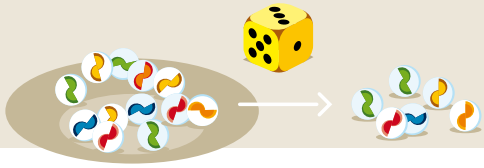
THE ALIGNMENT PUZZLE WHITE PAPER

Calculation Models for Supply Chain Alignment



1. Introduction to Calculation Models

In this chapter, we have grouped together a number of topics that, in our view, are very useful when working to improve alignment in the flow of goods, but which may be somewhat dense and dry for some readers. For that reason, we simplified or removed these topics from the book The Alignment Puzzle and instead published them as the white paper now before you.

These topics mainly concern calculation models related to the law of conservation of waiting time, formulas for determining batch sizes, inventory levels, and lead times. We also discuss a number of common mistakes in detail and present objections to well-known methods in order to help readers avoid pitfalls.

1.1 Lead Times, Waiting Times, Inventory Levels, and Batch Formation

Inventories form the buffers between the processes that supply goods and the processes that consume goods. They are essential building blocks of a well-aligned organization.

The task of inventory management is to ensure that the right goods are available in inventory, in the right quantities, at the right time. If manufacturers or trading companies maintain inventory levels that are too low, the service level is put at risk. When inventory levels are too high, they have spent too much money too early, with negative consequences for financing costs and risk.

Inventory can be viewed in different ways: as the waiting time of goods, as a capacity buffer, and as an insulating layer that shields production processes from the turbulence of the market.

If you view inventory as the waiting time of goods that prevents customers or suppliers from having to wait, it is useful to examine the law of conservation of waiting time more closely. How can you use it in calculations, and how can you use the results of those calculations to support decision-making? We will address these questions in the next section.

2. The Law of Conservation of Waiting Time

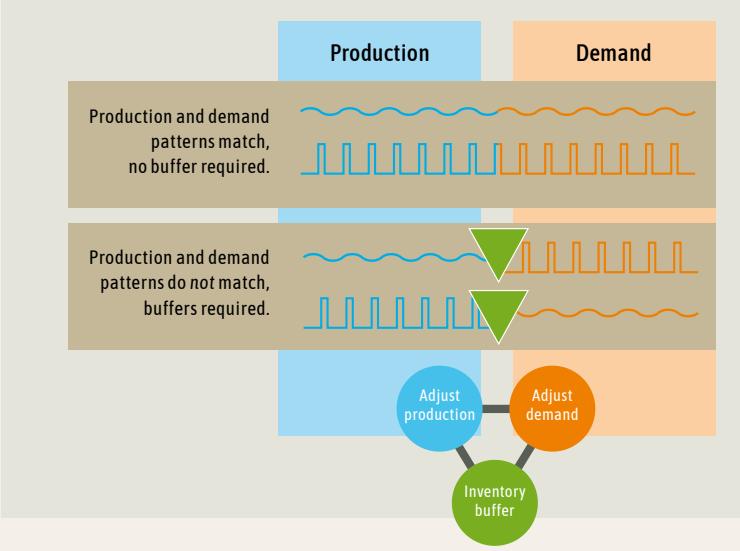
2.1 Who Waits When Supply and Consumption Do Not Align Seamlessly?

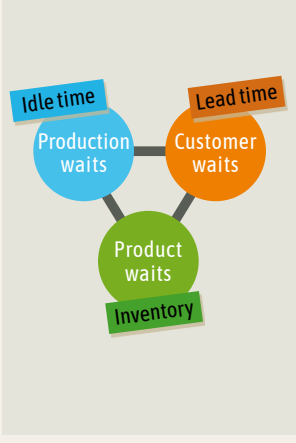
The law of conservation of waiting time states that something or someone must always wait when patterns of supply and demand do not align perfectly: either the customer, the supplier, or the product. Or a combination of these.

Both demand and supply may each have their own natural pattern. Consider, for example, a grain harvest that occurs only once a year, while demand for flour and bread remains steady throughout the year. The result is either inventory accumulation or waiting customers. Or consider a manufactured product for which there is demand every day, but which is produced only once per month in order to reduce setup times.

The Law of Conservation of Waiting Time

If the patterns of demand and production do not match perfectly, something or someone must always wait: either the customer, the product, or the production process. Or a combination of these.





2.2 Simple Numerical Example

A simple numerical example illustrates how total waiting time can be distributed among the product, the customer, and the supplier. Suppose that over a period of six weeks (or six days or six months—it does not matter), there is one inbound batch of six units and four sales orders: two orders of two units and two orders of one unit. We start with an inventory level of zero.

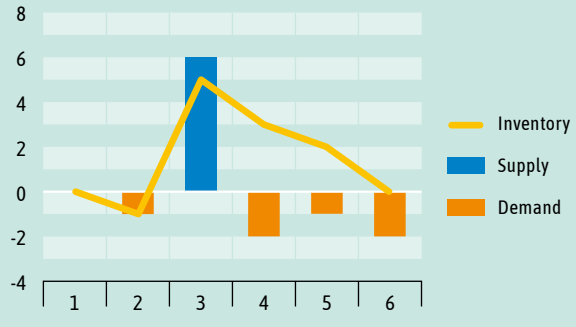
Period	1	2	3	4	5	6	Total
Demand		1		2	1	2	***6***
Supply			6				***6***
Inventory	0	-1	5	3	2	0	

When the first customer order arrives in week 2, we have no inventory, so the customer must wait until week 3 when a new production batch of six units arrives.

A negative inventory level in the table above represents backlogged demand, meaning that the customer is waiting. We then have five units remaining in inventory. In week 4, two units are demanded, reducing inventory to three units, and so on. We can now fairly easily calculate the total waiting time of customers and products. We express this in "Unit-Days" of waiting time. For example, if three products remain in inventory for two days, this amounts to six Unit-Days of waiting time.

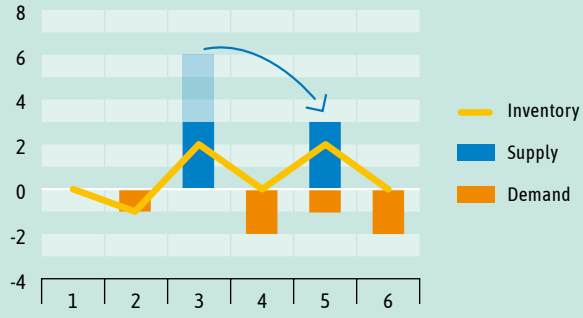
In the numerical example given, the total waiting time amounts to eleven Unit-Days. Specifically, ten Unit-Days of waiting for the product and one Unit-Day of waiting for the customer.

Period	1	2	3	4	5	6	Total
Demand		1		2	1	2	*6*
Supply			6				*6*
Inventory	0	-1	5	3	2	0	
Waiting times							
Customer waits	0	1	0	0	0	0	*1*
Product waits	*0*	*0*	5	3	*2*	*0*	*10*
Machine waits	0	0	0	0	0	0	*0*
Total unit-days	0	1	5	3	2	0	***11***



Machine waiting time is still zero because we assume the natural production supply pattern. We can now ask what happens if we choose to deviate from that supply pattern in order to align more closely with the demand pattern. For example, we could split the batch of six units into two batches of three units each, with one delivered in week 3 and the other in week 5. We postpone half of the production by two weeks. In other words, we let the supply side wait.

Period	1	2	3	4	5	6	Total
Demand		1		2	1	2	*6*
Supply			3		3		*6*
Inventory	0	-1	2	0	2	0	
Waiting times							
Customer waits	0	1	0	0	0	0	*1*
Product waits	*0*	*0*	*2*	*0*	*2*	*0*	*4*
Machine waits	0	0	3	3	0	0	*6*
Total unit-days	0	1	5	3	2	0	***11***

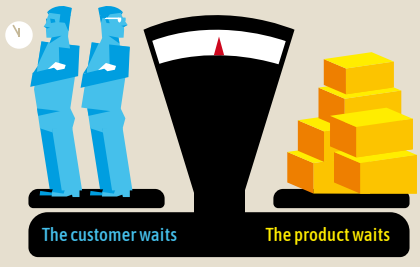


As a result of this adjustment, average inventory decreases, but the supply of three units must wait for two days. The calculation shows that the total waiting time remains eleven Unit-Days. It is simply distributed differently: one Unit-Day for the customer, four Unit-Days for the product, and six Unit-Days for the machine. Together, once again, eleven Unit-Days.

2.3 The Trade-off Between Letting Products Wait and Letting Customers Wait

Excessive inventory is undesirable because of financing costs, risk, and storage requirements. Inventory levels that are too low are undesirable because service levels may be negatively affected. Finding the optimal balance between these two extremes is one of the central challenges of supply chain management. Sometimes it is sensible to reduce inventory levels to the point where customers occasionally have to wait briefly for their goods.

The Law of Conservation of Waiting Time



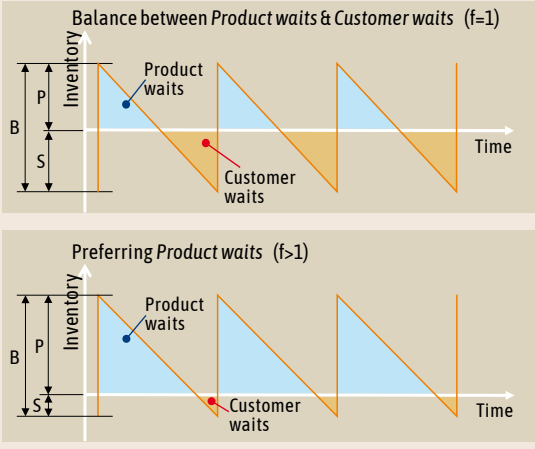
When we encounter a situation in which the supply pattern and demand pattern do not align well and cannot be adjusted, we must choose between two undesirable outcomes: do we let the customer wait, or do we let the product wait? We can influence this choice by selecting the initial inventory level and therefore also the average inventory level (assuming, of course, that this is possible).

With batch deliveries and uniform demand, inventory follows a sawtooth pattern.

Negative inventory means the customer is waiting. Positive inventory means the product is waiting. By choosing a higher or lower initial inventory level, we can shift the sawtooth pattern upward or downward and thereby influence the balance between Customer-Wait and Product-Wait.

It is very difficult to determine the precise impact on cash flows of a product spending one day in inventory versus a customer waiting one day for a product. To avoid making arbitrary assumptions about allocated costs, we introduce a weighting factor (f) that represents the ratio between these two effects. This weighting factor expresses how many times more undesirable it is to have a customer wait one day for a product than to have a product wait one day for a customer. If both situations are considered equally undesirable, then f equals 1. In most cases, however, f will be greater than 1 because organizations generally consider customer waiting time more undesirable than holding inventory.

The Law of Conservation of Waiting Time the inventory sawtooth



A batch delivery pattern combined with a uniform demand pattern produces a sawtooth inventory pattern. Negative inventory: Customer waits. Positive inventory: Product waits. The weighting factor *f* determines the optimal balance.

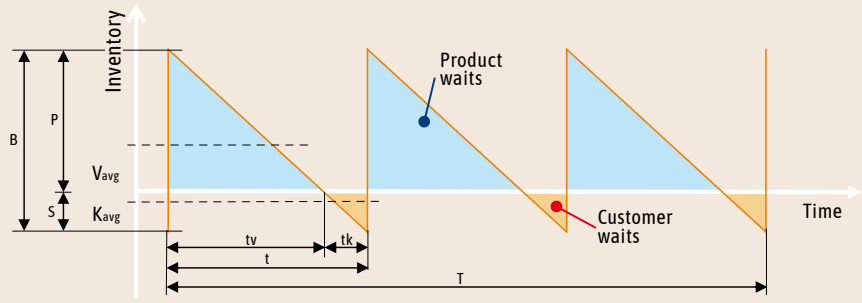
Variables

B	Batch size
S	The shortage at the moment the new batch arrives; i.e. the outstanding customer demand.
P	Remaining batch after satisfying the backlogged demand (P = B–S).
f	Weighting factor expressing how many times worse it is to let a customer wait for one day than to hold a product in inventory for one day.
C	Waiting-time cost. Calculated as the total number of product waiting days plus f times the total number of customer waiting days.

2.4 Formulas: The Optimal Balance Between Customer Waits and Product Waits

If we make a number of significant assumptions about the situation, it is relatively straightforward to derive the formula for the optimal equilibrium. We assume a calendar consisting solely of working days and perfectly predictable, uniform demand. We also assume that it is possible to determine a factor f indicating how many days we are willing to hold a product in inventory in order to prevent a customer from waiting one day.

Law of Conservation of Waiting Time Deterministic Version



Derivation of the optimal value of S
Determining the optimal balance between *customer waits* and *product waits* for a given number of batches, a given cost ratio, and perfectly uniform demand..

$c = V_{avg} + f * K_{avg} = \frac{1}{2}P^2/B + f * \frac{1}{2}S^2/B$

$B = Q/R$

$P = B - S = Q/R - S$

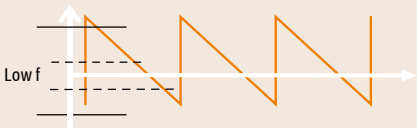
$c = \frac{1}{2}R * (Q/R - S)^2/Q + f * R * \frac{1}{2}S^2/Q$

$c = \frac{1}{2}R/Q * ((Q/R - S)^2 + f * S^2)$

$C = R * c = \frac{1}{2}R^2/Q * ((Q/R - S)^2 + f * S^2)$

$C' = R/Q * (1 + f) * S - 1$

$C' = 0$ IF **$S = Q/(R + R*f)$**



Variables used – Inputs

T	Length of the period under consideration.
Q	Demand during the period under consideration = total supply during the period under consideration, expressed in days of demand.
R	Number of batches in the chosen period.
f	Factor expressing how many times worse it is to let a customer wait for one day than to hold a product in inventory for one day.

Variables used – To be calculated

t	Cycle length between two batches ($t = T / R$).
B	Batch size, expressed in days of demand ($B = Q / R$ and $B = t$, expressed in days of demand).
S	The shortage at the moment the new batch arrives, i.e. the back-logged customer demand, expressed in days of demand. The optimal situation is obtained when $S = B / (1 + f) = Q / (R + R \cdot f)$.
P	Remaining batch after satisfying the backlogged demand ($P = B - S$ and $P = tv$, expressed in days of demand).
V_{avg}	Average inventory during T ($V_{avg} = \frac{1}{2}P^2/B$).
K_{avg}	Average customer waiting time per product during T ($K_{gem} = \frac{1}{2}S^2/B$).
c	Waiting-time cost per cycle: inventory waiting days plus f times customer waiting days ($c = V_{avg} + f * K_{avg}$).
C	Waiting-time cost during the period under consideration: inventory waiting days plus f times customer waiting days ($C = R * c$)..

Using the above set of formulas, it becomes easy to calculate, for example, the effect of increasing the setup budget or reducing batch sizes on average inventory and customer delivery performance. We assume that after adjusting batch sizes, the system is returned to the then-optimal balance between Material-Wait and Customer-Wait. Naturally, all previously stated assumptions regarding perfectly uniform demand and the absence of holiday schedules also apply.

2.5 Numerical Example: The Impact of More Frequent Setups on Inventory and Lead Time

We use the simplest possible example to illustrate the formulas presented above. Suppose we have a perfect sawtooth situation. We examine a calendar year of 365 days during which we produce six batches, all of equal size. Total production volume exactly matches total demand volume. We choose a weighting factor f of 20. In other words, we would rather have a product remain in inventory for 19 days than have a customer wait one day for it.

Calculation example for the Law of Conservation of Waiting Time Deterministic version

How much does inventory decrease and how much better are customers served if we produce smaller batches?

Variables used, data		Alt 0	Alt 1		Alt 2		Alt 3	
T	Length of the period under consideration	365	365		365		365	
Q	Demand in the period under consideration = total supply in the period under consideration, expressed in days of demand	365	365		365		365	
R	Number of batches in the selected period	6	9	50%	9	50%	12	100%
f	Factor expressing how many times worse it is to make a customer wait a day than to keep a product in stock for a day	20	20	0%	100	400%	20	0%

Variables used, to be calculated								
t	Cycle length between two batches	t = T/R	60.8	40.6		40.6		30.4	
B	Batch size, expressed in days of demand	B = Q/R B = t	60.8	40.6		40.6		30.4	
S	The shortage at the moment the new batch arrives (waiting customer demand) at optimal balance between *Customer waits* and *Product waits*	S = B / (1+f) S = Q / (R + R*f)	2.9	1.9		0.4		1.4	
P	Remainder of the batch after consumption of backlog demand	P = B-S	57.9	38.6		40.2		29	
V_{avg}	The average inventory during T	$V_{avg} = \frac{1}{2}P^2 / B$	**27.6**	**18.4**	**-33%**	**19.88**	**-28%**	**13.79**	**-50%**
K_{avg}	The average waiting time of a customer per product during T	$K_{avg} = \frac{1}{2}S^2 / B$	**0.07**	**0.05**	**-33%**	**0.00**	**-97%**	**0.03**	**-50%**
c	Cost of waiting days inventory plus f times waiting days customers per day	$c = V_{avg} + f*K_{avg}$	29.0	19.3		20.1		14.5	
C	Cost inventory waiting days plus f times customer waiting days in period T	C = T * c	10,573	7,049	-33,3%	7,328	-30.7%	5,287	-50.0%

The calculation results are shown in the column labeled Alt0. We achieve an optimal situation when the average inventory throughout the year is maintained at 27.59 units and customers experience an average delivery lead time of 0.07 days. This is achieved when a new batch arrives at the moment there are 2.9 units of backlogged demand. This backlog is relatively small because the weighting factor f is fairly large (20). The total waiting-time sacrifice amounts to 10,573 Unit-Days.

It is now interesting to examine the impact of a change in the setup budget. Suppose that, through a certain investment, we are suddenly able to perform 50% more setups. Instead of producing six batches, we can now produce nine. Demand remains uniform, so naturally the batches become 33% smaller. What is the impact on inventory levels and customer lead times?

The results are shown in the column labeled Alt1. Average inventory, average lead time, and total waiting-time sacrifice all decrease by 33%. Customers now wait an average of only 0.05 days. Put more plainly: 19 out of 20 customers can take the product immediately, and only one out of 20 customers experiences a one-day delivery time.

We can also examine the effect of the weighting factor f. If we increase it from 20 to 100, the balance shifts further in favor of reducing Customer-Wait. Customer waiting time decreases by another 97% and becomes too small to express with two decimal places. Inventory, however, increases. Customers wait less, but products wait more.

The column labeled Alt3 shows the impact of increasing the number of setups once again while returning the weighting factor to its original value. And so on. The spreadsheet is easy to recreate and can also be downloaded free of charge from www.alignmentpuzzel.nl, allowing experimentation with your own values.

3. Camp's Formula for Economic Order Quantity (EOQ)

3.1 The Derivation of the EOQ Formula

The best-known method for determining batch size is the so-called EOQ formula (Economic Order Quantity), also known as Camp's formula. Given the widespread familiarity of this formula, it is all the more concerning that the assumptions, prerequisites, and conditions under which it may—and especially may not—be used are much less well known. This is a source of many problems and instances of misalignment, which is why we address it here.

Let us begin with a brief explanation of the formula. It assumes a sawtooth pattern: batch replenishment combined with uniform demand. The formula balances inventory costs against setup costs. The larger the batches, the fewer setups are required per year and therefore the lower the setup costs.

The downside is that large batches result in higher inventory levels. Cycle stock is equal to half the batch size. This is therefore the quantity of goods that remains in inventory on average throughout the year, in addition to any safety stock.

Inventory costs form a straight-line function of batch size. Setup costs are inversely proportional to the number of setups and therefore form a hyperbola. Adding the two together produces a function with a minimum at the point where the two curves intersect.

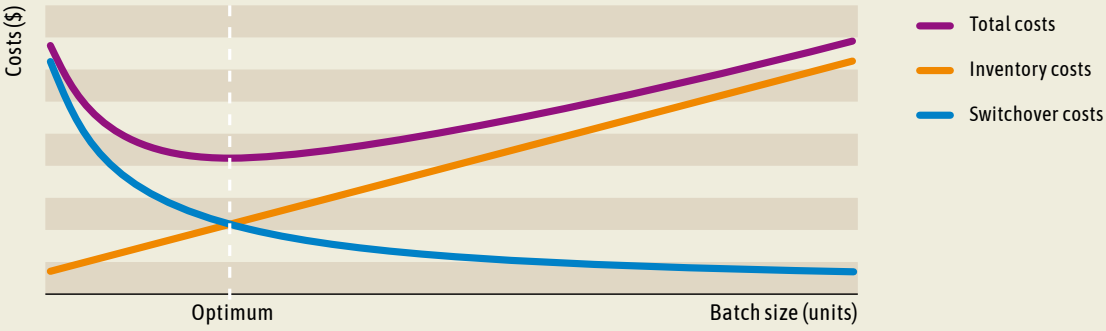
The figure illustrates how, by taking the derivative and setting it equal to zero, it can be shown that minimum cost is achieved when the batch size equals the square root of two times demand multiplied by setup cost, divided by inventory cost.

A simple example illustrates its use: suppose the setup cost S for performing one setup is $20, the inventory holding cost V for keeping one product in stock for one year is $1, and annual demand D equals 1,000 units. Then:

EOQ = SQRT(2 × 20 × 1,000 / 1) = 200.

According to this approach, it is therefore most economical to produce five batches of 200 units per year.

Calculation of the Economic Batch Size Q_{opt} (Camp's Formula)



Derivation of the formula
Determination of the optimal ratio between *customer waits* and *material waits* for a given number of batches, cost ratio, and (perfectly uniform) demand.

$V_{yr} = Q/2 \times V$ *(Avg. half a batch in stock)*

$S_{yr} = D/Q \times S$ *(Number of changeovers per year)*

$C_{jr} = V_{yr} + S_{yr} = Q/2 \times V + D/Q \times S$

$C'_{jr} = V/2 - S \times D/Q^2$

$C'_{jr} = 0$ IF $Q^2 = (2SD/V)$

$$Q_{opt} = \sqrt{2SD/V}$$

Variables used, data

D	pcs	Yearly demand
V	$	Voorraadkosten per stuk per jaar
S	$	Kosten per setup (omstelkosten)

Auxiliary variables

Q	pcs	Batch size
V_{yr}	$	Inventory costs per year
S_{yr}	$	Setup costs per year
C_{yr}	$	Total costs per year

Calculation result

Q_{opt}	pcs	The optimal batch size

Example
Perfectly uniform annual demand D = 1,000 units
Inventory costs V = $1 per unit per year
Changeover costs S = $20 per setup

$$Q_{opt} = \sqrt{2SD/V} = \sqrt{2 \times 20 \times 1{,}000 / 1} = 200 \text{ units}$$

3.2 Assumptions of the EOQ Formula

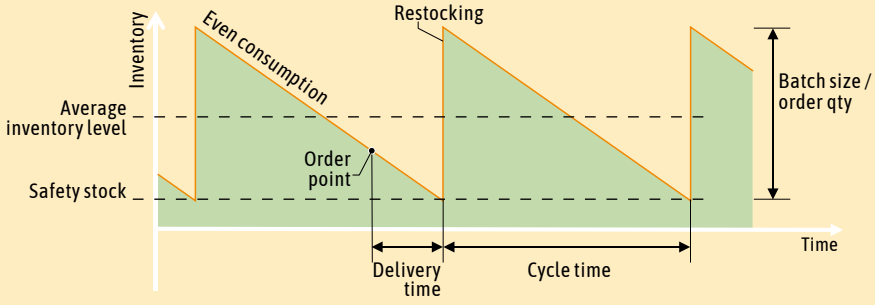
There are six important assumptions underlying the EOQ formula. It is advisable to understand them thoroughly before applying the formula.

1 There is perfectly uniform demand (the same quantity every day).
2 Setup costs are constant; they do not decrease as the number of annual setups increases through learning effects, and they are not dependent on the previous machine setting.
3 Inventory holding costs per product are constant. This means inventory costs are linear with inventory level and do not change in steps when inventory becomes very high or very low.
4 The allocated costs for inventory and setups accurately represent the real long-term economic impact on cash flows.
5 There are no other economic effects (such as quantity discounts or penalties for partially filled trucks) that influence batch-size decisions.
6 Replenishment lead time is shorter than the product's cycle time.

These six assumptions are discussed below.

1. There is perfectly uniform demand. Surprisingly, the most obvious assumption—the assumption of uniform demand—is often not mentioned at all in many textbooks currently used in education. This is highly inappropriate.
   Average inventory equals safety stock plus half the batch size, but only if the sawtooth patterns are perfect triangles. This occurs only when demand is very stable and exactly the same every day of the year.
   A small deviation may not be problematic, but what if demand follows a strong seasonal pattern, or worse, if annual demand consists of only four large orders? This is what we call lumpy demand.

Inventory level over time as a regular sawtooth pattern



Many situations involving lumpy demand do not originate from a volatile market but are created by planners themselves in production environments. Consider components used in a finished product. Suppose a particular finished product is produced twice per year in batches of 1,000 units. It obviously makes little sense to produce a unique component for that finished product three times per year in batches of 633 units. A perfect example of an alignment problem, entirely caused by the thoughtless application of a formula whose assumptions are not understood.

2. Setup costs are constant and do not change when the number of annual setups increases due to learning effects. The EOQ formula assumes a linear relationship between annual ordering costs and the number of orders, or in production environments, between annual setup costs and the number of setups.
   Thus, this assumption implies that performing a particular machine setup every month costs twelve times as much as performing it only once per year.
   In most cases, this will not be true. Due to learning effects, the task is likely to be performed increasingly efficiently, resulting in lower setup costs per setup as batch sizes decrease.
   On the other hand, suppliers often provide quantity discounts when large lots are purchased externally. According to this factor, ordering costs may be offset—or more than offset—as batch sizes increase.
   A purely linear relationship is therefore more often the exception than the rule.

3. Inventory holding costs per product are constant. This means inventory costs are linear with inventory levels and do not change in steps when levels become very high or very low. The EOQ model assumes a linear relationship between batch size and inventory costs.
   Let us take a closer look at the widely used—and even more widely misused—concept of "inventory costs."
   Inventory costs are often calculated as a cost factor multiplied by inventory value. Yet both factors are debatable. The inventory cost factor consists of multiple components and may reach as much as 25%. For example:

Risk of obsolescence and 'dead stock'	10 %
Capital costs WACC	5 %
Handling costs	4 %
Ordering costs	3 %
Storage costs	2 %
Insurance costs	2 %
Total inventory costs	26 %

Consider all these components and ask yourself whether they double when the batch size is doubled. And whether they are halved when the batch size is halved. For many of these components, such behavior simply does not apply.

Take the first component, the Risk of Obsolescence and Non-Moving Inventory. There is absolutely no linear relationship here. The higher the inventory levels, the greater the risk of obsolescence and non-moving stock. For products that are produced weekly, this risk is probably negligible, but for products that are produced only once a year, there is a real possibility that demand will suddenly decline after five months.

Handling costs depend on warehouse movements, and ordering costs depend on the number of orders. Moving a full pallet takes just as much time as moving half a pallet. Handling costs and ordering costs are not linearly correlated with either the inventory level or the length of time inventory remains in storage.

The behavior of storage costs can be highly complex. If you own a warehouse and have sufficient unused space, costs do not change with inventory levels. But if space is scarce, renting external storage or constructing a new warehouse can be very expensive.

In short, the assumption of a linear relationship between inventory levels and inventory costs is, at best, a significant simplification of reality and, at worst, simply incorrect.

4. The Allocated Costs Represent the Real Long-Term Economic Impact on Cash Flows. The question here is whether the cost figures genuinely reflect the impact on cash flows or whether they are merely allocations of sunk costs and other costs that will not change as a result of choosing a particular batch size. Economic reality consists of cash in the bank and the inflow and outflow of cash. Managerial accounting information concerning costs and values are merely constructed derivative measures of that reality.

   Suppose we have inventory valued at $8 million and we apply an inventory carrying cost factor of 25%. Suppose further that we decide to reduce all batch sizes substantially, so that within one year our inventory falls by half to only $4 million. Assume also that we can achieve this without increasing setup costs or ordering costs. Quantity discounts do not exist, and all other factors such as revenue and operating costs remain exactly the same. We therefore achieve a reduction in inventory costs of $1 million (= 25% × $4 million). Will we then see a profit increase of $1 million? Of course not! You must look at the real cash flows, not vague derivative concepts. Which cash flows will actually change, and by how much?

   Setup costs sometimes consist largely of machine hours. These hours are usually sunk costs, meaning they will not change depending on whether capacity is used for setup activities. One hour of setup on a machine with excess capacity costs nothing. One hour of setup on a bottleneck machine costs the value contribution that the company could otherwise have generated during that hour.

   The only change in the bank account is a delay in payments to suppliers for raw materials, because raw-material purchases are temporarily reduced until inventory levels reach a

new, lower equilibrium. This will result in lower financing costs, lower risk, and lower insurance costs associated with the purchase value of the materials—nothing more. Unless, of course, we can also operate with fewer employees, fewer buildings, and so forth, but that was not assumed in this example.

5. There Are No Other Economic Consequences. In some situations, other economic factors play an important role, such as quantity discounts, fixed order quantities (cartons, full truckloads), or maximum inventory constraints in facilities such as refrigerated storage. The EOQ formula does not take such effects into account. When using the formula, ensure that these aspects are not relevant in the situation under consideration.

6. The Replenishment Lead Time Is Shorter Than the Cycle Time. Camp's EOQ model assumes that the replenishment lead time is much shorter than the product's cycle time. But what happens when lead times are very long? For example, when a product is ordered from overseas?
   In such situations, it may occur that several orders are still in transit when the next order is placed. And when inventory is sitting aboard a ship in transit, how should concepts such as inventory carrying costs be interpreted?

4. Working with a Setup Budget

It is often more sensible to work with a setup budget for all products that must be produced on a particular machine. This ensures that all available capacity is utilized optimally.

If a machine has excess capacity, reducing the batch size of a single product may not be a problem. However, reducing the batch sizes of all products simultaneously may create a capacity issue. Calculating the optimal batch size for a single product therefore cannot lead to the overall optimum. You must consider the entire product portfolio and allocate the available setup capacity in the optimal manner, taking into account setup losses, labor costs, the value of components and raw materials, possible costs of flexible additional capacity, and so forth. The calculations required for this are somewhat more complex than the EOQ formula.

4.1 Formula for Determining Batch Sizes from a Setup Budget

The Veltman–Van Donselaar formula[1] can be used to allocate the available setup budget. It assumes a fixed capacity per period and a fixed portfolio of products to be produced.

The production quantities of the products are fixed. A portion of the available capacity can then be designated as the setup budget, possibly after reserving part of the capacity as "slack." Slack is capacity that is deliberately left unscheduled in order to absorb uncertainty and control lead times.

The question is how the time available within the setup budget can be distributed across products in such a way that the net cash-flow effect is optimized.

This approach differs from the EOQ formula. In particular, it assumes that performing setups does not have a direct impact on cash flow. This is the case when there are no material losses associated with setups and no use is made of flexible capacity. Under these conditions, the impact of batch size on cash flow consists solely of postponing raw-material purchases—or more precisely, the financing cost associated with the purchase value of those materials.

Additional assumptions are:

- There is perfectly uniform demand (the same quantity every day).
- The time required per setup is constant; it does not change as the number of annual setups increases through learning effects, and it is not dependent on the previous product configuration.
- There are no material losses associated with setups.
- There are no other economic effects (such as quantity discounts or penalties for partially filled trucks) that influence batch-size decisions.
- There are no inventory storage-space constraints.
- The replenishment lead time is shorter than the product cycle time.

Under these assumptions, the batch-sizing problem can be formulated as allocating the available setup budget in such a way that the raw-material value tied up in cycle stock is minimized. The variables that matter for each product are the setup time, the demand per period, and the purchase price of the raw materials.

Note that there are two variables that do not influence the outcome: the interest rate and the processing time per product. In Camp's approach, these values play an important role through

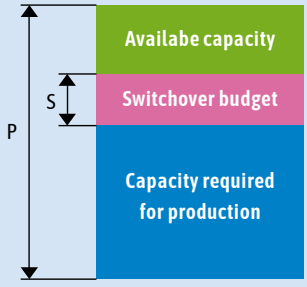
1 Hans Veltman (IPL-TNO), and Karel van Donselaar (TUE). Seriegroottebepaling in productiesituaties waar capaciteit een dominante factor is. *Bedrijfskunde, tijdschrift voor modern management*. 1993, issue 3.

the calculation of inventory value. That approach applies an interest rate to the product's cost price (including processing time).

According to Veltman and Van Donselaar, however, these values are not relevant to the decision-making process. Assuming that available capacity is fixed and that the associated expenditures are fixed as well, processing time per product, product cost value, and interest rate therefore play no role.

The figure illustrates the calculation based on the setup-budget approach.

Batch size based on changeover budget



Solution

$$Q_i = R / S * \sqrt{S_i * D_i / V_i}$$

Where $R = \sum \sqrt{D_i * V_i * S_i}$

Assignment
Based on a changeover budget (in 'hours per period'), choose batch sizes such that inventory holding expenses are minimized.

For all products i=1...n, choose the optimal batch size Q_i, such that:
- $\sum Q_i * V_i$ = minimum
- $\sum S_i * D_i / Q_i < S$

(where $\sum$ means: sum over all products i, 1...n)

Variables used – Data

D_i	pcs	Annual demand for product i during one period
V_i	$	Expenses to keep one product i in stock for one period
S_i	hours	Capacity to convert the production assets to product i
S	hours	Changeover budget, the available capacity to perform changeovers (for all products) during one period
n	pcs	The number of products in the portfolio
P	hours	The total available capacity of the production asset during one period for production, changeover, and free capacity

Calculation result

Q_i	pcs	The optimal batch size for product i
R		An auxiliary variable in the calculation, system constant

4.2 Numerical Example Using a Setup Budget

The table below contains a numerical example for a simple production environment with ten products. Management maintains a slack capacity of 10% (200 hours per year) to absorb uncertainty and control lead times. After deducting the required production capacity, a setup budget of 200 hours per year remains available.

The challenge is now to distribute this available setup time across the ten products in such a way that the net cash-flow effect is optimized. In the situation described, the only variable that can be influenced is the timing of raw-material purchases, and therefore the financing cost associated with deferred payments. In practical terms, this means minimizing the purchase value of the raw materials tied up in cycle stock.

Numerical example of batch size based on changeover budget I

General data		
Interest	5%	per year
Available capacity per year	2,000	hours per year
Required production capacity	1,600	hours per year
Slack in capacity	2,000 x 10% = 200	hours per year
Conversion budget C	200	hours per year

Data per product				Calculations			Expanded				
Product	Demand D (units per year)	Changeover time per batch S_i	Raw material purchase per unit ($)	Interest cashflow 1pc. 1jr. on inventory V_i ($)	QSRT (V * D * c)	Qopt = R/C * SQRT(Dc/V)	Number of batches per year	Average inventory (units)	Purchase value ($)	Outgoing cash flow ($)	Changeover time per year (hours)
1	1,000	1	$ 100	$ 5.00	70.7	92.3	10.8	46.2	$ 4,616	$ 231	10.8
2	1,000	2	$ 100	$ 5.00	100.0	130.5	7.7	65.3	$ 6,527	$ 326	15.3
3	1,000	1	$ 200	$ 10.00	100.0	65.3	15.3	32.6	$ 6,527	$ 326	15.3
4	1,000	1	$ 250	$ 12.50	111.8	58.4	17.1	29.2	$ 7,298	$ 365	17.1
5	1,000	1	$ 300	$ 15.00	122.5	53.3	18.8	26.6	$ 7,994	$ 400	18.8
6	1,000	1	$ 350	$ 17.50	132.3	49.3	20.3	24.7	$ 8,635	$ 432	20.3
7	1,000	1	$ 500	$ 25.00	158.1	41.3	24.2	20.6	$ 10,321	$ 516	24.2
8	1,000	2	$ 550	$ 27.50	234.5	55.7	18.0	27.8	$ 15,308	$ 765	35.9
9	1,000	2	$ 200	$ 10.00	141.4	92.3	10.8	46.2	$ 9,231	$ 462	21.7
10	1,000	2	$ 180	$ 9.00	134.2	97.3	10.3	48.7	$ 8,758	$ 438	20.6
					R=1,305.5		**153.3**	**367.9**	**$ 85,216**	**$ 4,261**	**200.0**

The available setup time is allocated across the ten products based on annual demand, setup time per batch, and inventory value. The higher the demand and setup time, the larger the batch size. The higher the raw-material price, the smaller the batch size. The allocation is based on the square root of this relationship. A system constant, R, is required to convert the relationship into actual batch sizes. In the example given, the system constant R equals 1305.5. (Note: R is the name we assigned to the term that was separated from the formula on the previous page.)

We see that a total of 153 batches are produced per year. Product 7 receives the highest number of batches (24) because of its short setup time and high raw-material cost.

The raw-material value of the average inventory amounts to $85,216.

It is now interesting to investigate how this cash flow is affected by different management decisions. Consider, for example, an investment in the machine that improves all production and setup times by 5%. We assume that all capacity released by this improvement is made available to the setup budget.

The relationship between setup budget and inventory value is linear. Do not be misled by the square root in the formula. The system constant R contains the same variables under the square root, causing the overall relationships to remain linear, at least when all setup times change by the same factor.

We could recalculate the spreadsheet using the new values, but we can also arrive at the result quickly by using the linear relationships. It is relatively easy to determine that reducing the required processing time from 1,600 hours to 1,520 hours allows the setup budget to increase from 200 to 280 hours per year. This results in batches that are 71% of their original size. Because setup times themselves are also multiplied by 95%, the resulting batch sizes become 71% × 95% = 68% of their original size. In other words, a 5% efficiency improvement results in batch sizes that are no less than 32% smaller.

The purchase value tied up in cycle stock will decrease by the same factor, releasing an amount of:

(1 – 68%) × $85,216 = $27,391

once all products have been produced at least once. This occurs quickly because every product is produced at least seven times per year. The investment is therefore beneficial.

If you would like to verify the calculations in detail, you can download the spreadsheet from www.alignmentpuzzel.nl. The starting values have already been entered, and you can experiment with the impact of different values on the results.

Numerical example of batch size based on changeover budget II

How does the batch size behave when the changeover budget and changeover times change?

Example

An investment of $10,000 reduces all processing and changeover times by 5%.

- The changeover budget decreases from 200 to 280 hours per year. After all: 2,000 hours - (0.95 * 1,600 production) - (10% * 2,000 downtime) = 280 hours per year.
- 5% more changeovers can be performed per hour.

Dus

- The average batch size decreases by a factor of 200/280 * 0.95 = 68%.
- The purchase price of raw materials for the batch size inventory becomes $85,216 * 68% = $57,825.
- The result of $27,391 is much greater than the investment of $10,000.

So the investment is favorable in this case and is more than recouped as soon as each product has been made once!

Note that we are dealing here with real cash flows, provided that raw-material purchases can be aligned precisely with production requirements. By reducing batch sizes, fewer raw materials are required, which reduces purchasing and creates an immediate cash-flow benefit. No vague inventory carrying percentages or speculative assumptions are used.

4.3 Further Analysis

It is interesting to investigate what happens to the flow of goods, inventory levels, and service levels when the ratio between spare capacity and setup budget changes. Spare capacity is needed to absorb irregularities and uncertainty and to keep lead times under control. When variability is low, more capacity can be allocated to the setup budget. When variability is high, the setup budget will need to be reduced, but this leads to larger batch sizes, which in turn contribute to increased variability.

This is a fascinating question, particularly when considered in conjunction with the law of conservation of waiting time and the selection of safety-stock levels. However, exploring it in depth would go beyond the scope of this white paper.

5. Calculating Safety Stock

The purpose of safety stock is to guarantee product availability and reliable delivery. It plays an important role in supply chain alignment.

If there is no uncertainty, there is no need for safety stock. Safety stock exists to ensure delivery despite uncertainty in demand and lead times. In practice, generic rules such as "10% of expected demand" are often used. This is incorrect because it does not take into account the probability distributions of demand and lead time that apply in specific situations.

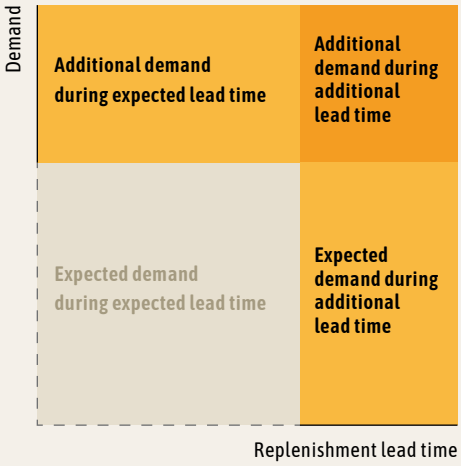
5.1 Safety Stock as Protection Against Uncertainty

Material availability is always managed based on an expectation of demand. For the expected normal flow of goods, the expected consumption during the expected replenishment lead time forms the basis. In SIC- or Kanban-controlled inventory replenishment systems, this is embedded in the reorder-point formulas.

However, uncertainty is always present, both in demand levels and in replenishment lead times. Each has a probability distribution with a certain degree of spread. Safety stock provides availability in cases where demand exceeds the 50th percentile (the expected value) of the demand distribution, and in cases where the replenishment lead time is longer than its expected value. It is often rather arbitrarily stipulated that the target should be 95% delivery reliability, interpreted as "on time, in full." The gap between the 50th percentile and the 95th percentile must be covered by safety stock.

5.2 Components of Safety Stock

The calculation of safety stock consists of three components: additional demand during the expected lead time, expected demand during additional lead time, and additional demand during the additional lead time.



5.3 Calculation of Safety Stock

The probability distributions of demand and lead time for a specific situation must be established through the analysis of observed data in practice. Broadly speaking, all components of safety stock can be added together. However, there is a substantial body of scientific literature covering the details of the theoretically exact calculation, particularly regarding the "additional demand during additional lead time" component.

5.4 Seasonal Effects on Safety Stock

In practice, only a single probability distribution for demand and lead time is usually quantified, even though the reality during peak season is completely different from that during the off-season. Demand is dramatically different, the intervals between replenishments (under an FTL transport policy) are dramatically different, and the probability distributions of the uncertainties are dramatically different.

5.5 Example: An Ice Cream Manufacturer

During the summer, average demand is high because everyone eats a lot of ice cream. Demand cannot rise much above that average because there is a practical limit to how much people will

consume. Safety stock as a fraction of average demand can therefore be kept relatively low during the summer. That is fortunate, because retailers' freezers are already filled with the daily flow of products required to meet demand.

During the late season and winter, on the other hand, average demand is low. However, everyone wants an ice cream when the weather suddenly becomes exceptionally pleasant for a weekend. This takes demand peaks to an entirely different level. Peak demand may then be as much as ten times normal winter demand. Safety stock as a fraction of average demand must therefore be extremely high. Fortunately, this is feasible because retailers' freezers are largely empty during this period.

In addition, we must consider that replenishment frequency (and therefore replenishment lead time) is drastically different in summer and winter. During the summer, there is almost always a truck already on its way, whereas in winter replenishment must be arranged specifically when needed.

In practice, none of these factors are taken into account, and this creates an amplifying bullwhip effect throughout the supply chain. At the start of the peak season, inventories are built up not only for expected consumption but also for safety stock (that standard 10%).

The supply chain becomes flooded.

As the transition to the off-season begins, all inventories are drawn down again, causing the supply chain to come almost to a complete standstill.

6. The Relationship Between Efficiency, Lead Time, and Work-in-Process Inventory

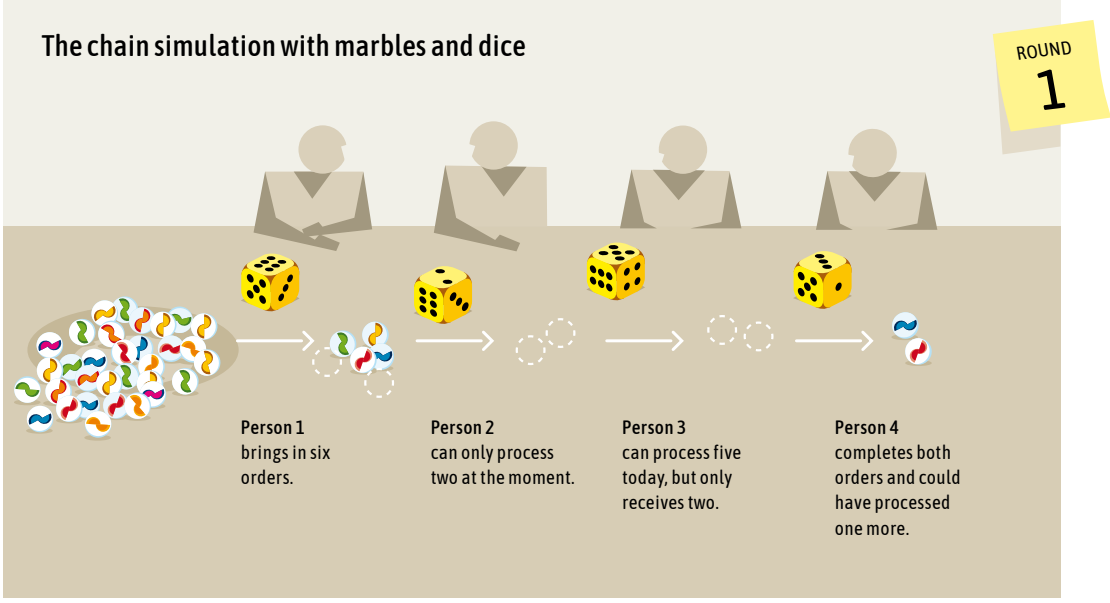
6.1 A Simple Experiment You Can Perform Yourself

To observe the relationship between the amount of work-in-process inventory (WIP), order lead times, and the efficiency of a production system, you can play a simple game with a few colleagues.

Take a long table, a large collection of marbles, and four dice (assuming four participants). With these, you can simulate a process chain.

The person sitting closest to the pile of marbles starts. He or she rolls the die and may take as many marbles from the initial container and place them in front of themselves as the number shown on the die. The marbles represent orders flowing through a production process, while the roll of the die represents one day of production. In the case of the first player, it represents daily order intake. One day six orders arrive, another day only one. On average, the number is 3.5, assuming fair dice are used.

The marbles lying in front of the first person can be viewed as the queue of work waiting for the second station. The second person then rolls the die. That person may remove from the first

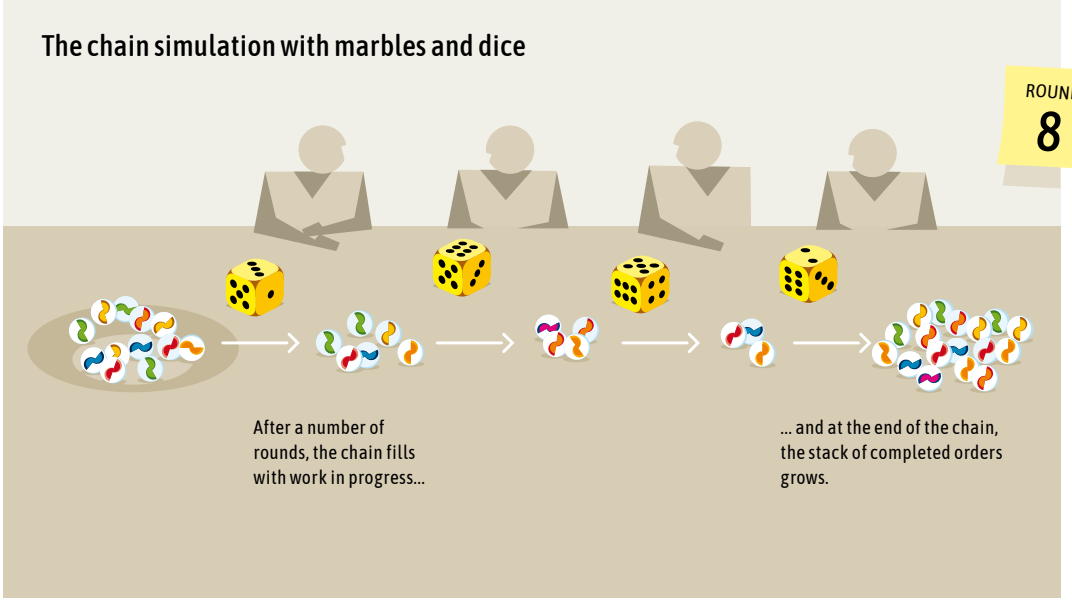


person's queue as many marbles as the number rolled. If more spots are rolled than there are marbles available, the excess spots are lost. This can be interpreted as idle capacity due to a lack of orders. The third person then rolls the die and may remove marbles from the second person, and so on. In this way, the marbles move across the table from player to player. A process chain with work-in-process buffers emerges. The stock of marbles that gradually accumulates in front of the last person can be viewed as the total output of the process.

You will notice that the process runs somewhat sluggishly at first because there are too few marbles in the chain, but over time the chain fills up.

We can now restrict the inflow of work by allowing the first person to take one marble fewer than the number rolled. This means fewer orders enter the system than the theoretical capacity would allow. The first person now takes an average of 2.5 marbles per turn from the container, while the others continue at an average of 3.5 marbles per turn. We are therefore feeding work into the chain at 71% of capacity (because 2.5 / 3.5 = 0.71).

The interesting observation is that, in the long run, the WIP adjusts itself such that average idle time at each workstation never exceeds approximately 29%. (29 = 100 – 71). At first glance this seems obvious, since work is being supplied at only 71% of capacity. But upon further reflection, it is actually quite fascinating. Within the dynamics of the process, the WIP apparently adjusts itself automatically so that enough die rolls are lost at each of the three workstations to result in exactly 29% lost capacity over time.

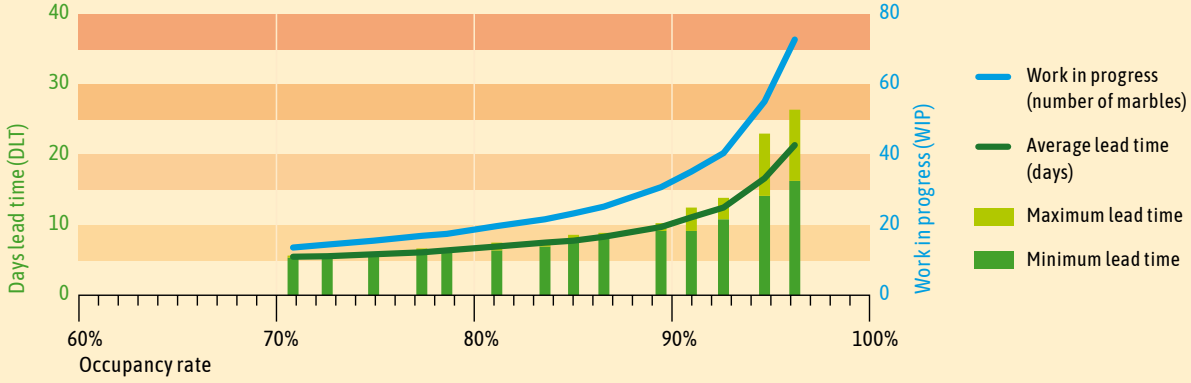


6.2 The Relationship Between Efficiency and Work-in-Process Inventory

After the initial warm-up period, it turns out that an average of 13 to 15 marbles accumulate in the three buffers at the workstations when input is set at 71%. The utilization rate should converge to approximately that level in the long run. You can verify this yourself.

By restricting the first station more or less aggressively, we can establish a relationship between utilization (or efficiency or loading) and WIP. For example, if we devise a new rule for the first player that results in a throughput rate of 91% at the first station, the amount of work-in-process inventory rises to approximately 35 marbles across the three stages. In this way, we can identify a relationship between WIP and utilization. The graph below shows an example of the results such a simulation game can produce. Note that the graph has two vertical axes: lead time is shown on the left (green) axis and work-in-process inventory on the right (blue) axis. The measured utilization rate is shown on the horizontal axis.

Relationship between WIP, utilization rate, and lead time



You can easily conclude that work-in-process inventory has a positive effect on efficiency but a negative effect on lead time and lead-time variability. The more orders that are waiting in WIP queues, the longer it takes for a new order to pass through the system.

Perhaps even more interesting is to reason in the opposite direction and conclude that a certain level of WIP is required in order for a production system to operate at a desired level of efficiency. We will return to this point later.

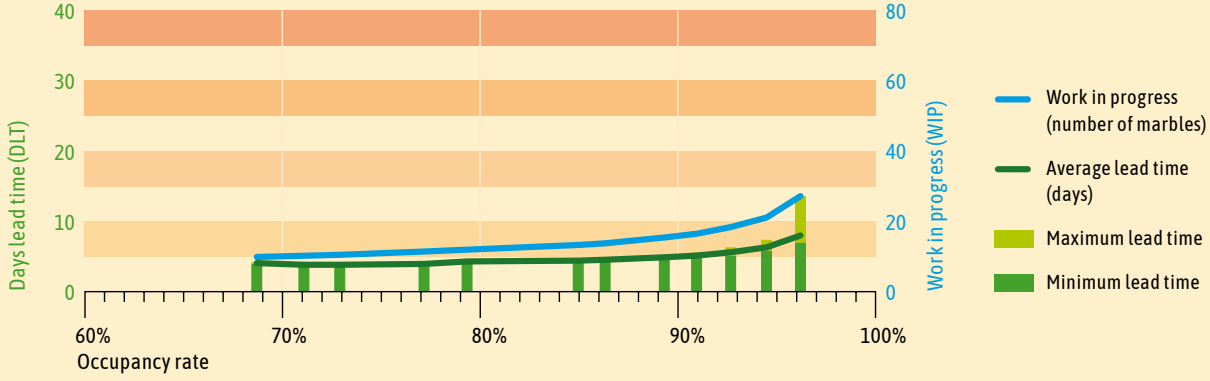
6.3 The Effect of Variability on the Curve

If we want to understand the effect of workload variability (that is, one day with a great deal of work and the next day with almost none), we can modify the rules of the game described above as follows:

- When rolling 1, 2, or 3, move three marbles.
- When rolling 4, 5, or 6, move four marbles.

Instead of drawing a value between one and six, we now obtain only three or four. This reduction in variability immediately creates a much more stable pattern. The effect on the amount of WIP required to achieve a particular utilization rate is remarkable. It turns out that the same efficiency can be achieved with much lower levels of WIP and therefore with much shorter lead times.

Relationship between WIP, utilization rate, and lead time with reduced variety



6.4 Using WIP as a Control Parameter

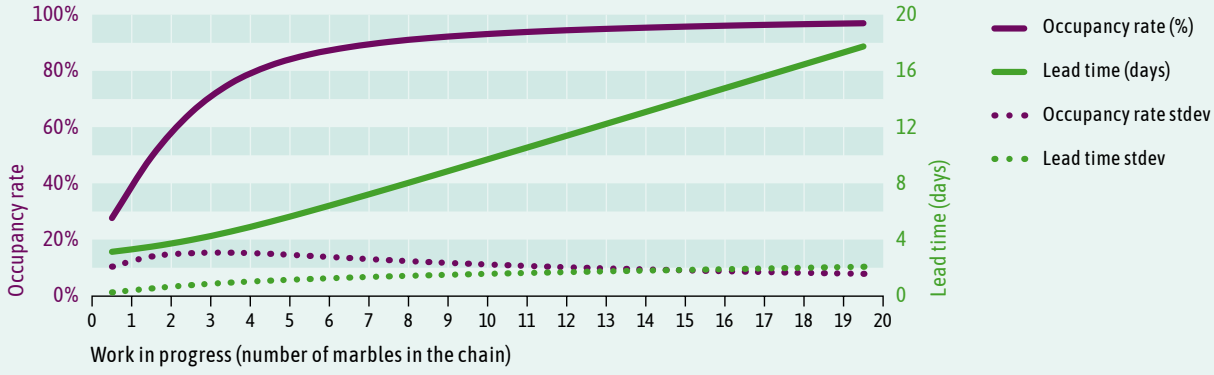
We can make another modification to our marble game by applying workload control[2]. In this version, the total number of marbles in the chain is kept constant throughout each experiment. We add exactly as many marbles to the beginning of the chain as leave it at the end. Each experiment starts with a predetermined number of marbles in the chain (naturally, a different number for each experiment). The last person rolls first, and the first person no longer has their own die but instead uses the value rolled by the last person.

2 Also see *The Alignment Puzzle*, paragraph 9.6.3

This allows us to observe the relationship between WIP and utilization while treating WIP as a controllable parameter.

We can now see even more clearly the effect of WIP levels on efficiency and lead time. This is shown in the graph above. The thick purple line represents production efficiency or utilization. It is strongly curved, approaching a 100% utilization asymptote. At low levels of WIP, efficiency falls well below 60%. Only when WIP reaches a minimum of eight or nine marbles does utilization exceed 90%. Achieving 100% efficiency requires an extremely large amount of WIP.

Capacity utilization and lead time as a function of work in progress



The thick green line represents lead time. Beyond a certain WIP level, lead time increases linearly with WIP. This is logical. The lead time of a particular marble through the process is determined primarily by the marbles waiting ahead of it in the queues. The more marbles there are in front of it, the longer it must wait before being processed.

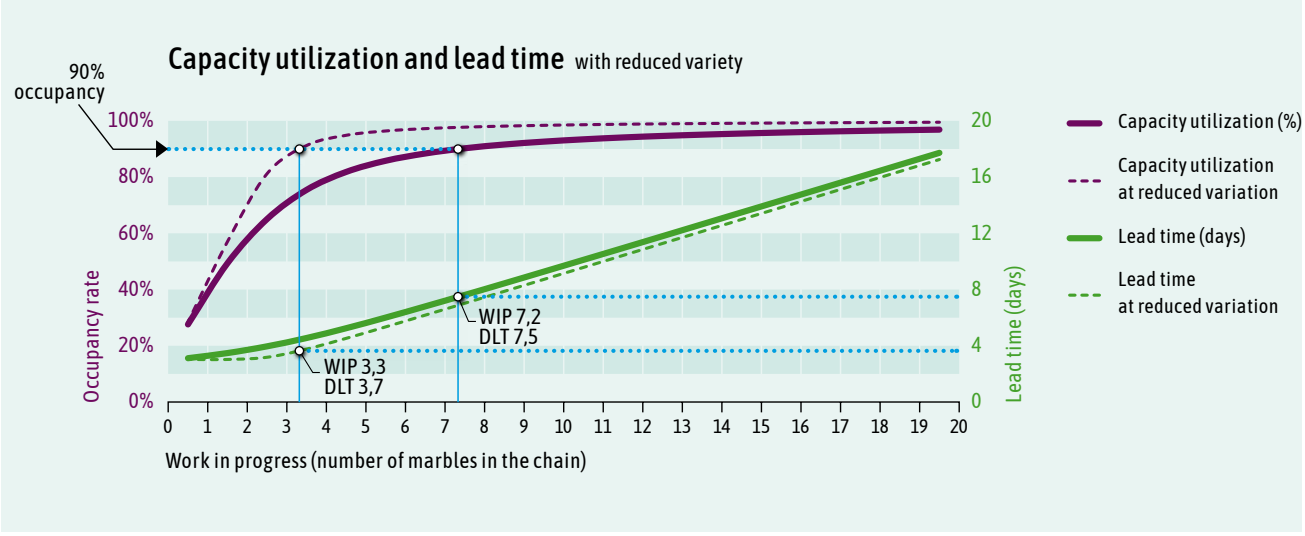
When efficiency is (virtually) constant, WIP is the only determining factor. At low WIP levels below six marbles, lead time is no longer proportional to WIP. This occurs because efficiency is low, reducing throughput and slowing the flow of work.

The two thin lines represent the standard deviation of lead time and efficiency. These remain relatively low in this simple linear arrangement where marbles (work orders) are not allowed to overtake one another.

Even in this workload-controlled situation, we can once again reduce variability by switching from values of 1–6 to values of only 3 or 4. We observe the same effect on efficiency and lead

time. The graph below shows that, in a dynamic situation with a WIP level of eight marbles, a utilization rate of 90% and an average lead time of 7.5 days were achieved.

After reducing variability, WIP can be reduced to four marbles while maintaining the same efficiency, resulting in an average lead time of only 3.7 days.



7. Extension of the Experiment: Complex Routings and Queueing Rules

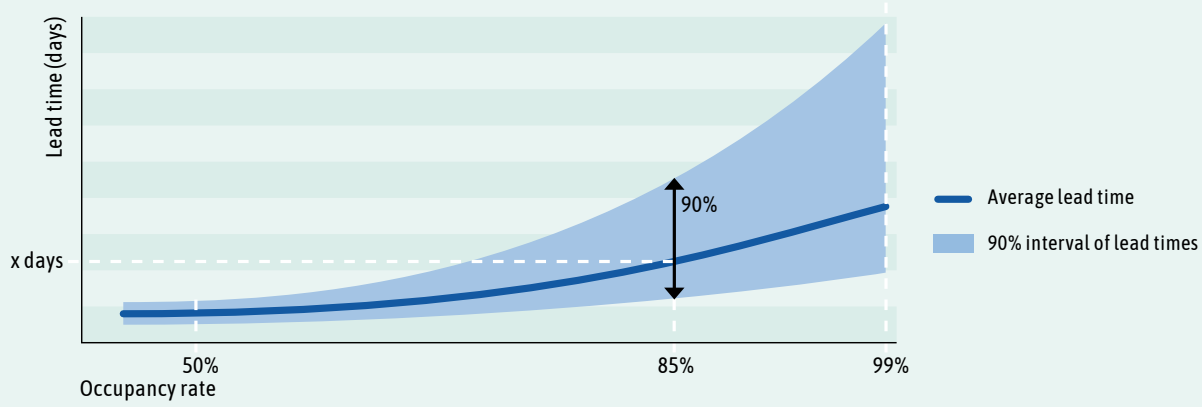
From the experiment described in the previous section, it can be concluded that there is a strong relationship between utilization and WIP. As WIP increases, utilization and efficiency increase along a curved relationship. The curve ultimately approaches 100% utilization asymptotically.

These principles apply not only to a simple four-stage production line but also to much more complex situations involving many workstations and a wide variety of crisscross routing patterns.

It is particularly interesting to examine the reliability of predicted lead times. For this purpose, we use lead-time variability as a measure. It turns out to depend heavily on the queueing rules applied. In a linear production flow where all orders follow the same routing and are assigned due dates in the same manner, the choice of queueing rule makes little difference. Orders simply move through the system one after another in sequence.

Relationship between lead time and utilization rate

in a situation where orders follow various routes past different workstations



However, in a complex environment with many intersecting routing paths, orders can overtake one another. In such situations, the decision rule applied at each workstation becomes highly important. Lead-time variability can increase dramatically if poor decisions are made regarding queue priorities. The degree of lead-time variation is critical when process predictability is important.

There are many different queueing rules. Several are listed in the table. The simplest is the First-In, First-Out (FIFO) rule. The order that has been waiting the longest is processed first. In environments with complex routing structures, this rule results in a high degree of lead-time variation.

A much better rule is the Critical Ratio rule. This ratio compares the remaining time until the promised delivery date with the remaining processing time. Orders that are close to their due date receive priority. The focus is not on planned dates for individual process steps but on the promised delivery date of the order as a whole. This is the best option for systems with reasonably balanced capacities, substantial demand volatility, and more than occasional rush orders. Under these conditions, lead-time variation is found to be at its lowest.

This rule helps ensure that orders are completed on time because delayed orders automatically receive higher priority. One of its major advantages is that it eliminates the need for special scheduling procedures for rush orders. Such orders simply receive delivery dates that are not far in the future and then naturally move through the system very quickly because the Critical Ratio rule assigns them high priority and continuously places them at the front of queues—until they once again fit within the normal workflow.

However, these rush orders will inevitably push many other orders further back in the queue. It therefore remains essential to monitor the total workload assigned to each week when promising delivery dates. It is entirely possible that allowing one rush order to move ahead could cause one hundred other orders to be delivered late.

Priority rules in queues

Rule		Effect
FIFO *First In, First Out*	First come, first served. The most obvious and most common rule. This rule stands for equivalence in queueing systems where customers are physically present and where it is psychologically difficult to deviate from queues.	Does not result in an improvement for system performance. Is not smart in production situations where orders have no feeling.
SPT *Shortest Processing Time*	The shortest processing time takes precedence. As a result, small orders move through the system faster. The SPT rule must be handled with care. Large orders can experience significant delays.	Is popular because the average waiting time per order becomes low. However, large orders can run into trouble.
WINQ *Work in Next Queue*	Select the order that has the shortest queue in the next operation. The idea is that it is pointless to send orders to a place where they end up in a queue anyway, but the reasoning is incorrect. A bottleneck must not stall and therefore must have a long queue. Non-bottlenecks have little queue time but inevitably have to stop every now and then.	Does not yield any efficiency advantage.
CR *Critical Ratio*	Select the order with the lowest ratio of remaining lead time to remaining processing time. Orders approaching their due date are given priority. Priority is based on the due date of the entire order, not on the due dates of individual process steps.	The best option for a system with reasonably balanced capacities, high demand volatility, and more than occasional rush orders.

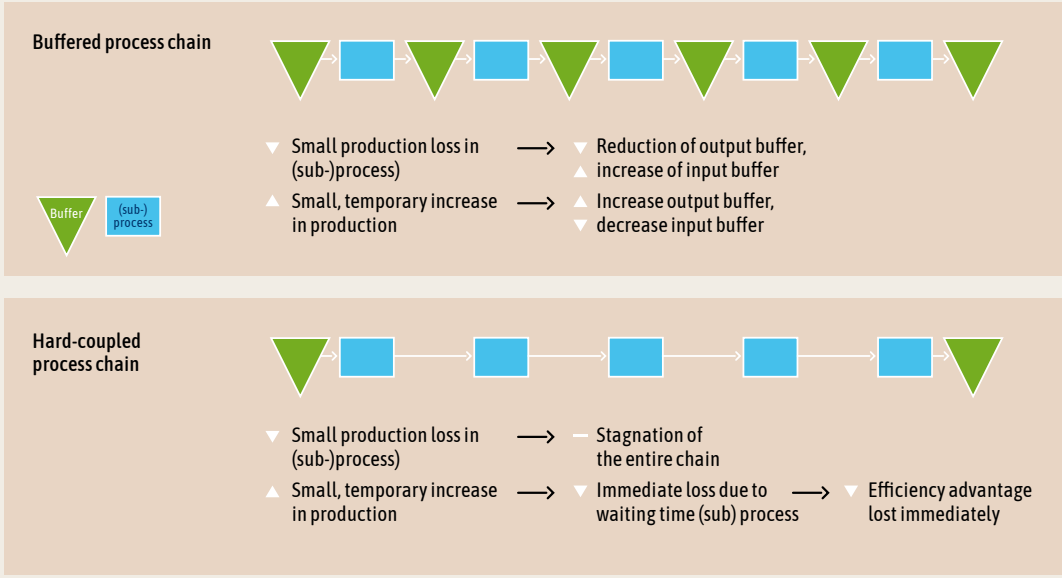
8. The Essential Role of Buffers and Work-in-Process Inventory (WIP)

8.1 Tight Coupling Leads to Loss of Gains and Accumulation of Losses

The title above deserves to be read twice. What we mean by it is the following. A chain of processes or subprocesses without intermediate buffers is referred to as a tightly coupled chain. In such a chain, a brief interruption at one stage immediately causes all other stages (both upstream and downstream) to stop as well. The resulting loss can no longer be recovered within normal working hours. Conversely, if something positive happens at one stage—for example, a task is completed faster than normal—that gain is lost because the stage must then wait for the next stage to become available before the product can move forward (since no buffers exist). In short: losses are permanent, while gains disappear.

When buffers are present, however, a brief stoppage or reduction in output at one stage does not immediately propagate throughout the entire chain. Likewise, a small gain results in an increase in the buffer after the subprocess, and that gain can later be consumed when a minor setback occurs elsewhere. Put differently: tight coupling between stages in a process chain causes favorable deviations to be lost and unfavorable deviations to accumulate.

Hard coupling leads to loss of profit and accumulation of loss



Consider also the law of conservation of waiting time[3]. If two processes or subprocesses are not perfectly aligned, something must always wait: material, capacity, or the customer. If no buffers exist, material can never wait, meaning that either capacity or the customer must do so instead. There are very few situations in which this is genuinely desirable.

8.2 Defying "Lean"

The popular Japanese Lean philosophy seeks to optimize production processes by eliminating all forms of "Muda" (waste). There are traditionally seven forms of Muda, four of which relate directly to inventory. In some cases, followers of this philosophy embark on what amounts to a crusade against anything resembling inventory within a production system.

This inventory primarily consists of work-in-process inventory waiting for the next processing step in the manufacturing process. Through so-called value stream mapping, the total time an order spends moving through the production process is analyzed, and waiting times caused by WIP are made visible.

The Lean approach regards the time that material spends waiting as pure waste. Consequently, many Lean improvement initiatives focus on reducing lead time, which effectively means reducing work-in-process inventory.

However, work-in-process inventory waiting in front of a production resource serves an important positive function. It decouples the resource from preceding and subsequent operations, thereby enabling the resource to operate efficiently.

At every stage of a process, some variation in processing time may occur. Tight coupling between operations causes a workstation to be unable to begin work until the preceding workstation has completed its task. The workstation therefore waits due to a lack of work and becomes less productive. Tight coupling with the following operation causes a workstation to wait as well, because it cannot release its completed work until the next workstation has finished processing the previous order.

3 See *The Alignment Puzzle*, paragraph 5.3

Tight coupling creates a domino effect for every delay. Even worse, delays cannot be recovered because tight coupling prevents any workstation from operating faster than the surrounding workstations allow. The bad luck of longer processing times accumulates throughout the chain, while the good luck of shorter processing times cannot be exploited. The result is a loss of efficiency.

Our advice is therefore: eliminating genuine forms of waste is excellent practice, but do not take it to extremes, and beware of tight coupling.

9. Optimizing the Process Chain

9.1 Reducing Utilization Is Expensive and Has Only a Limited Effect

It should be clear that at higher utilization levels—that is, when higher efficiency is required—there is a greater likelihood that the incoming workload will temporarily exceed the available capacity. This causes work-in-process inventory to grow. Traffic jams occur when roads are busy. Likewise, it should be clear that eliminating backlogs takes longer when average utilization is already high. WIP declines only slowly. Congestion takes a long time to clear when conditions remain busy.

Utilization can be reduced by increasing capacity and/or by allowing less work to enter the system. Increasing capacity with properly educated, trained, and experienced personnel (bearing in mind the required quality standards) and high-grade production resources generally requires permanently hiring additional employees and purchasing additional equipment. These additions reduce efficiency and entail significant costs. There are, of course, also flexible options for expanding capacity, such as overtime, temporary labor, and outsourcing.

If you want to achieve high efficiency, you cannot do so with insufficient WIP. In that case, operations often become chaotic. A great deal of time is lost because people are waiting for the previous workstation, only to discover at the end of the day that extensive overtime is suddenly required after all.

9.2 The Harmful Influence of Elephant Orders

One phenomenon that amplifies this effect is the so-called elephant order: an exceptionally large order that creates congestion within the process. Often, such elephant orders are created by the company itself. For example, several small orders may be combined into one large order in an attempt to save setup time. Such a decision is intended to improve efficiency but often produces the opposite effect because it introduces large lumps of work into the process. Only at a genuine bottleneck does such a decision make sense.

Elephant orders may also be caused by bonus or discount structures. A good example of a bonus/discount scheme that promotes misalignment is a confectionery manufacturer whose large foreign customer placed a single enormous order each year in order to qualify for a quantity discount. The customer's buyer was evaluated based on achieving that discount. The annual order repeatedly created major production problems. Further investigation revealed that the customer itself also suffered from the enormous volume delivered and experienced bottlenecks in its warehouse operations. Thus, both customer and supplier were adversely affected by an elephant order that they had artificially created through the discount system.

9.3 Reducing Variability in the Work Stream Delivers Significant Benefits

An effective way to reduce the amount of WIP required is to reduce variability in the capacity demand generated by the workflow. This can be managed externally by quoting lead times that depend on the current workload of the production system. Everyone understands that things take longer when it is busy, and some customers are even willing to accept this. Sales can also actively pursue orders that fit within the reality of the available capacity profile.

However, there are also much less intrusive internal measures, some of which are counterintuitive. A workflow operates most smoothly when all orders require approximately the same amount of work. A supermarket checkout line illustrates this perfectly: a single customer with an overflowing shopping cart causes many customers with only a few items to wait unreasonably long.

This means that the natural tendency to combine supposedly similar orders everywhere in order to save setup time should be resisted.

9.4 Increasing Responsiveness: Flexible Deployment of People and Resources

The counterpart to reducing variability in incoming work is increasing responsiveness to variability. After all, variability only becomes a problem when you are unable to cope with it effectively.

Responsiveness is best improved through the flexible deployment of available personnel and production resources. We are, after all, dealing with work that requires education, training, and experience in order to guarantee the required quality. Versatility is the key concept.

Human-resource development policies should focus on enabling employees to support multiple workstations, allowing them to assist wherever capacity is needed, whether in their own area or at a workstation experiencing temporary overload. To achieve efficient use of equipment, it is advisable to select production resources that are sufficiently universal to be used for multiple types of work.

Naturally, there are costs associated with cross-training permanent and loyal employees. There may also be additional setup activities, and from a commercial perspective, lead times that fluctuate with workload may be perceived negatively by customers. Careful trade-off analysis is therefore required when making decisions.

Finally, one additional observation regarding the application of these improvements for maximum benefit. Up to this point, we have discussed how improvements can be used to reduce WIP and shorten lead times. However, it is equally possible to maintain existing WIP levels and lead times while increasing the efficiency and loading capability of the production system to as much as 98% of available capacity, thereby enabling the organization to cope more effectively with periods of heavy demand.

More refreshing insights and techniques:
Read *The Alignment Puzzle*, the new standard work on alignment in organizations.

www.alignmentpuzzle.com

Authors: Hans Veltman, Jacques Adriaansen, Peter Morren, and Rob Kwikkers.
Design & production: Okapi Ontwerpers.
Images: Illustrations and infographics by Fons Moers / Okapi.