MODELS OF HEIJUNKA-LEVELLED KANBAN-SYSTEMS

KAI FURMANS
IFL, University of Karlsruhe (TH)

1. ABSTRACT

Heijunka is a key-element of the Toyota production system. It is used to level the release of production kanbans in order to achieve an even production program over all possible types of products thus reducing or eliminating the bullwhip effect. In this paper the dependencies between the parameters of the levelling process and the requirements in finished goods stocks are derived, resulting in an instrument for the sizing of heijunka controlled kanban loops.

2. PROBLEM DESCRIPTION

Queueing networks have been extensively used to model and study control policies for production systems. Especially prominent are the works about CONWIP- and kanban systems. Both keep the level of production and the work-in-process constant. While CONWIP keeps the WIP constant by releasing a new production order when one production order is finished, kanban also takes into account the demand process, releasing a new job when demand has occurred and the number of jobs in process has not reached its upper limit. Extensive work is available about models of these two control systems.

The Operations Research community has focused on these two basic types of production control. At Toyota and many of it's suppliers, kanban control is combined with a sequencing discipline, called heijunka. Heijunka is a part of the Toyota Production system which levels the production of different products evenly over a defined period, which could be a day, a shift or less. The goal is to achieve a constant flow of parts in a mixed model production which supplies one or more customer processes with a constant flow of different parts.

At the same time, a constant demand of parts is generated for the upstream processes, thus reducing or eliminating the need for spare capacity or stocks to cope with peaks of demand.

Introducing heijunka requires the determination of a suitable base period. Within this base period, it must be possible to produce all required parts, including changeover times, expected downtimes and scrap. This base period is called the EPEI (= Every Part Every Interval). The length of the EPEI is an indicator for the capability of the production process. An in depth explanation of the determination of the EPEI and the development of the heijunka plan can be found at [3].

The system works as follows: The customer requests parts in regular intervals, possibly with kanban cards. The requested parts are taken from a finished goods inventory (often called "supermarket") and are shipped to the customer. The same number of kanban cards (usually one per shipped container with finished goods) is sorted into the heijunka board. According to the number of parts which should be produced in every base period, spaces for kanban cards of the respective product are reserved in the heijunka board. If more

This work has been supported by the D. Hübner Stiftung.

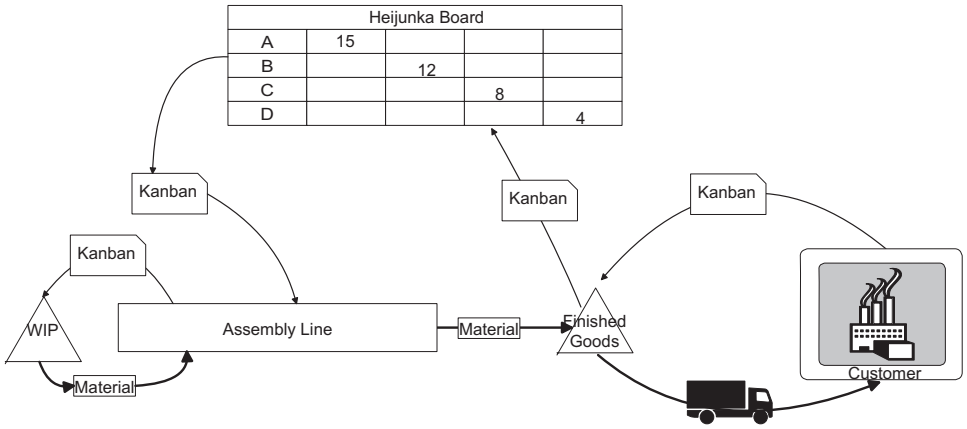

FIGURE 1. Heijunka as part of a manufacturing system

kanban cards are generated due to higher demand, the excessive cards are stored in an overflow location. If not all spaces for the kanban cards for a specific products can be filled, cards from the overflow location are added, if they are referring to the same finished product. Allocated space which can not be filled stays empty in order to avoid to produce goods, which are not requested.

A formal description of the heijunka system including the finished goods stock and the kanban system for the assembly line will be given in the subsequent sections.

3. HEIJUNKA FOR THE ONE PRODUCT CASE

Although Heijunka is intended to be used in an environment, where several different products are produced, the analysis begins with a very simplified case, where one product is assembled for one customer. In the beginning, the assembly line is treated in isolation, serving a market with independent demands (see figure 2). It will serve as a basic model for the next step, where the customer is using a kanban-system in order to control his parts supply.

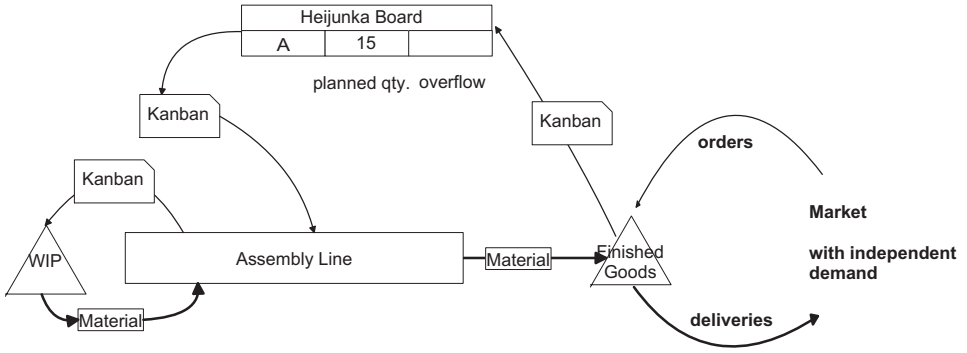

FIGURE 2. Simplified Heijunka-system for a merket with independent demands

3.1. **Basic model with unlimited capacity - no Heijunka**

It is assumed, that the products which are assembled, are distributed into a market, where demand is stochastic and the realizations of the demand per interval are independent of each other but identically distributed. The system is operating in intervals (i.e. days) and the system status is recorded for each interval. The interval has a length t_I.

The demand in interval n is given as a vector of probabilities $\vec{d_n}$, whose elements $d_{n,i}$ indicate the probability, that the demand in one interval is exactly i items. It is assumed, that the demand process is stable, so that the realizations of $\vec{d_n}$ are i.i.d. .

The fluctuations in demand are compensated by the shipping buffer for finished goods. Its dimensions are determined as follows: It is assumed, that the replenishment time t_r through the assembly time is guaranteed and it is measured in multiples of the basic interval t_I. If the capacity of the assembly line is not limited and it is controlled like a base stock system, where the demand Q_n in interval n is started at interval $n+1$ as a production lot, the finished goods stock will be replenished at interval $n+1+t_r$ with the quantity Q_n.

In each interval the finished goods stock therefore develops as follows: The stock decreases by the number of items Q_n which have been picked up and increases with the number of items, which have been started at interval $n-t_r+1$, which is identical with Q_{n-t_r}. Since the realizations Q_n are samples from the same distribution, the distribution of the stock development $\vec{s_n}$ in interval n can be described by:

$$s_{n,i} = \sum_{j=0}^{\infty} Q_{n-t_r,j+i} \cdot Q_{n,j} \qquad \forall i = -\infty, \ldots, \infty; i \in \mathbb{Z} \tag{3.1}$$

With the assumed independency of the realizations Q_n and thus of the replenishment quantities, it can be generalized as follows:

$$s_i = \sum_{j=0}^{\infty} d_{j+i} \cdot d_j \qquad \forall i = -\infty, \ldots, \infty; i \in \mathbb{Z} \tag{3.2}$$

Using the number of replenishment intervals t_r between delivery of a quantity of finished goods and its replenishment, the distribution of the cumulative stock development during that time is the t_r-fold convolution of $\vec{s}$.

$$\vec{S} = \underbrace{\vec{s} \otimes \vec{s} \otimes \cdots \otimes \vec{s}}_{t_r} \tag{3.3}$$

The necessary base stock level is then determined using $\vec{S}$. The components of $\vec{S}$ are summed up, starting with the most negative index until the guaranteed probability for the availability of goods has been reached. Then the absolute value of the respective index S_i equals the amount of base stock needed for finished goods. The base stock level is the absolute value of the largest integer l where

$$1 - \text{ServiceLevel}_\alpha \geq \sum_{i=-\infty}^{l} S_i \tag{3.4}$$

is still satisfied.

Driving the assembly line in this manner generates a demand for supplies, whose fluctuation equals the demand fluctuations of the market. The variability of the demand is neither increased nor decreased.

3.2. **Limited Capacity - Heijunka controlled**

Now it is assumed, that the capacity of the assembly line is limited and that these capacities are planned using a very simple heijunka board. In order to set the heijunka board up, the expected market demand is estimated and results in planned quantity m per interval. All kanbans released by the sale of a product are placed in the planned capacity slot up to m kanbans. All exceeding kanbans are placed in the overflow slot.

If in one interval less than the planned quantity has been sold, the remaining slots are filled by moving cards from the overflow slot to the planned capacity slot until the planned capacity slot is completely used or the overflow slot is empty.

The planned capacity slot contains the number of kanbans which are then released to production. Therefore the daily released production quantity never exceeds m but can be less. The system now resembles a $GI|D|1$-queueing system in discrete time. The sizing of the finished goods stock now has to take into account, that the replenishment time consists of the sojourn time of a started lot through the assembly line and a waiting time of the associated kanban card in the overflow slot.

From this point of view it is also easy to see, that the capacity c of the assembly line per interval must exceed the estimated daily demand m, otherwise the system will not reach a steady state. Each interval, a comparison between available capacity c and sold quantity Q_n can be done, resulting in the stochastic variable X_n.

$$X_n = Q_n - c \tag{3.5}$$

Note, that the expected value of X_n must be negative, when the system should be able to reach a steady state.

The number of kanbans in the overflow buffer W_{n+1} in interval $n+1$ is determined from the previous state using:

$$W_{n+1} = \max\{(W_n + X_n); 0\} \tag{3.6}$$

Equation 3.6 is equivalent to Lindley's equation for queueing systems in discrete time. Instead of treating the waiting time in equation 3.6 the number of kanbans which are placed in the overflow buffer are determined. Thus this formula can be considered as the dual formulation of the classic queueing formula.

The same methods which are used to compute the distribution of the waiting time can now be used to compute the distribution of the number of kanbans in the overflow slot (see for instance [2], [4]). Using the methods of Jain and Grassmann, the distribution of the number of waiting kanbans with the associated probability can be computed efficiently, resulting in a vector $\vec{w}$, where w_i is the probability that the number of waiting kanbans in the overflow is exactly i.

If the replenishment time t_r is one interval, then the probability distribution $\vec{b}$ of the number of not yet replaced items in the finished goods stock is given by

$$\vec{b} = \vec{d} \otimes \vec{w} \tag{3.7}$$

In order to reach the desired service level, the finished goods base stock is then sized as in the previous section. The base stock level is the absolute value of the largest integer l where

$$1 - \text{ServiceLevel}_\alpha \geq \sum_{i=-\infty}^{l} b_i \tag{3.8}$$

is still satisfied.

In most cases, where heijunka is used, it can be assumed, that the replenishment interval is one interval, if the replenishment time is longer than one interval the number of produced items and waiting items are not independent and an approximation must be used. This approximation can be based on the distribution of the replenishment times, which consists of the waiting time in the overflow location and the replenishment time in the assembly line.

The waiting time in the overflow slot can be computed by using methods, which have been developed for the computation of the number of customers in a $GI \mid G \mid 1$-queue in discrete time. They can be found for instance in [1]).

Heijunka based levelling with capacity restrictions leads to a reduction of the variability for products which are supplied to the assembly line, since the maximum number of consumed items is limited, and intervals of lower demand are compensated from the kanbans in the overflow location.

3.3. **Limited Capacity - Heijunka and Kanban controlled**

The customer is producing a finished product on his assembly line with a takt-time $t_{takt,customer}$. It is assumed, that the long term average of finished goods, which contain the product produced on the assembly line under investigation, is q, with $0 < q \leq 1$. Thus the necessary takt time at the assembly line t_{takt} is $t_{takt,customer}/q$ under the assumption, that both lines are operated in the same working time model, with the same availability and that the yield of the assembly line is 100%.

If the products are picked up by the customer (i.e. through his milkrun) regularly, with an interval between two successive pick-ups of t_I then on average at each pick-up

$$N = t_I/t_{takt} = \frac{t_I q}{t_{takt,customer}} \tag{3.9}$$

items are shipped to the customer. If the customer is not limiting or controlling the consumption of items, then the sizing of the finished goods stock can be done as in the preceding sections, using an estimated demand, which is derived from historic data.

If the pick-up quantity is controlled by a kanban-loop between the customer and his supplier the pick up quantity may vary between 0 and the number of kanban-cards in the loop between customer and supplier. If the number of kanban cards delivered per pick-up does not exceed the capacity c of the assembly line, and the replenishment time is one interval, then the finished goods stock has a base stock level which is equal to the number of kanbans in the loop with the customer. This way it is guaranteed that goods are always deliverable.

If the number of kanban cards exceeds the capacity limit, then an analysis of kanban controlled manufacturing systems can be done, using the relevant work of Dallery, DiMascolo, Frein as well as Shantikumar and Buzacotts book.

4. **HEIJUNKA FOR THE MULTIPLE PRODUCT CASE**

The multiple product case can be derived from the single product cases by analyzing the method which is used to allocate the production capacity on the different products.

When the EPEI is determined, the products are usually classified in categories, where the ”A”-products are produced each day, because the average daily consumption amounts at least to one handling unit. ”B”-products are, where the average weekly consumption amounts to at least one handling unit per week, but less than one handling unit per day. ”C”-products are the ones below the consumption level of the ”B”-products.

Each EPE-Interval, slots are allocated for all A-products, and the expected number of B- and C-products which have to be manufactured, because the last parts of a handling unit have been shipped.

The finished goods stock for the A-products can be determined as in section 3.2, since a fixed capacity for each product is allocated in each interval.

For the B-products, another approach has to be taken, since the B-products all share the same capacity slots. The queueing model that models the capacity situation for B-products has to combine the expected demand of all B-products. The resulting waiting time distribution in combination with the replenishment time can be used to determine the necessary stocks in the classic model with stochastic demand and stochastic replenishment times.

The same approach can be used for the C-products. For both categories however is the result less accurate as for the A-products and the stocks have to be sized with appropriate safety factors.

5. CONCLUSIONS

The introduction of heijunka levelling changes the character of the underlying system considerably. By using Lindley's equation for discrete time queueing networks and all derived work, an analysis of the system is partially feasible. More work should be directed towards the treatment of assembly systems with longer replenishment times and to the combination of this work with the relevant work on kanban-systems which are modelled in the continuous time domain.

Another interesting aspect is the integration of heijunka models in supply chain models. It is a possible method to decrease the bullwhip effect, but creates in turn finished goods stock, which is used to secure delivery when waiting times in the overflow location occur.

REFERENCES

1. A. Furmans, K. ; Zillus, *Modeling independent production buffers in discrete time queueing networks*.
2. J. L. Grassmann, W. K. ; Jain, *Numerical solutions of the waiting time distribution and idle time distribution of the arithmetic $gi \mid g \mid 1$ queue*, Operations Research **37** (1989), no. 1, 141–150.
3. R. Rother, M. Harris, *Creating continuous flow: An action guide for managers, engineers and production associates*, Lean Enterprise Institute (LEI), 2002.
4. P. Tran-Gia, *Analytische leistungsbewertung verteilter systeme*, Springer, Berlin, 1996.

E-mail address: `kai.furmans@ifl.uka.de`