Milk-run kanban system for raw printed circuit board withdrawal to surface-mounted equipment

Swee Li Chee, Mei Yong Chong, Jeng Feng Chin

Universiti Sains Malaysia (Malaysia)

rayley_winter06@hotmail.com, meiyongc@yahoo.com, chinjengfeng@end.usm.my

Abstract:

Purpose: The paper aims to present a case study and later simulation analysis on a kanban system that incorporating milk-run operation to draw in raw material to the process.

Design/methodology/approach: Data collection at the case study company for ten weeks followed by a process study called value stream mapping. The proposed kanban model is simulated to test its various performances including total output, average flow time, average work-in-process, SME utilization, and average waiting time. Response surface methodology is adopted to generate suitable representative regression models.

Findings: For all performance measures, simulation results showed that the proposed system consistently outperforms the push system currently practiced. Second, the system indicates the advantages of leveling, particularly in the event of machine failure and blockage. Third, operator in the proposed kanban system has a lower utilization, even with the additional material handling task.

Research limitations/implications: This study only begins to reveal the implication of leveling for production control on multi-machine scenario. The simulation of the system is solely based only the case study. The control parameters critical to the case study, were naturally used. The furtherance of the research should include generalizing the system and devising the respective methodology to facilitate wider applications.

Originality/value: The kanban system is proposed in the light of conflicting interests in handling the surface mounting and the related upstream processes. Such aspect is common to electronics assembly industry.

1. Introduction

The electronics assembly industry has today become increasingly challenging due to rapid technological changes, high material costs, and continuing trend of miniaturization (Tatsiopoulos, Avramopoulos, & Theoharis, 1997). In turn, high product variety and vastly different cycle times have given rise to the high keep of raw inventory and work-in-process (WIP).

Endeavours in reducing inventory, in the context of operation management, are often related to lean manufacturing. Lean manufacturing is about removing wastes from activities and processes that have no value-added on the products (Chen & Meng, 2010).

A distinct feature of lean manufacturing is that products are not pushed through the production but are rather pulled by customers (Liker, 2004). Along this line, demarcation can be established between push and pull systems. In a push system, jobs are queued in the system from entry to exit in order to process a complete product (Bonney, Zhang, Head, Tien, & Barson, 1999). This often entails an excessive amount of inventory, including WIP on the shop floor. On the other hand, the pull system is a method by which jobs are made when there is a request from the succeeding machine, effectively avoids overproduction.

The main component for the electronics assembly industry is the printed circuit board (PCB). Depending on the layout, different electronic components will be positioned and eventually bonded to PCB by surface-mounted equipment (SME). The handling of PCBs from warehouse to SME necessitates consideration of aspects appertaining to the PCBs. These include their brittleness, collective weight, and bulkiness. The duration of a PCB's exposure to ambient environment once unwrapped also has to be restricted to reduce environmental contamination. Generally, SME is the bottleneck process in an electronics assembly. Management constantly demands all raw materials, including raw PCBs, to be immediately available to SME through buffering. This ensures a continuous process at the equipment, although with a considerable higher inventory.

In this light, a unique pull system is conceived from a related industrial case study. The heart of the system is a machine-specific kanban system. The system incorporates a milk-run delivery operation to regulate both the kanban and material flow. The objective of the present paper is to present the inception of the pull system from the current case study and the analyses of system performance based on simulation study. The paper begins with a short review of pull systems and the milk-run method, followed by the research methodology. The industrial case study and a detailed description of the pull system are then presented. The

succeeding section presents the simulation models that have been developed. The performance results and analysis of the models are given next. Finally, the last section concludes the study.

2. Literature review

Pull system is one of the crucial elements in just-in-time (JIT) manufacturing (Ou & Jiang, 1997). JIT was introduced by Toyota in Japan as an effective production control system to produce the right quantity of parts for the right location at the right time with competitive cost (Savsar, 1997). In a pull system, the start of a production is triggered by the completion of another at the downstream station (Spearman, Woodruff, & Hopp, 1990). Therefore, a loading schedule is not required. The pull system heavily relies on visual management, which allows prompt detection of operational problems, e.g., machine failure.

The most basic form of a pull system is the supermarket pull system. Supermarket is a figurative term given to WIP storage with finite capacity between two processes. Each depleted buffer at the supermarket is replenished by the upstream process upon a withdrawal by the downstream process. The replenishment is just the amount required by the downstream process (Rother & Shook, 1999).

Kanban, a widely known pull system, is a card. Signals initiated by customer orders flow backward and pass the cards through each workstation (Spearman et al., 1990; Chan, 2001). An order is released only when the parts or raw materials have a card authorizing production (Spearman et al., 1990). Thus, the parts are being pulled through the system. Through this method, WIP between processes can be tightly controlled (Marek, Elkins, & Smith, 2001). For example, the WIP is limited to a certain number of kanban cards between the processes. No card will be authorized when a breakdown occurs in the downstream process.

By comparison, the Constant WIP (CONWIP) system (Spearman et al., 1990) is more flexible. The first workstation receives a card to produce a new batch of parts upon the completion of one batch when it exits the system. The new batch is then pushed to the subsequent workstations. The number of cards issued determines the WIPs in the system. The CONWIP system was studied by Bonvik, Cao and Chen (2005), and Takahashi, Myreshka and Hirotani (2005).

A Kanban–CONWIP hybrid proposed by Bonvik et al. (1997), combines the advantages of CONWIP with those kanban in order to limit the WIP level at each stage. Two authorization cards, kanban and CONWIP, are required. CONWIP cards control the number of orders processed in the system. Kanban cards maintain the inventory at each stage. The last stage of production does not involve kanban control because it is impossible for the quantity of the material to exceed the inventory allowed in the buffer of the whole line.

Kanban can be rather versatile as signal carrier and is not restricted to one format. In the industry, for two nonadjacent related process facilities where material handling times are significant, the dual-card kanban system is often adopted. The system consists of production kanban and withdrawal kanban (Matzka, Di Mascolo, & Furman, 2009). A withdrawal kanban defines the quantity that the following facility withdraws from the previous facility. A production kanban is triggered to replenish the quantity withdrawn in the previous facility. Pattern production imposes a fixed production sequence to replenish the input inventory of the downstream process up to a fixed level upon receiving authorization (Seidman & Holloway, 2002). Such a technique is suitable when an upstream process is a provider for multiple independent downstream processes. In a batch board system, parts are produced by batch with a display board containing an outlined shadow space for all part numbers. Triangle kanban is a triangle-shaped card with a specific display of information. It is generally placed at the trigger point in the inventory location. When it reaches the trigger point, it would be sent to the producing machine.

Milk-run is a quantity variable material conveyance and replenishment method, either time-fixed or event-triggered. The system is more well-known in logistics research. The idea is derived from the milk delivery business. Following a specific route, a milkman delivers milk contained in bottles to his customers. Upon delivery, an empty milk bottle is collected. The application of this system is prominent in auto manufacturing companies to minimize inventory and transportation costs. For example, a delivery is made more efficient when items such as empty containers are returned to the source on the return trip. In manufacturing operations, this method is often adopted for the material handling of two nonadjacent processes. The relationship between the two processes is reciprocal. They both need output from each other. The form of output can be varied, such as material or instruction. A milk-run example can be given in Horbal, Kagan and Koch (2008) model. The milk-run system complements the designed pull system for high-mix production environments. Material handlers are responsible for collecting kanban cards and, at the same time, replenishing raw material to all the work cells. Milk-run has been reported to be suitable for repetitive production where the component type is seldom changed. The heijunka kanban system modeled by Matzka et al. (2009) involves milk-runs between customer and supplier. The supplier supplies parts based on the number of kanban cards received. Unsatisfied demands then have to wait until the next milk-run for the parts to be available. In a different scenario, Sadjadi, Jafari and Amini (2009) optimized milk-run operations with genetic algorithm by first modelling them as a mixed integer problem.

3. Methodology

Figure 1 illustrates the steps undertaken for the research methodology. As a prerequisite, the purpose of the study is determined by the problem description provided in the introduction. Value stream mapping (VSM) is constructed to show and understand the flow of material and information.
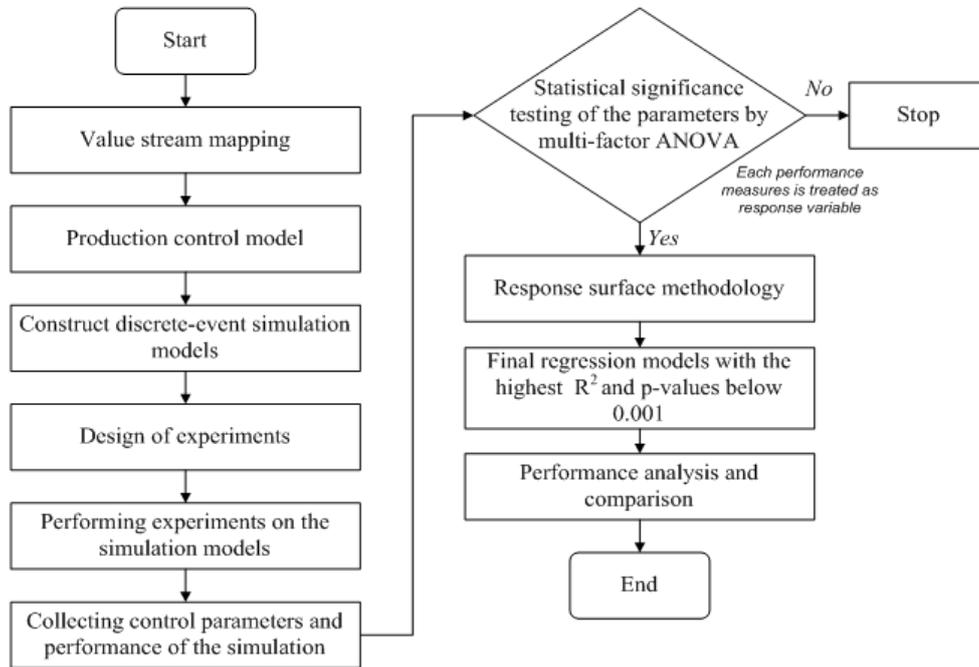

Figure 1. Flow chart of the research methodology

VSM is a visualization tool for analyzing the flow of materials and information through a shop floor; it focuses on one product family only. Two maps, current-state value stream and later future-state value stream, are drawn. A set of standardized icons is prescribed to represent each process in the shop floor. The current-state VSM is essentially for seeing and understanding the existing flow of material and information being done to produce the product family. The system is then analyzed to identify the value-added or nonvalue-added process. Next, a future-state VSM is created. It is the improved state of current-state value stream after the nonvalue-added processes have been removed.

A new production control model is designed. To test its performance, a discrete-event simulation is built. Control parameters and performance measures are decided based on their relevance and criticality to the real environment. Likewise, the selection of values and ranges of control parameters must be realistic.

In designing the experiments, full factorial experiments are applied. The experiments cover all possible combinations of values assigned to the control parameters. Such an approach requires

basic programming but obtains comprehensive results. However, running all the experiments is time consuming. The effects and interactions of control parameters on each performance measure are analyzed by multi-factor analysis of variance (ANOVA). A control parameter or interaction with statistical significance on a performance measure warrants further analysis with response surface methodology (RSM).

The control parameters and performance measures in RSM are called input variables and response variables, respectively. In RSM, several input and response variables are mapped onto a surface contour plot followed by the regression analysis. Low-order regression models, notably order of one and two, provide an approximation response function. The R^2 and p-values associated with the regression models are then computed. The regression model with the highest R^2 and acceptable p-values would be used in the performance analysis and comparison.

In the aspect of programming, the simulation model consists of an auto-executed experiment engine of three shells: innermost shell, middle shell, and outermost shell. Innermost shell provides the setting for predesigned discrete-event simulation models in Witness 2008. Middle shell runs the simulation models iteratively. The multi-factor ANOVA test on data extracted from the simulation models follows. The outermost shell seeks to obtain response functions by RSM. The middle shell and outermost shell are written in Visual Basic 6.

4. Case study

Company X is an electronics assembly plant in Malaysia. Two decades ago, it started production as an automotive electronic components developer and manufacturer. With a total of 592 employees, the plant operates three shifts per day, six days per week (overtime on Sunday, if there is backlog). Lean manufacturing has been practiced by the company for years.

4.1. Value stream mapping

Current-state VSM

The process flow and observation for current-stage VSM are illustrated in Figure 2. The raw PCBs are prepared by the warehouse personnel based on the schedule released by the scheduler. On average, eight orders are prepared per shift. The actual delivery of raw PCBs to the shop floor, however, is authorized by the kitting group leader based on a separate SME schedule. The schedule is adjusted with the SME current conditions and space available in the standby area. There are six SMEs on the shop floor. The lead time to deliver raw PCBs from the warehouse to the production shop floor is about three hours, although this may vary widely. The raw PCBs are kept in bins and delivered to the material acceptance station for verification, which is done by scanning the barcodes of the packages of raw PCBs. On average, a package contains approximately 60 raw PCBs. Two bin sizes are used, one sized 292 mm x 394 mm x 166 mm and another sized 394 mm x 592 mm x 166 mm. The smaller bins are used to keep marked PCBs, whereas the larger bins are used to keep raw PCBs. After verifying,

each bin is moved to the standby area. The raw PCBs are grouped and stacked together according to model, and the bins are labeled according to its model and raw PCB number.

One operator is responsible for both laser marking (LM) and manual marking (MM). Based on the SME schedule, the operator marks the raw PCBs of the next model. For the LM machine, the raw PCBs are loaded onto the PCB holders. Subsequently, a number of set-up steps are followed, e.g., installing fixture, setting the vacuum nozzles, and program loading. The marked PCBs are unloaded to empty magazines, each holding approximately 40 panels. For MM, labels are printed out and manually attached to individual raw PCBs. A marking code of a raw PCB is later scanned for record purposes. The marked raw PCBs will either be sent to the specific SME directly or stored in a nitrogen chamber.

To handle the raw PCBs, protective gloves or finger coats need to be worn to prevent oxidization by finger prints. Unless the raw PCBs are kept in the nitrogen chamber or dry cabinet, prolonged exposure to the environment causes further oxidization.

A high number of lean wastes are observed, particularly in the categories of overproduction and inventory. The details are given below:

- An excessive amount of raw PCBs are kept in the standby area. As raw PCBs are prepared much earlier than their actual use, standby time can be more than two days. Approximately 36 models are kept in 82 large bins in the standby area. In the few days of observation, a daily average of 14 models is drawn in by SMEs. Rearrangement of bins is necessary to accommodate the influx of inventories in the standby area. Searching for the needed model is thus time consuming and, sometimes, help from a second operator is required. Mishandling and misplacement of labels are common. In addition, a fully loaded large bin, much heavier than a small one, usually exceeds the safety weight limit of 20 kg

- Instead of the exact quantity needed, the warehouse prepares raw PCBs to the nearest packed quantity and delivers them wrapped. The rationale is to protect the raw PCBs until they are used. With varying models and quantities to produce, verification is needed. The process is laborious and error-prone. The unused raw PCBs are kept in the standby area temporarily, to be returned to the warehouse after back flushing

- The flow from marking to SMEs is asynchronous. Forward scheduling is applied in the marking process, resulting in frequent overproduction
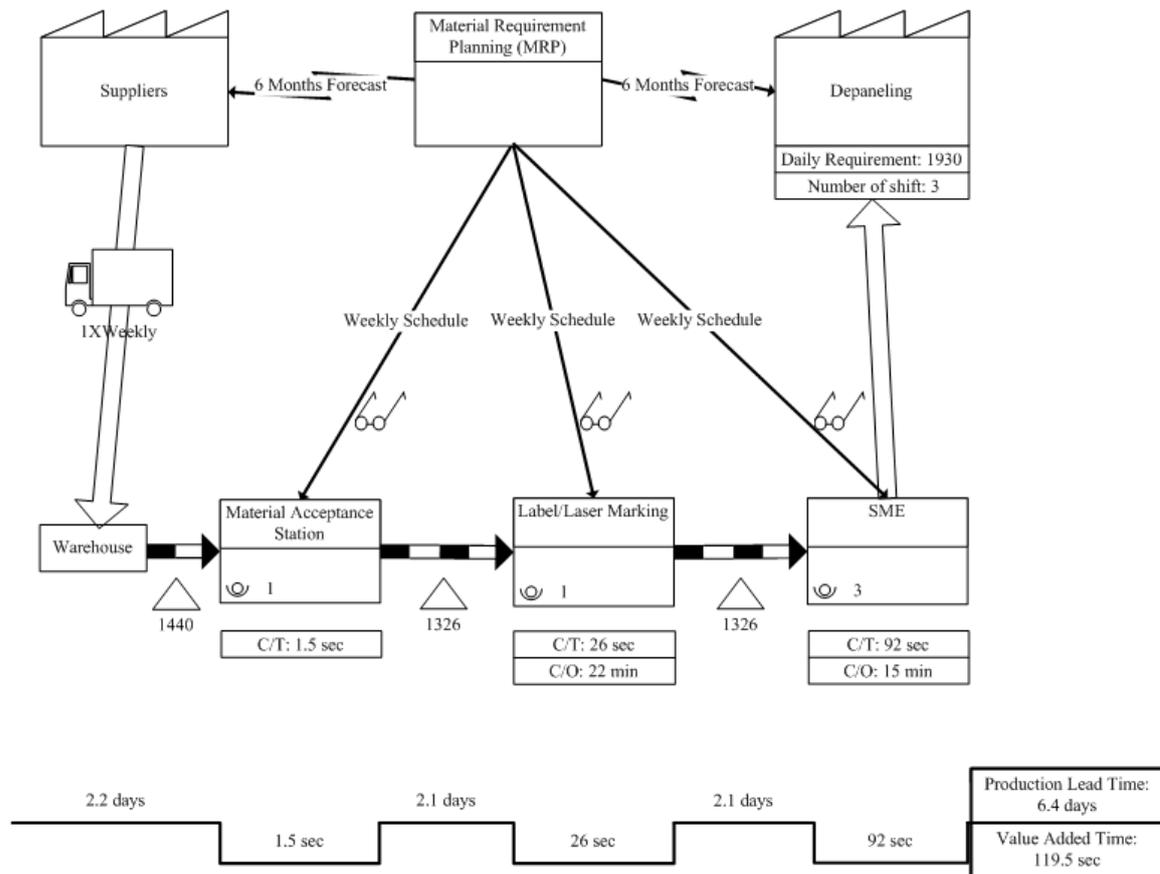
Note: C/T= cycle time; C/O=changeover time

Figure 2. Current-state VSM

Future-state VSM

A future-stage VSM is shown in Figure 3, where a pull system is proposed. The first requirement of the pull system is to regulate the flow of raw PCBs from the warehouse to the marking process, and, finally, to the SMEs. Since the marking process is not considered bottleneck, the feasibility of milk-run operation by the operators in the marking process is investigated. The details of the pull system are explained in the next section. In conjunction with this, the processing lot size is reduced and standardized, which prevents overproduction and reduces the inventory considerably. In turn, retrieval of raw PCBs and verification in the standby area are improved.

In the standby area, the raw PCBs are arranged according to the SMEs on the newly proposed roller racks by first-in-first-out (FIFO) basis, which eliminates searching and sorting when retrieving one particular model. For safety reasons, the small bins are used for storage. The labeling system is redesigned to minimize mishandling.

Supermarkets, which are installed with marking machines and SMEs, are separated into dedicated FIFO lines based on the number of machines. Thus, there are two FIFO lines for the marking process and six for the SMEs. A FIFO line serves as a buffer when the items stored are drawn on FIFO basis.

With this installation, overproduction would be eliminated and production flow would become smoother. The nitrogen chamber and dry cabinet can be removed entirely, resulting in more space and significant savings on maintenance cost.
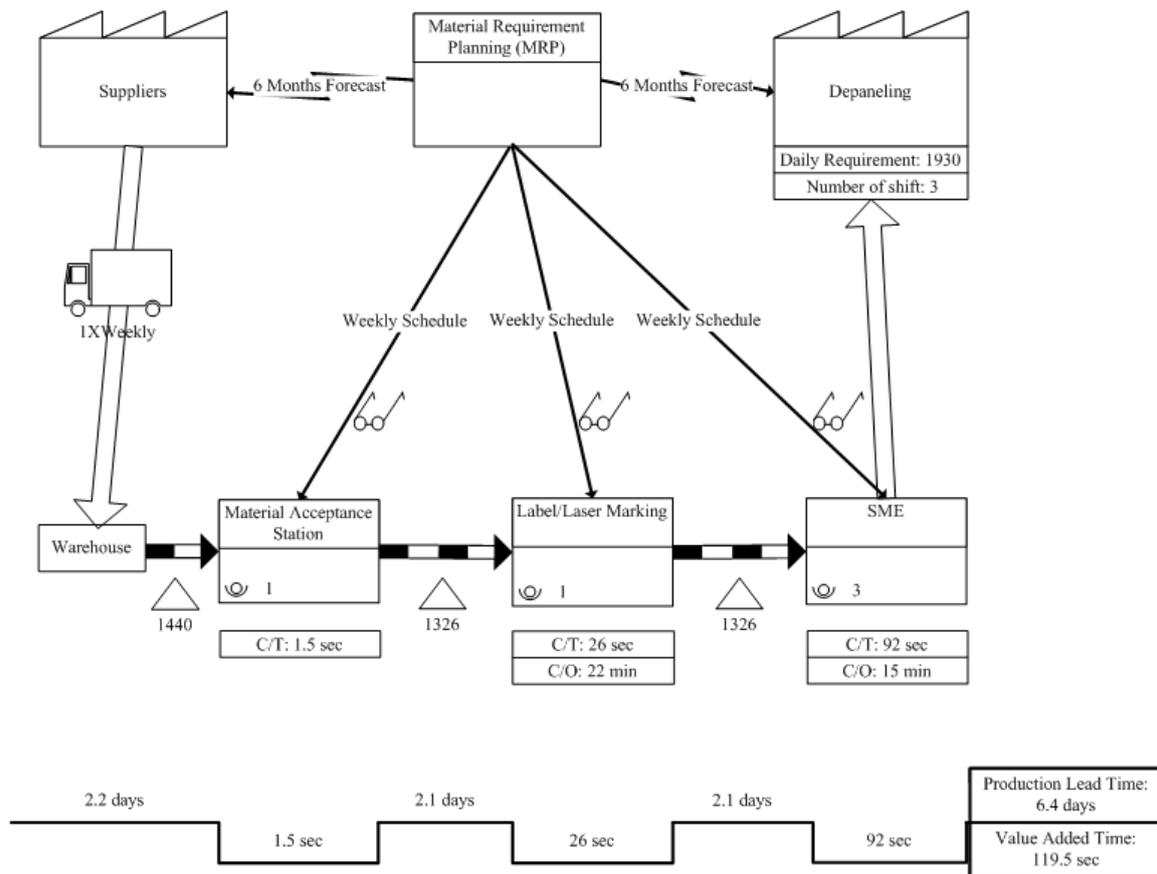
Note: C/T= cycle time; C/O=changeover time. The lead is estimated based on on n_k=2, n_r=2, n_s^{LS} =200, n_{op}=1 (notations are given at the later section).

Figure 3. Future-state VSM

The simulation study is conducted on the pull system to gain some understanding of its behaviors on the shop floor.

4.2. Detailed mechanism

A formal description of the milk-run kanban system (MRKS) has been provided. The manufacturing process (MP) consists of three distinct departments of operation, denoted respectively as D1, D2, and D3 (Figure 4). D1 consists of one warehouse, D2 consists of two heterogeneous parallel machines, denoted as LM and MM. D3 has six heterogeneous parallel machines, denoted as SME(1), SME(2), SME(3), SME(4), SME(5) and SME(6), for the surface-mounting process. Two supermarkets, B1 and B2, are installed as buffers. In B1, two FIFO lines are each responsible for a marking machine in D2. In B2, six FIFO lines are dedicated to individual SME machines. B1 acts as the input storage between D1 and D2, and B2 between D2 and D3. Although there is no capacity limit for FIFO lines, a limit is indirectly imposed through the number of kanban cards in circulation, as explained in the following section.

In MP, the flow of raw PCBs, in predetermined lot size, commences from D1, proceeds to one FIFO line in B1, and then to a selected marking machine in D2. Later, the lot is transferred to a FIFO line in B2, and finally reaches one SME in D3.

A single kanban card type, ks, is adopted. Each ks carries a standard lot size (nsLs). The movement of a lot must have an authorized ks attached to it. To explicate the flow of the kanban system, let us assume that FIFO lines in B2 have lots for individual SMEs in D3. Once an SME machine in D3 retrieves a lot from a FIFO line in B2, the ks attached to the lot is removed and collected. Upon collecting nr of kanban cards, they are immediately sent to D1 by an operator (highest priority). The kanban cards signal the preparation of the next lot based on the pending orders. The preparation requires time Tp for one lot independently. To replenish the base stocks from which the kanban cards are drawn, the lots of the orders chosen must go through the same particular SMEs. This restriction makes the kanban signal machine-specific. When a lot is ready, it will be attached with a kanban card and together they are waited for in D1.

A milk-run operation is incorporated in the process. On the return trip, the operator will retrieve any lots that have been prepared and place them at B1. When a prepared lot exceeds twice the duration of Tp, it will be pushed to B1. After the process in D2, the lot and the kanban card will pass to the designated FIFO lines in B2, therefore completing one circle of kanban flow.

The pull system prevents the accumulation of inventory in MP. If a machine breakdown occurs in a downstream machine, e.g., in D3, no kanban card will be issued from that machine, thus preventing further WIPs to be pushed to that location. If a breakdown occurs in the upstream machine, e.g., in D2, operations in D3 will continue as long as there is a kanban card received and lots are available in B2. In this scenario, the buffer in B2 provides crucial slack for the broken down machine to be restored. This prevents idling at the other machines.

5. Simulation and experiments

5.1. Simulation models

Two simulation models are constructed. The first model is based on the currently practiced push system. The second model is MRKS. The push system does not contain any kanban cards. Instead, orders are drawn from the upstream buffer for process. The completed lots are then sent to the downstream process. D1 transfers the lots to B1 soon after preparation without any help from the shop floor operator.

Each simulation model contains approximately 36 elements and has a run time of 15 days (three 8-hour shifts totalling 1,296,000 seconds). The warm up period is one day (86,400 seconds). A screen shot of the simulation model is given in Figure 4. All combinations of

control parameter values are investigated. A combination is a run with five iterations of different random seeds.
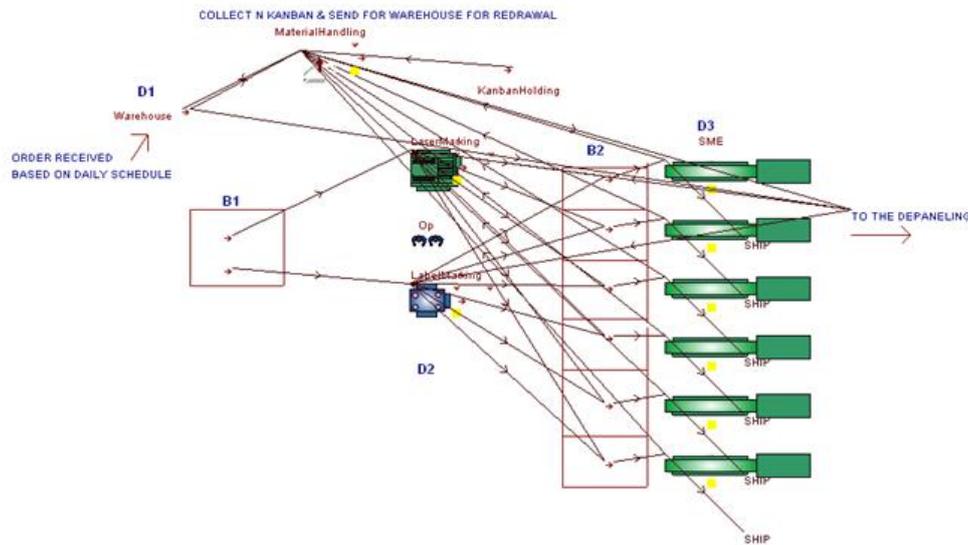

Figure 4. MRKS simulation model with two operators (n_{op}=2)

5.2. Determining the input parameters and performance measures

The control parameters are as follows:

1) n_s^{Ls} : Standard lot size is the fixed number of parts in one batch.

2) n_r: Number of kanban cards to be collected before delivering them to D1.

3) n_k: Number of kanban cards allocated for each machine in D3; it also implies the maximum number of lots that can be contained for each machine at B2 as WIP.

4) n_{op}: Total operators in the system.

5) n_{bf}: Maximum buffer size at B2 for individual SMEs in the number of orders held.

In the present paper, n_r and n_k are applied to MRKS only, whereas n_{bf} is applied to the push system only. The selected value sets for the control parameters are listed in Table 1.

Control parameters	Values
n_s^{Ls}	50, 100, 200
n_r	2, 4, 6
n_k	2, 4, 6
n_{op}	1, 2, 3
n_{bf}	2, 4, 6

Table 1. Control parameter value list

Other input information used are as follows:

1) T*he total units received daily* have to be set to prevent the MP from lacking orders, therefore creating a bias on the results later obtained. Every 8-hour cycle (one shift), new orders of approximately 3000 units would be sent to the MP. Afterward, the number of units in each order is determined. If the accumulated unit assigned to the orders at one point during the assignment exceeds 3000, the remaining unassigned orders are discarded. The orders represent products according to the demand (Table 3).

2) *Processing and set-up time* for each product is shown in Table 3. The set-up time is incurred at every product type change (as observed in the industry) even if the following lot is from the same product family.

3) *Machine breakdown and repair times* are modelled with Witness build-in distribution. Negative exponential distribution is used to simulate machine breakdown, whereas Erlang distribution with k = 3 is best fit for machine repair times. Both distributions are commonly recommended in literature for such purposes. The machines' individual mean time between failures (MTBF) and mean time to repair (MTTR) are given in Table 2.

4) T_p: Raw PCB preparation time at D1 which is fixed at five hours

5) T_s: Time required for an operator to send the kanban cards to D1, set at 60 seconds

6) T_m: Material handling time for operators travelling from D1 to B1, set based on the load carried by operators. In a single trip, a maximum of 500 PCB pieces can be transported, requiring 180 seconds. If no load is bought back to B1, Tm is set at 60 seconds; otherwise, Tm is set at 180 seconds at the start of 500 pieces of raw PCB, with an increment of 240 seconds for every subsequent 500 pieces of raw PCB.

Machines	MTBF	MTTR
MM	-	-
LM	47,520	20,760
SME(1)	29,088	2,878
SME(2)	41,328	3,014
SME(3)	27,936	1,849
SME(4)	37,728	2,511
SME(5)	39,600	2,948
SME(6)	34,560	3,429

Table 2. MTBF and MTTR

Product	Product demand (%)	D2				D3		
		machine	St	ct	lt	machine	St	ct
P1	0.03	MM	264	15	60	SME(1)	1,800	56
P2	0.38	LM	1,037	17	80	SME(2)	900	58
P3	0.03	MM	264	17	60	SME(2)	900	52
P4	0.15	LM	1,037	26	80	SME(3)	900	103
P5	0.15	MM	264	17	60	SME(3)	900	46
P6	0.07	LM	1,037	34	80	SME(4)	900	84
P7	0.02	MM	264	30	60	SME(4)	900	91
P8	0.10	LM	1, 037	26	80	SME(5)	900	88
P9	0.08	MM	264	14	60	SME(6)	1,800	53

Note: *st* = set-up time (in seconds); *ct* = cycle time (in seconds)
lt = loading time (in seconds)
Table 3. Processing and set-up times

The following assumptions are made of the models:

1) The change of shift or break does not result in machine stoppage or productivity loss.

2) The machines can only process a new batch of parts when all the parts in the previous batch have been sent to the buffer of downstream process.

3) There is no capacity limit at the warehouse.

4) There is nonstop production during the simulation runs (in real case, the factory is closed on Sundays).

To gauge the overall functionality of the models, several performance measures are chosen.

1) *Total output:* The number of parts which exited the system.

2) *AvFT(part):* Average flow time (average production flow time) is the average duration of the completion of a part. It is measured from the time the order is prepared at D1.

3) *AvWaitatB2(order):* Average waiting time of an order at B2.

4) *AvWIPatB2(order):* Average WIP at B2. The WIP is measured by the number of orders.

5) *Util(SME):* The percentage of time the SME is used over the course of the simulation run, excluding breakdowns, delays, and set-up times. The set-up time is included as part of the processing time due to technical difficulties of separating them.

6) *Util(Op):* Average utilization of operators.

6. Results and discussions

After the experiments, the results are analyzed by multi-factor ANOVA. Next, box plot analysis is used to compare the range and relative median between the current situation and MRKS. For comparability between box plots, the output parameters are filtered to leave behind data with a selected nr value. Regression models, presented in a two-dimension graph plot, are used to study the system behavior with changes in control parameters. In the graph plot, the control parameters common to the studied regression models with the highest value of coefficient is used as determinant. Others control parameters are homogenized with the median from the corresponding value set found in Table 1.

6.1 Multi-factor ANOVA

Tables 4 and 5 present the results from the multi-factor ANOVA. The observed effects for the factors are shown. The factors are control parameters, whereas the effects are performance measures. The hypothesis built for each observation is that there is no difference in the mean effect in the groups tested. A result leading to the rejection of the hypothesis statistically confirms the presence of effect with the changes in the selected factor or interaction. The main factors are mostly considered significant in the analysis. The interactions along the main factors, however, are mostly insignificant. The absence of interaction between the factors implies that the effect of one factor is the same for all levels of the other factor.

	n_r	n_k	n_s^{Ls}	n_{op}	$n_r \times n_k$	n_r^2	$n_k \times n_s^{Ls}$	$n_r \times n_{op}$	$n_k \times n_{op}$	$n_s^{Ls} \times n_{op}$	$n_r \times n_k \times n_s^{Ls}$	$n_r \times n_k \times n_{op}$	$n_r \times n_s^{Ls} \times n_{op}$	$n_k \times n_s^{Ls} \times n_{op}$	$n_r \times n_k \times n_s^{Ls} \times n_{op}$
TotalOutput	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	Yes	No	No	No	No
AvFT(part)	Yes	Yes	Yes	No	Yes	Yes	Yes	No	No	No	Yes	No	No	No	No
AvWIPatB2(order)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	No	No	No	No	No
Util(SME)	Yes	Yes	Yes	No	Yes	Yes	Yes	No	No	No	Yes	No	No	No	No
Util(Op)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
AvWaitatD1(order)	Yes	Yes	Yes	No	Yes	Yes	Yes	No	No	No	Yes	No	No	No	No
AvWIPatB1(order)	Yes	Yes	Yes	No	Yes	Yes	Yes	No	No	No	Yes	No	No	No	No
AvWaitatB2(order)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	No	No	No	No

Note: *Yes = Reject hypothesis; No= Accept hypothesis*

Table 4**.** Summary of multi-factor ANOVA tables for the MRKS

	n_{bf}	n_s^{Ls}	n_{op}	$n_{bf} \times n_s^{Ls}$	$n_{bf} \times n_{op}$	$n_s^{Ls} \times n_{op}$	$n_{bf} \times n_s^{Ls} \times n_{op}$
TotalOutput	Yes	Yes	Yes	No	No	No	No
AvFT(part)	Yes	Yes	Yes	No	No	No	No
AvWIPatB2(order)	Yes	Yes	Yes	No	No	No	No
Util(SME)	Yes	Yes	Yes	No	No	No	No
Util(Op)	Yes	Yes	Yes	No	Yes	Yes	No
AvWIPatB1(order)	Yes	Yes	Yes	No	No	No	No
AvWaitatB2(order)	Yes	Yes	No	Yes	No	No	No

Note: *Yes = Reject hypothesis; No= Accept hypothesis*

Table 5. Summary of multi-factor ANOVA tables for the push system

Total output

In Figure 5, MRKS exhibits both the highest range and the total output compared with the push system. The highest output obtained from the MRKS is 75,600 units (run with the values n_s^{Ls} = 200, n_r = 2, n_k = 6, n_{op} = 2, 3rd iteration). For the push system, the highest output recorded is 41,000 units (also run with the values n_s^{Ls} =200, n_k = 6, n_{op} = 2, 3rd iteration). The total output of the MRKS is evidently much higher than that of the push system. The finding challenges the common notion that the attainment of higher output in the push system in environments where incoming orders are continuous and that material handling times are independent of buffer size. Simulations were rerun manually and observed. In the push system, when the individual B2 FIFO line reaches its limit incidental to a completed order at D2 for that line, the affected machine is blocked at D2. This frequently leads to machine breakdown at D3. In real life, the incident can be avoided by having the operator in charge at D2 process the orders for the other D3 machines during that period. Nevertheless, this requires additional micro-controlling. Similar incidents, however, are not observed in the MRKS. From the investigation, when a B2 FIFO line of the MRKS is full, all the associated kanban cards would have been allocated. This prevents further processing of similar orders at D2, thus preventing subsequent blockage. Machines at D2 are redirected to process orders for unaffected D3 machines.

A separate simulation is conducted to find the optimum total output for a perfect push environment. Set-up requirements, breakdown probabilities for all machines, limitation on B2 buffer sizes, kitting times, and even operator requirement are either relaxed or made insignificant. Surprisingly, the total output consistently falls between 66,000 to 70,200, which is still lower than the best value of MRKS. Most orders, as WIPs, have been observed to accumulate at the B2 FIFO lines of SME(2) and SME(3). Both are considered a bottleneck due to either the high proportion needed to process orders or the longer average cycle time. For the MRKS, using a fixed number of kanban cards for each downstream machine, in the cases, SME, provides some form of regulation and levelling to distribute the order to the other machines. This innate attribute prevents other machines from idling. The finding is important because kanban systems, as such, allow effective path divergence to prevent blockage; this, therefore, consistently results in higher output.

Now that the MRKS is confirmed to be superior over the push system, the effects of its control parameters are studied. The graphs generated from the best regression models are plotted in Figure 6. From the figure, the total outputs for both the push system and the MRKS increase initially. Both start to decrease after reaching an optimum standard lot size. From n_s^{Ls} =80 onward, the MRKS dominates the push system by generating higher total output. Generally, for both systems, the increase in n_s^{Ls}, diminishes the effect of set-ups at B1 and B2. Nevertheless, the reason for the decrements in their total output afterward is varied. For the push system, it is believed that when the n_s^{Ls} is larger, the effect of machine failure at D3 is

more severe. Machines are likely to fail earlier with a smaller number of orders processed. Coupled with poor levelling and larger n_s^{Ls}, the system encounters blockage more often, hence, the lower total output is explained. For the MRKS, larger n_s^{Ls} entails a slower and more intermittent replenishment pattern. The detached kanban cards have to wait for a longer period of time before they are delivered to the warehouse for the next batch of orders.
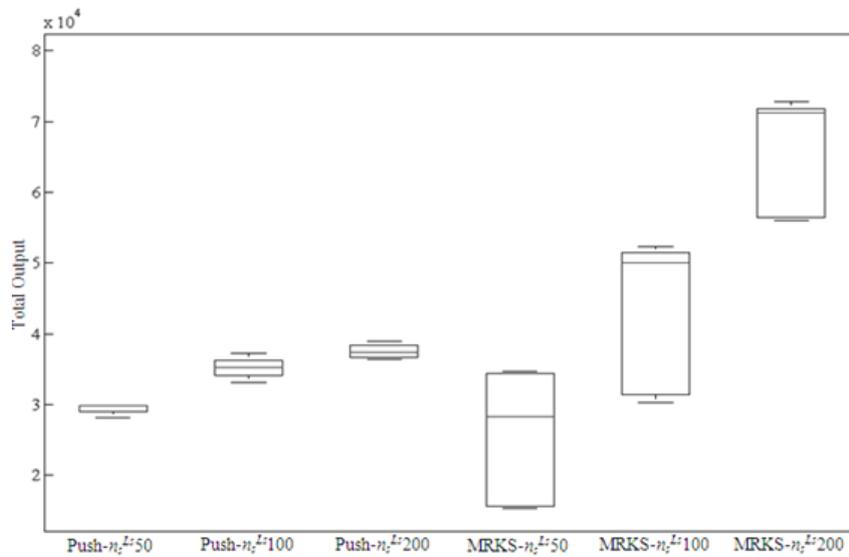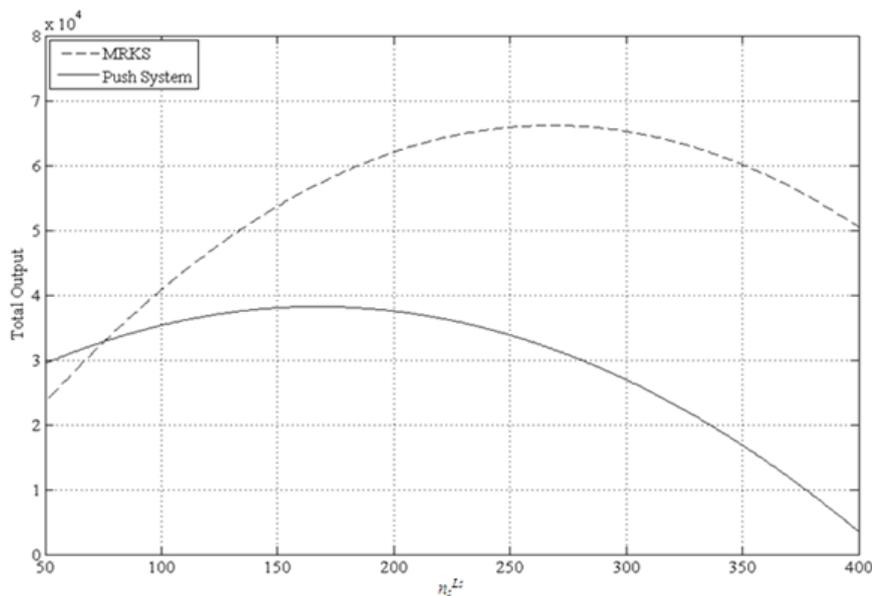

Figure 5. Box plot analysis on total output

Figure 6. Effect of n_s^{Ls} on total output

Average flow time per part [AvFT(part)]

In Figure 7, MRKS exhibits significantly lower *AvFT(part)* based on the location of the 50th percentile region. The lowest value obtained from MRKS is 307 seconds (run with the values $n_s^{Ls}=200$, $n_r = 2$, $n_k = 2$, $n_{op} = 1$, 3rd iteration). Their ranges, however, are largely indifferent. There is also a strong indication of declining *AvFT(part)* with bigger n_s^{Ls}. The trend is believed to be incomplete because the respective regression models (Figure 8) generated resembles a

catenary. The presence of a catenary commonly signifies the shifting of one dominant variable to the other. Initially, the *AvFT(part)* is reduced due to better machine utilization and the decreased effect of set-up over a larger lot. Eventually, the *AvFT(part)* is increased due to the long waiting time per unit in a lot.
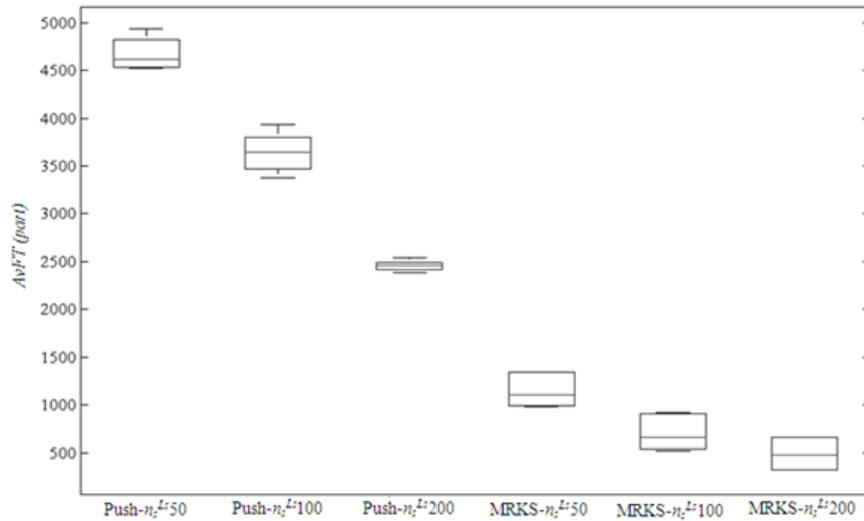
Figure 7**.** Box plot analysis on *AvFT(part)*
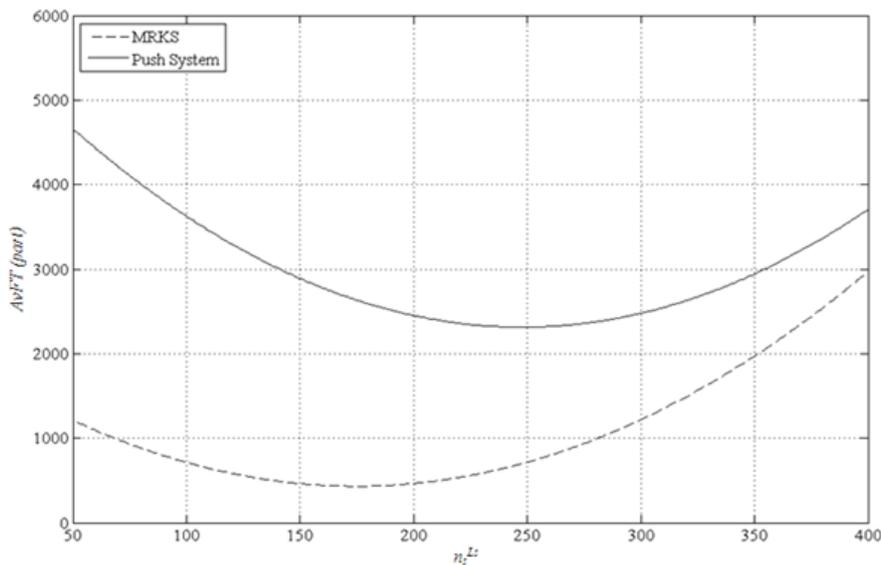
Figure 8. Effect of n_s^{Ls} on *AvFT(part)*

Average waiting time of an order at B2 [AvWaitatB2(order)]

AvWaitatB2(order) is a measurement to trace the average time exposure of the order to the environment. A part should not be exposed for more than 172,800 seconds (48 hours) to prevent oxidazation. All the systems do not exceed this limit. In Figure 9, the *AvWaitatB2(order)* in the MRKS is clearly lower than the push system. A similar finding is observed in Figure 10, where regression models approximating the systems are plotted. The MRKS consistently exhibits a much lower average waiting time for all n_s^{Ls} values.
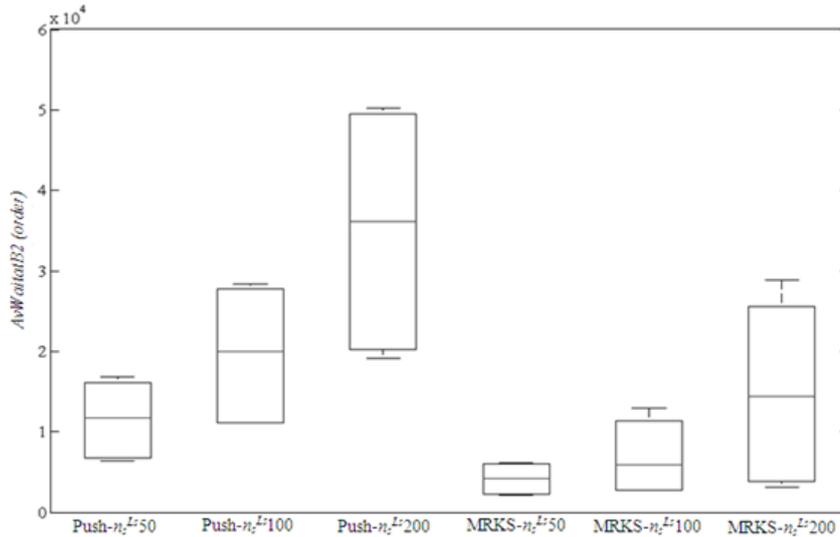
Figure 9. Box plot analysis on *AvWaitatB2(order)*
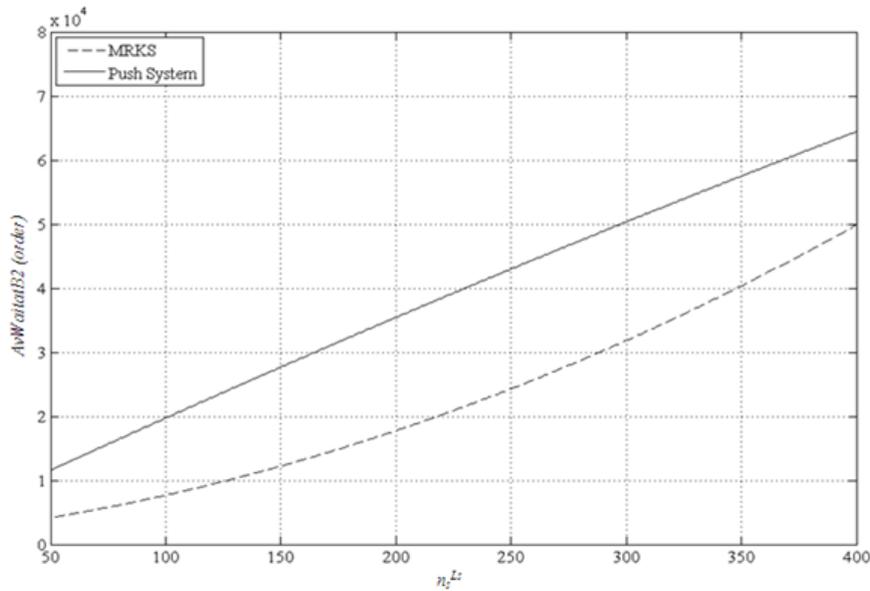
Figure 10. Effect of n_s^{Ls} on *AvWaitatB2(order)*

Average WIP at B2 [AvWIPatB2(order)]

In Figure 11, the region of the push system is entirely overlapped by the MRKS at $n_k = 4$ and $n_k = 6$. At $n_k = 2$, the *AvWIPatB2(order)* is lower compared with the push system. The lowest value of the MRKS is 0.4 (run with the values $n_s^{Ls} = 50$, $n_r = 6$, $n_k = 2$, $n_{op} = 3$, 1st iteration). An increase in n_k has a larger effect on the MRKS for the *AvWIPatB2(order)*. For the push system, the effect of a larger n_{bf} appears to be linear with *AvWIPatB2(order)*. In Figure 12, the plots from the respective regression models also reveal that the MRKS is producing more WIP at B2, approximately after the $n_k = 4$ (or $n_{bf} = 4$). In an isolated case, the findings may be surprising because the kanban card is generally an effective tool for reducing inventory levels. In line with the discussion in the total output, for the pull system, D2 will cease production if the machine is unable to send the output to B2's FIFO line which is full. This is common when

machine failure happens at D3. Although the average WIP at B2 is higher, MRKS is arguably more productive. Another finding on the corresponding n_k value is that the *AvWIPatB2(order)* is always far less than the maximum buffer limit assigned. For example, for n_k = 4, the maximum buffer limit at B2 is 24 orders (four kanban cards for each of the six SMEs). As shown in the graph, the *AvWIPatB2(order)* is about five orders.
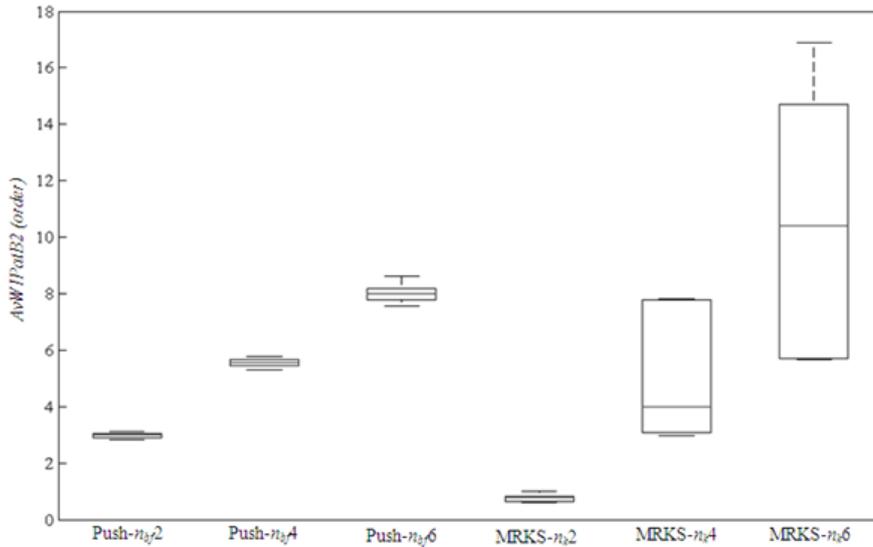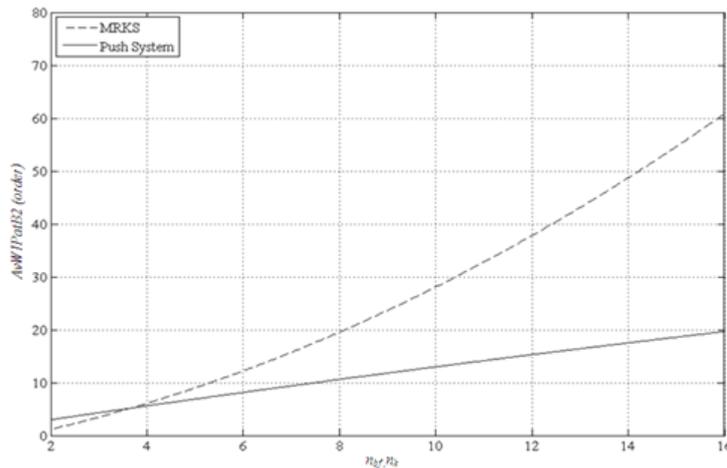

Figure 11. Box plot analysis on *AvWIPatB2(order)*

Figure 12. Effect of $n_{bf,}$ n_k on *AvWIPatB2(order)*

Average utilization of the machine for SME [Util(SME)]

In Figure 13, MRKS exhibits both the highest range positions and utilization of SME machines compared with that of the push system. The highest utilization of SME obtained by the MRKS is 73.95% (run with the values n_s^{Ls}= 200, n_r = 2, n_k = 6, n_{op} = 2, 3rd iteration). From the regression models obtained, the machine utilizations of the MRKS linearly increase with n_s^{Ls} (Figure 14). Machine utilization and its gradients are higher for the MRKS compared with that of the push system. Potentially contributing to this is the levelling effect inherent in MRKS, as aforementioned. Levelling results in a more effective and clever distribution that prevents

bottleneck production and machine failure. The results are consistent with the observations in total output and *AvFT(part)*.
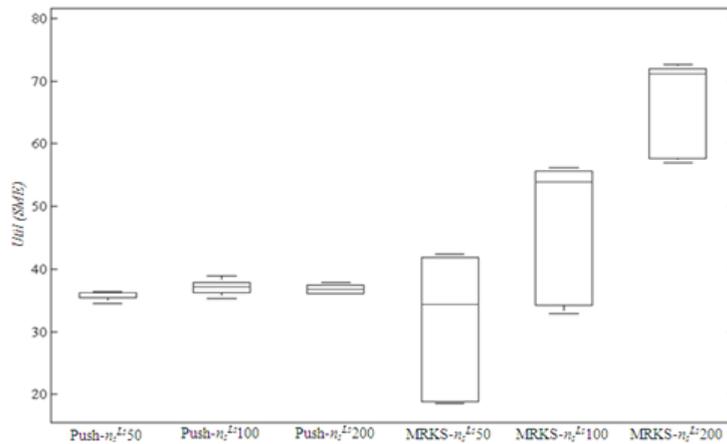

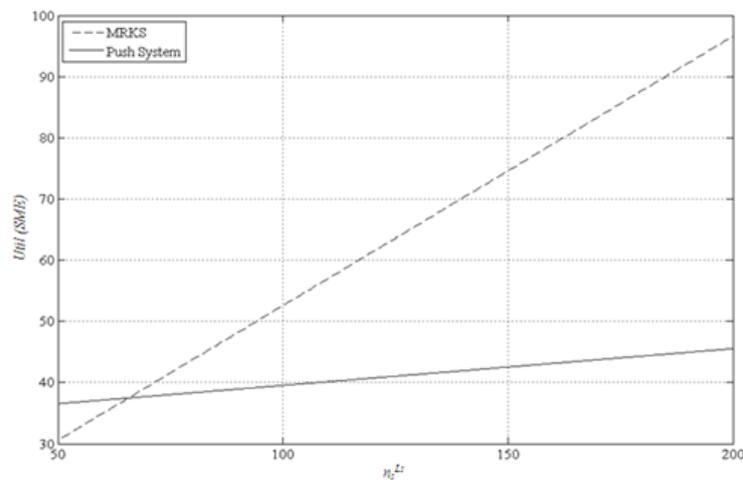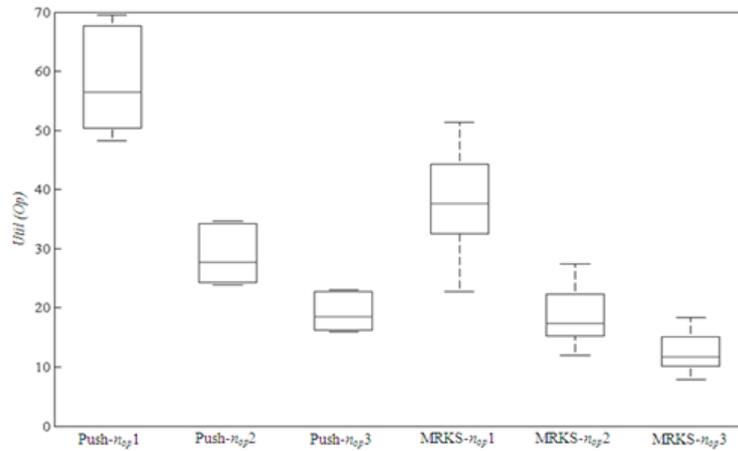Figure 13. Box plot analysis on utilization of SMEs

Figure 14. Effect of n_s^{Ls} on SME utilization

Average utilization of operators [Util(Op)]

In Figure 15, operator utilization is reduced when more operators are assigned to the system. The operator utilization of a push system is generally higher than that of the MRKS. The highest value of the MRKS is 52.9% (run with the values $n_s^{Ls} = 50$, $n_r = 2$, $n_k = 6$, $n_{op} = 1$, 5th iteration). In Figure 16, the plots from the respective regression models also reveal that the operator(s) in the push system is(are) marginally busier. The utilization of operator(s), however, does not translate into a higher total output. From the simulation rerun, the operator(s) is(are) observed to spend a significant amount of time producing orders that feed the bottleneck machine. While batches build up in front of the machine, the other machines are mostly idle. Although the order selection at D2 is purely on a FIFO basis, the quantity of orders that requires processing at the bottleneck machine is proportionally higher.

Figure 15. Box plot analysis on *Util(Op)*
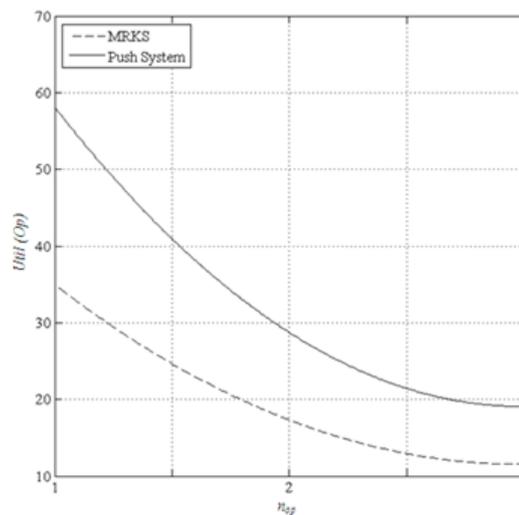
Figure 16. Effect of n_{op} on operator utilization

The effects of nr, nk, and nop

Most of the graph plots above, e.g. Figure. 6, 8, 10, 14 show the changes in performance measures by varying n_s^{Ls} as the prime control parameters. The effects of the less dominant control parameters, namely, n_r, n_k, and n_{op}, are studied here. Due to the similarity exhibited, a number of graphs are plotted on selective performance measures (Figure 17). For the MRKS, a higher total output can be attained with a smaller n_r, a larger n_k, or both. On the same regression model, n_{op} is absent, implying that its effect is negligible. When n_r is relatively small, an operator's trip to the warehouse is more intense. Thus, the waiting time of a kanban card is reduced, and a more vigorous flow of part preparation at the warehouse is achieved until D1 is replenished. In addition, the effect of the momentary process stoppage at D1 due to the absence of an operator is negligible. A larger n_k indicates a bigger WIP input capacity at D3, preventing the effects of machine starving. All these factors improve productivity. Furthermore, n_r has a milder effect compared with n_k because the plotted lines are closer to one another. As shown in Figure 17(e), n_{op} has exerted very little effect on the MRKS.
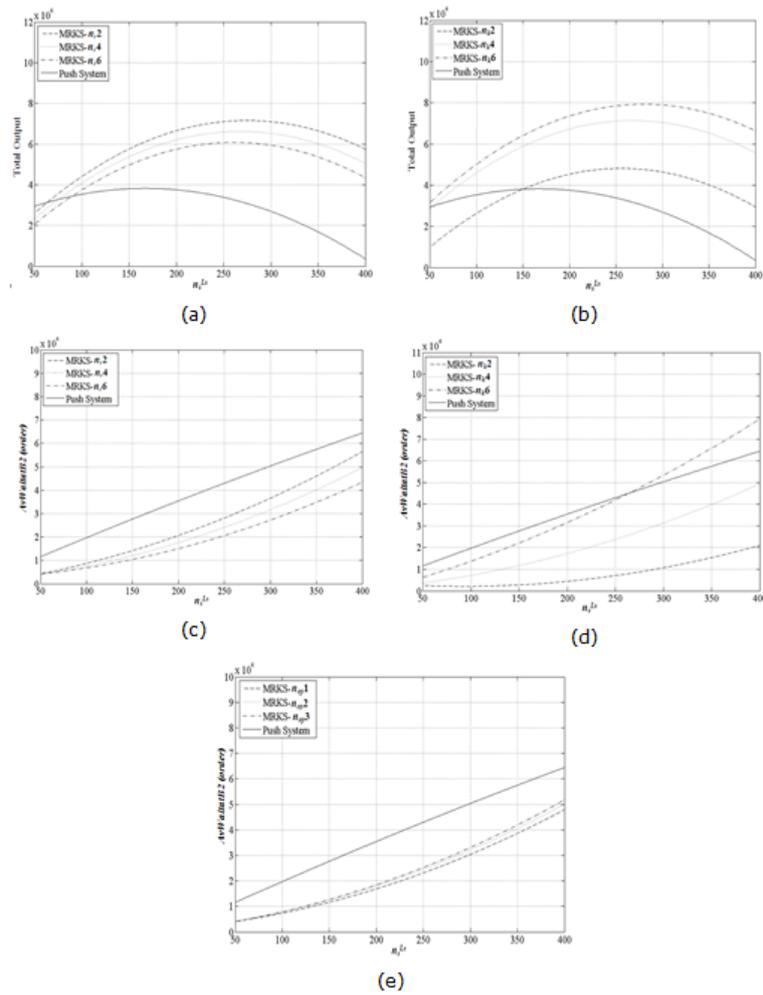
Figure 17. Effect of n_s^{Ls} on total output and *AvWaitatB2(order)*

Managing the MP is a complex production control problem involving a number of processes which vary in terms of cycle times, set-up times, operator requirements, and machine failure rates. Some processing locations are nonadjacent. The output and the utilization of bottleneck machines have to be maximized while keeping the WIP level low. Additional requirements include controlling the waiting time for WIP. With these requirements, the results obtained clearly indicate that the MRKS is superior to the push system, making it a highly attractive choice. The advantages of the MRKS stem from its ability to maximize production by diverting from the failed machines. Such characteristic provides better levelling, improved overall machine utilizations, and reduction on average flow time. Better levelling, in turn, provides better product mix. Another finding in the present study is that dedicating material handling jobs to operators running the MM and LM do not lead to noticeable degradations on the performance of the MRKS.

7. Conclusions

The proposed MRKS produced a number of insights on production management in the case study. First, the system is a more robust and superior method of production planning and

control than the present push system. Second, the system indicates the advantages of levelling, particularly in the event of machine failure and blockage. The current study also led to some observations on the applicability of simulation models for evaluating production planning and control systems. Future development of the system includes a study of the mechanism for varying the number of kanban cards to cater to the conditions of individual SME machines, and using a hybrid kanban system to handle urgent orders.

References

Bonney, M.C., Zhang, Z., Head, M.A., Tien, C.C., & Barson, R.J. (1999). Are push and pull systems really so different?. *International Journal of Production Economics*, 59(1-3), 53-64. http://dx.doi.org/10.1016/S0925-5273(98)00094-2

Bonvik, A.M., Couch, C.E., & Gershwin, S.B. (1997). A comparison of production-line control mechanisms. *International Journal of Production Research*, 35(3), 789-804. http://dx.doi.org/10.1080/002075497195713

Cao, D., & Chen, M. (2005). A mixed integer programming model for a two line CONWIP-based production and assembly system. *International Journal of Production Economics*, 95(3), 317-326. http://dx.doi.org/10.1016/j.ijpe.2004.01.002

Chan, F.T.S. (2001). Effect of kanban size on just-in-time manufacturing systems. *Journal of Material Processing Technology*, 116(2-3), 146-160. http://dx.doi.org/10.1016/S0924-0136(01)01022-6

Chen, L., & Meng, B. (2010). Application of value stream mapping based lean production system. *International Journal of Business and Management*, 5(6), 203-209.

Horbal, R., Kagan, R., & Koch, T. (2008). Implementing lean manufacturing in high-mix production environment. *IFIP (International Federation for Information Processing),* 257, 257-267. http://dx.doi.org/10.1007/978-0-387-77249-3_27

Liker, J.K. (2004). *The Toyota way: 14 management principles from the world's greatest manufacturer*. New York: McGraw-Hill.

Marek, R.P., Elkins, D.A., & Smith, D.R. (2001). Understanding the fundamentals of kanban and CONWIP pull systems using simulation. *Proceedings of the Winter Simulation Conference*, 2, 921-929.

Matzka, J., Di Mascolo, M., & Furman, K. (2009). Buffer sizing of a heijunka kanban system. *Journal of Intelligent Manufacturing*, 23(1), 49-60. http://dx.doi.org/10.1007/s10845-009-0317-3

Ou, J., & Jiang, J. (1997). Yield comparison of push and pull control methods on production systems with unreliable machines. *International Journal of Production Economics*, 50(1), 1-12. http://dx.doi.org/10.1016/S0925-5273(97)89131-1

Rother, M., & Shook, J. (1999*). Learning to see: value stream mapping to create value and elimination muda.* Massachusetts: The Lean Enterprise Institute.

Sadjadi, S. J., Jafari, M., & Amini, T. (2009). A new mathematical modeling and a genetic algorithm search for milk-run problem (an auto industry supply chain case study). *International Journal of Advanced Manufacturing Technology*, 44(1-2), 194-200. http://dx.doi.org/10.1007/s00170-008-1648-5

Savsar, M. (1997). Simulation analysis of a pull-push system for an electronic assembly line. *International Journal of Production Economics*, 51(3), 205-214. http://dx.doi.org/10.1016/S0925-5273(97)00055-8

Spearman, M.L., Woodruff, D.L., & Hopp, W.J. (1990). CONWIP: a pull alternative to kanban. *International Journal of Production Research*, 28(5), 879-894. http://dx.doi.org/10.1080/00207549008942761

Seidman, T.I., & Holloway, L.E. (2002). Stability of pull production control methods for systems with significant setups. *IEEE Transactions on Automatic Control*, 47(10), 1637-1647. http://dx.doi.org/10.1109/TAC.2002.803531

Tatsiopoulos, I.P., Avramopoulos, I., & Theoharis, I. (1997). A reference data model for production control in the electronics industry. *Computers in Industry*, 34(2), 221-231. http://dx.doi.org/10.1016/S0166-3615(97)00057-2

Takahashi, K., Myreshka, & Hirotani, D. (2005). Comparing CONWIP, synchronized CONWIP and kanban in complex supply chains. *International Journal of Production Economics*, 93-94, 25-40. http://dx.doi.org/10.1016/j.ijpe.2004.06.003