Beyond Kanban: Creating and analyzing lean shop floor control policies [1]

Asbjoern M. Bonvik • Stanley B. Gershwin

Operations Research Center, MIT, Cambridge, MA 02139
Laboratory for Manufacturing and Productivity, MIT, Cambridge, MA 01239

Abstract

We present a unified view of some common shop floor control policies for repetitive manufacturing, including kanban, basestock, and CONWIP control. This view focuses on the patterns of information flow in systems controlled by these policies. By combining the information flows from several policies, new hybrid policies can be created. These policies can attain the same throughput and service levels as traditional policies, while operating at significantly reduced inventory levels. We present simulation results that show how one of these hybrids outperforms kanban in a particular line. We also demonstrate that larger or more variable systems reap even larger benefits from these policies compared to kanban.

1 Introduction

There are several conflicting goals in operating a manufacturing system: It is desirable to have high levels of throughput and customer service, but at the same time, the inventory levels should be low. These goals are in conflict, because throughput or customer service can be improved by adding more buffer inventories to absorb process variability. A system that achieves high levels of throughput and/or customer service at low inventory levels is described as *lean* (Womack, Jones, and Roos 1990).

At the shop floor level, several control strategies have been proposed and implemented to achieve leanness. In Toyota Motor Company, which is often credited as the origin of lean manufacturing, a control system called *kanban* is used (Shingo 1989). This is often classified as a "pull" system of production control, since all production occurs in response to actual demand events. A Western invention is *Constant-Work-in-Progress* (CONWIP) control (Spearman, Woodruff, and Hopp 1990). Recent papers have investigated combinations and hybrids of two or more of these control policies (Buzacott and Shantikumar 1992; Van Ryzin, Lou, and Gershwin 1993; Dallery and Liberopoulos 1995; Bonvik, Couch, and Gershwin 1996). All of these policies can be implemented by circulating cards or similar tokens in the factory. This is discussed in Bonvik (1996).

Each control policy imposes a particular pattern of information flows on the factory. Since all the policies mentioned can be implemented by circulating cards or similar information carriers, these information flows are the paths followed by these cards. We differentiate between *global* and *local* information flows. Global information flows transmit information from the demand process to a particular machine, without going through any intermediate machines. Local information flows circulate between a machine and an adjacent buffer.

In this paper, we first investigate the information flows imposed on the factory by kanban, CONWIP, and a hybrid of these policies. We present a simulation study of a simple production line, comparing the

[1] To be presented at the 1996 Manufacturing and Service Operations Management Conference, Tuck School of Management, Dartmouth College, June 28, 1996

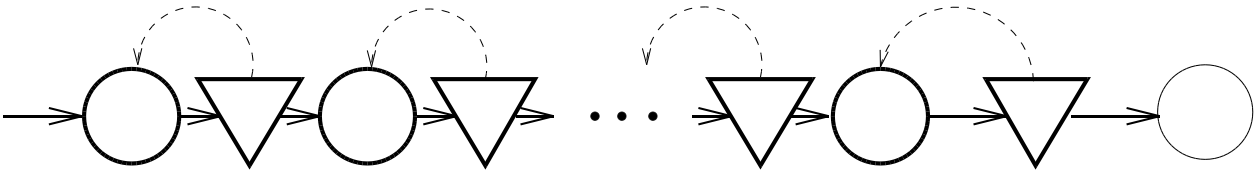

Figure 1: Kanban control

different policies in a fairly typical environment. This study is different from that of Bonvik, Couch, and Gershwin in three respects: The length of the line, the handling of unmet demand, and the behavior of the demand process. Finally, we draw some conclusions about the relative strengths and weaknesses of these control policies.

2 Kanban control

Kanban control contains only local information flows. The cards (kanbans) circulate between a buffer and the immediate upstream machine. The machine is *blocked* when all cards are attached to parts in the buffer, i.e., when the buffer is full. When a machine picks up raw materials to perform an operation, it also detaches the card that was attached to the material. The card is then circulated back upstream to signal the next upstream machine to do another operation. This way, a demand for a unit of finished goods percolates up the supply chain. Kanban control is illustrated in figure 1, where machines are drawn as circles and buffers as triangles. The solid arrows indicate material flow, and the dashed arrows information flow.

Kanban control ensures that parts are not made except in response to a demand. The analogy is to a supermarket: Only the goods that have been sold are re-stocked on the shelves. However, it has a major drawback: It uses the parts themselves as carriers of information. A machine is told to stop production when its output buffer is full. This is the same as requiring the machine to fill the buffer whenever possible.

The parts waiting in a buffer act as a buffer inventory, partially decoupling the operation of downstream machines from any interruptions of upstream production. If a machine fails, the machine downstream of it can continue production by consuming the parts that are already in the buffer. With luck, the upstream machine will be repaired before the buffer is empty, and the failure will not affect the downstream machine (or the customer on the downstream end of the chain).

Our kanban model is equivalent to tandem queues (Berkley 1991). Another often-used kanban model is called *minimal blocking*, where cards can bypass failed or busy machines (So and Pinault 1988). This is also a local information flow.

3 CONWIP control

CONWIP designates a control strategy that limits the total number of parts allowed into the system at the same time. Once the parts are released, they are processed as quickly as possible until they wind up in the last buffer as finished goods. One way to view this is that the system is enveloped in a single kanban cell: Once the consumer removes a part from the finished goods inventory, the first machine in the chain is

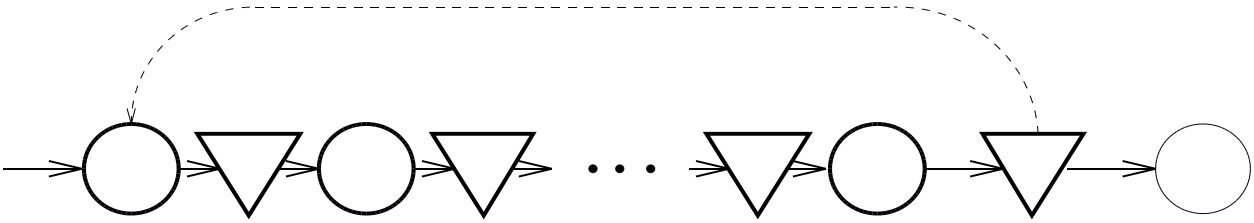

Figure 2: CONWIP control

authorized to load another part. This causes a global information flow from the demand process to the first machine, as shown in figure 2.

This leads to subtly different behavior from a kanban control. First of all, like kanban, the CONWIP system only responds to actual demands that have occurred, so it is still a "pull" type system. But unlike kanban, the resting state of the system has all buffers empty, except finished goods, which is full.

CONWIP attains partial decoupling of machines through part/hole duality. Inventory in a buffer protects the downstream portion of the line against the consequences of failures upstream. But it does not protect the upstream portion of the line against failures downstream. If a buffer is full, and the machine downstream of it fails, a kanban line will stop production upstream of the failure. When the failed machine is repaired, it will suddenly impose an increased workload on the upstream portion of the system, since it needs to catch up with the demand.

The mostly empty buffers in a CONWIP line contain useful (but inexpensive) empty space. This space is used to decouple the upstream portion of the line against failures downstream. If the last machine in the line fails, the customers will be served from the finished goods buffer, while new parts will be released to the line as usual and proceed to the buffer in front of the failed machine. There they wait for the repair. When the machine is repaired, it has a sufficiently large number of parts in its input buffer to catch up with demand and replenish the finished goods buffer.

4 Hybrid control

Sometimes, if the system is heavily utilized or there is a bottleneck in the line, the buffers towards the upstream end of a CONWIP line will have quite high levels. On the other hand, kanban control was designed to prevent individual buffer levels from exceeding designated limits.

Therefore, we construct a hybrid control policy where the CONWIP control is supplemented with secondary kanban cells. These detect problems in the line, and block release of parts to the line if they cannot be processed further. We do not need a separate kanban cell to block the last machine, since any material that has gotten this far surely will reach the finished goods buffer if the machine can do an operation. The resulting control policy acts mostly like CONWIP, but at decreased inventories when trouble occurs.

Note how similar this is to a kanban control: We circulate cards between the machines and buffers. The sizes of the buffers are determined by the number of cards in circulation. The only difference is that cards detached from finished goods are passed to the first machine instead of the last. From there, they follow the parts back to the finished goods buffer. This is shown in figure 3.

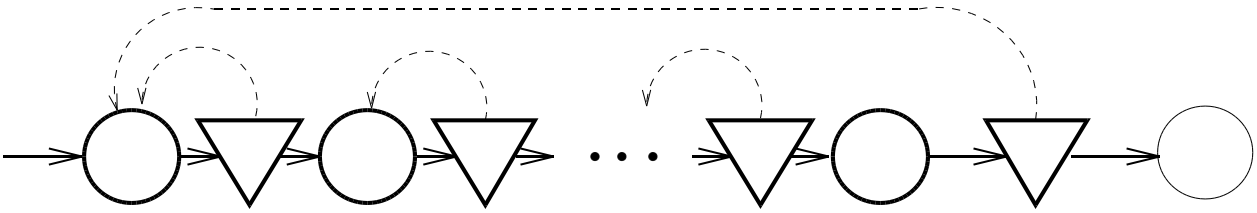

Figure 3: Hybrid control

5 Comparing the policies

To demonstrate that these policies have different characteristics, we perform the following experiment: Consider a six-machine tandem line serving a stochastic demand process. Each machine is unreliable, and has some operation time variability. There is space for buffer inventories between the machines, there is room for a finished goods inventory, and we allow unlimited backlogging of unmet demand.

We want to control the line to achieve three objectives: The production rate must match the demand rate. At least 98% of the arriving demand must be served from stock. (This measure is also known as the line *fill rate*.) And the inventories should be as low as possible. In addition, we are concerned about other measures of service performance, such as average backlog length and waiting time.

We let demand events occur only at integral times, with a certain probability for each integral time. The average demand rate is then equal to the probability of a demand event occurring. This represents a downstream assembly line using the parts produced by our line. We study three different demand rates: .65, .75, and .85. Each machine has a mean time to fail of 100 time units, and a mean time to repair of 10 time units. The failures and repairs are exponentially distributed. The operation times at each machine have a lognormal distribution with $\mu = 0$ and $\sigma = .1$, giving a mean of 1.005 and a standard deviation of .101. This gives each machine an isolated production rate of .905, and the demand rates can clearly be met with sufficiently large buffer inventories in the system.

We try several policies: **Kanban**: We search in internal buffer sizes from 4 to 10, combined with finished goods buffers sized 15 to 70. With internal buffer sizes of 4, this line has a saturated throughput of .683, so the demand rate of .65 is feasible, even with the smallest buffers. **CONWIP**: We try all inventory limits between 15 and 70. **Hybrid**: Combining kanban and CONWIP, we try all configurations with internal buffer sizes between 4 and 10, and inventory limits between 15 and 70. In the hybrid policy, the size of the finished goods buffer is always the same as the inventory limit for the system.

We evaluate these cases by discrete event simulation. Each case is run for 240,000 time units, with an initial warm-up period of 9,600 time units. If the time units are minutes, this is two years of operation with one 8-hour shift per day, 5 days a week. The warm-up period is then a month. Each case was started from a different random number seed. From the results, we find the parameter choice for each policy that met the 98% service target with the least inventory. See table 1. The best kanban policy has internal buffers of size 7 and a finished goods buffer of size 50. The best CONWIP policy has inventory limit 55. The best hybrid policy turned out to have as large buffers as possible, which made it identical to CONWIP. For this case, the CONWIP and hybrid policies had 26% less inventory than kanban.

We also repeated the experiment with the demand rate increased to .75, this time searching in internal

Policy	Throughput, parts/min.	Fill rate	Inventory, parts	Backlog, parts
Kanban	.650	98.1%	66.6	.13
CONWIP	.648	98.3%	49.3	.16
Hybrid	.648	98.3%	49.3	.16

Table 1: Performance measures, 6-machine line, demand rate .65

buffer sizes between 10 and 50, and finished goods buffer sizes between 30 and 120. The results are shown in table 2. The best kanban control used internal buffers of size 18 combined with a finished goods buffer of size 56. The CONWIP control used an inventory limit of 91. The best hybrid control used an inventory limit of 90, combined with internal buffers of size 50.

Policy	Throughput	Fill rate	Inventory	Backlog
Kanban	.749	98.3%	113.6	.36
CONWIP	.750	98.2%	82.4	.19
Hybrid	.750	98.2%	80.8	.23

Table 2: Performance measures, 6-machine line, demand rate .75

Finally, we increased the demand rate to .85 and repeated the experiment, searching in inventory limits/finished goods buffer sizes up to 300 and internal buffer sizes up to 150. The best kanban control used internal buffers of size 56 and a finished goods buffer of size 114. The best CONWIP policy used an inventory limit of 242. The best hybrid policy combined an inventory limit of 225 with internal buffers of size 110. The results are shown in table 3.

Policy	Throughput	Fill rate	Inventory	Backlog
Kanban	.852	98.2%	253.0	.28
CONWIP	.850	98.4%	223.1	.36
Hybrid	.849	98.4%	206.2	.28

Table 3: Performance measures, 6-machine line, demand rate .85

6 Discussion

We see that there is a large difference in performance between the policy with local information flows (kanban) and those with global information flow (CONWIP and hybrid). The difference in inventory between kanban and CONWIP at the same service level range from 27% at the lowest demand rate to 12% at the highest. This is a larger difference than that reported by Bonvik, Couch, and Gershwin (1996), who found a 7.8% reduction in inventory when going from kanban to CONWIP for a heavily utilized, low-variability four-machine line. We believe the larger difference between kanban and CONWIP has two reasons: This system is larger, so there are more buffers for kanban to fill up; and the system has higher variability, so more buffering is required to maintain the service level.

Introducing the additional inventory limits of hybrid control can reduce the inventories further. In this

study, there was no improvement over CONWIP at the lowest demand rate, 2% at the middle rate, and 8% at the highest rate. Bonvik, Couch, and Gershwin observed a 4.5% difference between CONWIP and hybrid control. This indicates that the importance of the finite buffers in the hybrid policy increases with system utilization, and possibly with system variability.

These results are consistent with a simulation study of a 10-machine line by Bonvik (1996), where similar differences between the policies were found. In that study, additional control policies (such as basestock, minimal blocking, and other hybrids) were tried, and the CONWIP/kanban hybrid achieved the service target with the least inventory.

7 Conclusions

We have found that a simple CONWIP control policy outperforms kanban with respect to average inventory levels, when subject to the same requirements on throughput and service level (fill rate). When the system operates close to capacity, the hybrid control combining CONWIP and kanban improves the inventory levels further.

References

Berkley, B. J. (1991). Tandem queues and kanban-controlled lines. *International Journal of Production Research 29*(10), 2057–2081.

Bonvik, A. M. (1996). *Performance Analysis of Manufacturing Systems Under Hybrid Control Policies.* Ph. D. thesis, Massachusetts Institute of Technology.

Bonvik, A. M., C. Couch, and S. B. Gershwin (1996). A comparison of production-line control mechanisms. Accepted for publication in the *International Journal of Production Research.*

Buzacott, J. A. and J. G. Shantikumar (1992). A general approach for coordinating production in multiple-cell manufacturing systems. *Production and Operations Management 1*(1), 34–52.

Dallery, Y. and G. Liberopoulos (1995). Extended kanban control system: A new kanban-type pull control mechanism for multi-stage manufacturing systems. Unpublished manuscript.

Shingo, S. (1989). *A Study of the Toyota Production System from an Industrial Engineering Viewpoint* (2nd ed.). Productivity Press.

So, K. C. and S. C. Pinault (1988). Allocating buffer storages in a pull system. *International Journal of Production Research 15*(12), 1959–1980.

Spearman, M. L., D. L. Woodruff, and W. J. Hopp (1990). CONWIP: a pull alternative to kanban. *International Journal of Production Research 28*(5), 879–894.

Van Ryzin, G., S. X. C. Lou, and S. B. Gershwin (1993). Production control for a tandem two-machine system. *IIE Transactions 25*(5), 5–20.

Womack, J. P., D. T. Jones, and D. Roos (1990). *The machine that changed the world : The story of lean production.* Rawson Associates.