549 LECTURE NOTES IN ECONOMICS AND MATHEMATICAL SYSTEMS

Georg N. Krieg

Kanban-Controlled Manufacturing Systems

Springer

Lecture Notes in Economics and Mathematical Systems 549

Georg N. Krieg

Kanban-Controlled Manufacturing Systems

Springer

Author

Dr. Georg N. Krieg
Chair of Production and Operations Management
Faculty of Business Administration and Economics
Catholic University of Eichstaett-Ingolstadt
85049 Ingolstadt, Germany
georg.krieg@web.de

Doctoral dissertation:
Faculty of Business Administration and Economics
Catholic University of Eichstaett-Ingolstadt, 2003
Referees: Professor Dr. Heinrich Kuhn and Professor Dr. Ulrich Küsters

Gedruckt mit Unterstützung des Förderungs- und Beihilfefonds Wissenschaft
der VG WORT

Library of Congress Control Number: 2004115950

ISSN 0075-8442
ISBN 3-540-22999-X Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springeronline.com

Printed in Germany

Typesetting: Camera ready by author
Cover design: *Erich Kirchner*, Heidelberg

Printed on acid-free paper 42/3130Di 5 4 3 2 1 0

To my academic teachers and mentors
Prof. Dr. Werner Delfmann,
Prof. Dr. Heinrich Kuhn, and
Prof. Dr. Horst Tempelmeier
with my sincere thanks for their guidance
and support during these past years

Contents

1

Subject and Outline

The production management approach Just-In-Time (JIT) gained worldwide prominence when the rest of the world noticed the increasing success of Japanese companies in the late 1970s and early 1980s. As one major operational element of JIT, the kanban control system became a popular topic in western research and industry (e.g., Sugimori et al. 1977, Monden 1981a–d, 1998; Kimura and Terada 1981, Schonberger 1982, Hall 1983, Ohno 1988, Shingo 1989). Manufacturing companies outside Japan began to use kanbans to control production and flow of material. Several empirical studies document that kanban control bears great potential to significantly improve operations (e.g., White, Pearson, and Wilson 1999, Fullerton and McWatters 2001, White and Prybutok 2001).

Some operational improvements that follow the implementation of kanban systems are commonly attributed to organizational changes rather than to the kanban principle itself. A company, however, may reap the full benefits of kanban control only after determining an optimal or near-optimal system configuration. Finding such a configuration requires methods that can determine key performance measures, such as average fill rates and average inventory levels. Computer simulation may generally be used to analyze the performance of a system, but to identify an optimal configuration, many different system variants may have to be evaluated. To finish the search in a reasonable amount of time, the evaluation method should be fast—reliable simulation, however, is usually very time-consuming. Analytical (mathematical) evaluation methods are therefore needed that can determine key performance measures quickly, even if these methods only approximate the true performance of the system.

Some analytical evaluation methods can be found in the literature, particularly for systems with a single product. Kanban systems in industrial operations, however, usually control the production of several different products produced on shared manufacturing facilities (e.g., Anupindi and Tayur 1998). For the analysis of such multi-product kanban systems, we propose a construction-kit approach that makes it possible to build stochastic analytical models of a large class of single- and multi-product kanban systems.

Outline. In the following two chapters, we describe different implementations of kanban control, and we review the literature on stochastic models of kanban systems. The review shows that most models published so far represent single-product systems. In Chapter 4, we introduce the center part of our research: a construction-kit approach that yields new models of single- and multi-stage kanban systems with single- and multi-product manufacturing facilities.

The details of the construction-kit approach are given in Chapters 5–8. First, we develop three different one-product models that are the basic building blocks ("components") of the construction kit (Chapter 5). Then, in Chapter 6, we describe two procedures to build modules ("subassemblies") consisting of several instances of the second and the third one-product model, respectively. The subassemblies are models of kanban-controlled multi-product manufacturing systems with one and two production stage(s). They may be used to build composite models of systems with multiple stages. The general technique for linking models of single- and two-stage (sub-)systems is explained in Chapter 7. Technical restrictions limit the applicability of the basic version of the model construction kit to systems without multi-product facilities in immediate succession. A modeling trick, however, may be used to work around this limitation so that the extended version of the construction kit may be used to build models of systems with multi-product facilities in series (Chapter 8).

Since most models built with the construction kit only approximate the true behavior of the modeled systems, the quality of the approximation is of primary concern. We conducted systematic tests to examine the approximation quality for several important modeling examples. The results of these tests are reported in Chapter 9. Heuristic procedures were used to identify plausible kanban configurations for the test instances. The algorithms of these procedures are given in the appendix. In Chapter 10, we demonstrate how models generated with the construction kit may be used to study the behavior of kanban systems. We give numerous examples for different system variants. Finally, in Chapter 11, we conclude with a summary and directions for future research. A comprehensive list of all symbols and abbreviations is provided after the appendix (pp. 223–230).

2

Kanban-Controlled Manufacturing Systems: Basic Version and Variations

2.1 Basic Kanban System

The least complex variant of a kanban-controlled manufacturing system with multiple products is a system with a single multi-product manufacturing facility. Besides the production facility, the system contains a scheduling board, an output store for finished products, containers to store and carry finished items, and one set of kanbans for each product in the system (Fig. 2.1).

Traditionally, a kanban is a tag-like card (*kanban* is Japanese meaning "card" or "visible record" [Schonberger 1982, p. 17]). One kanban must be attached to each container in the output store. The number of kanbans is limited, restricting the maximum amount of finished items in the system. When a container is withdrawn, the accompanying kanban is detached from the container and placed on the scheduling board. Alternatively, the kanban may be detached when the last item is removed from the container (this is equivalent to using a fixed number of containers to limit the maximum inventory of a product). Also, removed kanbans may be put in a kanban collection box located in the output store before they are transferred to the scheduling board, either when a given number of kanbans has accumulated or when a specified

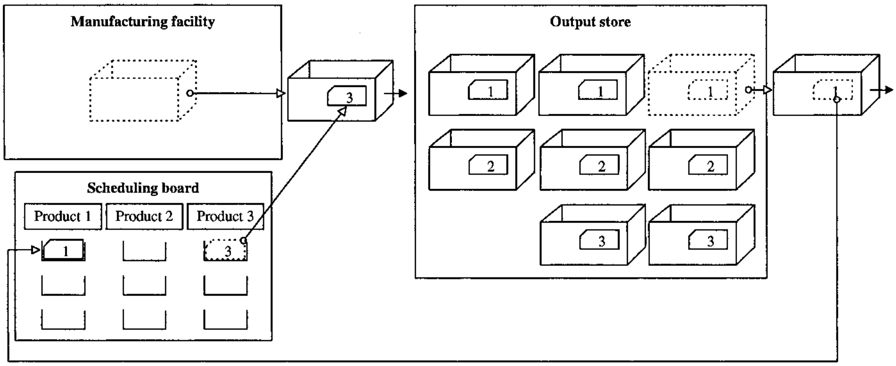

Fig. 2.1. Basic kanban system with three different products

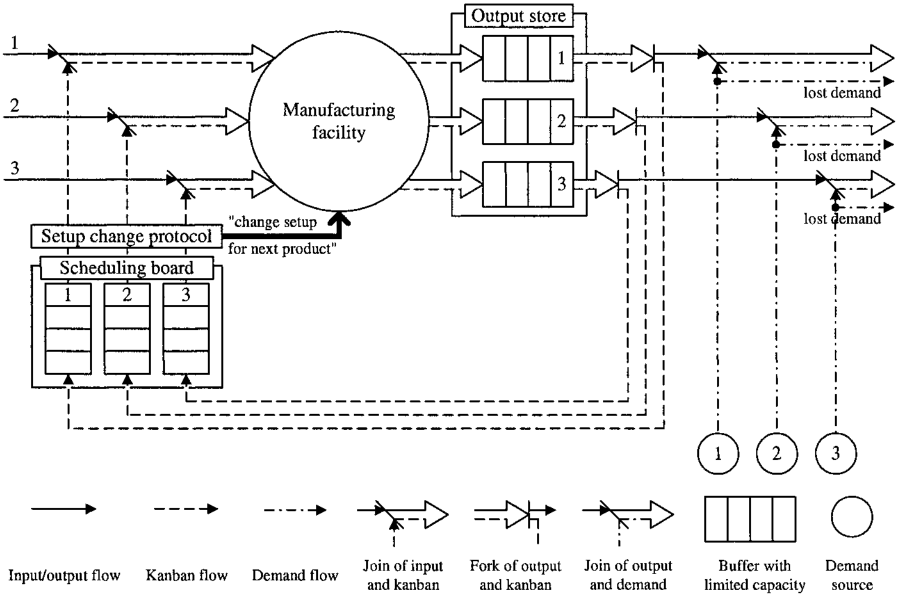

Fig. 2.2. Model of a single-stage multi-product kanban system without backorders (adapted from Mitra and Mitrani 1990, Fig. 3)

amount of time has elapsed from the last transfer. A detached or "active" kanban authorizes manufacture of one standard container of the product indicated on the card. When a container has been filled with the prescribed number of items, the now "inactive" kanban is affixed to the container and the container is transferred to the output store (Fig. 2.2).

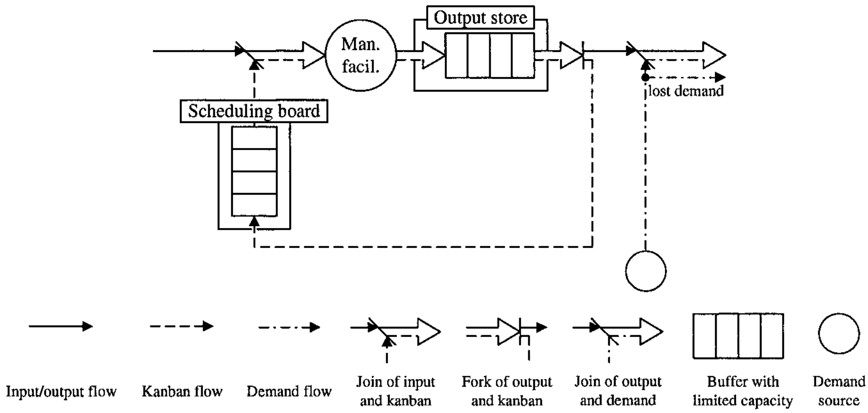

Fig. 2.3. Model of a single-stage single-product kanban system without backorders

Between the production of different products, setup changes must be performed that may consume a significant amount of time. A *setup change protocol* defines the rules for deciding when and to which other setup the current setup of the manufacturing facility should be changed next. One option is that items of a product are processed until there are no more active kanbans for this product on the scheduling board (*exhaustive processing*). Then the manufacturing facility is being set up for the next product according to a predetermined fixed *cyclic* setup sequence, for example, product 1, product 2, product 3 (repeated). Should no kanban be active for the next product, then this product is skipped. Should no kanban be active for any product, then the manufacturing facility idles. During the idle period, the current setup is maintained so that the manufacturing facility may immediately resume production if the next active kanban authorizes production of the product that was manufactured last. We refer to this setup change protocol as *cyclic-exhaustive processing*. Several other setup change protocols are suggested in the literature (e.g., Amin and Altiok 1997).

Single-product systems. A single-stage single-product kanban system is a special case of the basic kanban system. The manufacturing facility processes items as long as active kanbans are available and idles when all kanbans are inactive (Fig. 2.3). Setup changes and, hence, setup change protocols are not required.

2.2 Backorders

In the basic kanban system, customers whose demand cannot be filled from stock do not wait until the system can meet their demand. They either withdraw their request, or they turn to a different supplier who offers the same product, possibly at somewhat less favorable conditions (e.g., higher price, lower quality). As a result, no backorders accumulate in the basic kanban system.

In a different system, customers may have no alternative but to wait until their demand is satisfied. This is the standard situation for manufacturing stages that draw raw material or parts from a single supplier (the supplier may be the preceding manufacturing stage or an outside supplier). In those systems, the maximum number of backorders depends on the number of customers who generate requests. If there is only a single customer—the standard in serial manufacturing systems with one producing and one consuming manufacturing facility for each product—then the maximum number of backorders is one, and the demand source runs dry during the backorder situation. If there are several customers, then the maximum number of backorders equals the number of customers, and the average arrival rate of demand changes with the number of backorders. If customers have a local inventory of input material with a given target inventory level, then the maximum number of backorders a single customer can cause equals the target inventory level plus one (the additional backorder is due to the customer waiting for input material). If the number of customers is very large—a common situation for final products sold to private customers—then the number of backorders may become practically infinite, and the average arrival rate of demand is not significantly affected by the number of waiting customers.

It is also possible that a customer with outstanding orders continues generating new orders with constant average rate. This behavior, for example, may be observed in some assembly lines: rather than stopping the line, missing parts are added after the incomplete products have passed the last station and are waiting in a designated buffer area for unfinished products. In those systems, the number of backorders (for input material) is unlimited, even if there is only a single customer (the assembly system), and the average arrival rate of demand is independent of the number of backorders.

Finally, customers may accept a waiting time only if the total number of unfilled orders in the system is below a certain limit. Otherwise, they either withdraw their request or contact a different supplier. Assuming that all costumers consider the same number of backorders in the system prohibitive, then the maximum number of

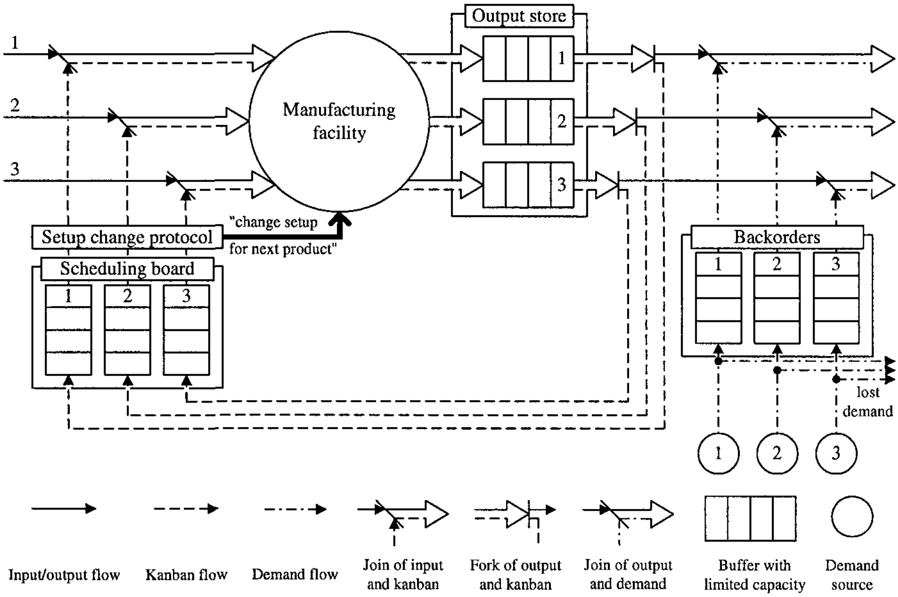

Fig. 2.4. Model of a single-stage multi-product kanban system with a limited number of backorders and lost demand

backorders equals the number of customers if the number of customers is less than the prohibitive number of unprocessed orders. Otherwise, the maximum number of backorders equals the prohibitive value. The average arrival rate of demand depends on the number of waiting customers, unless the number of customers is very large. If customers keep generating new requests irrespective of outstanding orders, then the maximum number of backorders always equals the prohibitive number of unprocessed orders, and the average arrival rate of demand is independent of the number of backorders in the system.

The model of a single-stage multi-product kanban system with a limited number of backorders and lost demand is illustrated in Figure 2.4.

2.3 Multiple Stages

In a kanban system with two or more production stages, the processed items of one stage are the (main) input material of the following stage. Unlike the manufacturing facility in the basic kanban system, the manufacturing facilities may experience shortage of input material (*starvation*). For modeling purposes, we define that in-

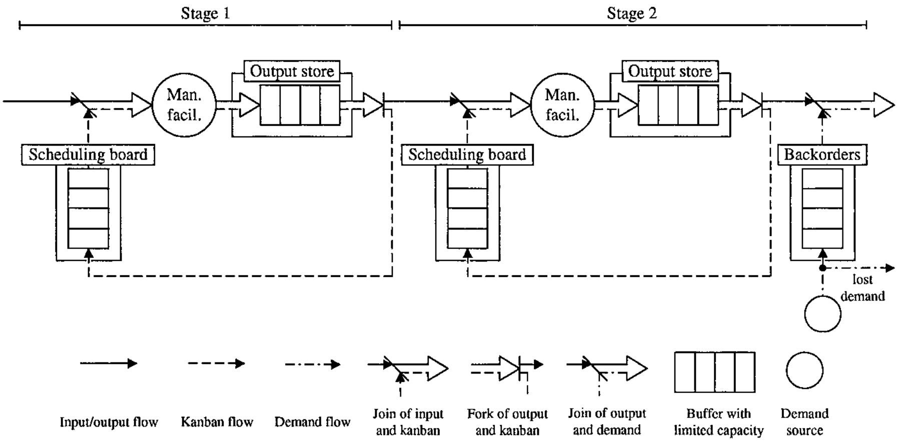

Fig. 2.5. Model of a two-stage single-product kanban system with type-1 material transfer, a limited number of backorders, and lost demand

put material for the first stage is always available in the required quantities, that is, manufacturing facilities in the first stage never starve. If manufacturing facilities in the first stage of the modeled system may experience shortage of input material, we declare that the first stage of the model corresponds to the procurement process of the input material for the first stage of the real system, and that the second stage of the model represents the first stage of the modeled system.

Production of each product in each stage is controlled by a distinct kanban loop with a fixed number of kanbans (Figs. 2.5 and 2.6). Immediately before the beginning of production, a container with input material is withdrawn from the output store of the preceding stage. Should no material be available, then the manufacturing facility either waits until new material arrives, or the setup is changed to process items of a different product.

The setup change protocol in multi-stage kanban systems must consider that, at some times, input material may not be available for products with active kanbans. One possible setup change protocol is *cyclic-exhaustive processing with limited input material.* With this setup change protocol, the precondition for a setup is that at least one active kanban and one container with input material must be available. Once the manufacturing facility has been set up for a specific product, it processes items of this product until either the number of active kanbans is zero, that is, all empty containers for the product have been filled, or the input material is depleted. Then the manufacturing facility is being set up for the next product that meets the setup con-

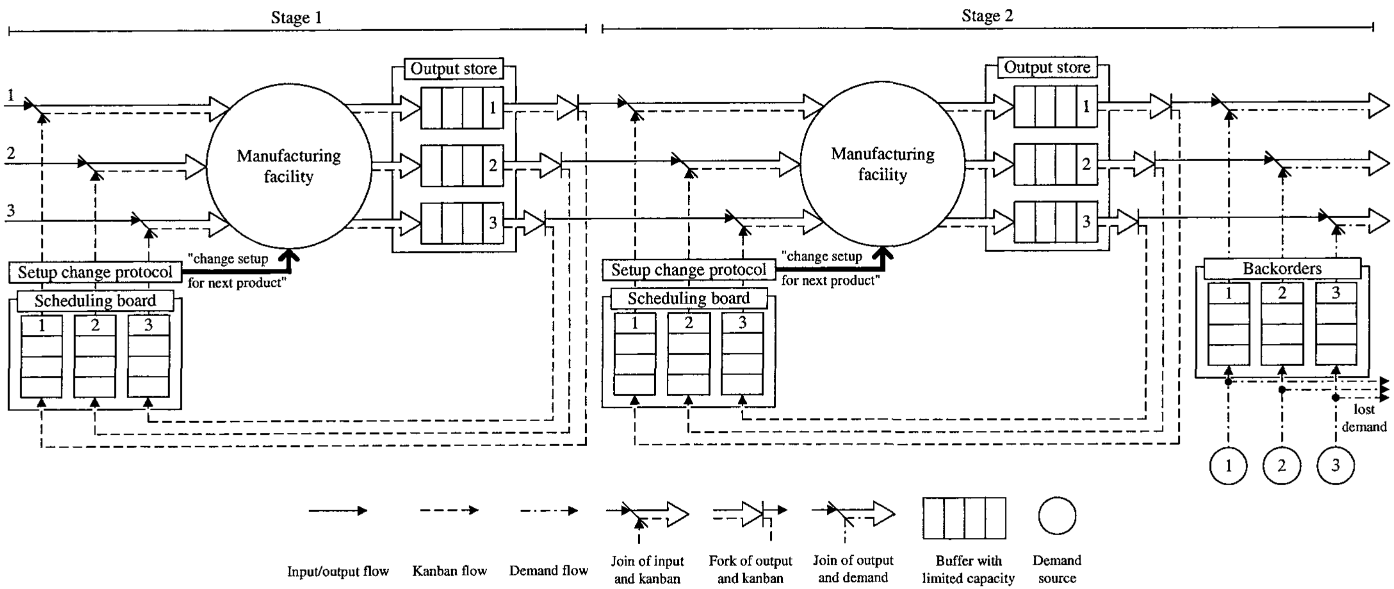

Fig. 2.6. Model of a two-stage multi-product kanban system with type-1 material transfer, a limited number of backorders, and lost demand

Table 2.1. Classification of Material Transfer Schemes

<table>
<tr><td>Output Store Stage m
$=$
Input Store Stage $m+1$</td><td colspan="3">Output Store Stage m
$\neq$
Input Store Stage $m+1$</td></tr>
<tr><td>Type 1
Withdrawal immediately before start of production</td><td>Type 2
Withdrawal immediately after activation of kanban</td><td>Type 3
Fixed quantity, variable withdrawal cycle</td><td>Type 4
Fixed withdrawal cycle, variable quantity</td></tr>
<tr><td colspan="2">One-Card System</td><td colspan="2">Two-Card System</td></tr>
</table>

dition. The order in which products are considered for production is stipulated by a predetermined fixed cyclic setup sequence, for example, product 1, product 2, product 3 (repeated). Should no product meet the setup condition, then the manufacturing facility idles until one product can offer at least one active kanban and one container with input material. If the first product to meet the setup condition is the same product that was manufactured last, then no setup activities are required and production may resume instantly. Otherwise, a setup change is initiated and production starts upon completion of the setup process.

2.4 Material Transfer Schemes

Multi-stage kanban systems may be classified by the rules for transferring containers from the output store of one stage, say stage m, to the input store of the following stage, say stage $m+1$. At least four different set of rules may be found in the literature. In some systems, the output store of a stage is also the input store of the next stage. Consequently, there is no need for a transfer mechanism. In systems with separate output and input stores, the transfer of containers from the output store to the input store may be executed at different points in time. In Table 2.1, we summarize four different material transfer schemes. Each type is shortly explained in the following paragraphs. A discussion of the different transfer types is provided by Berkley (1991).

Type-1 material transfer. The output store of a stage is also the input store of the following stage, and material is withdrawn from the store immediately before start of production. This scheme has been labeled "late material transfer" by Gstettner and Kuhn (1996).

Type-2 material transfer. The output store of a stage is physically separated from the input store of the next stage. The material from the output store of stage m is withdrawn immediately after activation of a kanban in stage $m+1$ (the active kanban authorizes the withdrawal of a container with input material). The kanban is attached to the container and both join the queue in front of the manufacturing facility of stage $m+1$. If the output store of stage m is empty upon activation of a kanban in stage $m+1$, then the transfer is delayed until the manufacturing facility of stage m completes a container with the appropriate parts. This scheme, first described by Mitra and Mitrani (1990), has been labeled "immediate material transfer" by Gstettner and Kuhn (1996).

Type-3 and type-4 material transfer. The output store of a stage is physically separated from the input store of the following stage, and an additional set of cards, called *withdrawal*, *conveyance*, *delivery*, *move*, or *transportation kanbans*, is used to organize the transfer of containers between the stages (e.g., Monden 1998, Chap. 2). These systems are commonly referred to as *two-card* or *dual-card kanban systems*, in contrast to *one-card* or *single-card kanban systems* that only use a single card type. A withdrawal kanban must be attached to each container in the input store of a stage. When a container is taken up for production, the withdrawal kanban is removed and put into a kanban collection box. Eventually, a carrier takes the withdrawal kanbans out of the box and moves to the output store of the preceding stage. There, he withdraws a full container for each withdrawal kanban in his possession, removes the regular kanban from each container, and attaches one of the withdrawal kanbans instead (the regular kanbans are often called *production kanbans* in two-card kanban systems). Then he carries the containers to the input store of stage $m+1$. The removed production kanbans are put into a box from which they are eventually collected by a worker who places them on the scheduling board of stage m. If the carrier finds fewer containers in the output store than he holds withdrawal kanbans in his possession, then he returns the extra kanbans to stage $m+1$ and puts them back into the kanban collection box in the input store of stage $m+1$.

The point in time when the carrier removes the withdrawal kanbans from the kanban collection box is determined by one of two different schemes: (1) fixed quantity, variable withdrawal cycle (type-3 material transfer), or (2) fixed withdrawal cycle, variable quantity (*periodic material handling*, type-4 material transfer). In the first scheme, the carrier removes the withdrawal kanbans when a predetermined fixed number of cards has accumulated. The length of the withdrawal cycle, that is, the time between consecutive material transfers, may therefore vary. With the second scheme, the carrier removes the kanbans periodically, following a predetermined fixed schedule. Here, the withdrawal cycle is fixed and the number of cards may vary. Note that type-3 material transfer is equivalent to type-2 material transfer if the fixed withdrawal quantity is set to one.

3

Literature Review: Models of Kanban Systems

In this chapter, we review the literature on analytical stochastic models of kanban-controlled manufacturing systems. The review reveals that several mathematically tractable analytical models exist, however, almost exclusively for the evaluation of single-product systems. For multi-product kanban systems, only a small number of models may be found.

Other reviews are provided by Uzsoy and Martin-Vega (1990), Berkley (1992b), Singh and Brar (1992), Groenevelt (1993), Huang and Kusiak (1996), and Akturk and Erhun (1999). They also include simulation studies and deterministic models.

Unless explicitly stated otherwise, the manufacturing facilities are assumed to be perfectly reliable, input material (first stage) is always available in the quantities needed, containers contain only good parts, no parts are scrapped, transportation times between stages and withdrawal times are negligible, all processes are stochastic, all random variables are mutually independent, each product is produced and demanded by one manufacturing facility only, and a single (main) input is permitted for each processing operation (other inputs are admissible if they are always available). Systems that deviate from these assumptions are listed as "different systems."

In the systems that we consider here as kanban systems, several single- and/or multi-product manufacturing facilities (in parallel) may belong to a production stage, but a set of kanbans always controls the production of exactly one product in exactly one manufacturing facility, and the manufacturing facility can only process one item at a time. Other authors have a different, more general understanding of kanban systems. Schömig (1997, Section 4.5), for example, analyzes systems in which the same set of kanbans is used for all products of a stage. Di Mascolo, Frein, and Dallery (1992, 1996) study systems in which one set of kanbans is used to control the production of one product in one production stage, but there may be several manufacturing facilities and intermediate buffers in the production stage, arranged in series or in more complex formations (also De Araújo, Di Mascolo, and Frein 1993, Duri, Frein, and Di Mascolo 1995, Di Mascolo 1996, and Baynat et al. 2001). Hence, more than one item of a product may be being processed at the same time. In the review, we consider the results presented in these papers as far as they apply to systems with one manufacturing facility per stage.

3.1 Single-Product Kanban Systems

3.1.1 Single-Stage Systems

In this section, we review articles on analytical evaluation procedures for single-product kanban systems with one kanban-controlled production stage. A comprehensive presentation with additional system variations may be found in Buzacott and Shanthikumar 1993 (Section 4.3) under the label "single-stage single-product-type produce-to-stock systems."

Limited number of backorders. Jordan (1988) notes that closed-form results for queueing networks with exponential and Erlang distributions may be used to obtain steady-state performance measures for single-stage single-product kanban systems in which the demand is generated by a downstream manufacturing facility that is never blocked (it is always producing, provided input material is available). In this system, the maximum number of backorders is one ($B^{\max} = 1$) because the consuming manufacturing facility stops generating additional demands—as it stops producing (and thus consuming inputs)—when it lacks one container of input material (this is the situation where the number of backorders is one for the kanban-controlled manufacturing facility). For the case that the processing times of the producing and the consuming manufacturing facility are exponentially distributed with identical container processing rates, he gives equations for the average production rate and for the average number of full containers in the system (including the container in the downstream facility). The results are exact.

Karmarkar and Kekre (1989) model single-stage single-product kanban systems in which withdrawal kanbans are used to control the removal of full containers from the output store of the manufacturing facility. The removed containers are transported to a different, physically separated storage area from which external demand is met. The maximum number of backorders is given by the number of withdrawal kanbans. This implies that no customer waits for a full container if none is available at the time of his arrival. Since at least one withdrawal kanban must be present in the system, the maximum number of kanbans is always greater than zero ($B^{\max} \geq 1$). Demand arrivals are assumed to be Poisson, and processing times are exponential. The authors observe that their model is a truncated (finite) birth-and-death process, and they derive closed-form expressions for the steady-state probabilities of the system states (Equation (7) in the paper should be revised to read $p(i) = \left[\rho^{N-i}(1-\rho)\right]/(1-\rho^{N+M+1})$, $i = -M, -M+1, \ldots, 0, 1, \ldots, N$). The results are exact. The model of the kanban system is equivalent to the model of the standard $M/M/1/N$ queueing system: the customers in the queueing system correspond to the active kanbans and backorders in the kanban system, and the maximum number of customers in the queueing system, N, is equal to the sum of the number of kanbans and the maximum number of backorders, $K + B^{\max}$, in the kanban system.

Unlimited number of backorders. Karmarkar and Kekre (1989) observe that single-stage single-product kanban systems with exponential processing times, Poisson demand arrivals, and an unlimited number of backorders may be described exactly by a semi-infinite birth-and-death process. The authors provide a closed-form expression for the steady-state probability distribution of the system. The results are exact. This particular infinite-state Continuous-Time Markov Chain (CTMC) is equivalent to the CTMC of the standard $M/M/1$ queueing system and, hence, the same closed-form expressions for the steady-state probabilities of the system states apply.

Kim and Tang (1997) employ Erlang-k distributed times between the activation of kanbans where the number of Erlang phases, k, is equal to the container size as they assume that demand for single *items*, as opposed to demand for single *containers*, arrives according to a Poisson process. Each container is filled with k items. The processing time for a container of items (including a setup time) is assumed to be exponentially distributed. The model of this kanban system is equivalent to the model of a standard $E_k/M/1$ queueing system. This type of queueing system is also known as "bulk service system" since the server provides service to bulks of size k. Closed-form expressions exist for the steady-state probabilities of the states (e.g., Kleinrock 1975, Section 4.4). The results are exact.

Seki and Hoshino (1999) use the equivalence to an $M/E_k/1$ queueing system ("bulk arrival system") to model single-product single-stage kanban systems with Erlang-k processing times and Poisson demand arrivals. The authors assume that demand occurs for single items, and that a kanban is attached to each single finished item. The same queueing model may be employed to analyze single-product single-stage kanban systems with demand occurring for containers of size k and identically distributed exponential *item* processing times (hence, Erlang-k container processing times). Closed-form expressions exist for the steady-state probabilities of the states (e.g., Kleinrock 1975, Section 4.3). The results are exact. Besides the stationary behavior of the kanban system, Seki and Hoshino (1999) mainly focus on the *transient* system behavior.

Different systems. So and Pinault (1988) consider single-stage single-product kanban systems in which a minimum number of kanbans must be active before processing may start (threshold policy). They assume exponential container processing times, Poisson demand arrivals (for containers), and an unlimited number of backorders. The authors observe that their system is equivalent to an $E_k/M/1$ queueing system. For systems with general demand interarrival and processing times, they propose to use a scaling factor in the calculations based on the coefficients of variation of the processing time and the demand interarrival time. They suggest an expression for this scaling factor based on empirical results. In addition, the authors describe further modifications to include the possibility of machine breakdowns and multiple parallel manufacturing facilities in the stage. The results are exact for systems with one manufacturing facility, no breakdowns, exponential container processing times, and Poisson demand arrivals (for containers). For all other systems, the results are approximate.

Jordan (1988) sketches a discrete-time Markov chain for a basic assembly system with two kanban-controlled supplying manufacturing facilities and one assembly facility that is never blocked. As soon as one supplier is out of stock, the demand process is interrupted for both supplying manufacturing facilities. Hence, the maximum number of backorders is one for each supplier. The results are exact if the processing times of the modeled system follow a geometric distribution.

Wang and Wang (1990, 1991a, 1991b) model single-stage single-product kanban systems for which demand is generated by a manufacturing facility that is never blocked. In contrast to most other kanban systems, production is not controlled by cards, but by a limited number of containers, although production and withdrawal kanbans do circulate in the system. The manufacturing facility may start processing items only if an empty container is available. The consuming stage releases a container only after the last item has been removed. In this type of kanban system, the maximum number of full containers in the output store equals the number of containers minus one. As long as at least one container is partially filled, this container resides in the consuming stage. A backorder condition exists when all containers are empty. The authors assume exponential processing times in both stages and model the system as a CTMC. They determine the steady-state probabilities of the system states by solving the balance equations of the CTMC. The results are exact. Nori and Sarker (1998) note that the model is equivalent to the standard $M/M/1/N$ queueing model. They suggest to use the closed-form expressions available for this queueing model to find the steady-state probabilities of the states of the system.

Wang and Wang (1990, 1991b) also describe a CTMC for assembly systems with several kanban-controlled manufacturing facilities supplying parts or subassemblies to one assembly facility. The processing times are assumed to be exponentially distributed. The authors solve the balance equations numerically to obtain the steady-state probabilities of the system states. The results are exact.

3.1.2 Two-Stage Systems

Type-1 material transfer, no backorders. Karmarkar and Kekre (1989) illustrate a finite-state two-dimensional CTMC for two-stage single-product kanban systems with exponential processing times and Poisson demand arrivals. Demand that cannot be filled immediately is lost (no backorders), and the second stage withdraws material from the output store of the first stage just before start of production (type-1 material transfer). Since no closed-form solution exists for this CTMC, the authors