Planning and control

10.1 What is planning and control?

Planning and control is concerned with the activities that attempt to reconcile the demands of the market and the ability of the operation's resources to deliver. It provides the systems, procedures and decisions that bring different aspects of supply and demand together.

The difference between planning and control

Planning = a formalization of what is intended to happen at some time in the future

Control = the process of coping with the types of change such as the reasons as of why a plan's expectations don't go as expected

Control activities make the adjustments that allow the operation to achieve the objectives that the plan has set even when the assumptions on which the plan was based do not hold true.

Long-, medium- and short-term planning and control

Long-term planning = operations managers make plans concerning what they intend to do, what resources they need, and what objectives they hope to achieve. Emphasis is on planning rather than control since there is little to control. Uses forecasts of likely demand described in aggregated terms.

Medium-term planning and control = looks ahead to assess the overall demand which the operation must meet in a partially disaggregated manner. Must distinguish between different types of demand

Short-term planning and control = many of the resources will have been set and it will be difficult to make large changes. If things are not going to plan, short-term interventions are possible. Demand will be assessed on a totally disaggregated basis

In making short-term interventions and changes to the plan, operations managers will be attempting to balance the quality, speed, dependability, flexibility and costs of their operation dynamically on an ad hoc basis.

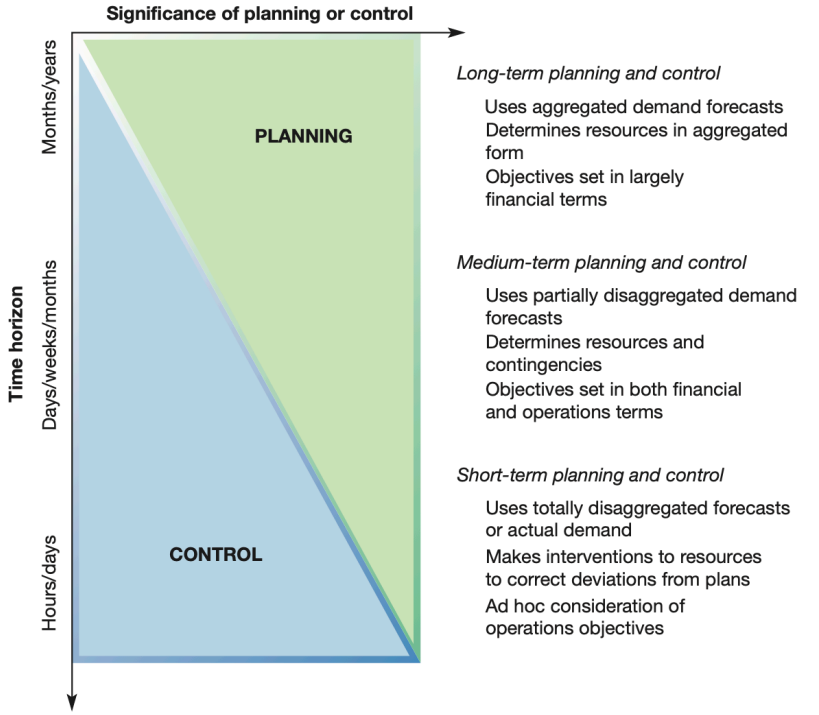


Figure 10.2 The balance between planning and control activities changes in the long, medium and short terms

The volume-variety effect on planning and control

Operations that produce a high variety of services or products in relatively low volume will have customers with different requirements and use different processes form operations that create standardized services or products in high volume

Volume	Variety	Customer responsiveness	Planning horizon	Major planning decision	Control decisions	Robustness
Low	High	Slow	Short	Timing	Detailed	High
↓	↓	↓	↓	↓	↓	↓
High	Low	Fast	Long	Volume	Aggregated	Low

The individual decisions that are taken in the planning process will usually concern the timing of activities and events

Control decisions will concern aggregated measures of output

10.2 How do supply and demand affect planning and control?

Uncertainty in supply and demand

Sometimes the supply of inputs to an operation may be uncertain, planned activities may take longer than expected, and demand may be unpredictable.

Dependent and independent demand

Dependent demand = Some operations can predict demand with relative certainty because demand for their services or products is dependent upon some other factor which is known. Other operations act in a dependent-demand manner because of the nature of the service/product they provide

Planning and control in dependent-demand situations is largely concerned with how the operation should respond when demand has occurred.

Independent demand = need to supply future demand without knowing exactly what the demand will be, do not have firm **forward visibility** of customer orders.

Independent demand planning and control makes best guesses concerning future demand, attempts to put the resources in place that can satisfy the demand and attempts to respond quickly if actual demand does not match the forecast

Responding to demand

Create and deliver to order (Make to order in manufacturing**) operation** = standard services/products are not created until the customer has chosen which particular service/product to have

Partially create and deliver to order operation (Assemble to order in manufacturing) **operation** = services/products are so predictable that they can start to create them before specific customer orders arrive

Created to stock (Make to stock) order = standardized services/products can be created before demand is known

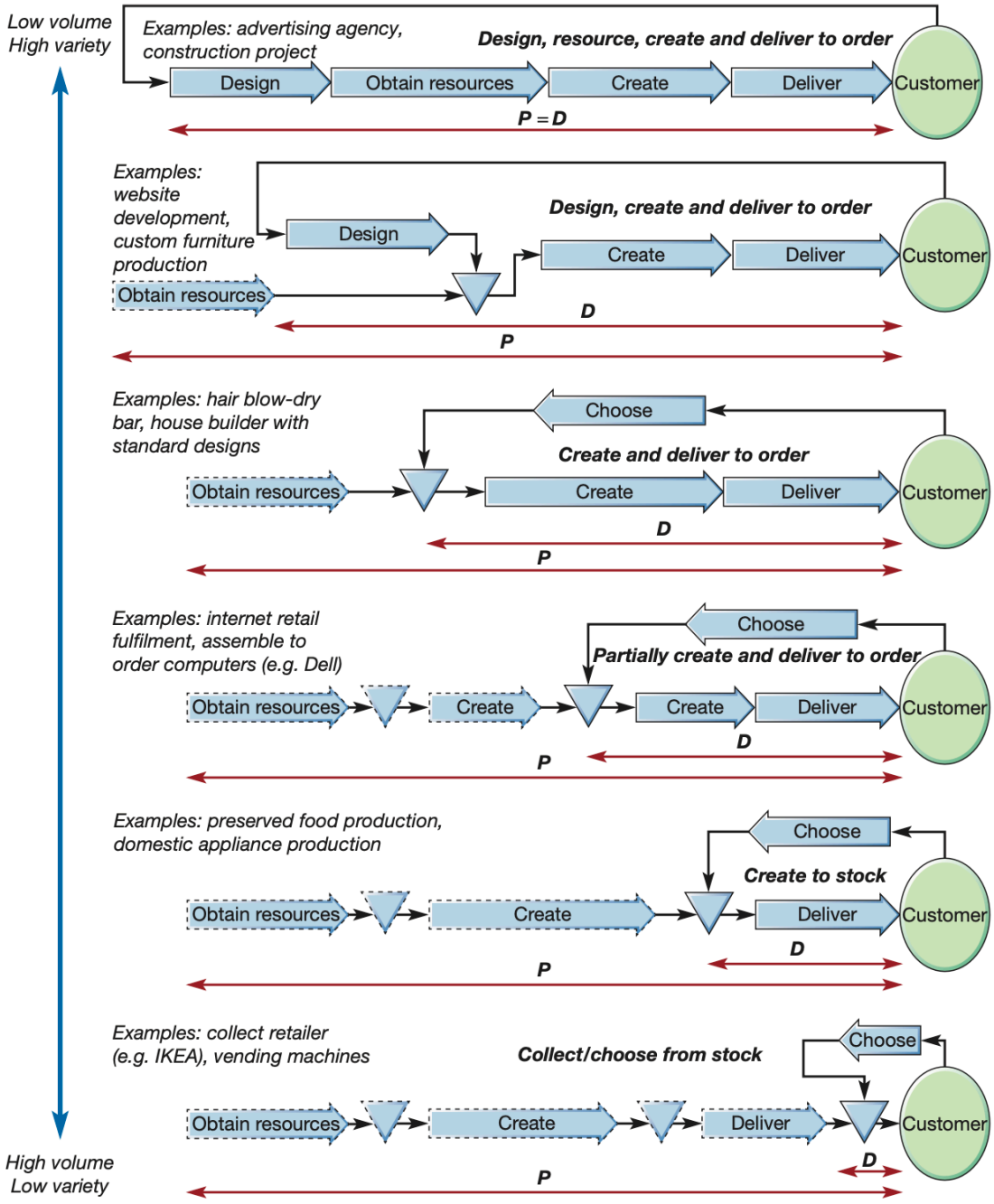


Figure 10.4 The *P:D* ratio of an operation indicates how long the customer has to wait for the service or product as compared with the total time needed to carry out all the activities to make the service or product available to the customer

Design, resource, create and deliver to order operations are intended for low-volume and high-variety businesses

P:D ratios

P:D ratio = contrasts the toal length of time customers have to wait between asking for the service or product and receiving it. **D = demand time**, **P = total throughput time from start to finish**

Throughput time = how long the operation takes to design the service or product, obtain the resources, create and deliver it

P and D times depend on the operation

The ratio of P to D gets larger as operations move from “design, resource, create and deliver to order” to “Collect/choose from stock”. That means that as one moves down the spectrum towards the “Create to stock” and “Collect/choose from stock” end the operation has anticipated customer demand and already created the services and products even though it has no guarantee that the anticipated demand will really happen.

The larger the P:D ratio, the more speculative the operation’s planning and control activities will be. By reducing the P:D ratio, operations reduce their degree of speculative activity and also reduce their dependence on forecasting

When the P:D ratio approaches 1, not all uncertainty is eliminated since the volume of demand may be known but not the time taken to perform each order.

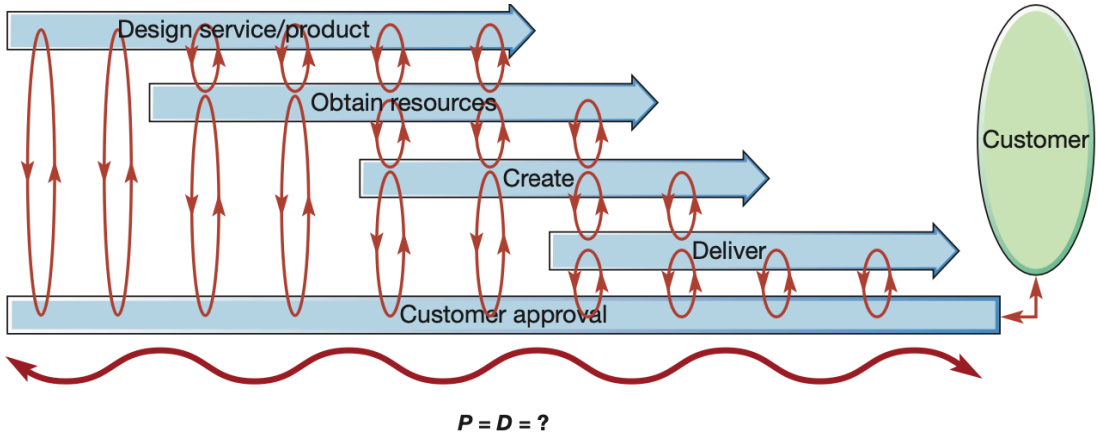


Figure 10.5 The relationship between stages in some ‘Design, resource, create and deliver to order’ operations, such as an advertising agency, can be complex with frequent consultation and unpredictable recycling

Sales and operations planning (S&OP)

One of the problems with traditional planning and control is that although several functions were often routinely involved in the process, each function could have a very different set of objectives.

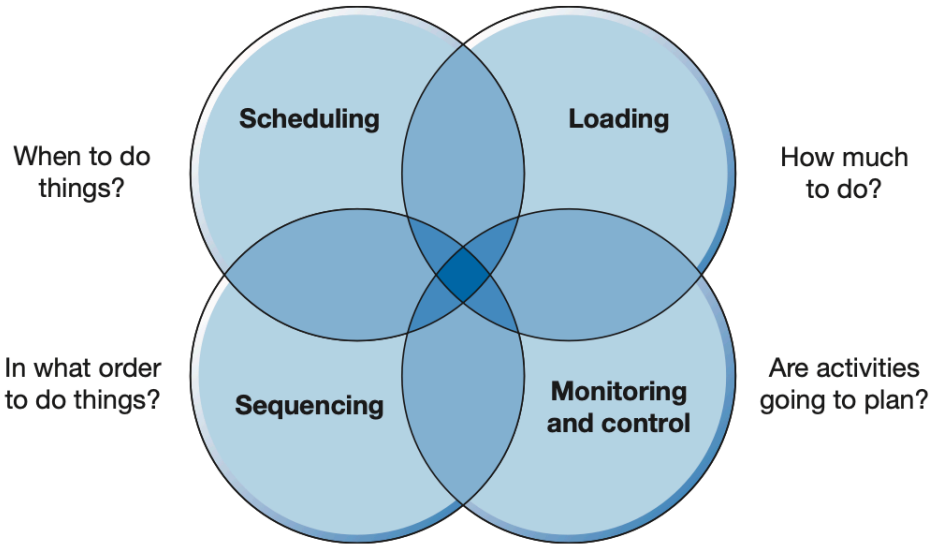
Functions are impacted by operations planning decisions and are probably involved in their own planning processes that partly depend on the output from the operations planning process.

Sales and operations planning = planning process that attempts to ensure that all tactical plans are aligned across the business’s various functions and with the company’s longer-term strategic plans. It is a formal business process that looks over a period of 18 to 24 months ahead. An aggregated process that does not deal with detailed activities but rather focuses on the overall volume of output

The activities of planning and control

Planning and control requires the reconciliation of supply and demand in terms of volumes, timing and quality.

There are 4 overlapping activities that plan and control volume and timing: **loading, sequencing, scheduling, and monitoring and control**



10.3 What is loading?

Loading = the amount of work that is allocated to a work centre

A machine can be available for X hours a week, but that does not mean that X hours of work can be loaded onto that machine. For some periods the machine cannot be worked i.e. during weekends, and of the time the machine is available for work, other losses further reduce the available time i.e. when changing components.

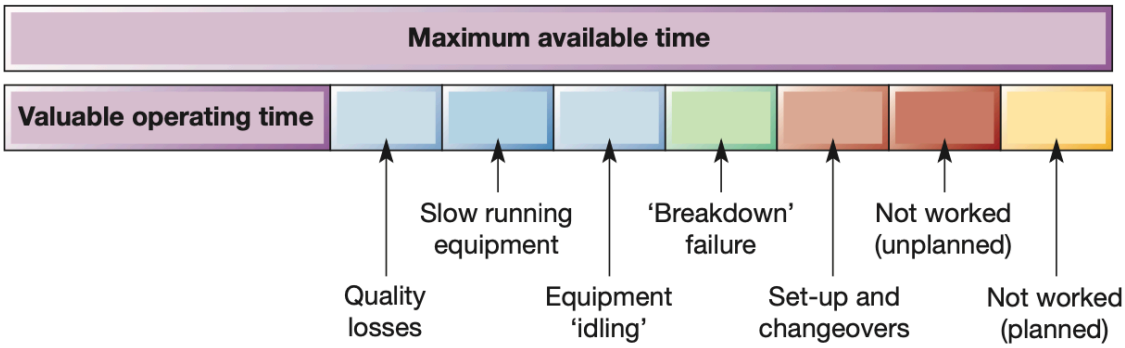


Figure 10.7 The reduction in the valuable operating time available

Finite and infinite loading

Finite loading = approach that only allocates work to a work centre (person, machine) up to a set limit which is the estimate of capacity for the work centre. Work over and under this capacity is not accepted

Finite loading is relevant for operations where:

- It is possible to limit the load
- It is necessary to limit the load
- The cost of limiting the load is not prohibitive

In complex planning and control activities where there are multiple stages, each with different capacities and with a varying mix of arriving at the facilities, the constraints imposed by finite loading make loading calculations complex and **not worth** the considerable computational power that would be needed

Infinite loading = approach to loading work that does not limit accepting work but instead tries to cope with it

Infinite loading is relevant for operations where:

- It is not possible to limit the load
- It is not necessary to limit the load
- The cost of limiting the load is prohibitive

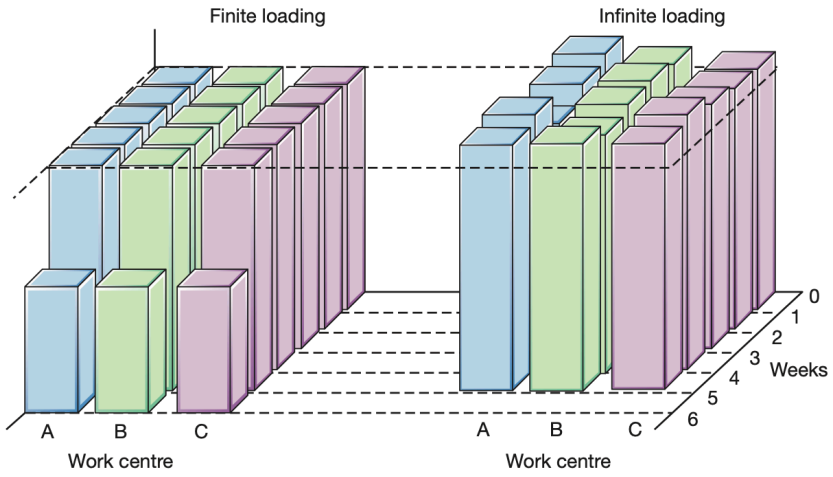


Figure 10.8 Finite and infinite loading of jobs on three work centres A, B and C. Finite loading limits the loading on each centre to its capacity, even if it means that jobs will be late. Infinite loading allows the loading on each centre to exceed its capacity to ensure that jobs will not be late

10.4 What is sequencing?

Sequencing = decisions taken on the order in which the work will be tackled. Priorities given to work in an operation are determined by a predefined set of rules:

Physical constraints = the physical nature of the inputs being processed may determine the priority of work, for example lighter colors might be sequenced before darker colors. Also jobs that fit together physically may be scheduled together to reduce waste

Customer priority = allows an important or aggrieved customer/item to be processed prior to others irrespective of the order of arrival of the customer or item. Typically used by operations whose customer base is **skewed** containing a mass of

small customers and a few large very important customers (i.e. banks). It may erode the service given to many others and lower the overall performance of the operation if work flows are disrupted to accommodate important customers.

Due date (DD) = work is sequenced according to when it is due for delivery, irrespective of the size of each job or the importance of each customer. DD usually improves the delivery dependability and average delivery speed but may not provide optimal productivity as a more efficient sequencing of work may reduce total costs. Can be flexible when new, urgent work arrives at the work centre.

Last in First out LIFO = for example unloading an elevator

First in First out FIFO = customers are served in exactly the sequence they arrive in.

Longest operation time LOT = the longest jobs are sequenced first. Advantage of occupying work centres for long periods. Keeps utilization high but does not take into account delivery speed, reliability or flexibility

Shortest operation time first SOT = in a stage where operations become cash constrained the sequencing rules may be adjusted to tackle short jobs first. These can be invoiced and payment received to ease cash-flow problems. Improves delivery performance, if the unit of measurement of delivery is jobs but may affect total productivity and can damage service to larger customers

Judging sequencing rules = the 5 performance objectives could be used to judge the effectiveness of sequenceing rules. The following performance objectives are often used:

- Meeting due date promised to customer (**dependability**)
- Minimising the time the job spends in the process, also known as flow time (**speed**)
- Minimizing the work-in-progress inventory (**element of cost**)
- Minimizing idle time of work centres (**element of cost**)

Table 10.2 Comparison of five sequencing decision rules

Rule	Average time in process (days)	Average lateness (days)
FIFO	12	6.4
DD	8.4	3.2
SOT	7.6	3.2
LIFO	8.4	3.8
LOT	12.8	7.4

SOT rule results in both the best average time in process and the best in terms of average lateness

10.5 What is scheduling?

Scheduling = Some operations require a detailed timetable showing at what time or date jobs should start and when they should end. They are familiar statements of volume and timing in many consumer environments

Schedules of work are used in operations where some planning is required to ensure that customer demand is met

Rapid-response service operations cannot schedule the operation in a short-term sense since they only can respond at the time demand is placed upon them since customers arrive in an unplanned way.

The complexity of scheduling

Scheduling is complex due to many reasons:

1. First schedulers must deal with several different types of resource simultaneously (machines have different capabilities and capacities, staff have different skills)
2. The number of possible schedules increases rapidly as the number of activities and processes increases, for n jobs there are n! factorial ways of scheduling the jobs through a single process
3. When there are many machines however, their sequencing is independent of each other so the formula for the number of possible schedules is $\text{Number of possible schedules} = (n!)^m$ where n = number of jobs, m = number of machines

Forward and backward scheduling

Forward scheduling = starting work as soon as it arrives

Backward scheduling = starting jobs at the last possible moment to prevent them from being late

Table 10.3 The effects of forward and backward scheduling

Task	Duration	Start time (backwards)	Start time (forwards)
Press	1 hour	3.00 pm	1.00 pm
Dry	2 hours	1.00 pm	11.00 am
Wash	3 hours	10.00 am	8.00 am

Table 10.4 Advantages of forward and backward scheduling

Advantages of forward scheduling	Advantages of backward scheduling
High labour utilisation – workers always start work to keep busy	Lower material costs – materials are not used until they have to be, therefore delaying added value until the last moment
Flexible – the time slack in the system work to be loaded	Less exposed to risk in case of schedule change by the customer
	Tends to focus the operation on customer due dates

Gantt charts

Gantt chart = simple device which represents time as a bar or channel on a chart. The start and finish times for activities can be indicated on the chart and sometimes the actual progress of the job is also indicated

Advantage is that they provide a simple visual representation both of what should be happening and of what actually is happening in the operation

Scheduling work patterns

Where the dominant resource in an operation is its staff, then the schedule of work times effectively determines the capacity of the operation itself.

Rostering = Main task of scheduling in that case is to make sure that sufficient numbers of people are working at any point in time to provide a capacity appropriate for the level of demand at that point in time

High-visibility operations need to schedule the working hours of their staff with demand in mind. Those operations cannot store their outputs in inventories and must respond directly to customer demand

2 timescales must be considered, during the day working hours need to be agreed with individual staff members and during the week, days off need to be agreed. All must be scheduled such that:

- Capacity matches demand
- The length of each shift is neither excessively long nor too short to be attractive to staff
- Working at unsocial hours is minimized
- Days off match agreed staff conditions
- Vacations and other time-off blocks are accommodated
- Sufficient flexibility is built into the schedule to cover for unexpected changes in supply (staff illness) and demand (surge in customer calls)

Theory of constraints (TOC)

Theory of constraints = focuses scheduling effort on the bottleneck pars of the operation. By identifying the location of constraints, working to remove them, and then looking for the next constraint, an operation is always focusing on the part that critically determines the pace of output

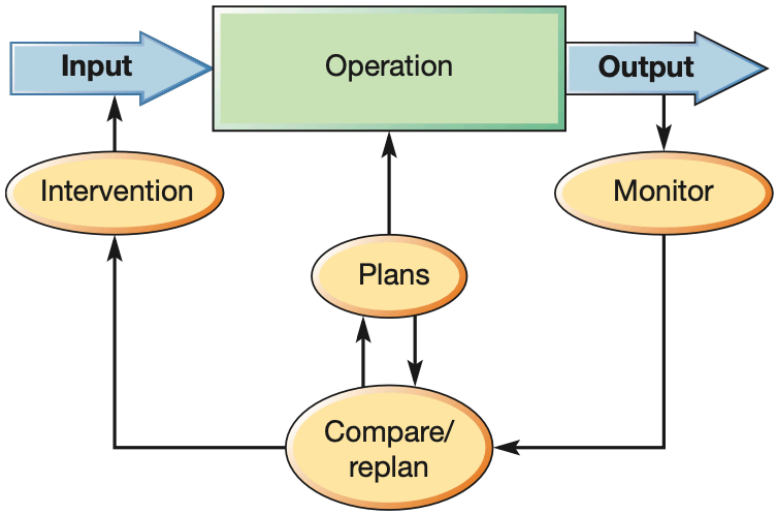
Optimised technology OPT = approach that uses the idea of theory of constraintns. Helps to schedule production systems to the pace dictated by the most heavily loaded resources (bottlenecks). If the rate of activity in any part of the system exceeds that of the bottleneck, then items are being produced that cannot be used. If the rate of working falls below the pace at the bottleneck, then the entire system is underutilized

OPT principles

1. Balance flow, not capacity. It is more important to reduce throughput time rather than achieving a notional capacity balance between stages or processes.
2. The level of utilisation of a non-bottleneck is determined by some other constraint in the system, not by its own capacity. This applies to stages in a process, processes in an operation, and operations in a supply network.
3. Utilisation and activation of a resource are not the same. According to the TOC, a resource is being *utilised* only if it contributes to the entire process or operation creating more output. A process or stage can be *activated* in the sense that it is working, but it may only be creating stock or performing other non-value-added activity.
4. An hour lost (not used) at a bottleneck is an hour lost forever out of the entire system. The bottleneck limits the output from the entire process or operation, therefore the underutilisation of a bottleneck affects the entire process or operation.
5. An hour saved at a non-bottleneck is a mirage. Non-bottlenecks have spare capacity anyway. Why bother making them even less utilised?
6. Bottlenecks govern both throughput and inventory in the system. If bottlenecks govern flow, then they govern throughput time, which in turn governs inventory.
7. You do not have to transfer batches in the same quantities as you produce them. Flow will probably be improved by dividing large production batches into smaller ones for moving through a process.
8. The size of the process batch should be variable, not fixed. The circumstances that control batch size may vary between different products. (See discussion of the EBQ model in Chapter 13.)
9. Fluctuations in connected and sequence-dependent processes add to each other rather than averaging out. So, if two parallel processes or stages are capable of a particular average output rate, in parallel, they will never be able to produce the same average output rate.
10. Schedules should be established by looking at all constraints simultaneously. Because of bottlenecks and constraints within complex systems, it is difficult to work out schedules according to a simple system of rules. Rather, all constraints need to be considered together.

10.6 What is monitoring and control?

Each part of the operation has to be monitored to ensure that planned activities are indeed happening. Any deviation from the plans can then be rectified through some kind of intervention in the operation, which itself will probably involve some re-planning.



Push and pull control

Periodic intervention into the activities of the operation is one element of control. The key distinction is between intervention signals which push work through the processes within the operation, and those which pull work only when it is required.

Push system of control = activities are scheduled by means of a central system and completed in line with central instructions. Each work centre pushes out work without considering whether the succeeding work centre can make use of it. Inventory and queues often characterize push systems

Pull system of control = the pace and specification of what is done are set by the customer workstation which pulls work from the preceding supplier workstation. Customer acts as the only trigger for movement. If a request is not passed back from the customer to the supplier, the supplier cannot produce anything or move any materials.

The inventory consequences of push and pull

Pull systems are less likely to result in inventory build-up and are favored by lean operations.

A push system is represented by an operation, each stage of which is on a lower level than the previous stage. When items are processed by each stage, gravity pushes them down the slope to the next stage. Any delay or variability in processing time at that stage will result in the items accumulating as inventory.

In the pull system items cannot naturally flow uphill so they can only progress if the next stage along deliberately pulls them forward and inventory cannot accumulate as easily under such circumstances.

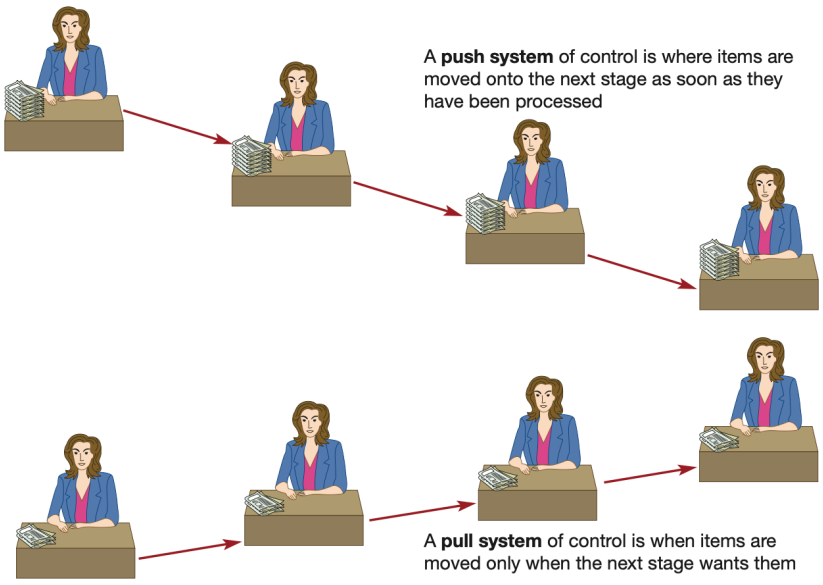


Figure 10.14 Push versus pull: the gravity analogy

Drum, buffer, rope

Drum, buffer, rope concept = idea that helps decide exactly where in a process control should occur.

Most do not have the same amount of work loaded onto each separate work centre (not perfectly balanced). → there is likely to be a part of the process that is acting as a bottleneck on the work flowing through the process.

Drum = TOC argues that the bottleneck in the process should be the control point of the whole process, it sets the beat for the rest of the process to follow

Since it does not have sufficient capacity, a bottleneck is working all the time and therefore it is sensible to keep a **buffer** of inventory in front of it to make sure that it always has something to work on

Rope = form of communication between the bottleneck and the input to the process that is needed to make sure that activities before the bottleneck do not overproduce