

# KANBAN: THE OLD AND THE NEW

John Bicheno, FIOM, Cardiff Business School

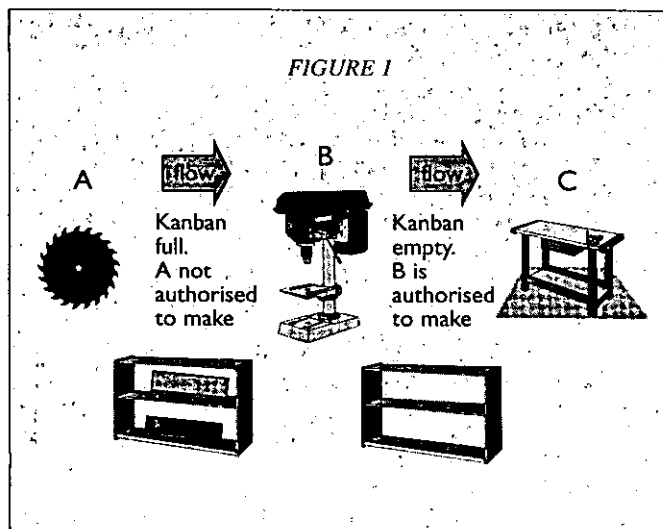
Kanban has been around for many years and is well established in a large number of companies. It is the classic signalling device for production pull systems. Nevertheless there remain uncertainties about types and quantities. For a few, kanban is still a new concept. Others may have decided that pull systems are not for them; they should think again in the light of new types of pull systems. This article attempts a summary of the old and the new [1] [2].

## SINGLE CARD KANBAN

Traditional kanban is suitable in all stable-manufacturing environments where there is repetitive production. In practice, the single card kanban category is by far the most popular type. It is easy to understand, easy to see and reasonably easy to install. Single card kanban means that a single card (or pull signal) operates between each pair of workstations. Although there may be several single-card kanbans in a loop between a pair of workstations, each kanban is the authorisation to both make a part or product and to move it. Two major categories are product kanban and generic kanban.

## PRODUCT KANBAN

Product kanban (or 'replacement kanban') is the simplest form of pull system. With this type, wherever a product is called for, it is simply replaced. If there is no call, there is no authorisation so there is no production. In practice, the variations of this type include kanban squares. (A vacant square is the authorisation to fill the square with another similar part), cards (which are returned to the feeding workstation to authorise it to make a replacement quantity as specified on the card) and other variations such as 'faxban' or 'e-ban' (which operate in exactly the same way as cards, except that the pull signals are not physical). These are shown in Figure 1.



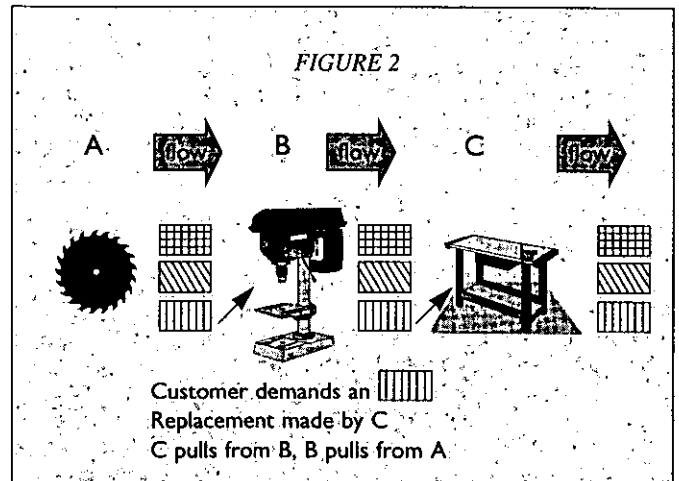
## Product Kanban with Multiple Products

### (a) Sequential Operations

In sequential operations, having several different products, product kanban can be used between stations provided there are not too many products. Here, one partly completed product of each type is placed as a buffer between each workstation. If product A is called for at the end of the line, this

triggers sequential pull signals to make a replacement A. The other products do not move, until they are called for. See Figure 2.

This system allows a quick response build from a limited selection of products, but of course has the penalty of holding intermediate buffers of part completed products of each type. Hence this system becomes impractical for more than a handful of products. The generic kanban type should then be used, as explained below.



### b) Assemble to Order Operations

A variation that is employed in several assemble-to-order operations (for instance, personal computer 'make to order') involves simply having shelves with at least one or two of all parts and sub-assemblies available surrounding the final assembly area.

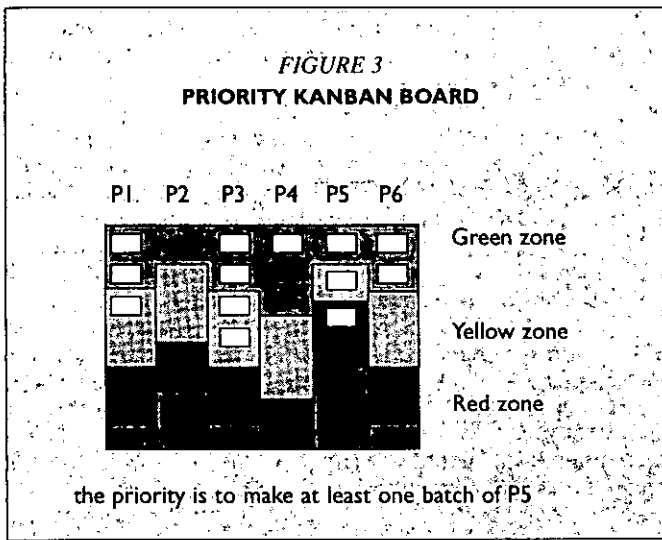
When an order comes in, it is simply configured from the appropriate shelves. This therefore creates a blank space on the shelf that is the signal for sub-assembly areas to replace that sub-assembly. Sub-assembly areas are themselves arranged into cells that pull parts from the shelf, and hence back to the store. In this way, literally millions of different configurations can be made under a pull system, provided the final products are designed to accommodate a variety of different but similarly sized modules.

### (c) Operations where there is Changeover and/or Changing Priorities

Where there is changeover, a priority kanban board is the usual option. A priority kanban board displays all the products that go through the work centre. For each product a number of kanbans are hung on the board, (see below for calculation of the number of kanbans).

The kanbans themselves are colour coded or alternatively there are bands of green, yellow, and red painted on the board to indicate priority. A kanban hanging in the green zone is a low priority authorisation to make a replacement part. A kanban hanging in the yellow zone indicates higher priority and a kanban hanging in the red zone would normally indicate that that product should be the very next one that should be made. The board therefore indicates at a glance the backlog situation. Kanbans are hung on the board for each product in the order green, yellow, red, but are replenished in the order red, yellow, and green.

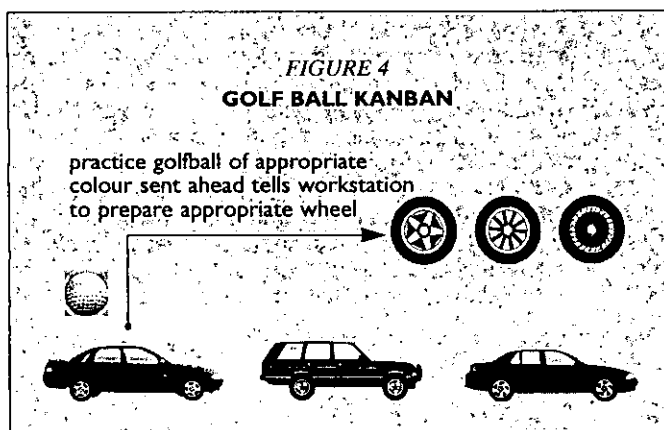
An operator looking at the board is authorised to make a batch quantity covering the red zone, or if there is more time the red plus the yellow zone, or if there is even more time the red plus the yellow plus the green zone. In slack times, the operator may decide to just replenish the green kanbans. See Figure 3.



**(d) Product Kanban with Synchronised Operations**

Where there are several legs in a bill of material or assembly structure, synchronisation can be achieved by variations of so-called 'golf ball' kanban. Here, as the main build progresses, signals are sent to areas producing supporting assemblies 'just in time' to meet up with the main build as it progresses along the line.

Different colour 'golf balls' are moved (often blown by air or sent electronically) to the sub-assembly stations to signal them to prepare the exact required sub-assembly. This form of kanban can be used internally (say to prepare different wind-screens or coloured bumpers) to go onto particular cars, or externally (for instance when sent to external seat suppliers to prepare the exact sequence of seats to meet up with a particular sequence of cars). See Figure 4.



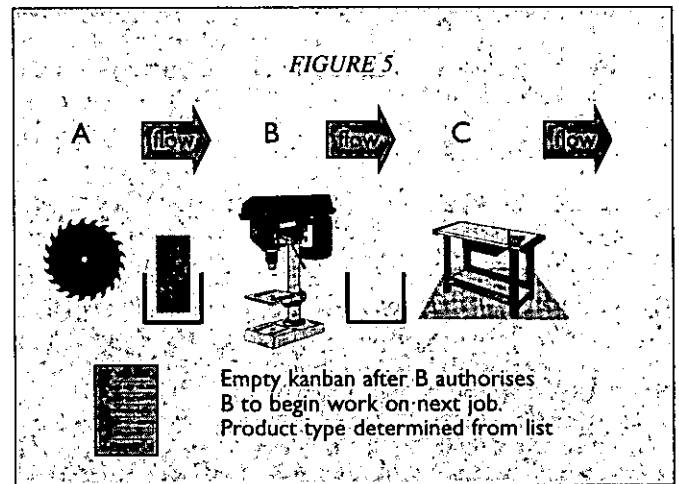
**Emergency Kanban**

Emergency kanban is a 'special event' kanban that is inserted in kanban loops to compensate for unusual circumstances. Such kanban cards are of a different colour so that they may be distinguished easily. Such kanbans automatically go to the head of the queue so their requirements are dealt with as soon as possible. Having produced the additional quantity, emergency kanbans are withdrawn.

A variation is additional kanbans that are inserted to meet seasonal demand or to compensate for transport disruption, such as rail disruption or poor weather. These cards are also withdrawn as soon as possible.

**GENERIC KANBAN**

Generic (or 'capacity') kanban authorise feeding work centres to make a part, but does not specify what part is to be made. The part to be made is specified via a manifest or a 'broadcast' system. It is therefore the preferable pull system where there are a large number of products, all of which have similar routings and fairly similar time requirements at each workstation. See Figure 5.



Comparing Figure 2 with Figure 5 notice that the generic type has far less WIP inventory, but has longer response time.

**Rules of Kanban**

- Downstream operations make to demand. Never over produce.
- Demands are placed on upstream operations by means of cards or other signals.
- Authorisation to produce is by card (or signal) only.
- Each kanban card circulates between a particular pair of workstations only.
- Quality at source is a requirement. Only good items are sent downstream.
- The number of kanbans should be reduced as problems decrease.

**DUAL CARD KANBAN**

Toyota and a few other manufacturers use 'dual card' kanban. This can be a tighter form of pull system. Instead of one all-purpose card, the dual card system has both 'production' kanbans (with authorise production) and 'conveyance' kanbans (which authorise movement). Production kanbans stay at a particular work centre and alternate from input board to finished goods container with production kanban attached. Conveyance kanbans stay between a particular pair of workstations and alternate between conveyance card mailbox and full container (with conveyance card attached) moved between workstations.

A material handler collects up the conveyance kanbans from the mailbox (at Toyota a tune sounds to warn of this event) and takes them to the appropriate feeder workstation. There, the material handler detaches the production kanban from the full container and attaches the conveyance kanban to the container. The production kanbans are returned to the kanban board of the workstation. When works starts on the batch, as authorised by the production kanban, the operator detaches the conveyance kanban from the container and hangs it on the mailbox for the upstream conveyance.

# NUMBERS OF KANBAN CARDS

## GENERAL

In line with lean manufacturing, the correct answer to the question of the number of kanban cards should generally be 'less than last time'! The well-known water and rocks analogy applies. That is, reduce the inventory levels by removing a kanban (or by reducing the kanban quantity) and 'expose the rocks'. Note that the philosophy of gradually reducing inventory by removing kanbans is 'win-win' approach; either nothing will happen in which case you have 'won' because you have found that you can run a little tighter or you 'hit a rock' in which case you have also 'won' because you have hit not just any old rock, but the most pressing rock or constraint. This is what Toyota has done for decades.

The general rule on kanbans is therefore to start 'loose', with a generous amount of safety stock and to move towards 'tight kanban' gradually, but steadily.

Of course, it would be unwise to jump into this policy. There are a number of pre-requisites for successful kanban. These include reasonably stable demand, reasonably capable processes, preventative maintenance or TPM to ensure that breakdown stoppages are not endemic, probably reasonably short changeover times and a disciplined workforce, including management, to ensure that kanban rules are maintained. Given that all of these are in place and that the company will move from 'loose' to 'tight' kanban, it is still necessary to be able to establish the approximate number of kanban cards.

## CALCULATING THE NUMBER OF CARDS

In general, kanban works like the traditional two-bin system. In the two-bin system the re-order point ROP is calculated thus:

$$ROP = D \times LT + SS$$

where D=demand during the lead time LT between placing and order and receiving delivery and SS is the safety stock. This familiar formula is the basis for all kanban calculations.

If the container or stillage quantity is Q, then the number of kanbans is simply:

$$N = (D \times LT + SS) / Q$$

where N should be rounded up and should be at least 2 unless the production system is very responsive.

### 1. Number of Cards for Assembly Operations or from Suppliers

In repetitive assembly operations where there is no changeover, the demand is expressed in units per day and the lead time LT in the fraction of the day required to go through all the necessary steps between 'placing the order' (hanging the kanban on the board) and receiving it. This would normally include the usual lead time elements of run + wait + move. Note that run time should be the time to fill the container, wait time should include both pre- and post-waiting for movement and waiting on the kanban board or mailbox before the order is actioned.

Where parts are obtained from an external supplier, the lead time would be the expected lead time for delivery as used in any inventory calculation. Demand and lead time should always be expressed in compatible units; say demand per week and lead time in weeks.

Safety stock should also be allowed. This would reflect the many uncertainties in delivery, quality, breakdown or other disruption. Note two points. First, the principle of moving from 'loose' to 'tight' pull and second, the fact that safety

stock has usually already been somewhat allowed for in the rounding up calculation to calculate the number of cards.

### 2. Number of Kanbans where there is Changeover

Where there are multiple products moving through a work centre and where there is changeover involved, the calculations become more complex. However, the basic formula remains valid.

In a press shop, for example, multiple products must go through a press. There may be a priority kanban board governing the control of sequencing. Some pressings would be repeaters, perhaps made everyday, but others more like 'strangers' made regularly but much less frequently. The number of kanbans hanging on the board must be capable of responding to the demands placed on the shop.

The correct or optimal number of changeovers in a sequence or 'campaign' of making all products is dealt with elsewhere [3] but here it will be assumed that these relative frequencies and batch sizes have been determined. (It should be noted that most press shops adopt (hopefully!) sensible rules of thumb rather than trying to optimise).

The requirement is to first determine the lead time for each product. This will reflect the frequency with which each product is run. The frequency of runs should take into account the demand and the costs of the parts and of course the capacity of the shop. This is given for product y by

$$LT_y = \text{total campaign length} / (\text{number of set-ups for } y)$$

The formula for the number of kanbans for product y is:

$$N_y = (D_y \times LT_y + SS_y) / Q_y$$

The container quantity  $Q_y$  should be a 'nice round number' taking into account practical considerations (such as a preference for human-movable containers), but also the balance of takt time (shift time divided by demand per shift) and changeover time. The ideal batch size is that for which the sum of changeover plus run times for all products exactly matches demand in the same length of time. For example, if there are 2 products and the changeover time is 30 minutes, run time is 1 minute per product and demand rate is 15 per hour for each product (or a takt time of 4 minutes). Ideal batch size would be 30. (Notice that the EOQ is not used!) Container quantity should be a convenient factor of 30, such as 10, 15, 30, but not more than 30 because this would necessitate waiting for the next changeover.

Example: There are 3 products A, B and C that go through a press. Container quantity is 50 for products A and C and 100 for B. All changeovers take 30 minutes. Total campaign length is 2 days. Normal planned production is A:B:A:C.

	A	B	C
demand/day	300	500	100
set-ups per campaign	2	1	1
safety stock	100	50	50
and			
LT <sub>y</sub>	1 day	2 days	2 days
and			
Number of kanbans	8	10.5 = 11	5

These kanbans would be hung on a board, perhaps with priorities arranged as follows.

Red zone	2	2	1
Yellow zone	4	6	3
Green zone	2	3	1

This is the case for 'loose' kanban. For 'tight' kanban the actual campaign length should be calculated. This should include the sum of set up plus run times for all products and allow for collection of kanban cards and delivery to the point of requirement. This total time should be divided by OEE (overall equipment effectiveness = availability x speed utilisation x quality rate). A check should be made for feasibility, that is whether production during the campaign length so calculated is less than actual average demand during this length of time.

In the preceding sections, traditional kanban has been explained. The weakness of traditional kanban is that it assumes repetitive production (even where generic kanban is used) and also a fairly level schedule. Where the schedule is not level, quite significant buffer inventories between the various stages may be idle for lengthy periods, waiting to be pulled. This is of course 'muda'. Further complications are routings that may vary significantly between products, and variation in processing times resulting in imbalanced lines and temporary 'bottlenecks'. In such circumstances traditional kanban systems can sometimes have more inventory than MRP push systems. Some variations have been developed to overcome these limitations.

## CONWIP

CONWIP or 'constant work in progress' (Hopp and Spearman, [4]) links the last process with the first by means of a job kanban card. Cards do not operate between each pair of workstations as in traditional kanban, but instead cards follow the product or batch through all stages in a section. As a product or batch is completed at the last process, the card is sent to the first process thereby authorising the start of a new batch. The CONWIP card loop is a whole assembly line, a cell, or a whole factory. CONWIP cards are not product specific. They authorise the start of production of a batch or product of whatever type is required. In this sense they are like generic or capacity kanbans.

CONWIP automatically results in accumulations of inventory in front of temporary bottlenecks, which is just where it is required. There is a similarity with the OPT type 'drum, buffer, rope' which links bottleneck work centre to gateway work centre and places a 'time buffer' in front of the bottleneck. However, with CONWIP the bottleneck does not have to be identified (indeed it may shift) and the time buffer does not need to be calculated.

The number of cards is calculated directly from WIP, which is:

$$\begin{aligned} \text{WIP} &= \text{feasible cycle time} \times \text{rate of production} \\ &= \text{minutes} \times \text{jobs per minute} \end{aligned}$$

The feasible cycle time is the total lead time through the series of processes covered by the CONWIP loop. Begin with a generous estimate and reduce. The usual minimum value is the sum of set up plus run plus move, except where overlapping of batches is the norm.

$$\text{minimum number of cards} = \text{WIP} / \text{average batch size}$$

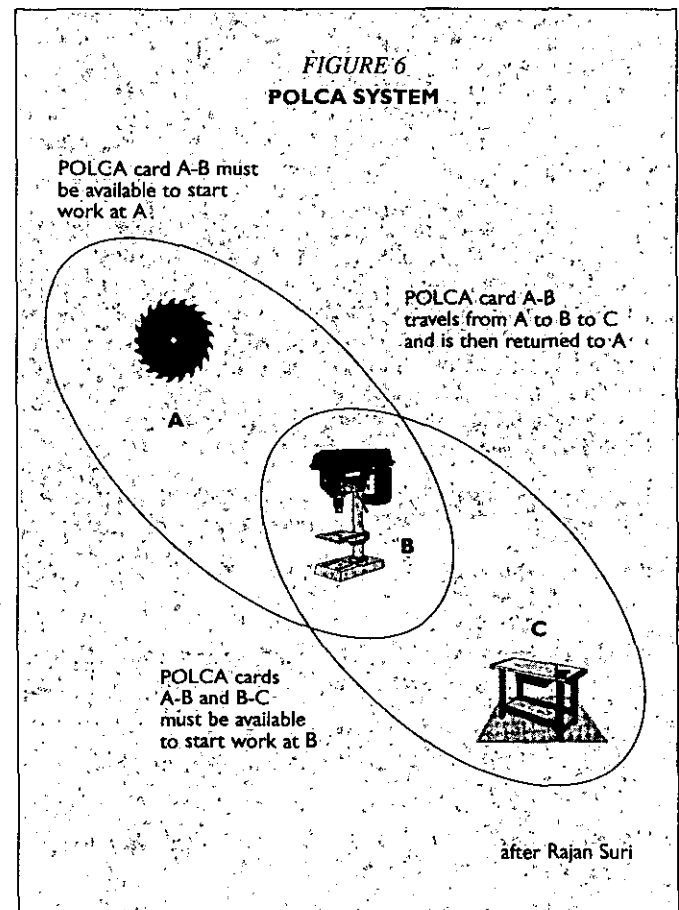
Electronic signals, faxban and so on, of course, can replace CONWIP 'cards'. A priority kanban board (or in this case a priority CONWIP board) as described earlier, can also be used at the start of a CONWIP line.

A further extension allows two categories of priority. Normally first in first out (FIFO) is used, but jobs having 'hot' CONWIP cards are allowed to jump the queue at the first operation or at buffers set up between CONWIP lines.

Why bother with kanban when there is CONWIP? Because, where production stages are well balanced, it is easier to pick up problems faster with kanban. Kanban is a tighter material control system. However, kanban requires strict operating conditions and higher linearity of flow. Of course it is possible to use both kanban and CONWIP simultaneously; CONWIP to control the main flow and kanban to control parts delivery to the line.

## POLCA

CONWIP like kanban requires stable routings. POLCA or 'paired cell overlapping loops of cards with authorisation' (Suri, [5]) is a push-pull system which allows different routings and job-shop type operations. The workings of a POLCA system is as follows (refer to Figure 6).



- Planned release of jobs, bearing in mind capacity limitations, is done on an MRPII/finite scheduling system. This determines the time before which jobs may not start, NOT when they are due to start.
- Each job has a card that travels with it through all workstations. This details all workstations to be visited.
- Each pair of workstations has a number of POLCA cards assigned to them. The number of cards is calculated by

$$\begin{aligned} &((\text{ave lead time through 1} + \text{average lead time through 2}) \times \text{jobs 1-2}) / t \\ &\text{where} \\ &\text{jobs 1-2 is the number of jobs planned to visit work centres 1-2 during } t \\ &t \text{ is the planning horizon, for example 20 days.} \end{aligned}$$

- At the first workstation there needs to be a suitable POLCA card available before work can start. ('suitable' means a POLCA card that has the required routing; for example card 1-2 would do, card 1-4 would not).
- The POLCA card travels with the job during the next THREE workstations (covering two overlapping loops). After completion at the third work centre the POLCA card is returned to the first work centre
- After the work at the first workstation is complete, the job and the POLCA card are moved to the second workstation. At the second workstation, work can only start when there is a suitable POLCA card for workstations 2 and 3. So at all workstations except the first, there must be two cards available to authorise production; one from the first loop and one from the second loop. This ensures that no job is started before there is available capacity at the workstation after next.

POLCA is a fairly complex system compared with traditional kanban, and should not be used where there is linear repetitive production. Both CONWIP and POLCA can handle non-linear demands and changeover operations. POLCA could be considered where routings are irregular or repeat only at infrequent intervals.

## CONCLUSION

There are many types of kanban, the choice of which should suit the system. There are 'kanbans for all seasons' - one should be for you. Kanban is not only a visible up-to-date production control system that helps avoid over production, but its use should foster continuous improvement. New types of kanban have extended the versatility of this powerful tool.

## REFERENCES

- [1] Yashiro Monden, "Toyota Production System", Second Edition, Chapman & Hall, London, 1994
- [2] William Sandras, "Just in Time: Making it Happen", Oliver Wight, 1989
- [3] John Bicheno and Jens Niesmann, "Lean Batch Sizing, Working Paper": Lean Enterprise Research Unit, Cardiff Business School, 1999
- [4] Wallace Hopp and Mark Spearman, "Factory Physics", Irwin, Chicago, 1996
- [5] Rajan Suri, "Quick Response Manufacturing", Productivity Press, Portland OR, 1998

## About the author

John Bicheno FIOM divides his time between the University of Buckingham where he is a Reader in Operations Management and Cardiff Business School where he is working at the Lean Enterprise Research Centre. Before moving to the UK he had 12 years experience in Operations Management and was an Associate Professor of Industrial Engineering. John has consulted and lectured widely on JIT, Quality and Productivity Improvement in the UK, South Africa and Germany.

He is a Fellow of the Institute of Operations Management and the South African Production & Inventory Control Society and a Certified Fellow in Production & Inventory Management. John is the author of three guide books - "Cause & Effect JIT", "Quality 60" and "The Lean Toolbox".

# INVEST IN TRAINING NOW

for success in the future

## PROGRAMME OF SHORT COURSES

- Advanced Shop Floor Control  
20th-21st October 1999
- Basic Techniques for PAC  
15th-16th September 1999
- Costs and Financial Basics for Operations Personnel  
8th-9th December 1999
- Creating a Continuous Improvement Programme  
28th-29th September 1999
- Creating a Customer Service Organisation and Culture  
2nd-3rd November 1999
- Effective Stores Management  
25th-26th November 1999
- Getting the Best from Customer/Supplier Partnerships  
30th September 1999
- Getting the Measure of Your Business  
26th-27th October 1999
- Inventory Management  
9th-11th November 1999
- Inventory Record Accuracy  
13th October 1999
- Lean Manufacturing  
14th-15th December 1999
- Master Planning  
5th-7th October 1999
- Material and Capacity Requirements Planning  
19th-21st October 1999
- MRP: A Hands-On PC Tutorial  
29th September 1999
- MRPII: An Introduction  
7th-9th September 1999 / 7th-9th December 1999
- Negotiation Skills in Operations Management  
10th-11th November 1999
- Optimising the Supply Chain: A Practical Approach  
1st-2nd December 1999
- Re-Engineering the Business through Activity Based Management  
6th October 1999
- Removing and Avoiding Excess and Obsolete Inventory  
8th-9th September 1999
- Shop Floor Control/PAC  
16th-18th November 1999
- Stock Control for Spare Parts  
23rd-24th November 1999
- Strategic Management Workshop: Transforming a Whole Business  
17th-18th November 1999
- Supply Chain Logistics  
12th October 1999
- The Lean Tool Box  
21st-23rd September 1999

ALL OF THE ABOVE PUBLIC COURSES CAN ALSO BE OFFERED ON AN IN-COMPANY BASIS.

For further information, please contact:  
**The Institute of Operations Management**  
 Tel: (02476) 692266 Fax: (02476) 692305  
 Email: [iom@iomnet.org.uk](mailto:iom@iomnet.org.uk)  
 or see our web site: <http://www.iomnet.org.uk>