.37472	0.49914	0.49928	0.58252	0.62330	0.74559	0.74644
$\lambda = 4.00$	0.25000	0.25000	0.33333	0.37500	0.50000	0.50000	0.58333	0.62500	0.75000	0.75000

$\rho = 0.70$										$T = 4.00$
Demand Arrival Rate	Number of Kanbans allocated at each stage									
	$K = 2$		$K = 3$		$K = 4$		$K = 5$		$K = 6$	
	MTR	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**
$\lambda = 0.25$	0.13465	0.13799	0.17828	0.17784	0.20538	0.20332	0.22171	0.21946	0.23172	0.22981
$\lambda = 0.50$	0.19562	0.20367	0.27173	0.28221	0.33919	0.34442	0.38974	0.39139	0.42597	0.42535
$\lambda = 1.00$	0.24313	0.24416	0.32616	0.35940	0.45953	0.46726	0.54383	0.56551	0.63877	0.65299
$\lambda = 2.00$	0.24991	0.24992	0.33322	0.37470	0.49905	0.49920	0.58243	0.62309	0.74501	0.74596
$\lambda = 4.00$	0.25000	0.25000	0.33333	0.37500	0.50000	0.50000	0.58333	0.62500	0.75000	0.75000

$\rho = 0.80$										$T = 4.00$
Demand Arrival Rate	Number of Kanbans allocated at each stage									
	$K = 2$		$K = 3$		$K = 4$		$K = 5$		$K = 6$	
	MTR	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**
$\lambda = 0.25$	0.12855	0.13102	0.16986	0.16835	0.19591	0.19243	0.21233	0.20828	0.22294	0.21913
$\lambda = 0.50$	0.18836	0.19684	0.26273	0.27147	0.32631	0.33005	0.37411	0.37418	0.40894	0.40653
$\lambda = 1.00$	0.24079	0.24223	0.32373	0.35487	0.44948	0.45905	0.53372	0.55296	0.62225	0.63590
$\lambda = 2.00$	0.24988	0.24989	0.33318	0.37460	0.49876	0.49895	0.58213	0.62246	0.74335	0.74463
$\lambda = 4.00$	0.25000	0.25000	0.33333	0.37500	0.50000	0.50000	0.58333	0.62500	0.75000	0.75000

$\rho = 0.90$										$T = 4.00$
Demand Arrival Rate	Number of Kanbans allocated at each stage									
	$K = 2$		$K = 3$		$K = 4$		$K = 5$		$K = 6$	
	MTR	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**	**MTR**	**AMTR**
$\lambda = 0.25$	0.12275	0.12435	0.16153	0.15888	0.18599	0.18104	0.20173	0.19583	0.21233	0.20628
$\lambda = 0.50$	0.18138	0.18989	0.25341	0.26044	0.31282	0.31491	0.35714	0.35542	0.38953	0.38507
$\lambda = 1.00$	0.23750	0.23956	0.32037	0.34900	0.43774	0.44893	0.52123	0.53794	0.60362	0.61562
$\lambda = 2.00$	0.24981	0.24983	0.33310	0.37437	0.49805	0.49836	0.58145	0.62108	0.73984	0.74185
$\lambda = 4.00$	0.25000	0.25000	0.33333	0.37500	0.50000	0.50000	0.58333	0.62500	0.74999	0.74999

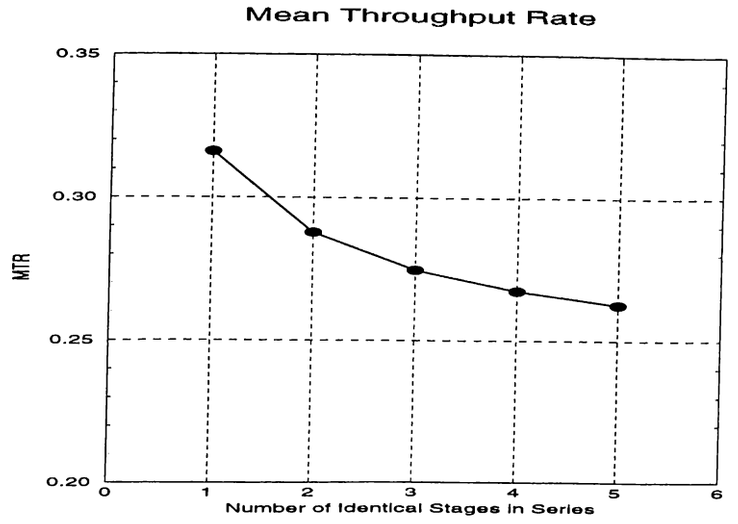


Figure A.1: Increasing the number of identical stages in series, decreases the mean throughput rate of the system with a decreasing rate.

(Fixed parameters of the system: mean demand arrival rate $\lambda = 1.0$; transfer/review period length $T = 1.0$; mean production rate at each stage $\mu = 1.0$; number of kanbans at each stage $K = 1$; maximum level for allowed backorders $B_{FP} = 0$.)

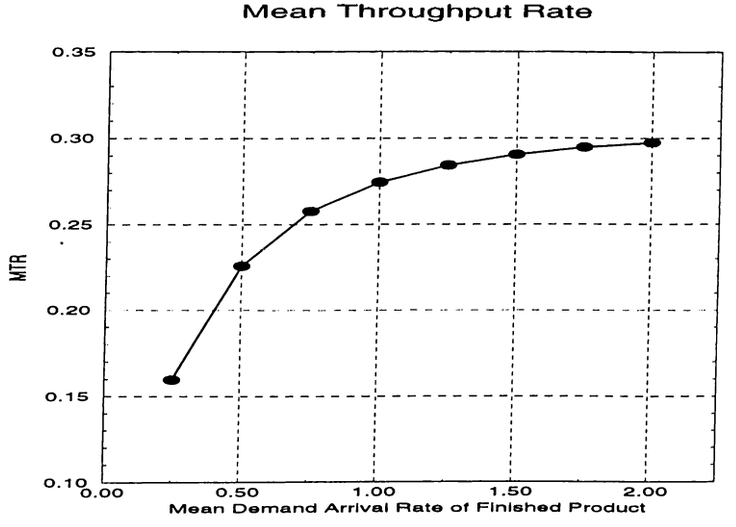


Figure A.2: Increasing mean demand arrival rate of finished product, increases the mean throughput rate of the system with a decreasing rate.

(Fixed parameters of the system: number of stages in the system $N = 3$; transfer/review period length $T = 1.0$; mean production rate at each stage $\mu = 1.0$; number of kanbans at each stage $K = 1$; maximum level for allowed backorders $B_{FP} = 0$.)

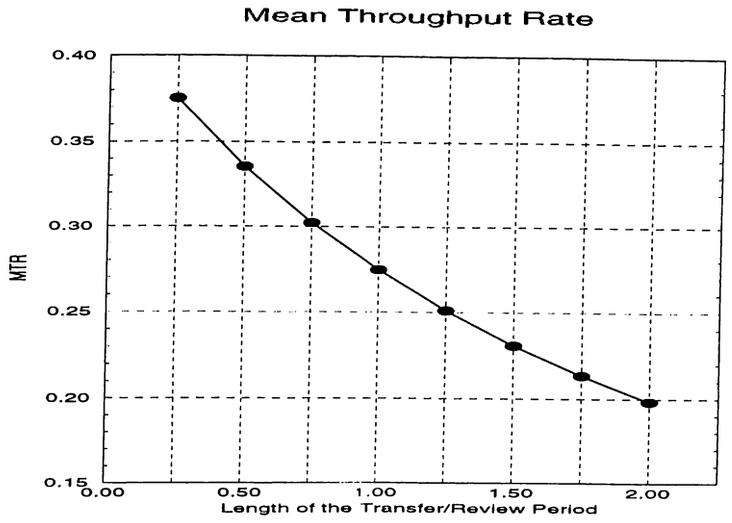


Figure A.3: Increasing the length of the transfer/review period, decreases the mean throughput rate of the system with a decreasing rate.

(Fixed parameters of the system: number of stages in the system $N = 3$; mean demand arrival rate $\lambda = 1.0$; mean production rate at each stage $\mu = 1.0$; number of kanbans at each stage $K = 1$; maximum level for allowed backorders $B_{FP} = 0$.)

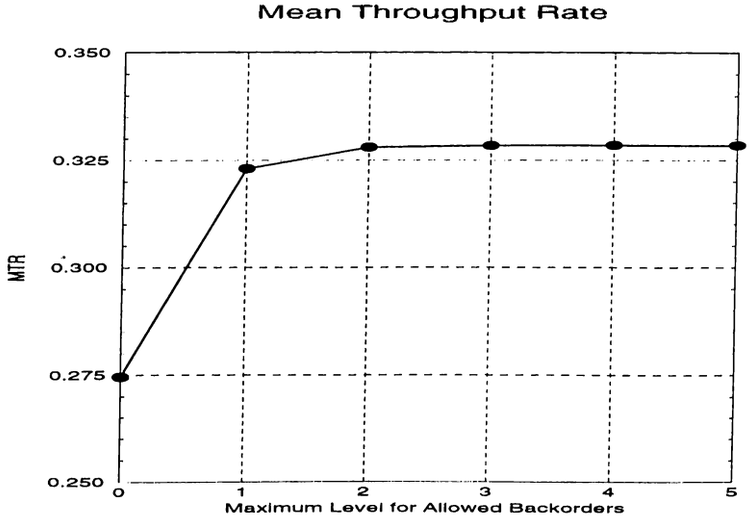


Figure A.4: Increasing the maximum level for allowed backorders, asymptotically increases the mean throughput rate of the system.

(Fixed parameters of the system: number of stages in the system $N = 3$; mean demand arrival rate $\lambda = 1.0$; transfer/review period length $T = 1.0$; mean production rate at each stage $\mu = 1.0$; number of kanbans at each stage $K = 1$.)

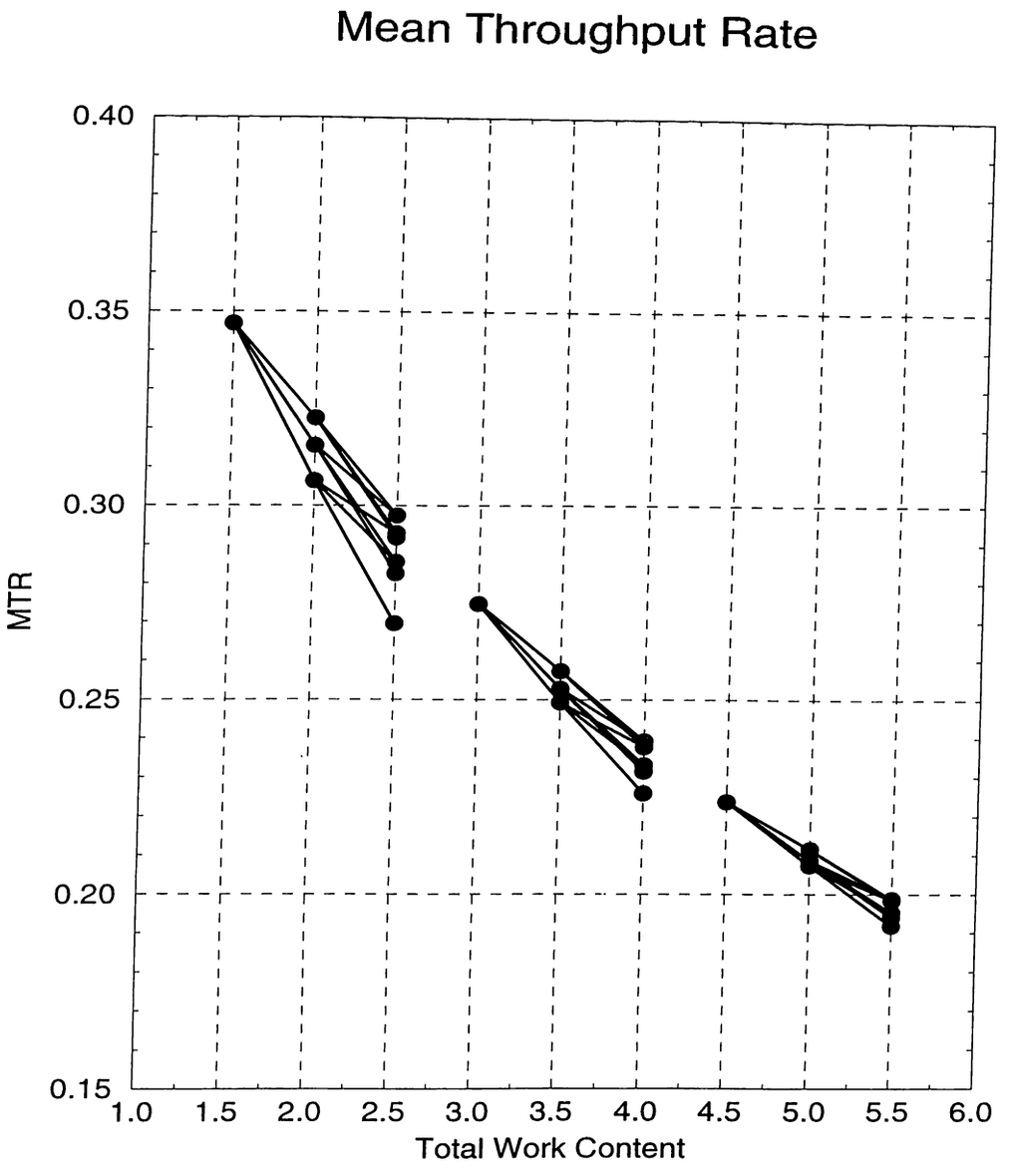


Figure A.5: Increasing total work content, decreases the mean throughput rate of the system with a decreasing rate.

(Fixed parameters of the system: number of stages in the system $N = 3$; mean demand arrival rate $\lambda = 1.0$; transfer/review period length $T = 1.0$; number of kanbans at each stage $K = 1$; maximum level for allowed backorders $B_{FP} = 0$.)

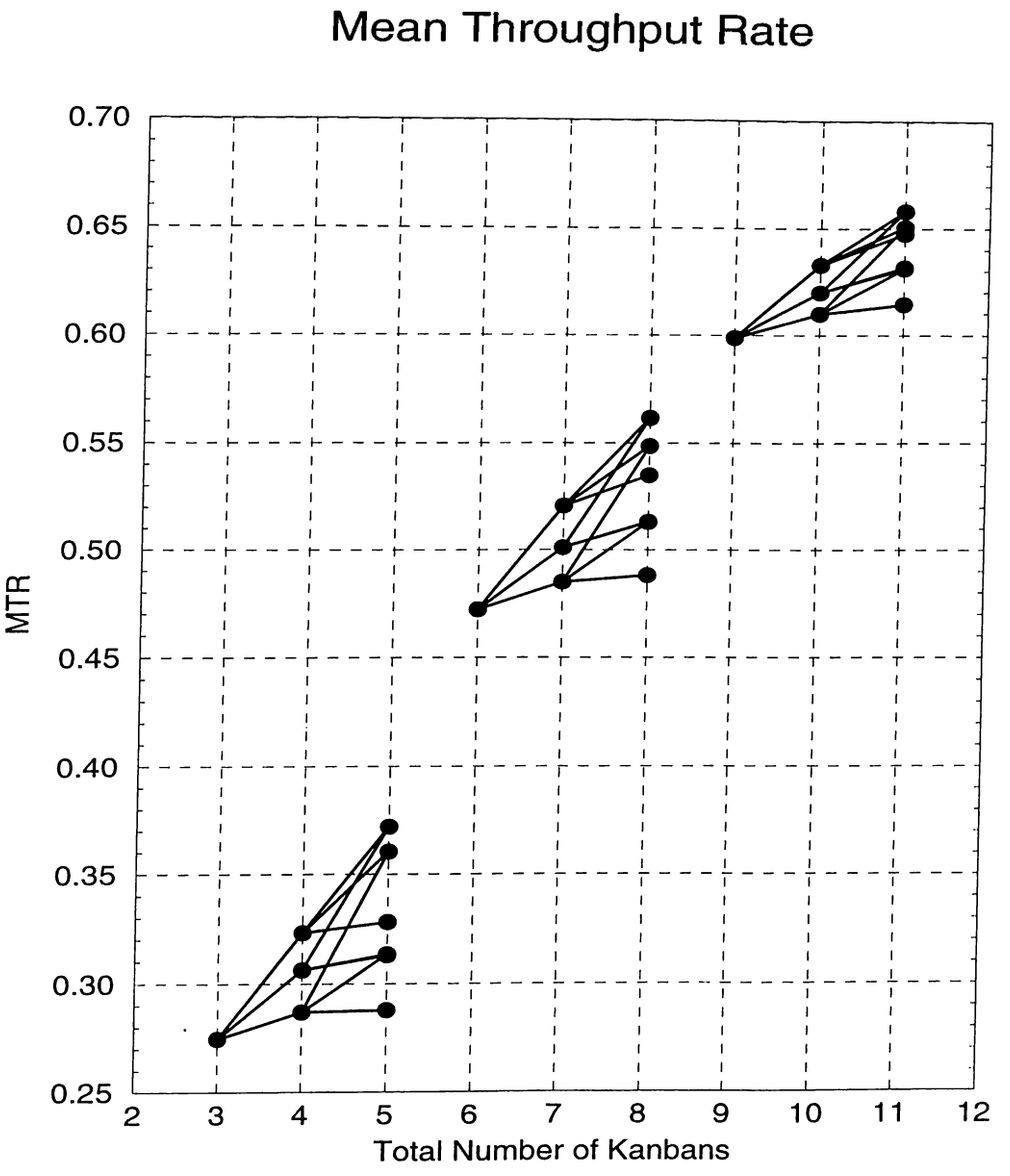


Figure A.6: Increasing total number of kanbans, increases the mean throughput rate of the system with a decreasing rate.

(Fixed parameters of the system: number of stages in the system $N = 3$; mean demand arrival rate $\lambda = 1.0$; transfer/review period length $T = 1.0$; mean production rate at each stage $\mu = 1.0$; maximum level for allowed backorders $B_{FP} = 0.$)

Table A.7: Experimental framework designed for investigating the workload and kanban allocation problem in two-stage pull systems. The first term in each cell denotes the number of possible kanban allocations and the second term denotes the number of possible workload allocations with degree of imbalance is less than or equal to 5 (that means both $DI_w \leq 5$ and $DI_k \leq 5$).

EXPERIMENTAL FRAMEWORK						
DEMAND ARRIVAL RATE $\lambda = 1.0$						
	CONTINUOUS approximated by $T = 0.0001$			PERIODIC with $T = 1.0$		
	TWC			**TWC**		
TNK	1.00	1.50	2.00	1.00	1.50	2.00
2	1x5	1x5	1x5	1x5	1x5	1x5
3	2x5	2x5	2x5	2x5	2x5	2x5
4	3x5	3x5	3x5	3x5	3x5	3x5
5	4x5	4x5	4x5	4x5	4x5	4x5
6	5x5	5x5	5x5	5x5	5x5	5x5
7	6x5	6x5	6x5	6x5	6x5	6x5
8	5x5	5x5	5x5	5x5	5x5	5x5
9	6x5	6x5	6x5	6x5	6x5	6x5
TOTAL: 960 **MTR** evaluations						

Table A.8: The independent factors determining the mean throughput rate of a two–stage pull system.

INDEPENDENT FACTORS

λ – Demand Arrival Rate
Level(s):
1.0000

T – Transfer/Review Period Length
Level(s):
0.0001
1.0000

TWC – Total Work Content	
TWC = $1/\mu_1 + 1/\mu_2$	
Level(s):	Workload Allocation Variables: $1/\mu_1$, $1/\mu_2$
1.0000	0.4000, 0.4500, 0.5000, 0.5500, 0.6000
1.5000	0.6000, 0.6750, 0.7500, 0.8250, 0.9000
2.0000	0.8000, 0.9000, 1.0000, 1.1000, 1.2000

TNK – Total Number of Kanbans	
TNK = $K_1 + K_2$	
Level(s):	Kanban Allocation Variables: K_1, K_2
2	1
3	1,2
4	1,2,3
5	1,2,3,4
6	1,2,3,4,5
7	1,2,3,4,5,6
8	2,3,4,5,6
9	2,3,4,5,6,7

Table A.9: Correlation analysis of the factors affecting the mean throughput rate of a two–stage system.

CORRELATION ANALYSIS		
	Dependent Factor: **MTR**	
Factors	CONTINUOUS approximated by $T = 0.0001$	PERIODIC with $T = 1.0$
TWC	-0.6491	-0.3037
$1/\mu_1$	-0.5202	-0.2540
$1/\mu_2$	-0.6229	-0.2808
TNK	0.5295	0.6933
K_1	-0.0062	0.2188
K_2	0.6157	0.5793

Table A.10: The summary of factorial regression models between the independent factors and the mean throughput rate of a two–stage pull system.

	FACTORIAL REGRESSION MODELS					
	CONTINUOUS approximated by $T = 0.0001$			PERIODIC with $T = 1.0$		
	Linear	Quadratic	**MTR**	**MTR**	Quadratic	Linear
Mean	0.7616	0.7616	0.7616	0.5831	0.5831	0.5831
St. deviation	0.1322	0.1408	0.1427	0.1813	0.1742	0.1430
Variance	0.0175	0.0198	0.0204	0.0329	0.0304	0.0205
CV	17.3547	18.4893	18.7359	31.0944	29.8799	24.5277
Skewness	0.0290	-0.1113	-0.2307	0.0552	0.3057	-0.2017
Kurtosis	-0.5507	-0.6135	-0.8185	-1.1406	-0.7358	-0.5792
Minimum	0.4500	0.4107	0.4269	0.2830	0.2452	0.2232
Maximum	1.0842	1.0419	0.9907	0.9411	1.0049	0.8900
Corl. coefficient	0.9263	0.9868	1.0000	1.0000	0.9609	0.7888
R-square	0.8580	0.9739	1.0000	1.0000	0.9234	0.6222
SS (error)	1.3850	0.2550	0.0000	0.0000	1.2059	5.9478
MS (error)	0.0029	0.0005	0.0000	0.0000	0.0026	0.0125
F-Value	717.5100	1237.2100	∞	∞	400.4200	195.5900
DF	4	14	480	480	14	4

Table A.11: The estimated values of the parameters used in factorial regression models for the mean throughput rate of a two-stage pull system.

	FACTORIAL REGRESSION MODELS			
	COEFFICIENT ESTIMATES			
	CONTINUOUS approximated by $T = 0.0001$		PERIODIC with $T = 1.0$	
Terms	Linear	Quadratic	Linear	Quadratic
a_0	0.8546672130	0.6718958945	0.3743663323	0.1192727596
a_1 * $1/\mu_1$	-0.1564861853	0.0020344691	-0.1114317500	0.0034219975
a_2 * $1/\mu_2$	-0.2968350647	-0.3237805074	-0.1580442500	0.0267544172
a_3 * K_1	0.0189694592	0.0109085469	0.0495299637	0.0862142283
a_4 * K_2	0.0577509472	0.1725336646	0.0780968651	0.1508285408
$a_{1,1}$ * $1/\mu_1 * 1/\mu_1$		-0.1618412765		-0.0578698268
$a_{1,2}$ * $1/\mu_1 * 1/\mu_2$		0.1383963978		0.0850718074
$a_{1,3}$ * $1/\mu_1 * K_1$		0.0285080909		0.0141196606
$a_{1,4}$ * $1/\mu_1 * K_2$		-0.0292585430		-0.0414995780
$a_{2,2}$ * $1/\mu_2 * 1/\mu_2$		-0.0464140861		-0.0498246548
$a_{2,3}$ * $1/\mu_2 * K_1$		-0.0077537412		-0.0325046974
$a_{2,4}$ * $1/\mu_2 * K_2$		0.0034130208		-0.0208720188
$a_{3,3}$ * $K_1 * K_1$		-0.0023813099		-0.0128808434
$a_{3,4}$ * $K_1 * K_2$		0.0024579544		0.0252641979
$a_{4,4}$ * $K_2 * K_2$		-0.0144679059		-0.0133012905

Table A.12: Optimal workload and kanban allocation results obtained through enumerating the allocation vectors around balanced allocation of two–stage pull systems in order to maximize the mean throughput rate (**MTR**).

(Fixed parameters of the two–stage system: mean demand arrival rate (λ) = 1.0; transfer/review period length T = 0.0001).

TWO–STAGE CONTINUOUS PULL SYSTEMS (Approximated by T = 0.0001)								
OPTIMAL UNBALANCED ALLOCATION							•	BALANCED
TWC	$1/\mu_1$	$1/\mu_2$	TNK	K_1	K_2	**MTR**		**MTR**
2.0000	1.2000	0.8000	2	1	1	0.4560		0.4444
2.0000	1.1000	0.9000	3	1	2	0.5661		0.5641
2.0000	1.2000	0.8000	4	2	2	0.6209		0.6133
2.0000	1.1000	0.9000	5	2	3	0.6724		0.6705
2.0000	1.1000	0.9000	6	3	3	0.7054		0.7003
2.0000	1.1000	0.9000	7	3	4	0.7348		0.7340
2.0000	1.1000	0.9000	8	4	4	0.7573		0.7543
2.0000	1.0000	1.0000	9	4	5	0.7767		0.7767
1.5000	0.9000	0.6000	2	1	1	0.5388		0.5233
1.5000	0.9000	0.6000	3	1	2	0.6781		0.6741
1.5000	0.8250	0.6750	4	1	3	0.7451		0.7204
1.5000	0.8250	0.6750	5	2	3	0.8004		0.7953
1.5000	0.8250	0.6750	6	2	4	0.8402		0.8190
1.5000	0.8250	0.6750	7	3	4	0.8690		0.8635
1.5000	0.8250	0.6750	8	3	5	0.8946		0.8771
1.5000	0.8250	0.6750	9	3	6	0.9126		0.9060
1.0000	0.6000	0.4000	2	1	1	0.6514		0.6315
1.0000	0.6000	0.4000	3	1	2	0.8158		0.8078
1.0000	0.5500	0.4500	4	1	3	0.8887		0.8410
1.0000	0.5500	0.4500	5	1	4	0.9294		0.9160
1.0000	0.6000	0.4000	6	2	4	0.9566		0.9266
1.0000	0.5500	0.4500	7	2	5	0.9744		0.9616
1.0000	0.5500	0.4500	8	2	6	0.9847		0.9651
1.0000	0.5500	0.4500	9	2	7	0.9907		0.9820
					Average:	0.7817		0.7695

Table A.13: Optimal workload and kanban allocation results obtained through enumerating the allocation vectors around balanced allocation of two–stage pull systems in order to maximize the mean throughput rate (**MTR**).

(Fixed parameters of the two–stage system: mean demand arrival rate (λ) = 1.0; transfer/review period length T = 1.0).

TWO–STAGE PERIODIC PULL SYSTEMS (Periodic with T = 1.0)								
OPTIMAL UNBALANCED ALLOCATION						•	BALANCED	
TWC	$1/\mu_1$	$1/\mu_2$	TNK	K_1	K_2	**MTR**		**MTR**
2.0000	1.2000	0.8000	2	1	1	0.2894		0.2878
2.0000	0.8000	1.2000	3	1	2	0.3757		0.3646
2.0000	1.1000	0.9000	4	2	2	0.4922		0.4892
2.0000	1.0000	1.0000	5	2	3	0.5616		0.5616
2.0000	1.1000	0.9000	6	3	3	0.6219		0.6179
2.0000	1.0000	1.0000	7	3	4	0.6673		0.6673
2.0000	1.1000	0.9000	8	4	4	0.7027		0.6990
2.0000	1.0000	1.0000	9	4	5	0.7316		0.7316
1.5000	0.9000	0.6000	2	1	1	0.3205		0.3186
1.5000	0.6000	0.9000	3	1	2	0.4150		0.4038
1.5000	0.9000	0.6000	4	2	2	0.5549		0.5501
1.5000	0.6750	0.8250	5	2	3	0.6434		0.6433
1.5000	0.9000	0.6000	6	3	3	0.7104		0.7033
1.5000	0.8250	0.6750	7	3	4	0.7732		0.7724
1.5000	0.7500	0.7500	8	3	5	0.8115		0.8011
1.5000	0.8250	0.6750	9	4	5	0.8512		0.8479
1.0000	0.6000	0.4000	2	1	1	0.3548		0.3533
1.0000	0.4000	0.6000	3	1	2	0.4528		0.4443
1.0000	0.6000	0.4000	4	2	2	0.6200		0.6150
1.0000	0.4500	0.5500	5	2	3	0.7270		0.7266
1.0000	0.6000	0.4000	6	3	3	0.7947		0.7859
1.0000	0.5500	0.4500	7	3	4	0.8697		0.8678
1.0000	0.5000	0.5000	8	3	5	0.9132		0.8874
1.0000	0.6000	0.4000	9	4	5	0.9411		0.9363
					Average:	0.6332		0.6282

Table A.14: Experimental framework designed for investigating the workload and kanban allocation problem in three–stage pull systems. The first term in each cell denotes the number of possible kanban allocations and the second term denotes the number of possible workload allocations with degree of imbalance is less than or equal to 5 (that means both $DI_w \leq 5$ and $DI_k \leq 5$).

	EXPERIMENTAL FRAMEWORK					
	DEMAND ARRIVAL RATE					
	$\lambda = 1.0$					
	CONTINUOUS approximated by $T = 0.0001$			PERIODIC with $T = 1.0$		
	TWC			**TWC**		
TNK	1.50	2.25	3.00	1.50	2.25	3.00
3	1x31	1x31	1x31	1x31	1x31	1x31
4	3x31	3x31	3x31	3x31	3x31	3x31
5	6x31	6x31	6x31	6x31	6x31	6x31
12	31x31	31x31	31x31	31x31	31x31	31x31
13	30x31	30x31	30x31	30x31	30x31	30x31
14	30x31	30x31	30x31	30x31	30x31	30x31
	TOTAL: 18786 **MTR** evaluations					

Table A.15: The independent factors determining the mean throughput rate of a three–stage pull system.

INDEPENDENT FACTORS

λ – Demand Arrival Rate
Level(s):
1.0000

T – Transfer/Review Period Length
Level(s):
0.0001
1.0000

TWC – Total Work Content	
TWC = $1/\mu_1 + 1/\mu_2 + 1/\mu_3$	
Level(s):	Workload Allocation Variables: $1/\mu_1$, $1/\mu_2$, $1/\mu_3$
1.5000	0.3500, 0.4000, 0.4500, 0.5000, 0.5500, 0.6000, 0.6500
2.2500	0.5250, 0.6000, 0.6750, 0.7500, 0.8250, 0.9000, 0.9750
3.0000	0.7000, 0.8000, 0.9000, 1.0000, 1.1000, 1.2000, 1.3000

TNK – Total Number of Kanbans	
TNK = $K_1 + K_2 + K_3$	
Level(s):	Kanban Allocation Variables: K_1, K_2, K_3
3	1
4	1,2
5	1,2,3
12	1,2,3,4,5,6,7
13	1,2,3,4,5,6,7
14	2,3,4,5,6,7,8

Table A.16: Correlation analysis of the factors affecting the mean throughput rate of a three–stage system.

CORRELATION ANALYSIS		
	Dependent Factor: **MTR**	
Factors	CONTINUOUS approximated by $T = 0.0001$	PERIODIC with $T = 1.0$
TWC	-0.7416	-0.4279
$1/\mu_1$	-0.5751	-0.3396
$1/\mu_2$	-0.6205	-0.3627
$1/\mu_3$	-0.6769	-0.3782
TNK	0.5006	0.6278
K_1	0.0964	0.1770
K_2	0.1452	0.2272
K_3	0.4872	0.5104

Table A.17: The summary of factorial regression models between the independent factors and the mean throughput rate of a three–stage pull system.

	FACTORIAL REGRESSION MODELS					
	CONTINUOUS approximated by $T = 0.0001$			PERIODIC with $T = 1.0$		
	Linear	Quadratic	**MTR**	**MTR**	Quadratic	Linear
Mean	0.7959	0.7959	0.7959	0.6551	0.6551	0.6551
St. deviation	0.1317	0.1374	0.1394	0.1732	0.1592	0.1381
Variance	0.0173	0.0189	0.0194	0.0300	0.0253	0.0191
CV	16.5524	17.2764	17.5282	26.4444	24.3055	21.0772
Skewness	-0.3210	-0.3173	-0.3959	-0.4129	-0.3417	-1.0701
Kurtosis	-0.3495	-0.4161	-0.5999	-0.6379	-0.4214	1.2460
Minimum	0.3946	0.3657	0.3968	0.2666	0.2565	0.1740
Maximum	1.0692	1.0342	0.9965	0.9638	0.9912	0.9148
Corl. coefficient	0.9443	0.9856	1.0000	1.0000	0.9191	0.7970
R-square	0.8918	0.9715	1.0000	1.0000	0.8448	0.6353
SS (error)	19.7617	5.2076	0.0000	0.0000	43.7497	102.7979
MS (error)	0.0021	0.0006	0.0000	0.0000	0.0047	0.0110
F-Value	12887.5800	11813.0200	∞	∞	1887.6800	2724.7300
DF	6	27	9393	9393	27	6

Table A.18: The estimated values of the parameters used in factorial regression models for the mean throughput rate of a three–stage pull system.

	FACTORIAL REGRESSION MODELS			
	COEFFICIENT ESTIMATES			
	CONTINUOUS approximated by $T = 0.0001$		PERIODIC with $T = 1.0$	
Terms	Linear	Quadratic	Linear	Quadratic
a_0	0.8565484076	0.5827241186	0.4308403946	0.1274007674
$a_1 * 1/\mu_1$	-0.1043070657	0.1110643034	-0.0874848155	0.0277271038
$a_2 * 1/\mu_2$	-0.1639678300	0.0765738780	-0.1252007223	0.0456764512
$a_3 * 1/\mu_3$	-0.2381940216	-0.2680796642	-0.1504406583	-0.0130468644
$a_4 * K_1$	0.0164634341	0.0006810667	0.0302945355	0.0400000461
$a_5 * K_2$	0.0197128229	0.0225298833	0.0344482260	0.0558162019
$a_6 * K_3$	0.0425290163	0.1534690116	0.0578864409	0.1195401918
$a_{1,1} * 1/\mu_1 * 1/\mu_1$		-0.1827325280		-0.0814183720
$a_{1,2} * 1/\mu_1 * 1/\mu_2$		-0.0216172871		0.0286002104
$a_{1,3} * 1/\mu_1 * 1/\mu_3$		0.1573230013		0.0897991465
$a_{1,4} * 1/\mu_1 * K_1$		0.0198780323		0.0229492137
$a_{1,5} * 1/\mu_1 * K_2$		-0.0052634765		-0.0130030766
$a_{1,6} * 1/\mu_1 * K_3$		-0.0208407372		-0.0284214115
$a_{2,2} * 1/\mu_2 * 1/\mu_2$		-0.1867305971		-0.0905698310
$a_{2,3} * 1/\mu_2 * 1/\mu_3$		0.1473482610		0.0961356227
$a_{2,4} * 1/\mu_2 * K_1$		0.0054222539		-0.0062979680
$a_{2,5} * 1/\mu_2 * K_2$		0.0125778621		0.0116310474
$a_{2,6} * 1/\mu_2 * K_3$		-0.0266545259		-0.0350325691
$a_{3,3} * 1/\mu_3 * 1/\mu_3$		-0.1318705263		-0.0808204875
$a_{3,4} * 1/\mu_3 * K_1$		-0.0174784408		-0.0240078790
$a_{3,5} * 1/\mu_3 * K_2$		-0.0005761688		-0.0122865211
$a_{3,6} * 1/\mu_3 * K_3$		0.0153442854		-0.0021439029
$a_{4,4} * K_1 * K_1$		-0.0007359557		-0.0095221394
$a_{4,5} * K_1 * K_2$		0.0014621663		0.0089114461
$a_{4,6} * K_1 * K_3$		0.0008061401		0.0122194491
$a_{5,5} * K_2 * K_2$		-0.0020666380		-0.0098969007
$a_{5,6} * K_2 * K_3$		-0.0006845384		0.0113491603
$a_{6,6} * K_3 * K_3$		-0.0109469800		-0.0117021530

Table A.19: Optimal workload and kanban allocation results obtained through enumerating the allocation vectors around balanced allocation of three-stage pull systems in order to maximize the mean throughput rate (**MTR**).

(Fixed parameters of the three-stage system: mean demand arrival rate (λ) = 1.0; transfer/review period length T = 0.0001).

THREE-STAGE CONTINUOUS PULL SYSTEMS (Approximated by $T = 0.0001$)									
OPTIMAL UNBALANCED ALLOCATION								•	BALANCED
TWC	$1/\mu_1$	$1/\mu_2$	$1/\mu_3$	TNK	K_1	K_2	K_3	**MTR**		**MTR**
3.0000	1.2000	1.1000	0.7000	3	1	1	1	0.4395		0.4236
3.0000	1.2000	0.9000	0.9000	4	1	1	2	0.5171		0.5146
3.0000	1.1000	0.9000	1.0000	5	1	2	2	0.5593		0.5587
3.0000	1.1000	1.0000	0.9000	12	4	4	4	0.7345		0.7318
3.0000	1.1000	0.9000	1.0000	13	4	4	5	0.7486		0.7477
3.0000	1.0000	1.0000	1.0000	14	4	5	5	0.7605		0.7605
2.2500	0.9000	0.8250	0.5250	3	1	1	1	0.5302		0.5064
2.2500	0.9000	0.6750	0.6750	4	1	1	2	0.6380		0.6324
2.2500	0.8250	0.6750	0.7500	5	1	1	3	0.6870		0.6810
2.2500	0.8250	0.6750	0.7500	12	3	3	6	0.8908		0.8690
2.2500	0.8250	0.7500	0.6750	13	4	3	6	0.9046		0.8954
2.2500	0.8250	0.7500	0.6750	14	4	4	6	0.9179		0.9047
1.5000	0.6000	0.5500	0.3500	3	1	1	1	0.6542		0.6207
1.5000	0.6500	0.4500	0.4000	4	1	1	2	0.7960		0.7837
1.5000	0.6000	0.4500	0.4500	5	1	1	3	0.8609		0.8246
1.5000	0.5500	0.5000	0.4500	12	2	3	7	0.9914		0.9646
1.5000	0.6000	0.5000	0.4000	13	3	3	7	0.9942		0.9815
1.5000	0.5500	0.5000	0.4500	14	3	3	8	0.9965		0.9829
							Average:	0.7567		0.7435

Table A.20: Optimal workload and kanban allocation results obtained through enumerating the allocation vectors around balanced allocation of three-stage pull systems in order to maximize the mean throughput rate (**MTR**).

(Fixed parameters of the three-stage system: mean demand arrival rate (λ) = 1.0; transfer/review period length T = 1.0).

THREE-STAGE PERIODIC PULL SYSTEMS (Periodic with $T = 1.0$)									
OPTIMAL UNBALANCED ALLOCATION								•	BALANCED
TWC	$1/\mu_1$	$1/\mu_2$	$1/\mu_3$	TNK	K_1	K_2	K_3	**MTR**		**MTR**
3.0000	1.3000	0.9000	0.8000	3	1	1	1	0.2769		0.2746
3.0000	0.9000	0.8000	1.3000	4	1	1	2	0.3353		0.3231
3.0000	0.7000	1.1000	1.2000	5	1	2	2	0.3936		0.3718
3.0000	1.1000	1.0000	0.9000	12	4	4	4	0.6828		0.6787
3.0000	1.1000	0.9000	1.0000	13	4	4	5	0.7025		0.7014
3.0000	1.0000	1.0000	1.0000	14	4	5	5	0.7185		0.7185
2.2500	0.9750	0.6750	0.6000	3	1	1	1	0.3111		0.3081
2.2500	0.6750	0.6000	0.9750	4	1	1	2	0.3812		0.3683
2.2500	0.5250	0.8250	0.9000	5	1	2	2	0.4378		0.4149
2.2500	0.7500	0.7500	0.7500	12	3	4	5	0.8096		0.7919
2.2500	0.8250	0.7500	0.6750	13	4	4	5	0.8374		0.8323
2.2500	0.8250	0.7500	0.6750	14	4	4	6	0.8586		0.8485
1.5000	0.6500	0.4500	0.4000	3	1	1	1	0.3498		0.3469
1.5000	0.4500	0.4000	0.6500	4	1	1	2	0.4331		0.4224
1.5000	0.3500	0.5500	0.6000	5	1	2	2	0.4754		0.4584
1.5000	0.5500	0.4500	0.5000	12	3	3	6	0.9260		0.8857
1.5000	0.5000	0.5000	0.5000	13	3	4	6	0.9477		0.9332
1.5000	0.6000	0.5000	0.4000	14	4	4	6	0.9638		0.9409
							Average:	0.6023		0.5900

Table A.21: Experimental framework designed for investigating the workload and kanban allocation problem in four–stage pull systems. The first term in each cell denotes the number of possible kanban allocations and the second term denotes the number of possible workload allocations with degree of imbalance is less than or equal to 4 (that means both $DI_w \leq 4$ and $DI_k \leq 4$).

	EXPERIMENTAL FRAMEWORK			
	DEMAND ARRIVAL RATE $\lambda = 1.0$			
	CONTINUOUS approximated by $T = 0.0001$		PERIODIC with $T = 1.0$	
	TWC		**TWC**	
TNK	2.00	4.00	2.00	4.00
4	1x93	1x93	1x93	1x93
5	4x93	4x93	4x93	4x93
6	10x93	10x93	10x93	10x93
7	20x93	20x93	20x93	20x93
8	35x93	35x93	35x93	35x93
	TOTAL: 26040 **MTR** evaluations			

Table A.22: The independent factors determining the mean throughput rate of a four–stage pull system.

INDEPENDENT FACTORS

λ – Demand Arrival Rate
Level(s):
1.0000

T – Transfer/Review Period Length
Level(s):
0.0001
1.0000

TWC – Total Work Content	
TWC = $1/\mu_1 + 1/\mu_2 + 1/\mu_3 + 1/\mu_4$	
Level(s):	Workload Allocation Variables: $1/\mu_1$, $1/\mu_2$, $1/\mu_3$, $1/\mu_4$
2.0000	0.3500, 0.4000, 0.4500, 0.5000, 0.5500, 0.6000, 0.6500
4.0000	0.7000, 0.8000, 0.9000, 1.0000, 1.1000, 1.2000, 1.3000

TNK – Total Number of Kanbans	
TNK = $K_1 + K_2 + K_3 + K_4$	
Level(s):	Kanban Allocation Variables: K_1, K_2, K_3, K_4
4	1
5	1,2
6	1,2,3
7	1,2,3,4
8	1,2,3,4,5

Table A.23: Correlation analysis of the factors affecting the mean throughput rate of a four–stage system.

CORRELATION ANALYSIS		
	Dependent Factor: **MTR**	
Factors	CONTINUOUS approximated by $T = 0.0001$	PERIODIC with $T = 1.0$
TWC	-0.8404	-0.6843
$1/\mu_1$	-0.7466	-0.6169
$1/\mu_2$	-0.7589	-0.6249
$1/\mu_3$	-0.7712	-0.6292
$1/\mu_4$	-0.8003	-0.6344
TNK	0.1802	0.2539
K_1	-0.1377	-0.1271
K_2	-0.0910	-0.0359
K_3	-0.0082	0.0194
K_4	0.4170	0.3975

Table A.24: The summary of factorial regression models between the independent factors and the mean throughput rate of a four–stage pull system.

	FACTORIAL REGRESSION MODELS					
	CONTINUOUS approximated by $T = 0.0001$			PERIODIC with $T = 1.0$		
	Linear	Quadratic	**MTR**	**MTR**	Quadratic	Linear
Mean	0.6175	0.6175	0.6175	0.3613	0.3613	0.3613
St. deviation	0.1397	0.1466	0.1472	0.0586	0.0552	0.0476
Variance	0.0195	0.0215	0.0217	0.0034	0.0030	0.0023
CV	22.6157	23.7377	23.8321	16.2121	15.2747	13.1834
Skewness	0.0768	0.4705	0.4896	0.7632	0.2761	0.0983
Kurtosis	-1.2138	-0.9269	-1.0123	0.7648	-0.6770	-0.9921
Minimum	0.3743	0.3638	0.3854	0.2584	0.2124	0.2683
Maximum	0.9580	0.9389	0.9088	0.6167	0.4854	0.4796
Corl. coefficient	0.9490	0.9960	1.0000	1.0000	0.9422	0.8132
R-square	0.9905	0.9921	1.0000	1.0000	0.8877	0.6613
SS (error)	28.0465	2.2288	0.0000	0.0000	5.0157	15.1293
MS (error)	0.0022	0.0002	0.0000	0.0000	0.0004	0.0012
F-Value	14723.5900	37008.6700	∞	∞	2330.9800	3174.8600
DF	8	44	13020	13020	44	8

Table A.25: The estimated values of the parameters used in factorial regression models for the mean throughput rate of a four–stage pull system.

	FACTORIAL REGRESSION MODELS			
	COEFFICIENT ESTIMATES			
	CONTINUOUS approximated by $T = 0.0001$		PERIODIC with $T = 1.0$	
Terms	Linear	Quadratic	Linear	Quadratic
a_0	0.7936631797	0.5192887308	0.3722276302	0.2755627673
$a_1 * 1/\mu_1$	-0.0671833238	0.1085928773	-0.0307020549	0.0151783863
$a_2 * 1/\mu_2$	-0.0978783119	0.0813317285	-0.0386232073	0.0025854375
$a_3 * 1/\mu_3$	-0.1285528357	-0.0276325096	-0.0428826597	-0.0148462827
$a_4 * 1/\mu_4$	-0.2010641000	-0.3908594062	-0.0481039644	-0.0303838768
$a_5 * K_1$	0.0051089087	0.0003898498	0.0060658897	-0.0051042942
$a_6 * K_2$	0.0107164913	0.0109853726	0.0104274168	0.0048119481
$a_7 * K_3$	0.0206690233	0.0480746366	0.0130687698	0.0208477458
$a_8 * K_4$	0.0717529858	0.2414146205	0.0311533289	0.0835979706
$a_{1,1} * 1/\mu_1 * 1/\mu_1$		-0.0745845444		-0.0185870126
$a_{1,2} * 1/\mu_1 * 1/\mu_2$		-0.0666015448		-0.0068323095
$a_{1,3} * 1/\mu_1 * 1/\mu_3$		-0.0114274720		0.0031145482
$a_{1,4} * 1/\mu_1 * 1/\mu_4$		0.0633173028		-0.0014105381
$a_{1,5} * 1/\mu_1 * K_1$		0.0175536817		0.0149456536
$a_{1,6} * 1/\mu_1 * K_2$		0.0065718007		0.0013640742
$a_{1,7} * 1/\mu_1 * K_3$		-0.0039167638		-0.0050026336
$a_{1,8} * 1/\mu_1 * K_4$		-0.0377445221		-0.0186924908
$a_{2,2} * 1/\mu_2 * 1/\mu_2$		-0.0790581407		-0.0183296576
$a_{2,3} * 1/\mu_2 * 1/\mu_3$		-0.0179731928		-0.0008414325
$a_{2,4} * 1/\mu_2 * 1/\mu_4$		0.0898985496		0.0016542802
$a_{2,5} * 1/\mu_2 * K_1$		0.0047046242		-0.0012703642
$a_{2,6} * 1/\mu_2 * K_2$		0.0184644019		0.0172411913
$a_{2,7} * 1/\mu_2 * K_3$		-0.0003705912		-0.0013038598
$a_{2,8} * 1/\mu_2 * K_4$		-0.0477860935		-0.0192686951

Table A.26: (Continued) The estimated values of the parameters used in factorial regression models for the mean throughput rate of a four–stage pull system.

	FACTORIAL REGRESSION MODELS (continued)			
	COEFFICIENT ESTIMATES			
	CONTINUOUS approximated by $T = 0.0001$		PERIODIC with $T = 1.0$	
Terms	Linear	Quadratic	Linear	Quadratic
$a_{3,3}$ * $1/\mu_3 * 1/\mu_3$		-0.0641752241		-0.0198141434
$a_{3,4}$ * $1/\mu_3 * 1/\mu_4$		0.0857245960		-0.0011263440
$a_{3,5}$ * $1/\mu_3 * K_1$		-0.0035111437		-0.0047421619
$a_{3,6}$ * $1/\mu_3 * K_2$		0.0021939812		-0.0027053583
$a_{3,7}$ * $1/\mu_3 * K_3$		0.0186148255		0.0199589240
$a_{3,8}$ * $1/\mu_3 * K_4$		-0.0391740122		-0.0118938360
$a_{4,4}$ * $1/\mu_4 * 1/\mu_4$		0.0060265887		-0.0180244209
$a_{4,5}$ * $1/\mu_4 * K_1$		-0.0095489671		-0.0044502908
$a_{4,6}$ * $1/\mu_4 * K_2$		-0.0140754731		-0.0063162769
$a_{4,7}$ * $1/\mu_4 * K_3$		-0.0097450178		-0.0036150522
$a_{4,8}$ * $1/\mu_4 * K_4$		0.0114917220		0.0199554239
$a_{5,5}$ * K_1*K_1		-0.0016811858		-0.0020408076
$a_{5,6}$ * K_1*K_2		-0.0016086535		0.0000749271
$a_{5,7}$ * K_1*K_3		0.0003351231		0.0034914875
$a_{5,8}$ * K_1*K_4		0.0052957324		0.0086338901
$a_{6,6}$ * K_2*K_2		-0.0032537168		-0.0038962730
$a_{6,7}$ * K_2*K_3		-0.0021031673		0.0002308363
$a_{6,8}$ * K_2*K_4		0.0073546464		0.0114513596
$a_{7,7}$ * K_3*K_3		-0.0063174994		-0.0054703226
$a_{7,8}$ * K_3*K_4		0.0010260753		0.0036148339
$a_{8,8}$ * K_4*K_4		-0.0221220675		-0.0135497103

Table A.27: Optimal workload and kanban allocation results obtained through enumerating the allocation vectors around balanced allocation of four–stage pull systems in order to maximize the mean throughput rate (**MTR**).

(Fixed parameter of the four–stage system: mean demand arrival rate (λ) = 1.0).

FOUR–STAGE CONTINUOUS PULL SYSTEMS
(Approximated by T = 0.0001)

OPTIMAL UNBALANCED ALLOCATION											•	BALANCED
TWC	$1/\mu_1$	$1/\mu_2$	$1/\mu_3$	$1/\mu_4$	TNK	K_1	K_2	K_3	K_4	**MTR**		**MTR**
4.0000	1.2000	1.1000	0.9000	0.8000	4	1	1	1	1	0.4277		0.4130
4.0000	1.2000	0.9000	0.9000	1.0000	5	1	1	1	2	0.4892		0.4858
4.0000	1.1000	1.0000	1.0000	0.9000	6	1	2	1	2	0.5222		0.5191
4.0000	1.1000	0.9000	1.0000	1.0000	7	1	2	2	2	0.5534		0.5531
4.0000	1.2000	1.0000	0.9000	0.9000	8	2	2	2	2	0.5832		0.5758
2.0000	0.5500	0.5500	0.5500	0.3500	4	1	1	1	1	0.6502		0.6168
2.0000	0.6000	0.5500	0.4500	0.4000	5	1	1	1	2	0.7840		0.7703
2.0000	0.6000	0.5000	0.4500	0.4500	6	1	1	1	3	0.8427		0.8117
2.0000	0.6000	0.4500	0.4500	0.5000	7	1	1	1	4	0.8778		0.8295
2.0000	0.5500	0.4500	0.5000	0.5000	8	1	1	2	4	0.9088		0.8348
									Average:	0.6639		0.6410

FOUR–STAGE PERIODIC PULL SYSTEMS
(Periodic with T = 1.0)

OPTIMAL UNBALANCED ALLOCATION											•	BALANCED
TWC	$1/\mu_1$	$1/\mu_2$	$1/\mu_3$	$1/\mu_4$	TNK	K_1	K_2	K_3	K_4	**MTR**		**MTR**
4.0000	1.2000	1.0000	1.0000	0.8000	4	1	1	1	1	0.2697		0.2672
4.0000	0.9000	0.9000	0.9000	1.3000	5	1	1	1	2	0.3101		0.3004
4.0000	0.8000	1.2000	0.8000	1.2000	6	1	2	1	2	0.3530		0.3412
4.0000	0.7000	1.1000	1.1000	1.1000	7	1	2	2	2	0.3988		0.3738
4.0000	1.2000	1.0000	0.9000	0.9000	8	2	2	2	2	0.4670		0.4625
2.0000	0.6000	0.5500	0.4500	0.4000	4	1	1	1	1	0.3470		0.3433
2.0000	0.4500	0.4500	0.4500	0.6500	5	1	1	1	2	0.4157		0.4065
2.0000	0.4000	0.4000	0.6000	0.6000	6	1	1	2	2	0.4531		0.4366
2.0000	0.3500	0.5500	0.5500	0.5500	7	1	2	2	2	0.4812		0.4618
2.0000	0.6000	0.5500	0.4500	0.4000	8	2	2	2	2	0.6167		0.6059
									Average:	0.4112		0.3999

Bibliography

[1] Akyildiz, I.F. and C. Huang, "Exact analysis of queueing networks with multiple job classes and blocking–after–service", *Queueing Systems*, Vol. 13, 427–440, (1993).

[2] Altıok, T., "On the phase-type approximation of general distributions", *IIE Transactions*, Vol. 17, No. 2, 110–116, (1985).

[3] Altıok, T., "(R,r) production/inventory systems", *Operations Research*, Vol. 37, No. 2, 266–276, (1989).

[4] Altıok, T., "Approximate analysis of queues in series with phase-type service times and blocking", *Operations Research*, Vol. 37, No. 4, 601–610, (1989).

[5] Altıok, T. and H.G. Perros, "Open networks of queues with blocking: split and merge configurations", *IIE Transactions*, Vol. 18, 251–261, (1986).

[6] Altıok, T. and G.A. Shiue, "Single–stage, multi–product production/inventory systems with backorders", *IIE Transactions*, Vol. 26, No. 2, 52–61, (1994).

[7] Altıok, T. and J.R. Stidham, "The allocation of interstage buffer capacities in production lines", *IIE Transactions*, Vol. 17, No. 2, 110–116, (1983).

[8] Andijani, A.A. and G.M. Clark, "Kanban allocation to serial production lines in a stochastic environment", In *Just-in-time Manufacturing Systems: Operational Planning and Control Issues*, A. Şatır (ed.), Elsevier Science Publishers B.V.(North–Holland), 175–190, (1991).

[9] Aneke, N.A.G. and A.S. Carrie, "A comprehensive flowline classification scheme", *International Journal of Production Research*, Vol. 22, No. 2, 281–297, (1984).

[10] Askin, R.G., G. Mitwasi and J.B. Goldberg, "Determining the number of kanbans in multi–item just–in–time systems", *IIE Transactions*, Vol. 25, No. 1, 89–98, (1993).

[11] Badinelli, R.D., "A model for continuous-review pull policies in serial inventory systems", *Operations Research*, Vol. 40, No. 1, 142–156, (1992).

[12] Baker, K.R., S.G. Powell and D.F. Pyke, "Buffered and unbuffered assembly systems with variable processing times", *Journal of Manufacturing and Operations Management*, Vol. 3, 200–223, (1990).

[13] Baker, K.R., S.G. Powell and D.F. Pyke, "Optimal allocation of work in assembly systems", *Management Science*, Vol. 39, No. 1, 101–106, (1993).

[14] Bard, J.F. and B. Golany, "Determining the number of kanbans in a multiproduct, multistage production system", *International Journal of Production Research*, Vol. 29, No. 5, 881–895, (1991).

[15] Baruh, H. and T. Altıok, "Analytical perturbations in markov chains", *European Journal of Operational Research*, Vol. 51, 210–222, (1991).

[16] Berkley, B.J., "Tandem queues and kanban-controlled lines", *International Journal of Production Research*, Vol. 29, No. 10, 2057–2081, (1991).

[17] Berkley, B.J., "A decomposition approximation for periodic kanban-controlled flow shops", *Decision Sciences*, Vol. 23, 291–311, (1992).

[18] Berkley, B.J., "A review of the kanban production control research literature ", *Productions and Operations Management*, Vol. 1, No. 4, 393–411, (1992).

[19] Bitran, G.R. and L. Chang, "A mathematical programming approach to a deterministic kanban system", *Management Science*, Vol. 33, No. 4, 427–441, (1987).

[20] Bitran, G.R. and S. Dasu, "A review of open queueing network models of manufacturing systems", *Queueing Systems*, Vol. 12, 95–134, (1992).

[21] Bitran, G.R. and D. Tirupati, "Multiproduct queueing networks with deterministic routing: Decomposition approach and the notion of interference", *Management Science*, Vol. 34, No. 1, 75–100, (1988).

[22] Bitran, G.R. and D. Tirupati, "Approximations for product departures from a single–server station with batch processing in multi–product queues", *Management Science*, Vol. 35, No. 7, 851–878, (1989).

[23] Brandwajn, A., "Equivalence and decomposition in queuing systems — A unified approach", *Performance Evaluation*, Vol. 5, 175–186, (1985).

[24] Brandwajn, A. and Y.L. Jow, "An approximation method for tandem queues with blocking", *Operations Research*, Vol. 36, No. 1, 73–83, (1988).

[25] Buxey, G.M., N.D. Slack and R. Wild, "Production flow line system design — A review", *AIIE Transactions*, Vol. 5, No. 1, 37–48, (1973).

[26] Buzacott, J.A., "Queuing models of kanban and MRP controlled production systems", *Engineering Costs and Production Economics*, Vol. 17, 3–20, (1989).

[27] Buzacott, J.A. and J.G. Shanthikumar, "Design of manufacturing systems using queueing models", *Queueing Systems*, Vol. 12, 135–214, (1992).

[28] Buzacott, J.A., S.M. Price, and J.G. Shanthikumar, "The Performance of Kanban Controlled Serial Production Systems", In *Operations Research in Production Planning and Control*, G. Fandel, T. Gulledge and A. Jones (ed.), Springer–Verlag, (Berlin - Heidelberg), 71–88, (1993).

[29] Carnall, C.A. and R. Wild, "The location of variable work stations and the performance of production flow lines", *International Journal of Production Research*, Vol. 14, No. 6, 703–710, (1976).

[30] Chang, T.M. and Y. Yih, "Generic kanban systems for dynamic environments", *International Journal of Production Research*, Vol. 32, No. 4, 889–902, (1994).

[31] Conway, R., W. Maxwell, J.O. McClain and L.J. Thomas, "The role of work-in-process inventory in serial production lines", *Operations Research*, Vol. 36, No. 2, 229–241, (1988).

[32] Dallery, Y. and S.B. Gershwin, "Manufacturing flow line systems: A review of models and analytical results", *Queueing Systems*, Vol. 12, 3–94, (1992).

[33] De Koster, M.B.M., "Approximate analysis of production systems", *European Journal of Operational Research*, Vol. 37, 214–226, (1988).

[34] Deleersnyder, J.L., T.J. Hodgson, H. Müller(-Malek) and P.J. O'Grady, "Kanban type controlled pull systems: An analytic approach", *Management Science*, Vol. 35, No. 9, 1079–1091, (1989).

[35] Deleersnyder, J.L., T.J. Hodgson, R.E. King, P.J. O'Grady and A. Savva, "Integrating kanban type pull systems and MRP type push systems: Insights from a markovian model", *IIE Transactions*, Vol. 24, No. 3, 43–56, (1992).

[36] Duenyas, I. and W.J. Hopp, "Estimating the throughput of an exponential CONWIP assembly system", *Queueing Systems*, Vol. 14, 135–157, (1993).

[37] El-Rayah, T.E., "The efficiency of balanced and unbalanced production lines", *International Journal of Production Research*, Vol. 17, No. 1, 61–75, (1979).

[38] El-Rayah, T.E., "The effect of inequality of interstage buffer capacities and operation time variability on the efficiency of production line systems", *International Journal of Production Research*, Vol. 17, No. 1, 77–89, (1979).

[39] Gershwin, S.B., "An efficient decomposition method for the approximate evaluation of tandem queues with finite storage space and blocking", *Operations Research*, Vol. 35, No. 2, 291–305, (1987).

[40] Gershwin, S.B., "Assembly/disassembly systems: An efficient decomposition algorithm for tree-structured networks", *IIE Transactions*, Vol. 23, No. 4, 302–314, (1991).

[41] Golhar, D.Y. and B.R. Sarker, "Economic manufacturing quantity in a just–in–time delivery system", *International Journal of Production Research*, Vol. 30, No. 5, 961–972, (1992).

[42] Golhar, D.Y. and C.L. Stamm, "The just-in-time philosophy: A literature review", *International Journal of Production Research*, Vol. 29, No. 4, 657–676, (1991).

[43] Gordon, W.J. and G.F. Newell, "Cyclic Queueing Systems with Restricted Length Queues", *Operations Research*, Vol. 15, 266–277, (1967).

[44] Gravel, M. and W.L. Price, "Using the kanban in a job–shop environment", *International Journal of Production Research*, Vol. 26, No. 6, 1105–1118, (1988).

[45] Grünwald, H., P.E.T. Striekwold and P.J. Weeda, "A framework for quantitative comparison of production control concepts", *International Journal of Production Research*, Vol. 27, No. 2, 281–292, (1989).

[46] Hillier, F.S. and R.W. Boling, "The effect of some design factors on the efficiency of production lines with variable operation times", *Journal of Industrial Engineering*, Vol. 17, 651–658, (1966).

[47] Hillier, F.S. and R.W. Boling, "Finite queues in series with exponential or erlang service times: A numerical approach", *Operations Research*, Vol. 15, 286–303, (1967).

[48] Hillier, F.S. and R.W. Boling, "On the optimal allocation of work in symmetrically unbalanced production line systems with variable operation times", *Management Science*, Vol. 25, No. 8, 721–728, (1979).

[49] Hillier, F.S. and K.C. So, "The effect of the coefficient of variation of operation times on the allocation of storage space in production line systems", *IIE Transactions*, Vol. 23, No. 2, 198–206, (1991).

[50] Hillier, F.S. and K.C. So, "The effect of machine breakdowns and interstage storage on the performance of production lines", *International Journal of Production Research*, Vol. 29, No. 10, 2043–2055, (1991).

[51] Hillier, F.S., K.C. So and R.W. Boling, "Notes: Toward characterizing the optimal allocation of storage space in production line systems with variable processing times", *Management Science*, Vol. 39, No. 1, 126–133, (1993).

[52] Hodgson, T.J. and D. Wang, "Optimal hybrid push/pull control strategies for a parallel multistage system: Part I", *International Journal of Production Research*, Vol. 29, No. 6, 1279–1287, (1991).

[53] Hodgson, T.J. and D. Wang, "Optimal hybrid push/pull control strategies for a parallel multistage system: Part II", *International Journal of Production Research*, Vol. 29, No. 7, 1453–1460, (1991).

[54] Hong, Y., C.R. Glassey and D. Seong, "The analysis of a production line with unreliable machines and random processing times", *IIE Transactions*, Vol. 24, No. 1, 77–83, (1992).

[55] Hopp, W.J. and M.L. Spearman, "Throughput of a constant work in process manufacturing line subject to failures", *International Journal of Production Research*, Vol. 29, No. 3, 635–655, (1991).

[56] Jafari, M.A. and J.G. Shanthikumar, "An approximate model of multistage automatic transfer lines with possible scrapping of workpieces", *IIE Transactions*, Vol. 19, No. 3, 252–265, (1987).

[57] Jafari, M.A. and J.G. Shanthikumar, "Determination of optimal buffer storage capacities and optimal allocation in multi stage automatic transfer lines", *IIE Transactions*, Vol. 21, No. 2, 130–135, (1989).

[58] Jordan, S., "Analysis and approximation of a JIT production line", *Decision Sciences*, Vol. 19, 672–681, (1988).

[59] Kalkunte, M.V., S.C. Sarin and W.E. Wilhelm, "Flexible manufacturing systems: A review of modelling approaches for design, justification and operation", In *Flexible Manufacturing Systems: Methods and studies*, A. Kusiak (ed.), Elsevier Science Publishers B.V.(North–Holland), 3–25, (1986).

[60] Karmarkar, U.S. and S. Kekre, "Batching policy in kanban systems", *Journal of Manufacturing Systems*, Vol. 8, No. 4, 317–328, (1989).

[61] Kim, T., "Just-in-time manufacturing system: A periodic pull system", *International Journal of Production Research*, Vol. 23, No. 3, 553–562, (1985).

[62] Kimura, O. and H. Terada, "Design and analysis of Pull Systems: A method of multi-stage production control", *International Journal of Production Research*, Vol. 19, No. 3, 241–253, (1981).

[63] Kırkavak, N. and C. Dinçer, "Performance evaluation models for single–item periodic pull production systems", *Journal of Operational Research Society*, (forthcoming).

[64] Lee, L.C., "Parametric appraisal of the JIT system", *International Journal of Production Research*, Vol. 25, No. 10, 1415–1429, (1987).

[65] Li, A. and H.C. Co, "A dynamic programming model for the kanban assignment problem in a multi–stage multi–period production system", *International Journal of Production Research*, Vol. 29, No. 1, 1–16, (1991).

[66] Luss, H. and M.B. Rosenwein, "A lot–sizing model for just–in–time manufacturing", *Journal of Operational Research Society*, Vol. 41, No. 3, 201–209, (1990).

[67] Magazine, M.J. and G.L. Silver, "Heuristics for determining output and work allocations in series flow lines", *International Journal of Production Research*, Vol. 16, No. 3, 169–181, (1978).

[68] Meral, S., "A design methodology for just–in–time production lines", Ph.D. Dissertation, Department of Industrial Engineering, Middle East Technical University, Ankara, Turkey, (1993).

[69] Mitra, D. and I. Mitrani, "Analysis of a kanban discipline for cell coordination in production lines, I", *Management Science*, Vol. 36, No. 12, 1548–1566, (1990).

[70] Mitra, D. and I. Mitrani, "Analysis of a kanban discipline for cell coordination in production lines, II: Stochastic demands", *Operations Research*, Vol. 39, No. 5, 807–823, (1991).

[71] Muth, E.J., "The production rate of a series of work stations with variable service times", *International Journal of Production Research*, Vol. 11, No. 2, 155–169, (1973).

[72] Muth, E.J., "The reversibility property of production lines", *Management Science*, Vol. 25, No. 2, 152–158, (1979).

[73] Muth, E.J., "Stochastic processes and their network representations associated with a production line queuing model", *European Journal of Operational Research*, Vol. 15, 63–83, (1984).

[74] Muth, E.J. and A. Alkaff, "The bowl phenomenon revisited", *International Journal of Production Research*, Vol. 25, No. 2, 161–173, (1987).

[75] Muth, E.J. and A. Alkaff, "The throughput rate of three–station production lines: A unifying solution", *International Journal of Production Research*, Vol. 25, No. 10, 1405–1413, (1987).

[76] Onvural, R.O. and H.G. Perros, "On equivalencies of blocking mechanisms in queuing networks with blocking", *Operations Research Letters*, Vol. 5, 292–297, (1986).

[77] Payne, S., N. Slack and R. Wild, "A note on operating characteristics of 'balanced' and 'unbalanced' production flow lines", *International Journal of Production Research*, Vol. 10, No. 1, 93–98, (1972).

[78] Philipoom, P., L.P. Rees, B.W. Taylor and P.Y. Huang, "An investigation of the factors influencing the number of Kanbans required in the implementation of the JIT technique with kanbans", *International Journal of Production Research*, Vol. 25, No. 3, 457–472, (1987).

[79] Philipoom, P., L.P. Rees, B.W. Taylor and P.Y. Huang, "A mathematical programming approach for determining workcenter lotsizes in a just–in–time system with signal kanbans", *International Journal of Production Research*, Vol. 28, No. 1, 1–15, (1990).

[80] Philippe, B., Y. Saad and W.J. Stewart, "Numerical methods in markov chain modeling", *Operations Research*, Vol. 40, No. 6, 1156–1179, (1992).

[81] Pike, R. and G.E. Martin, "The bowl phenomenon in unpaced lines", *International Journal of Production Research*, Vol. 32, No. 3, 483–499, (1994).

[82] Price, W., M. Gravel and A.L. Nsakanda, "A review of optimisation models of kanban–based production systems", *European Journal of Operational Research*, Vol. 75, 1–12, (1994).

[83] Pyke, D.F. and M.A. Cohen, "Push and Pull in Manufacturing and Distribution Systems", *Journal of Operations Management*, Vol. 9, No. 1, 24–43, (1990).

[84] Rao, N.P., "A generalization of the 'bowl phenomenon' in series production systems", *International Journal of Production Research*, Vol. 14, No. 4, 437–443, (1976).

[85] Sarker, B.R., "Some comparative and design aspects of series production systems", *IIE Transactions*, Vol. 16, No. 3, 229–239, (1984).

[86] Sarker, B.R. and R.D. Harris, "The effect of imbalance in a just–in–time production system: A simulation study", *International Journal of Production Research*, Vol. 28, No. 5, 879–894, (1988).

[87] Sarker, B.R. and G.R. Parija, "An optimal batch size for a production system operating under a fixed–quantity, periodic delivery policy", *Journal of Operational Research Society*, Vol. 45, No. 8, 891–900, (1994).

[88] Seidmann, A., "Regenerative pull (kanban) production control policies", *European Journal of Operational Research*, Vol. 35, 401–413, (1988).

[89] Sheskin, T.J., "Allocation of interstage storage along an automatic production line", *AIIE Transactions*, Vol. 8, No. 1, 146–152, (1976).

[90] Siha, S., "The pull production system: Modelling and characteristics", *International Journal of Production Research*, Vol. 32, No. 4, 933–949, (1994).

[91] Smith, J.M. and S. Daskalaki, "Buffer space allocation in automated assembly lines", *Operations Research*, Vol. 36, No. 2, 343–358, (1988).

[92] So, K.C., "On the efficiency of unbalancing production lines", *International Journal of Production Research*, Vol. 27, No. 4, 717–729, (1989).

[93] So, K.C. and S.C. Pinault, "Allocating buffer storages in a pull system", *International Journal of Production Research*, Vol. 26, No. 12, 1959–1980, (1988).

[94] Soyster, A.L., J.W. Schmidt and M.W. Rohrer, "Allocation of buffer capacities for a class of fixed cycle production lines", *AIIE Transactions*, Vol. 11, No. 2, 140–146, (1979).

[95] Spearman, M.L. and M.A. Zazanis, "Push and pull production systems: Issues and comparisons", *Operations Research*, Vol. 40, No. 3, 521–532, (1992).

[96] Spearman, M.L., D.L. Woodruff and W.J. Hopp, "Conwip: A pull alternative to kanban", *International Journal of Production Research*, Vol. 28, No. 5, 879–894, (1990).

[97] Springer, M.C., "A decomposition approximation for finite–buffered flow lines of exponential queues", *European Journal of Operational Research*, Vol. 74, 95–110, (1994).

[98] Srinivasan, M.M. and H. Lee, "Random review production/inventory systems with compound poisson demands and arbitrary processing times", *Management Science*, Vol. 37, No. 7, 813–833, (1991).

[99] Stecke, K.E. and T.L. Morin, "The optimality of balancing workloads in certain types of flexible manufacturing systems", *European Journal of Operational Research*, Vol. 20, 68–82, (1985).

[100] Stidham, S. and R. Weber, "A survey of markov decision models for control of networks of queues", *Queueing Systems*, Vol. 13, 291–314, (1993).

[101] Sugimori, Y., K. Kusunoki, F. Cho and S. Uchikawa, "Toyota production system and kanban system: Materialization of just-in-time and respect for human system", *International Journal of Production Research*, Vol. 15, No. 6, 553–564, (1977).

[102] Tayur, S.R., "Properties of serial kanban systems", *Queueing Systems*, Vol. 12, 297–318, (1992).

[103] Tayur, S.R., "Structural properties and a heuristic for kanban–controlled serial lines", *Management Science*, Vol. 39, No. 11, 1347–1368, (1993).

[104] Thompson, W.W. and R.L. Burford, "Some observations on the bowl phenomenon", *International Journal of Production Research*, Vol. 26, No. 8, 1367–1373, (1988).

[105] Veatch, M.H. and L.M. Wein, "Optimal control of a two–station tandem production/inventory system", *Operations Research*, Vol. 42, No. 2, 337–350, (1994).

[106] Villeda, R., R. Dudek and M.L. Smith, "Increasing the production rate of a just–in–time production system with variable operation times", *International Journal of Production Research*, Vol. 26, No. 11, 1749–1768, (1988).

[107] Wang, H. and H. Wang, "Determining the number of kanbans: A step toward non-stock production", *International Journal of Production Research*, Vol. 28, No. 11, 2101–2115, (1990).

[108] Wang, H. and H. Wang, "Decomposition and optimal design of kanban systems", In *Just-in-time Manufacturing Systems: Operational Planning and Control Issues*, A. Şatır (ed.), Elsevier Science Publishers B.V.(North–Holland), 165–174, (1991).

[109] Wolisz, A., "Production rate optimization in a two–stage system with finite intermediate storage", *European Journal of Operational Research*, Vol. 18, 369–376, (1984).

[110] Yamazaki, G., T. Kawashima and H. Sakasegawa, "Reversibility of tandem blocking queueing systems", *Management Science*, Vol. 31, 78–83, (1985).

[111] Yamazaki, G., H. Sakasegawa and J.G. Shanthikumar, "On optimal arrangement of stations in a tandem queueing system with blocking", *Management Science*, Vol. 38, No. 1, 137–153, (1992).

[112] Yao, D.D., "Some properties of throughput function of closed networks of queues", *Operations Research Letters*, Vol. 3, No. 6, 313–317, (1985).

[113] Yao, D.D. and J.A. Buzacott, "The exponentialization approach to flexible manufacturing system models with general processing times", *European Journal of Operational Research*, Vol. 24, 410–416, (1986).

[114] Yeralan, S. and E.J. Muth, "A general model for a production line with intermediate buffer and station breakdown", *IIE Transactions*, Vol. 19, No. 2, 130–139, (1987).

[115] Yu, K.C. and D.L. Bricker, "Analysis of a markov chain model of a multistage manufacturing system with inspection, rejection and rework", *IIE Transactions*, Vol. 25, No. 1, 109–112, (1993).

[116] Zipkin, P.H., "Models for design and control of stochastic, multi–item batch production systems", *Operations Research*, Vol. 34, No. 1, 91–104, (1986).

Vita

Nureddin Kırkavak was born in Gemlik, Bursa, Turkey, in 1965. He attended the Department of Industrial Engineering, Middle East Technical University, in September 1983 and graduated with high honors in July 1987. In September 1987, he joined to Department of Industrial Engineering, Bilkent University, as a research assistant. From that time to the present, he worked with Dr. Cemal Dinçer for his graduate study at the same department. He got his M.S. degree in June 1990 with the thesis titled as *"Analytical Loading Models and Control Strategies in Flexible Manufacturing Systems: A Comparative Study"*. During his Ph.D. study with Dr. Cemal Dinçer, he worked on *"Performance Evaluation Models of Pull Production Systems"*. Currently, he is an instructor at Department of Industrial Engineering, Bilkent University.