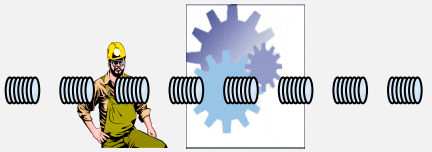
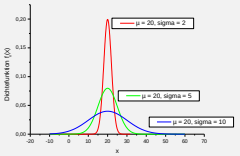
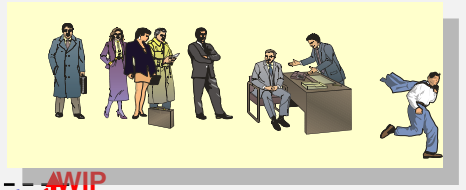
Chapter 7 : Factory Dynamics

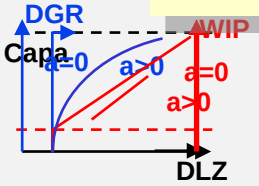
This chapter deals with the derivation of the basic parameters and equations to describe "productivity"

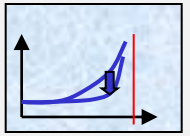
short cycle times
stable cycle times
high throughput
high utilization
low inventory











Customer Satisfaction Forces Logistical Parameters

Optimization of a production system by logistical parameters (Productivity Controlling)

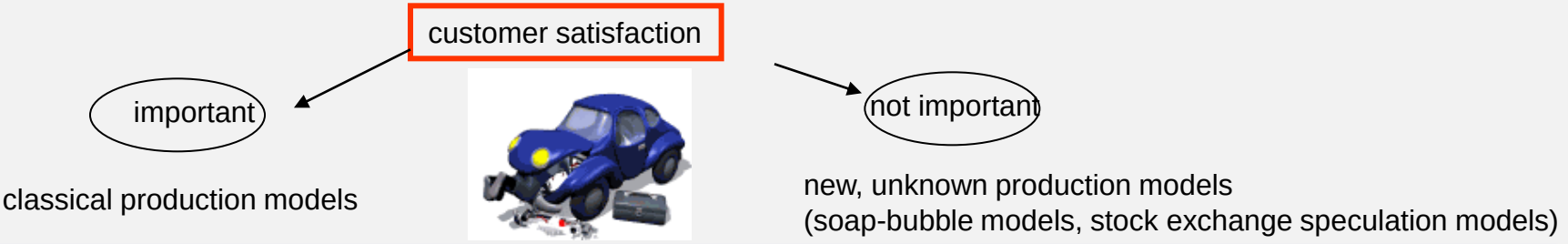
Productivity Controlling is the method to acquire, to valuate and to control the actual performance of a production system (Factory Dynamics)

Which parameters have to be optimized in a production system ?
Let us study a few very simple cases:

Case 1: In an anyhow working factory completely new products are fabricated (e.g. the first mobile tape recorders, called "walkman").
Always a lot of rich customers exist, who will pay a high price to be one of the first users with this product.
The factory may demand any price -> also an "unorganized" factory will make profit.

Case 2: If all rich customers will have a "walkman", a new product has to be developed
or the selling price must be reduced to open a new "poorer" (but larger) clientele.
Low prices and high output force a reduction and optimization of fabrication costs to remain in the profit zone.

If the factory in case 2 does not provide the consumer completely itself, but is dependent on component suppliers (e.g. motor or housing for the walkman), the producer of the end device becomes the customer of the supplier.
Usually the number of suppliers (of simple components) is larger than the number of producers of final products.
Even if not, the best guarantee for the supplier to ensure a long-term, stable success is the **customer satisfaction**.



The **satisfaction of customers** and **optimization of profit** force an evaluation of the production system

Customer Satisfaction

Example for an (extreme!) speculation model:

A few speculators (without owning any money!) setup a company to collect money, which is named "MONEYSAVE".
(zu deutsch: Wirtschafts- und Finanzberatungsgesellschaft, engl. **Investment House**).

By clever arguments money spending people can be found, who are hoping for a high interest rate of their invested capital.
Exceeds the capital a case-dependent minimum amount, the company "MONEYSAVE" now transforms itself to a **Venture-Capital House**, which finances the start-up of a new, **trendy enterprise**, for example "GENTRONIX".

With the money a representive GENTRONIX enterprise building will be rented (or leased), employees will be hired, a promotion campaign will be started and a great future will be forecasted. The enterprise will go public (e.g. in the segment New Market NEMAX), the Venture-Capital House will remain a part of the stocks (e.g. 40%) for itself.

But the enterprise GENTRONIX is doing not much: looking good, showing incredible development chances and quickly setting a few products (in the case of GENTRONIX devices called "gen sorters with voice-controlled electronics") on the market.
The stocks are rising from the starting value of 20 € to 87 € (which is about 300% profit).
The Venture-Capital House observes the new enterprise GENTRONIX, the market and the trend.

But now, completely surprising, the enterprise GENTRONIX fails to deliver devices, which were sold at exhibitions, to the promised date (official statement: in some devices software errors were detected). Previous customers are annoyed, new customers are not willing to pay.
MONEYSAVE recognizes the upcoming downturn (stocks falling to 54 €), calculates the actual selling profit of their stocks (also possible capital connections with GENTRONIX). A remarkable profit remains, MONEYSAVE sells its stocks.

GENTRONIX is not able to serve its financial obligations, can not find new money spenders, the stocks are dropping and GENTRONIX files for insolvency.
MONEYSAVE does not care much about this, MONEYSAVE pays the promised return of invest to its own money spenders and keeps the rest of the profit.

This model does not take care of long-term success and the classical customer satisfaction

Customer Satisfaction and Logistical Parameters

Factors of success for customer satisfaction in classical production models:

with relevance for the customer:		Logistical parameter of production:
short delivery time (today ordered, tomorrow delivered)	→	short cycle times (from order to delivery)
high delivery reliability (meet the promsed delivery date)	→	stable cycle times; low scrap rates
Attractive selling prices	→	competitive production costs → high throughput; high utilization (cost of ownership and employees)
Large flexibility	→	short cycle times; low inventory; high flexibility in capacity

The simultaneous optimization of all logistical parameters usually is not easy

In this chapter we will:

- quantifize the logistical parameters and their interaction
- develop models to describe the productivity
- discuss procedures to increase productivity

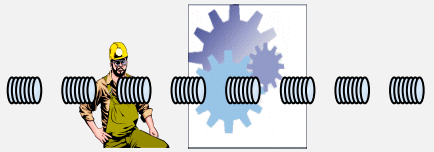
7 Factory Dynamics

This chapter deals with the derivation of the basic parameters and equations to describe "productivity"

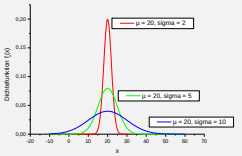
Performance parameters of a production line

short cycle times
high throughput
stable cycle times
high utilization
low inventory

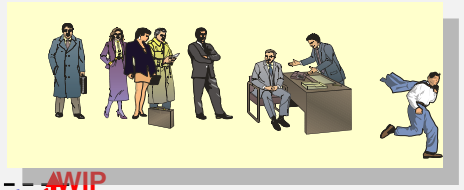
4-Partner Model



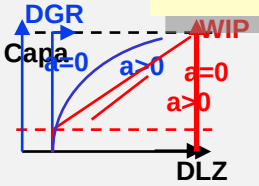
The variability



Queuing theory

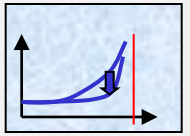


The operation curve



Calculation of the operation curve in a semiconductor fab

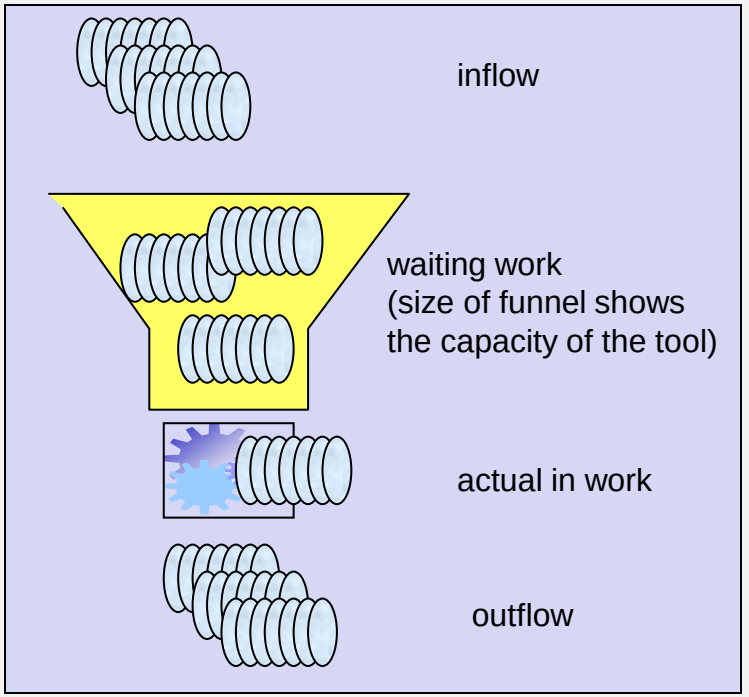
Optimization potentials from operation curves



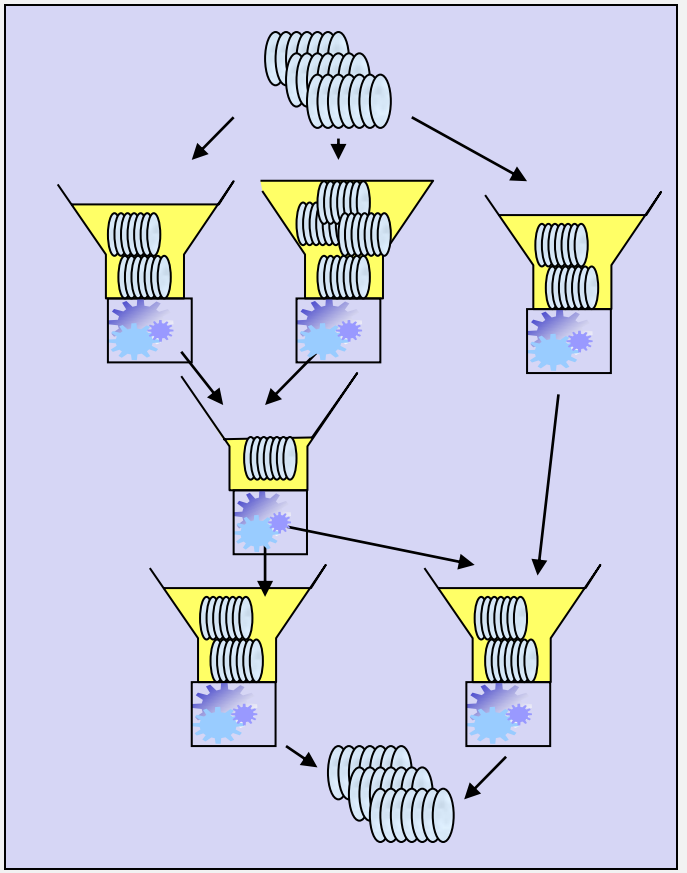
Schematic Model of a Process Line

The Funnel Model

Basic unit = production tool



Formation of a production line



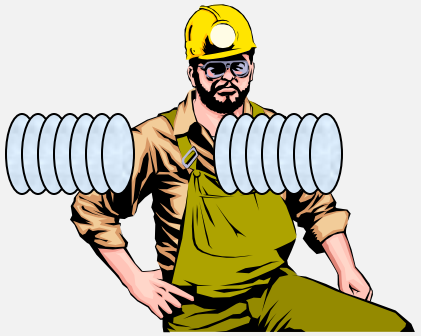
But:

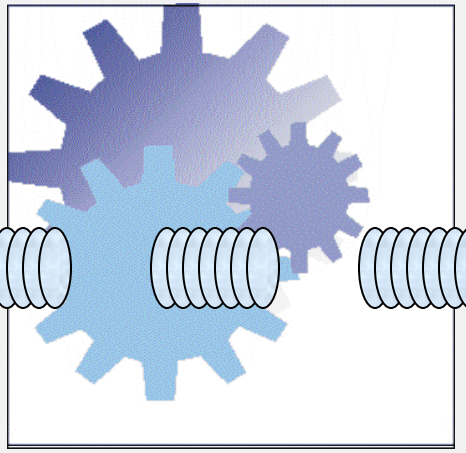
Funnel models consider only 2 components of the system:

tool and Material in the system = **work in progress**, **wip** -> wafer in progress, wip

Closer to practice is the addition of 2 more components -> this leads to the **4 Partner Model**

Four-Partner Model of a Process Line





Material (WIP)

Man (Operator)

Machine (tool)

Method (Process)

The 4 Partners
or
The 4-Ms

Production Tool PT

Production Unit PU

Production System PS

Production is only possible, if all 4 partners are simultaneously at the same place available

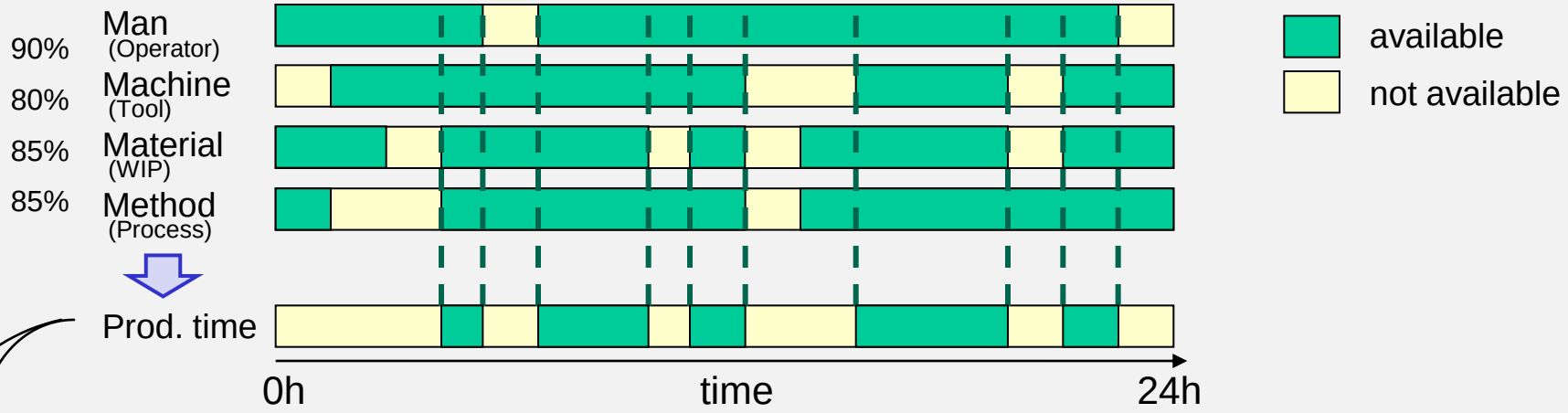
Productive Time of a Process Line

The production system PS consists of:

1 machine
1 process
1 operator
+ WIP

The availability A of the single partners should be very high:

Example:



If the 4 Partners are **statistically independent**:
Productive time per day = 24h * 0,90 * 0,80 * 0,85 * 0,85 = 12,5h

Despite the high availability of each partner the overall availability is significant lower

If the down-times of the 4 Partners are **synchronized**:
Productive time per day = 24h * Min{0,90; 0,80; 0,85; 0,85} = 19,2h

By synchronization the overall availability is significantly increased

Synchronization of the single partners increases productivity

Quantitative Calculation of Productive Time

But: "Productive time" is not a countable parameter, because this parameter only indicates, that production would be possible.
A quantifized logistical parameter can be derivated by introducing a new parameter, called **throughput**.
Frequently, the throughput can be normalized by a time periode (for example 1 day), then this time-related throughput is called **going rate**

$$Goingrate = time\ \ period \cdot process\ speed\left[\frac{wafer}{time}\right] \cdot partner\ \ availability$$

Daily-Going-Rate
(Tagesproduktion)

$$DGR[units] = 24h \cdot process\ speed\left[\frac{wafer}{time}\right] \cdot partner\ \ availability$$

Hourly-Going-Rate
(Stundenproduktion)

$$HGR[units] = 1h \cdot process\ speed\left[\frac{wafer}{time}\right] \cdot partner\ \ availability$$

The goingrate must be calculated different for synchronized and not-synchronized production systems:

Non-synchronized production systems:

$$Goingrate = time\ \ period \cdot process\ speed\left[\frac{wafer}{time}\right] \cdot A_{op} \cdot A_{wip} \cdot A_{proc} \cdot A_{tool}$$

The goingrate can be increased, if the availability of any partner can be increased.

Synchronized production systems:

$$Goingrate = time\ \ period \cdot process\ speed\left[\frac{wafer}{time}\right] \cdot Min\left\{A_{Op}; A_{WIP}; A_{Proc}; A_{tool}\right\}$$

The goingrate can be increased, if the availability of the worst partner can be increased.

Optimization of Productive Time

The optimization potential of a production system depends on the already realized degree of synchronization:

As lower the synchronization degree, as higher the potential to increase productivity

Optimization strategy:

1 Analysis of the synchronization degree

2 Increase of availablity of the single partners

3 Increase of synchronization

4 Increase of availability of the worst partner

new loop

Example:

Car fabrication on a conveyor belt:

The fabrication of cars on a conveyor belt is due to the many, simple process steps well suited for synchronization:

because: the method is simple (process can be divided in small portions), e.g. fixing of doors, drilling screws, ...
the tools are simple (e.g. hands of workers, screwdriver, ...)
the material is simple(e.g. screws, cables, doors, ...)
the worker due to the simple requirements has a high availability (street workers may be sick variuos time)

The multitude of many, simple work allows working units, that each working station can work in common mode

Production of **semiconductor chips**:

Semiconductor lines are (up today) due to the high complexity of the processes not well synchronized:

because: the method (technology) is difficult, e.g. adjustments on a nanometer scale
the tools are quite complex and process times are quite different
the material is sophisticated (e.g. extreme cleanlyness, various process requirements)
the operator has to move to the material and the tools

The multitude of various, complex processes is hardly adjustable to a common mode for various products

Optimization of Productive Time

Effects of the 4-Partner-Model

The 4 Partners influence the dynamical (= time-varying) performance of the whole production line by:

- the absolut value of their **availability**
- the **variation of timely fitting** of their availability (degree of **synchronization**)
- the single partner **availability variation**

We have to discuss:

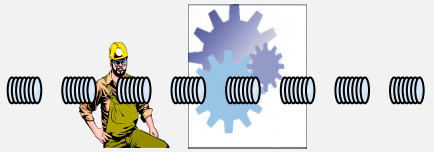
? Why influences the variation of availability the performance ? → see cueuing theory

This chapter deals with the derivation of the basic parameters and equations to describe "productivity"

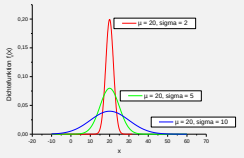
Performance parameters of a production line

- short cycle times
- high throughput
- stable cycle times
- high utilization
- low inventory

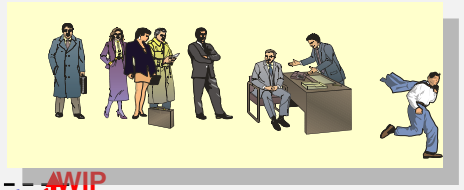
4-Partner Model



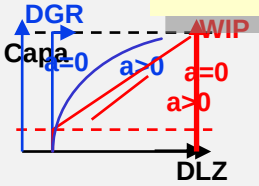
The variability



Queuing theory

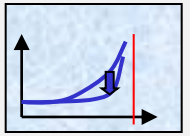


The operation curve



Calculation of the operation curve in a semiconductor fab

Optimization potentials from operation curves



The Variability

Variability

With the parameter **variability** a measurement of the non-uniformity of events is defined

Example:
if the measurement of 10 fabricated screws delivers 10 times the same result (within the accuracy of measurement, say 1 µm) of the length, than the sample shows no variability.
If the precision of measurement is increased (say to 1 nm) than variations of length may be detected. Now the sample shows a variability.

In production systems many factors exist, which show variability, e.g. dimensions of screws, time to tool-breakdown, ...

Variability is closely connected to the term "randomness", but it is not the same.

Controlled variability:

If a printed circuit board (PCB) depending on the product plan is equipped one time with 100 parts (consumer products), next time with 300 parts (high end products), than the number of parts on the board (and also other parameters like assembly time) show a variability. But this variability is adjustable and not random.

Random variability:

If the adjustment of a parameter defies our human control, than we talk about a random variability.
In this sense the breakdown times of a tool can not predicted and they are merely adjustable. If the breakdown times are measured and displayed, a "random distribution" of the measured times will be the result.

* in the following the random variability is used and shortly abbreviated just as variability

Remember from Statistics

If the probability distribution of a random variable x is known, so-called "expectation values" for this random variable can be predicted. These expectation values are called **moments** of the random variable.

mean:

(1. common moment)

variance

(2. central moment)

discrete	continous
$\mu = \sum_{i=1}^{n} x_i \cdot f(x_i)$	$\mu = \int_{-\infty}^{\infty} x \cdot f(x)dx$
$\sigma^2 = \sum_{i=1}^{n} (x_i - \mu)^2 \cdot f(x_i)$	$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x)dx$
with rearrangement better suited for samples:	with rearrangement better suited for samples:
$\sigma^2 = \sum_{i=1}^{n} x_i^2 \cdot f(x_i) - \mu^2$	$\sigma^2 = \int_{-\infty}^{\infty} x^2 \cdot f(x)dx - \mu^2$

standard deviation

$$\sigma = \sqrt{\sigma^2}$$

coefficient of variation

$$c_v = \frac{\sigma}{\mu}$$

Measurement of Variability

The **standard deviation** is an absolute measurement of scattering of single events from a fixed value (usually the mean value).
This absolute value does not allow a comparison between different distributions.

Example:
If we would place a house on a scale of 10µm, we would think, this is very exactly.
But if we would adjust an interconnect line on a semiconductor chip to the same absolute value of 10 µm, this would be unacceptable.

For comparison of the width of 2 distributions the variation coefficient is introduced:

$$c_v = \frac{\sigma}{\mu} \cdot 100\%$$

Example from Olympic Games:
A member of the Greek javelin team reaches the width´s: 80,14 m, 78.32 m, ...
A member of the Animals jumper team reaches the width´s: 7.12 m, 7.10 m, ...
A member of the US sprinter team runs the 100m in the time of: 9.91 sec, 10.04 sec...



	value1	value 2	value 3	value 4	mean value	standard deviation	C_v
javelin	80,14 m	78,32 m	84,44 m	81,87 m	81,19 m	2,60 m	3,2 %
jumper	7,12 m	7,10 m	7,55 m	7,58 m	7,34 m	0,26 m	3,6 %
runner	9,91 sec	10,04 sec	10,13 sec	9,95 sec	10,01 sec	0,1 sec	1,0 %

the result:
Although the absolute value of variation is larger from the javelin compared to the jumper and in addition the values of the jumper looks more constant, the variation coefficient shows, that the javelin is the "more constant" athletic.
The best value is achieved by the runner, whose values are now comparable (also in a different system ([sec] instead of [m]) with other athletics.

Valuation of Variability

Usually we have a good estimation for the valuation of the mean value of a probability distribution:
if a 100-m sprinter achieves times around 10 sec, another sprinter around 12 sec, we can fairly good estimate, who is faster.

For the valuation of variability our feeling is not so good:
from the last example we can not estimate quite well, if the javelin athletic or the jumper is the more constant athletic.
Even harder is the valuation, if it is more cost saving for the owner of a car (or tool) to have frequent preventative maintenance or to wait until the car breaks down unexpected.

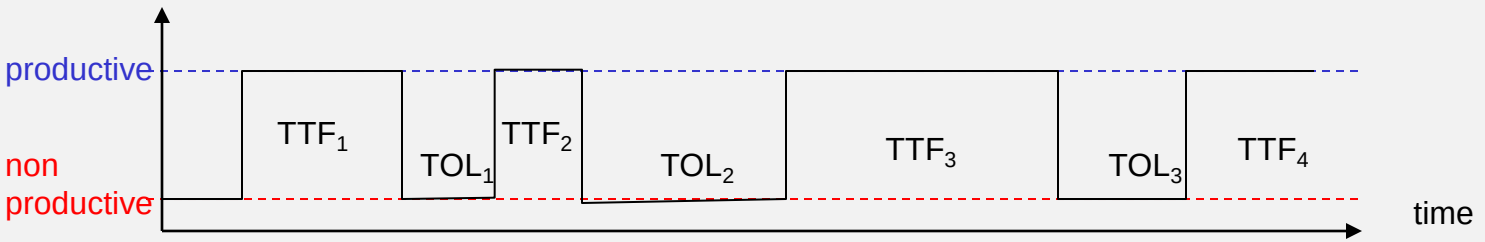
For the valuation of variability the following classes were introduced:
if the standard deviation σ of a distribution has the same value as the mean value μ (=> C_v ~1), than the variability is called moderate;
if the scattering of the values is smaller than the mean value, the variability is low, in the opposite case high.

$$c_v = \frac{\sigma}{\mu}$$

class of variability	variation coefficient c_v	Typical situation
low	c < 0.75	tool process times without breakdowns
moderate	0.75 < c < 1.33	process times with short interrupts (e.g. new setups)
high	c > 1.33	process times with long break downs

Variability in a Production System

The availability of a production system PS:



non productive time = time-off-line TOL

mean value: $MTOL = \overline{TOL} = \frac{1}{n} \cdot \sum_{i=1}^{i=n} TOL_i$

Mean time-off-line

Productive time = time-to-failure TTF

mean value: $MTTF = \overline{TTF} = \frac{1}{n} \cdot \sum_{i=1}^{i=n} TTF_i$

Mean time-to-failure

we recognize:

The length of the individual time-off-lines TOL variies

Can the TOLs be described by a statistical distribution ?

Is it possible to receive from this distribution performance parameters (e.g. throughput) and forecasts ?

The lengths of the individual productive times (= server times) TTF variies

Can the TTFs be described by a statistical distribution ?

Is it possible to receive from this distribution performance parameters (e.g. throughput) and forecasts ?

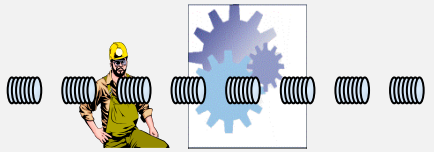
These questions will be answered in the queuing theory

This chapter deals with the derivation of the basic parameters and equations to describe "productivity"

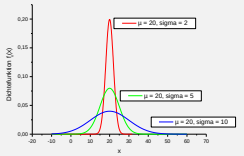
Performance parameters of a production line

short cycle times
high throughput
stable cycle times
high utilization
low inventory

4-Partner Model

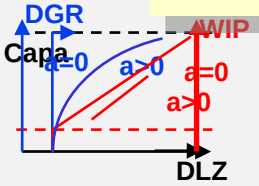


The variability

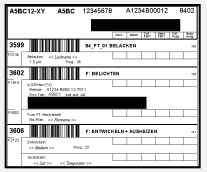


Queuing theory

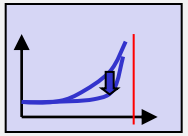
The operation curve



Calculation of the operation curve in a semiconductor fab



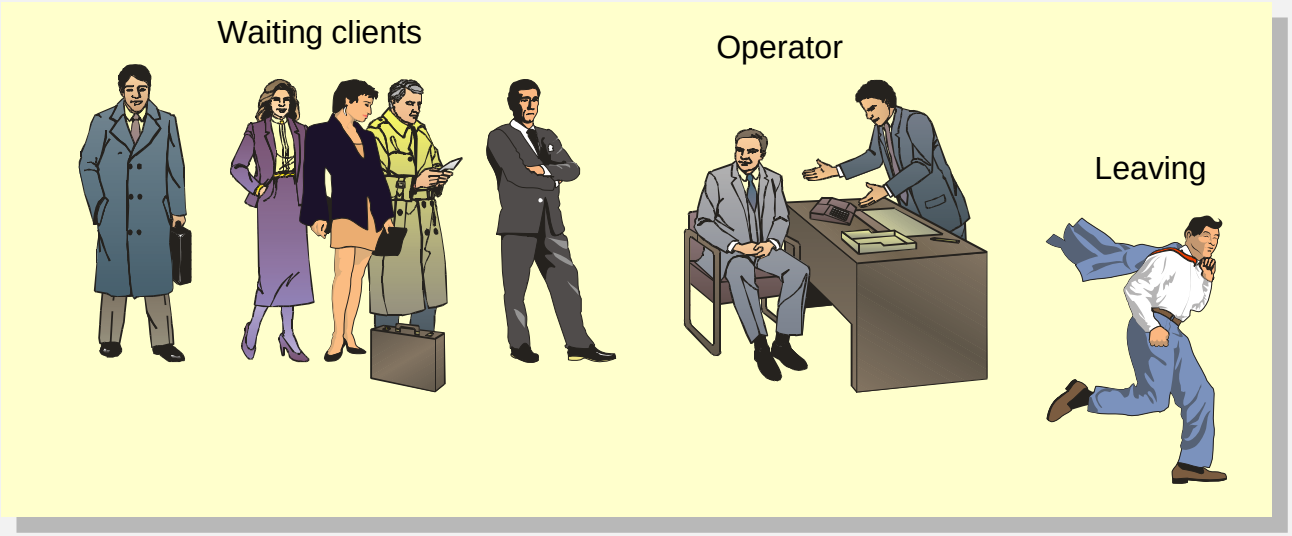
Optimization potentials from operation curves



Queuing Systems are Everywhere

Production is a dynamical throughput problem
Throughput problems are handled in the so-called "Cueueing System Theory"

Typical situation:



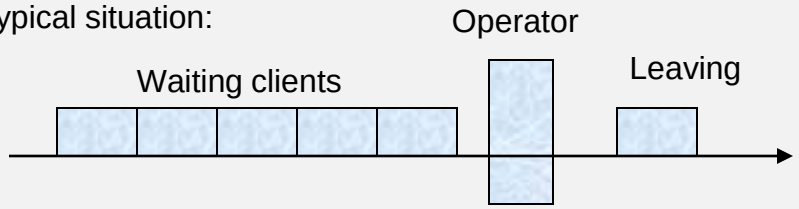
Question:

From which parameters
the cueueing is dependent ?

Parameters of Cueueing Systems

Production is a dynamical throughput problem
Throughput problems are handled in the so-called "Cueueing System Theory"

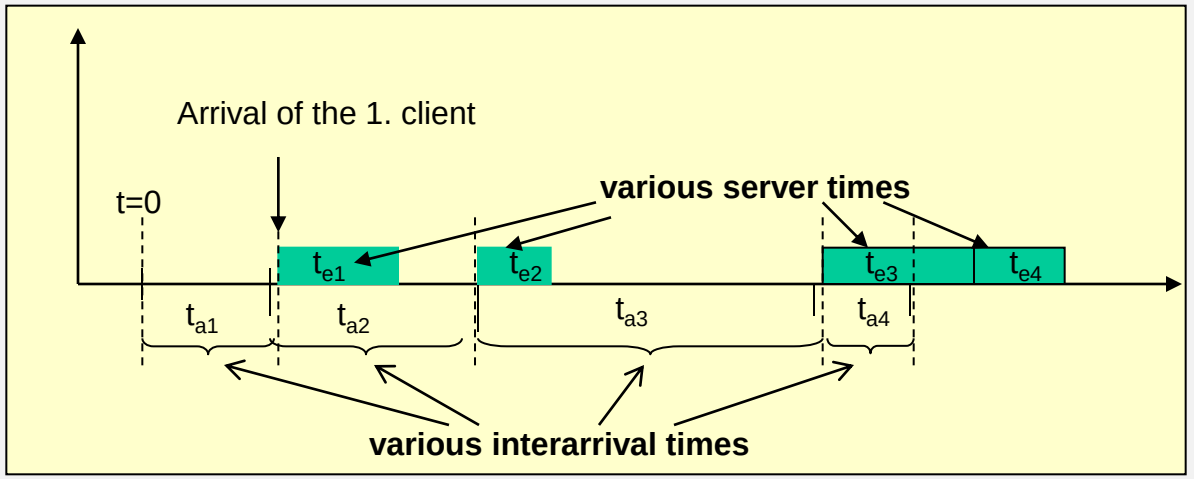
Typical situation:



Question:

From which parameters
the cueueing is dependent ?

Mathematical description of the system:



The interarrival times and the server times
are statistically distributed.
Independent of the distribution some parameters
can be defined:

mean interarrival time $\overline{t_A}$
(*a: arrival)

mean interarrival rate $\bar{r}_a = \frac{1}{\overline{t_a}}$

mean server time $\overline{t_e}$
(*e: execution)

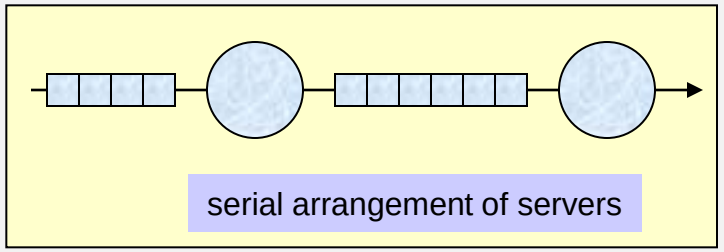
mean server rate $\bar{r}_e = \frac{1}{\overline{t_e}}$

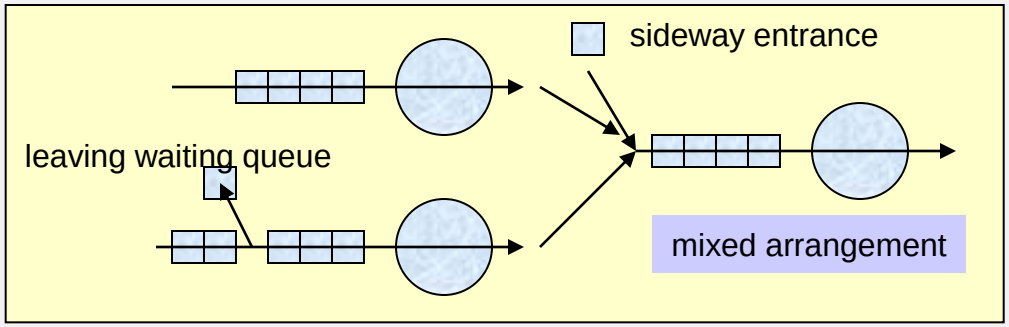
$$Utilization\ U = \frac{mean\ server\ time}{mean\ inter\ \ arrival\ time} = \frac{arrival\ rate}{server\ rate}$$

The system is only stable, if U < 1 , this means it is not allowed that more clients will come than can be served

Parameters of Cueueing Systems

Examples for queueing systems:





description of the queueing system:

1. the inter-arrival times and the server times are statistical distributed and can be described by probability distributions

2. Complexe queueing systems can be created.
In **open systems** waiting clients can entrance or leave the system

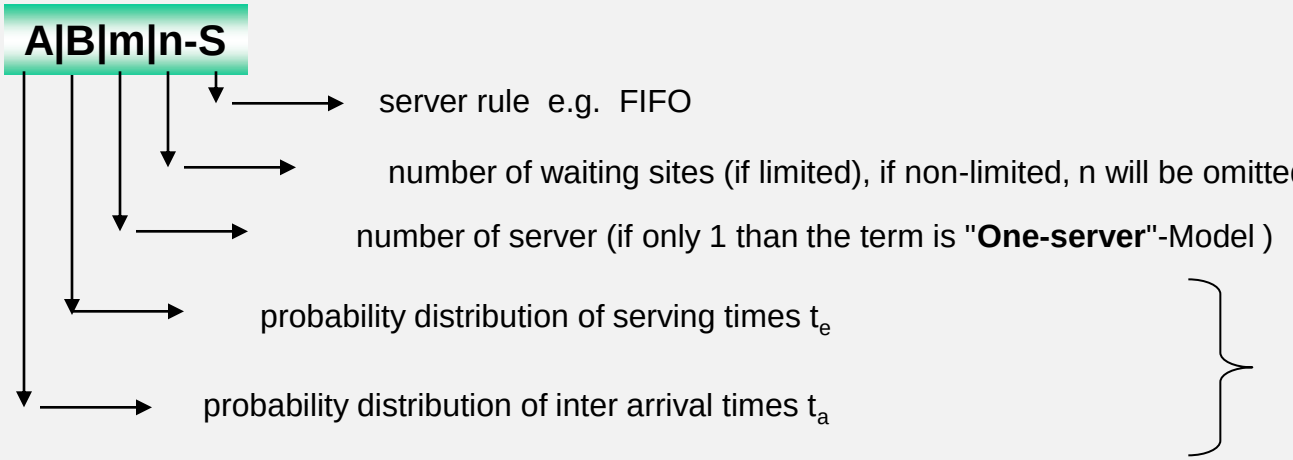
3. For the excecution **serving rules** must be introduced

Examples of serving rules:

if all clients (jobs) have the same priority:
FIFO (firstin, first out): * serving in the order of arrival -> usually low variation of waiting times
LIFO (last in, first out): * serving in opposite order of arrival -> usually large variations in waiting times
SIRO (served in random order): * serving in random order independent of arrival -> usually large variations in waiting time
RR (round Robin): * fixed serving time per client (job), then back in the waiting queue
if clients have different priority:
* allocation of priority status (statically), then separation in different waiting queues (e.g. server for First-Class and Business-Class passengers)
* priorities are changing (dynamically) -> with interrupting priority (called **pre-emptive**, the running job will be interrupted)
-> non-interrupting priority (**non pre-emptive**, the running job will be finished)

Parameters of Cueueing Systems

Kendall-Notation for waiting systems



Exponential distribution

the length of time intervals between the occurence of random events is usually exponential distributed

Poisson distribution

if the server is dedicated for a special process (e.g. mail office, fabrication process) the server times are usually normal distributed (or Poisson, if only a few).

Parameters of Cueueing Systems

Performance parameters of queueing systems:

mean inter-arrival time: $\bar{t}_A$ **mean arrival rate:** $\bar{r}_a = \frac{1}{\bar{t}_a}$

mean server (excecution) time: $\bar{t}_e$ **mean server rate:** $\bar{r}_e = \frac{1}{\bar{t}_e}$

$$Utilization\ U = \frac{mean\ server\ time}{mean\ interarrival\ time} = \frac{arrival\ rate}{server\ rate}$$

throughput TP : mean number of clients, which are served per time (or leaving the system per time)

duration time T_D: the whole time of a client in the system = waiting time + server time

number of waiting clients N_W: number of waiting clients

filling N_F : total number of clients in the system

state probability p(k) probability, that k clients are in the system

For all queueing systems is valid:

The Law of Little $N_F = TP \cdot T_D$

(basically not a law, but a definition: throughput = jobs/time)

Parameters of Cueueing Systems

For all queueing systems the following relations hold:

Utilization

$$Utilization \quad U = \frac{arrival\ rate}{server\ rate} = arrival\ rate \cdot server\ time$$

$$U = \frac{r_a}{r_e} = \frac{r_a \cdot t_e}{m}$$

m: number of parallel servers

Duration time T_D

$$Duration\ time = Waiting\ time + server\ time$$

$$T_D = T_w + t_e$$

Little´s Law

$$N_F = TP \cdot T_D$$

modified to

because the system to be stable is not allowed, that $r_a > r_e$.
So the TP is limited by r_a

unit: r_a [units/time]

but:

this performance parameter is measured after execution
-> + **real, actual value**
- timely behind production

unit: TP [units/time]

$$N_F = r_a \cdot T_D$$

for the total system (waiting + server)

$$N_W = r_a \cdot T_w$$

but also just for the waiting queue

this performance parameter is measured before execution
(-> e.g. how many wafers are introduced in the fab per day)
-> + **can be used to control the system by contolling r_a**
- is a little speculative (not taking distortions into account)

this kind of parameters and equations are called "**dynamically**" and several of this kind of equations are used to control production and productivity

Queuing Systems - M | M | 1 -FIFO

The **M | M | 1 -FIFO** queue is the easiest system for analysis:

* exponential distributed arrival times (that is usually true)
* exponential distributed server times (that is usually not true, but may be a good compromise for easy calculations)
* 1 unlimited queue and 1 server
* FIFO serving rule

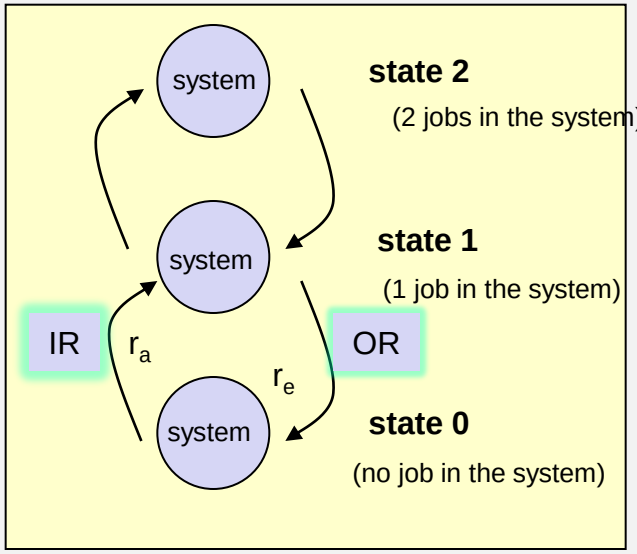
Analysis:

1) If the system consists of k clients, this fact is called **state p(k)**

2) The system can change this state p(k), if a client step in the sytem (k -> k+1) or a client leaves the system (k -> k-1)

3) the transfer rate from state k->k+1 corresponds to the arrival rate r_a, the transfer rate from k->k-1 corresponds to the server rate r_e

4) the probability to find the system in the state p(k) is noted as **probability p_k**



So the step-in rate IR is: $IR = p_{k-1} \cdot r_a$

and the out rate: $OR = p_k \cdot r_e$

With this system description the calculation of the performance parameters can be done in the easiest way:

the results are shown next page

Queuing Systems - M | M | 1 -FIFO

For Queueing systems - M | M | 1 -FIFO:

exponential interarrival times, called M (from Markov)
exponential server times, called M
1 server, called 1
server rule: first in first out, called FIFO

State probability

$$p_k = (1-U)\cdot U^k$$

$$Utilization \quad U = \frac{arrival\ rate}{server\ rate} = arrival\ rate \cdot server\ time$$

Mean waiting time T_w (in the cueue)

$$\overline{T_W}(M\,|\,M\,|1) = \frac{\frac{U}{r_e}}{1-U} = \frac{U}{1-U}\cdot \overline{t_e}$$

Mean duration time T_D (total)

$$\overline{T_D}(M\,|\,M\,|1) = \frac{\frac{1}{r_e}}{1-U} = \frac{\bar{t}_e}{1-U}$$

Mean number of clients

$$\overline{N}(M\,|\,M\,|1) = \frac{U^2}{1-U}$$

mean filling k : (utilization)

$$\bar{k}(M\,|\,M\,|1) = \frac{U}{1-U}$$

$$k = \sum_{k=0}^{\infty}(k\cdot U_k) = (1-U)\sum_{k=0}^{\infty}(k\cdot U^k) \rightarrow \cdots \frac{U}{1-U}$$

Queuing Systems - G | G | 1 -FIFO

In complicated cueueing systems the calculations are not so easy.

For the general **G | G | 1 -FIFO** System it was shown from Kingman in 1961, that the mean waiting time T_w in the cueue can be described by:

$$\bar{T}_w(G|G|1) = \left(\frac{c_a^2 + c_e^2}{2}\right) \cdot \left(\frac{U}{1-U}\right) \cdot \bar{t}_e$$

Kingman equation

or the so-called

$$\bar{T}_w(G|G|1) = V \cdot U \cdot T$$

VUT - equation

Kingman, J. F. C. (1961). The single server queue in heavy traffic. Proc. Cambridge Philos. Soc. ***57*** *902--904.*

* for the M|M|1-System this equations is exact, because in the exponential distribution $c_a=c_e=1$ and therefore V=1

The waiting time in a queue depends on 3 factors: the variability V, which is a measurement of fluctuations
the utilization U, which corresponds to the filling of the system
the mean server time t_e

If the variability V will be > 1 (the arrival times and server times are varying strongly), than the waiting time in the cueue will increase

The Kingman-equation from the queueing theory :

$$\bar{T}_w(G|G|1) = \left(\frac{c_a^2 + c_e^2}{2}\right) \cdot \left(\frac{U}{1-U}\right) \cdot \bar{t}_e$$

proves the proposal from the 4-Partner-Model, that also the variation of the availabilities (= server times) the performance of a production system (e.g. throughput) influences.

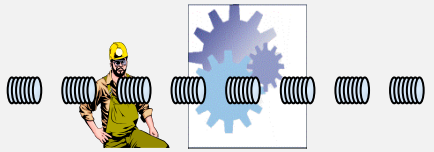
Factory Dynamics

This chapter deals with the derivation of the basic parameters and equations to describe "productivity"

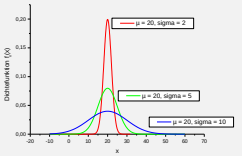
Performance parameters of a production line

- short cycle times
- high throughput
- stable cycle times
- high utilization
- low inventory

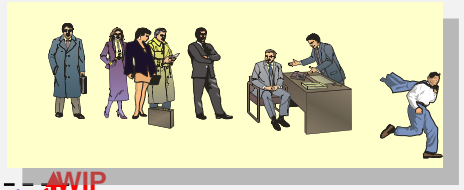
4-Partner Model



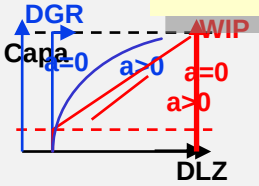
The variability



Queuing theory

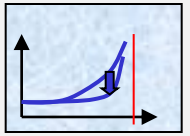


The operation curve



Calculation of the operation curve in a semiconductor fab

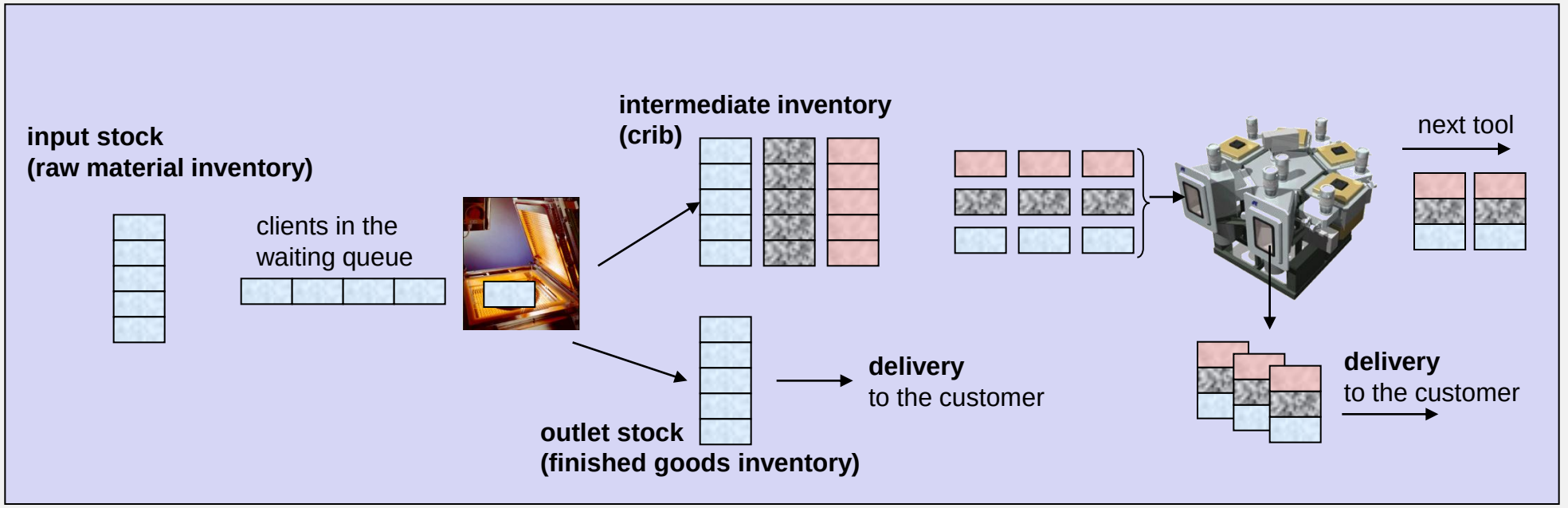
Optimization potentials from operation curves



Parameters of Productivity Control

The following definitions are used:

Work station	this is a tool (maschine, equipment)
Work station group (center, module)	is a group of several tools that perform the same fabrication process (e.g. oxidation, polishing,...)
Routing	describes the path of a product from starting to the end through the complete production line
Job	this is the execution rule (methode, process)



Work in progress, WIP (wafer in production)

frequently different defined.
Usually WIP is used for the number of products between (and including) the input stock and outlet stock and are not waiting (for various reasons) in intermediate stocks.

Parameters of Productivity Control

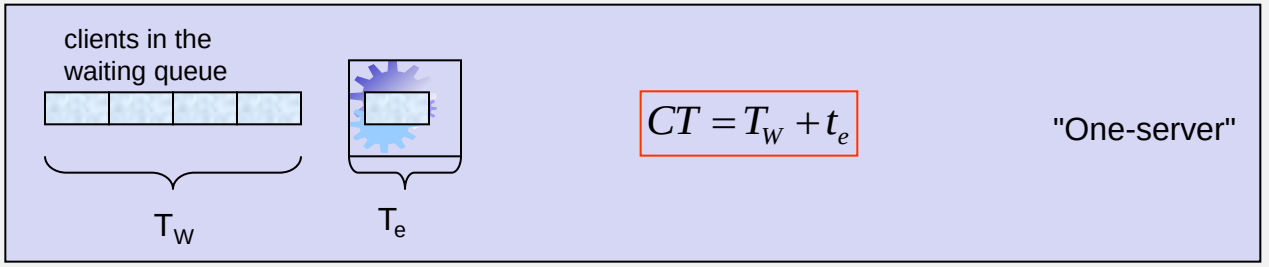
Cycle time CT

for a "one-server" from definition the cycle time CT is the time used for a client from entrance in the waiting queue until reaching the next waiting queue of the following server.
Cycle time CT is the sum of waiting time and server time.

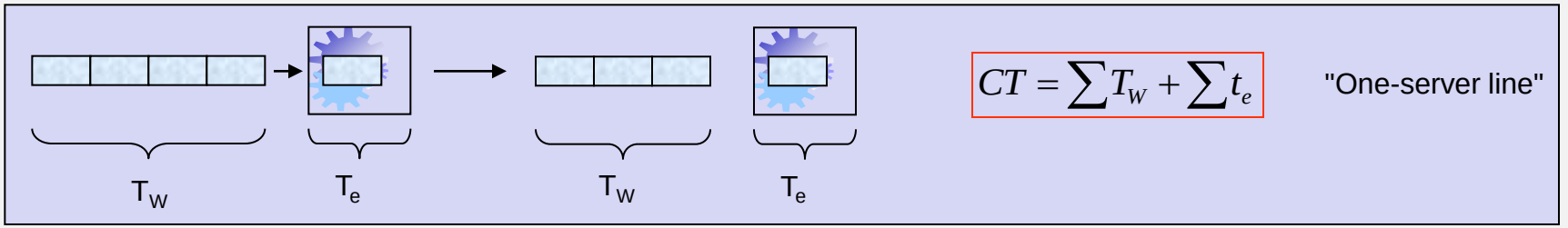
$$CT = T_W + t_e$$

In complexe production lines with various products the calculation of cycle time may be complicated

Example 1:

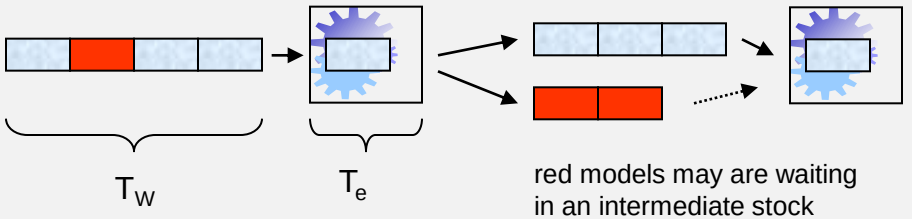


Example 2:



Example 3:

(special cases may be handeled different in various factories and may cause confusions, if cycle times are compared)



$CT = ?$

various possibilities:

CT is measured for
all products after finishing

CT is calculated separately
for each product

intermediate stock is considered → for special products the waiting time is longer

intermediate stock is not considered → because special products may have longer t_e this can be calculated now

Care has to be taken for cycle time CT calculations in complexe production lines with various products

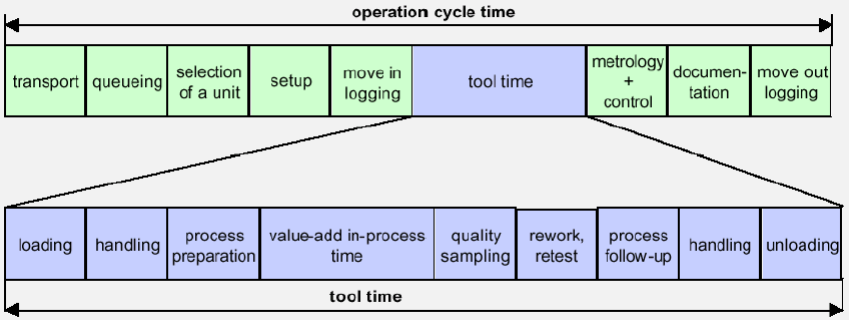
Parameters of Productivity Control

Raw cycle time RCT

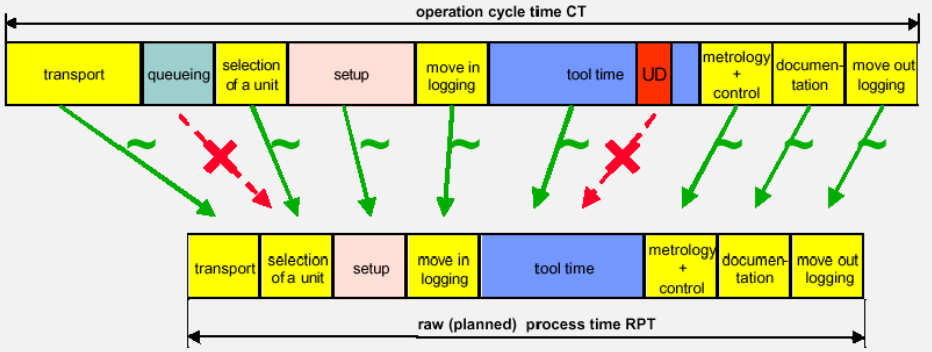
Cycle time without any waiting time. This is the physical shortest fabrication time for a product and is just the sum of all server times.

$$RCT = \sum t_e$$

sometimes in the literature the RCT is also named as "Raw Process Time RPT"



* source: Baur et al, GaAs Mantech Inc., 2001



Example for the definition of **CT** at Infineon in a (former) GaAs line*

the CT includes here:
transport time to the tool
+ time in queue + tool setup time
+ process time in tool + measurement + time for documentation (run sheets and data base)
= cycle time for one **operation**

Example for the definition of **RCT** at Infineon in a (former) GaAs line*

(from the cycle time as defined left only the queueing time and always existing tool breakdowns (-unscheduled downs UD) are subtracted, because all other times are unavoidable neccessary and therefore counted as RCT.
This definition is different from RCT as defined usually (only tool time!).

Flow factor FF

The ratio of cycle time CT and raw cycle time RCT describes how longer the fabrication time is compared to the best value.
With this normalization different tools and production lines can be compared.

$$FF = \frac{CT}{RCT}$$

Parameters of Productivity Control

Throughput TP

sometimes also called

Going rate GR (output)

the average output of products per time of a tool or a complete production line.
At this point there is no definition, if throughput is counted only for good units or all fabricated units (including scrap).
- Sales manager define as throughput the number of sold products per time,
- Production manager define as throughput the portion of sellable (= good) products per time.

Throughput TP [units/time] can be measured or calculated using Little´s Law:

$$TP\left[units/time\right] = \frac{fabricated\ \ units}{time}$$

$$TP\left[units/time\right] = \frac{WIP\left[units\right]}{CT\left[time\right]}$$

Little´s Law

Attention: To **control !** productivity in production lines sometimes special definitions of throughput are used, where the units are no longer [units/time], but only [units] and usually lead to irritations by mixing up equations.

The reason is, that in practice a big part of throughput is fixed by the **capability of the process tools**, which is given as a throughput or better as a process speed in [units/time] (typical [wafer/hour]).
For the line control management, this value is fixed, it does not help to characterize or control the real throughput. For this reason a more flexible definition is chosen:

$$tool\ speed\left[\frac{wafer}{time}\right] \cdot A_{Op} \cdot A_{WIP} \cdot A_{Proc} \cdot A_{tool} \bullet 24h = Daily\ Going\ Rate\left[units\right]$$

Daily Going Rate DGR

fixed

control abibility = availability of the partners in period of investigation

Output (Ausstoß) [units] in period under observation

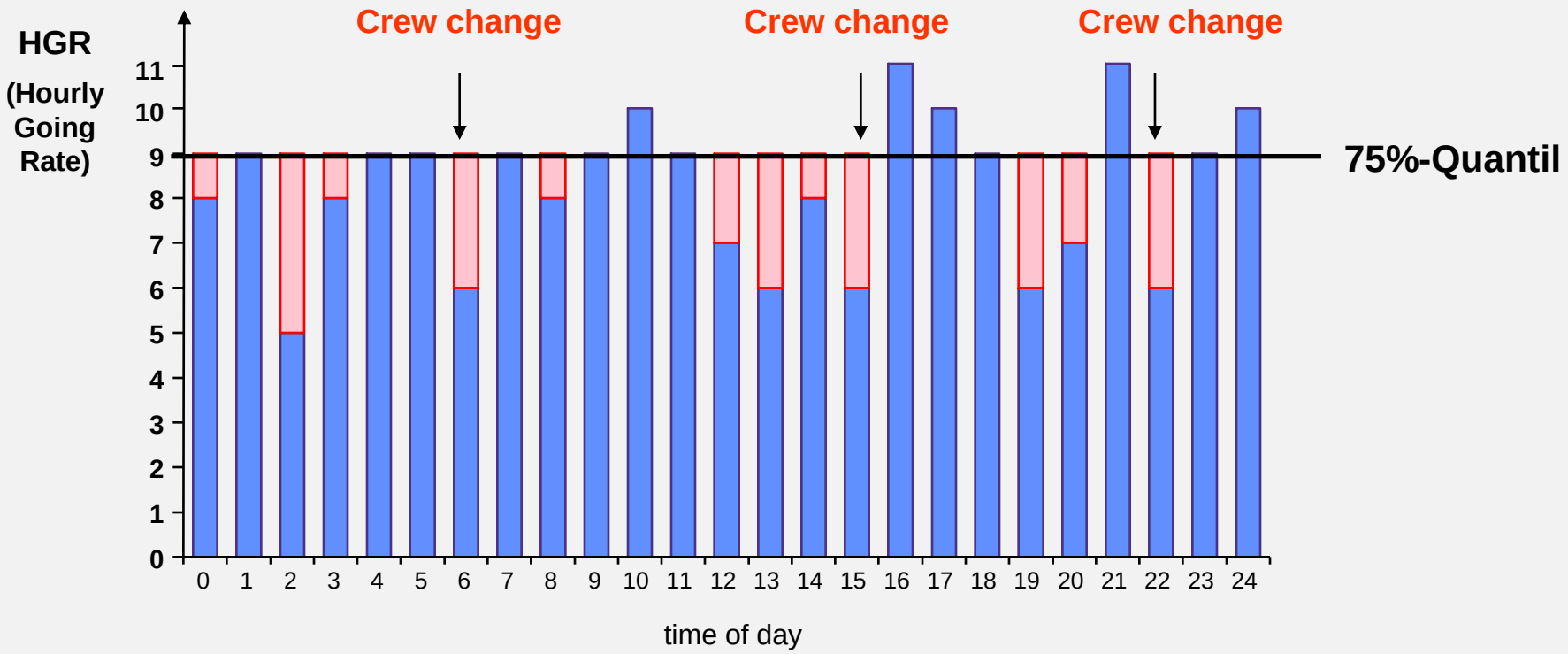
The Daily Going Rate DGR usually is applied for a production period (e.g. one day) and used to characterize improvements in productivity for a longer period (for 1 year).
For actual (dynamically) control shorter periods are used (**hourly going rate HGR, minutely going rate MGR).**
By electronically login and logout of the lots (=wip) at the tools, the throughput of the last 60min will be visualized the control management at their PC on their office desk. If interrupts in HGR occur, the line controller immediately runs into the production line to find out, which partner causes the interrupt and tries to start counter measures.

But attention again!

In practice frequently the units of DGR are mixed, sometimes given in [units], sometimes in [units/time]

Parameters of Productivity Control

Typical **cycle chart** of a semiconductor fab



Parameters of Productivity Control

Capacity Capa

is also like throughput a complicated term with the same puzzling of units.
Usually with capacity the maximum throughput is defined, which could only be achieved, if all 4- Partners would be available for 100%:

$Capa_{max}$ [units/time]

$$Capa_{max} = process\ speed\left[\frac{wafer}{time}\right] \bullet 24h\ [units]$$

Frequently this $Capa_{max}$ with 100% partner availabilty is not choosen, but corrected for experience data for availability of the partners:

Capa= $Capa_{max}$ * (established partner availability)

Capacity of a production tool PT:

$$Capa_{PT} = 24h \cdot tool\ \ process\ speed\left[\frac{wafer}{time}\right] \cdot A_{Proc} \cdot A_{tool} = 24h \cdot tool\ \ process\ speed\left[\frac{wafer}{time}\right] \cdot uptime$$

Capacity of a production unit PU:

$$Capa_{PU} = 24h \cdot process\ speed\left[\frac{wafer}{time}\right] \cdot A_{Proc} \cdot A_{tool} \cdot A_{Op}$$

In contradiction to the throughput, these values of capacity are taken as <u>fixed</u> and not corrected every hour or day.

Utilization U

The utilization is a dynamically performance parameter, which relates the actual throughput TP to the maximum possible throughput, namely the capacity Capa:

$$U = \frac{TP}{Capa} \qquad U[\%] = \frac{TP[units/time]}{Capa[units/time]} \cdot 100\%$$

The Operation Curve

Starting with the Kingman-equation and using the terms of production we achieve:

mean waiting time in queue — variability — mean number of utilization — mean serving time

Kingman-equation:

$$\bar{T}_w(G|G|1) = \left(\frac{c_a^2 + c_e^2}{2}\right) \cdot \left(\frac{u}{1-u}\right) \cdot \bar{t}_e$$

Just an approximation !

with: $\left(\frac{c_a^2 + c_e^2}{2}\right) = \alpha$

$$T_W(G|G|1) = \alpha \cdot \left(\frac{U}{1-U}\right) \cdot \bar{t}_e$$

adding t_e on both sides:

$$T_W(G|G|1) + \bar{t}_e = \alpha \cdot \left(\frac{U}{1-U}\right) \cdot \bar{t}_e + \bar{t}_e$$

and using: $CT = T_W + \bar{t}_e$

$$CT(G|G|1) = \left(\alpha \cdot \left(\frac{U}{1-U}\right) + 1\right) \cdot \bar{t}_e$$

using: $\bar{t}_e = RCT$

$$CT(G|G|1) = \left(\alpha \cdot \left(\frac{U}{1-U}\right) + 1\right) \cdot RCT$$

using: $FF = \frac{CT}{RCT}$

$$FF(G|G|1) = \alpha \cdot \left(\frac{U}{1-U}\right) + 1$$

Operation Curve

(advantage: universal use for any production line)

Parameters of the Operation Curve

$$FF(G|G|1) = \alpha \cdot \left(\frac{U}{1-U} \right) + 1$$

operation curve

showing

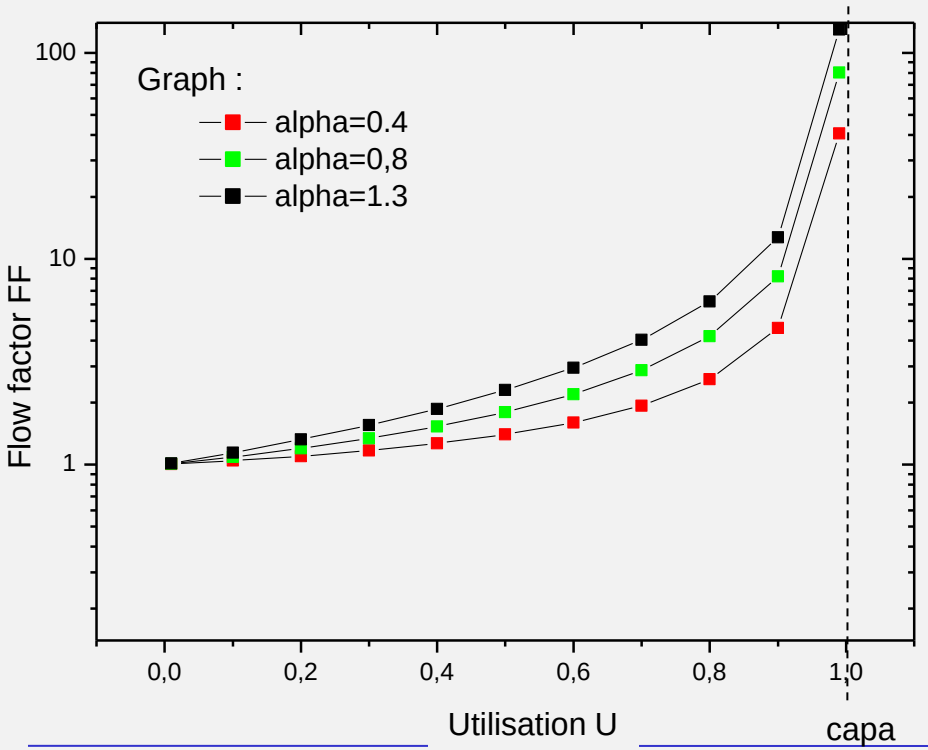
basic parameters

$$\frac{CT}{RCT} = \alpha \cdot \frac{\frac{TP}{Capa}}{1 - \frac{TP}{Capa}} + 1$$

Kingman-equation

$$CT = \frac{WIP}{TP}$$

Little´s Law



3 Parameters are used in the Kingman-equation, which may be changed in a production line only on a **long-term scale**:

raw cycle time	RCT [time]
Maximum capacity	$CAPA_{max}$ [units/time]
Variability	α [0 - ∞]

The other 3 parameters may be changed on a **short-term scale**:

cycle time	CT [time]
number of products	wip [units]
throughput	TP [units/time]

Because 2 parameters are connected by equations (Kingman, Little), in reality only one parameter is free to change, both other parameters are dependent.

Various Forms of the Operation Curve

Collection of equations:

$$\underbrace{FF = \frac{CT}{RCT}}_{\text{Definition of FF}} \qquad \underbrace{U = \frac{TP}{Capa}}_{\text{Definition of utilization}} \qquad \underbrace{CT = \frac{WIP}{TP}}_{\text{Little's Law}} \qquad \underbrace{FF = \alpha \frac{U}{1-U} + 1}_{\text{Operation curve (One-Server-Approximation)}}$$

Various Forms of the Operation Curve (dependent of the choosen free variable):

Variable → / dependent parameter ↓	**TP**	**CT**	**WIP**
TP		$TP = \frac{Capa}{\frac{\alpha \cdot RCT}{CT - RCT} + 1}$	$TP = \frac{Capa}{\frac{\frac{RCT \cdot Capa}{WIP} + 1}{2} + \sqrt{\left(\frac{\frac{RCT \cdot Capa}{WIP} + 1}{2}\right)^2} + \alpha \cdot \frac{RCT \cdot Capa}{WIP}}$
CT	$CT = \alpha \cdot \frac{\frac{TP}{Capa}}{1 - \frac{TP}{Capa}} \cdot RCT + RCT$		$CT = \frac{RCT + \frac{WIP}{Capa}}{2} + \sqrt{\left(\frac{RCT - \frac{WIP}{Capa}}{2}\right)^2} + \alpha \cdot \frac{RCT \cdot WIP}{Capa}$
WIP	$WIP = \alpha \cdot \frac{\frac{TP}{Capa}}{\frac{TP}{Capa} - 1} \cdot RCT + TP \cdot RCT$	$WIP = \frac{\left(\frac{CT}{RCT} - 1\right) \cdot Capa \cdot CT}{\alpha + \left(\frac{CT}{RCT} - 1\right)}$	

Which form is best suited is dependent on the problem:
- are the cycle times CT to long, - to much products (wip) in stock, ...?

Various Forms of the Operation Curve

1 Output controlled production ⟶ throughput will be changed

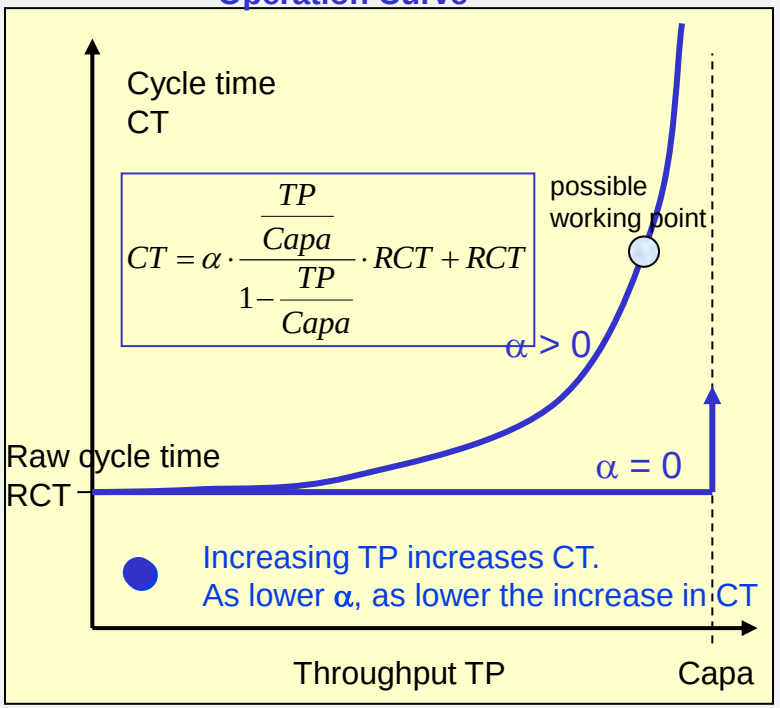
Task: Produce more products/time on the market as your competitor

If it is assumed, that: - RCT = constant (because given by the process speed of the tools. This is true, if only one product is fabricated with no changes. If the properties of products change (for example in wafer fabrication due to shrinking dimensions the adjustment time increases) or the product spectrum changes (e.g. more complicated products), the RCT will be changed as well.

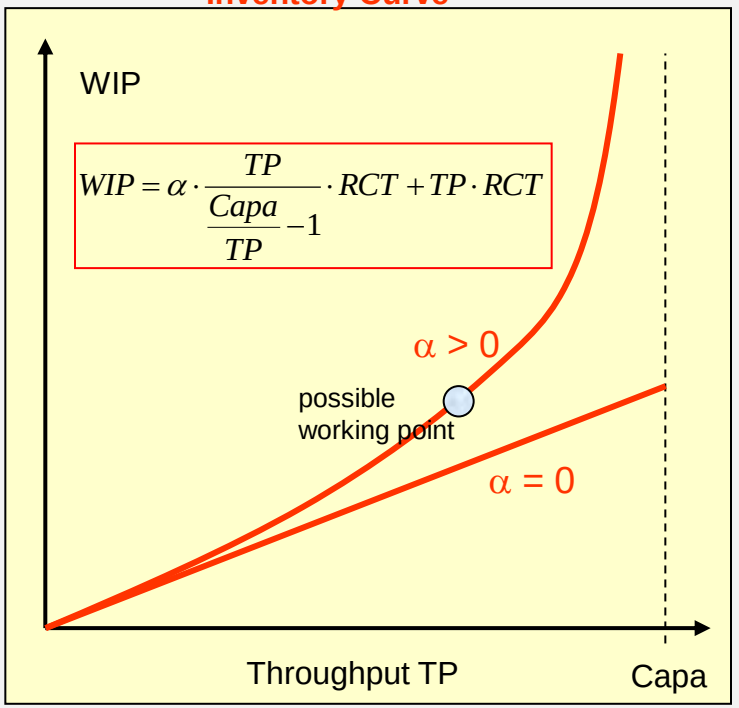
- and Capa = const (because given by the installed tools. But if RCT changes due to one of the above reasons, than Capa changes as well).
- and α = const , (this is usually not true on a short-term scale (e.g. HGR), but may be true, if long-term average values are used)

than: throughput can be varied by varying input (number of starting products per time)

Operation Curve



Inventory Curve



- An increase in TP always forces an increase of WIP (and capital commitment) (slope = RCT !)
- An improvement of inventory (= less) with increasing TP is only achievable by changing long-term parameters RCT, α and Capa !!

Various Forms of the Operation Curve

2 Cycle time controlled production ⟶ Cycle time will be changed

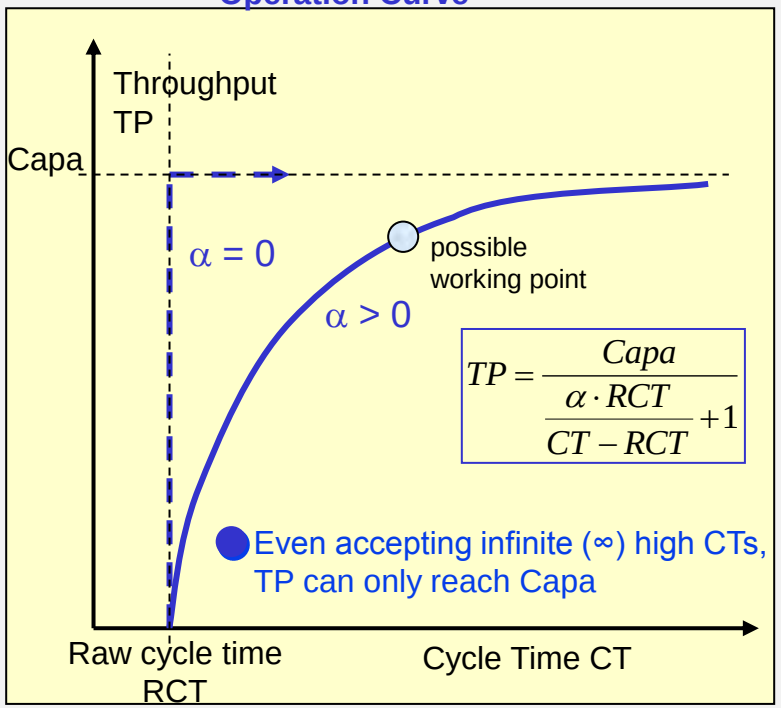
Task: Be faster on the market as your competitor

If it is assumed, that: - RCT = constant (because given by the process speed of the tools. This is true, if only one product is fabricated with no changes. If the properties of products change (for example in wafer fabrication due to shrinking dimensions the adjustment time increases) or the product spectrum changes (e.g. more complicated products), the RCT will be changed as well.

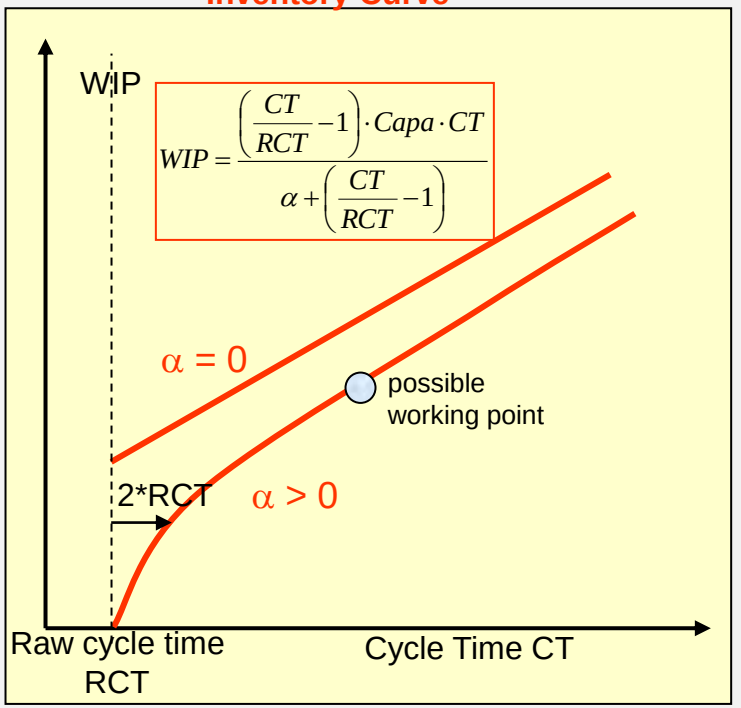
- and Capa = const (because given by the installed tools. But if RCT changes due to one of the above reasons, than Capa changes as well).
- and α = const , (this is usually not true on a short-term scale (e.g. HGR), but may be true, if long-term average values are used)

than: cycle time can be varied by varying input (number of starting products per time)

Operation Curve



Inventory Curve



- For real systems with a>0 and working points larger twice the RCT, the WIP is nearly linear in increasing CT
- CT is not a good control parameter for small α

Various Forms of the Operation Curve

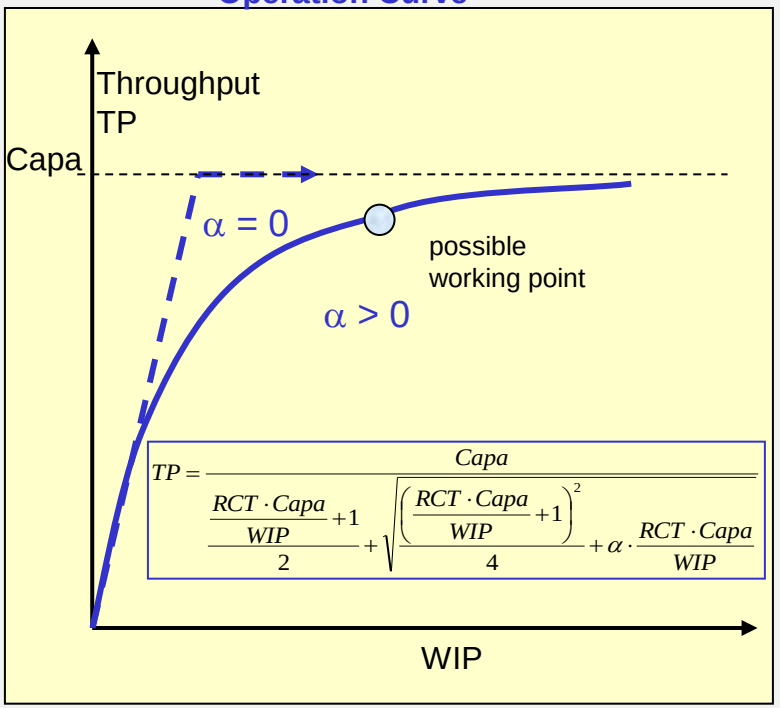
3 WIP controlled production ⟶ WIP will be changed

Task: Reduce WIP to create free money to buy-out your competitor

- RCT = constant (because given by the process speed of the tools. This is true, if only one product is fabricated with no changes. If the properties of products change (for example in wafer fabrication due to shrinking dimensions the adjustment time increases) or the product spectrum changes (e.g. more complicated products), the RCT will be changed as well.
- and Capa = const (because given by the installed tools. But if RCT changes due to one of the above reasons, than Capa changes as well).
- and α = const , (this is usually not true on a short-term scale (e.g. HGR), but may be true, if long-term average values are used)

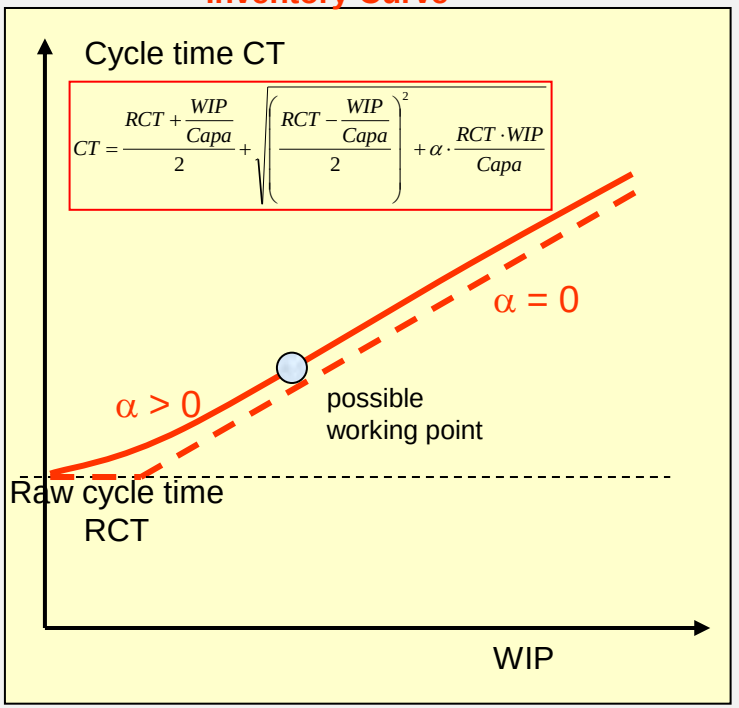
than: WIP can be varied by varying input (number of starting products per time)

Operation Curve



$$TP = \frac{Capa}{\frac{RCT \cdot Capa}{WIP \cdot 2} + 1 + \sqrt{\left(\frac{RCT \cdot Capa}{WIP \cdot 4} + 1\right)^2} + \alpha \cdot \frac{RCT \cdot Capa}{WIP}}$$

Inventory Curve



$$CT = \frac{RCT + \frac{WIP}{Capa}}{2} + \sqrt{\left(\frac{RCT - \frac{WIP}{Capa}}{2}\right)^2} + \alpha \cdot \frac{RCT \cdot WIP}{Capa}$$

- changing working points in the flat region of operation curve induces only minor changes in output

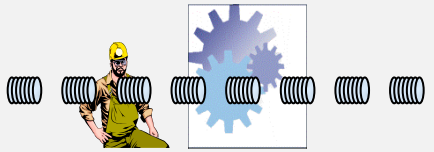
but much improvement in CT
As lower α, as more the effect

This chapter deals with the derivation of the basic parameters and equations to describe "productivity"

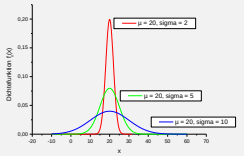
Performance parameters of a production line

short cycle times
high throughput
stable cycle times
high utilization
low inventory

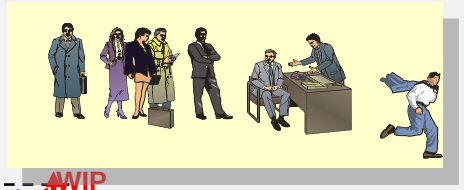
4-Partner Model



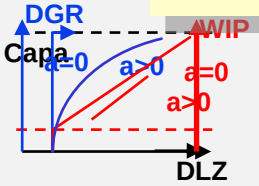
The variability



Queuing theory

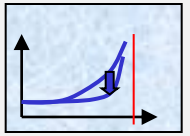


The operation curve



Calculation of the operation curve in a semiconductor fab

Optimization potentials from operation curves



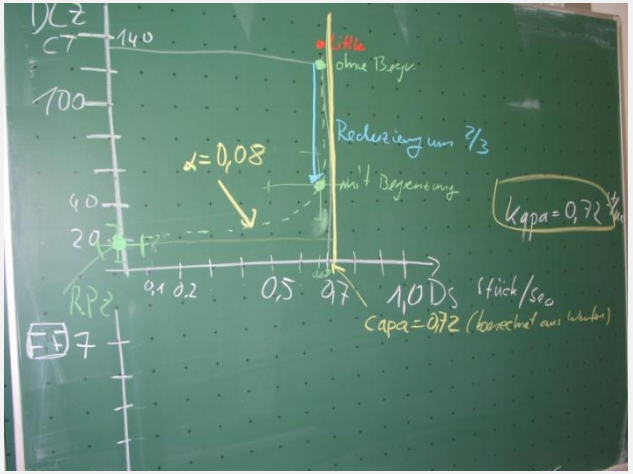
Exercise: "Chip-Factory"

Of course we can not represent a real Semiconductor fab in our hands-on training.
Our workaround is a simplified production line that has to set aside some typical attributes, but can nevertheless show the theory of **Operation Curve Management**.

Every player represents a production machine. The machining of the goods WIP is simulated by dicing.
The machine produces chips equal to the number on the dice, resulting in a variability V for every machine.
Thus it comes to accumulation and idle time in front of the machines (queuing).

The input of fabrication line

The managers



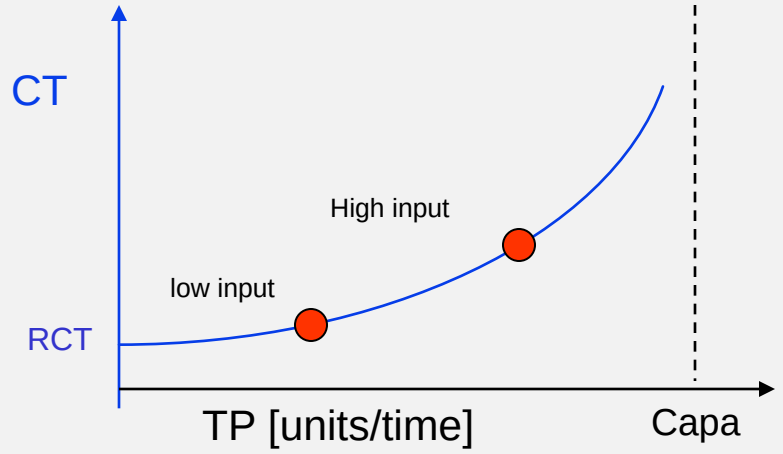
The operation curve

How do we find the operation curve? ?

We have to assume that the dicing speed of every player remains constant during the whole experiment. Otherwise it would result in an increasing capacity and therefore odd conclusions.

1. We start with a try run so that every player can find his dicing speed
2. We start new and send a single chip through the production line. With this chip we measure the **raw cycle time RCT**.
3. We start new. In the first step we limit the input to a maximum of 3 chips/throw. We continue a few minutes to wait for the production line to settle.
Then we send a marked chip into the line. This chip has to enter the queue in front of every machine. A manager measures the **cycle time CT** for the marked chip. Another player measures the **throughput TP** during the time window. Finally the total number of chips in the production line is counted. This number is called **work-in-progress WIP**. With this values we get our first point of the **operation curve**.
With the three values **CT, TP** and **WIP** we can check **Little's law**.

4. Now we want to utilize our line to full extent. We raise the input to 1-6 chips/throw. After a settling time we do a new measurement of **CT, TP** and **WIP**. Now we can enter a second point to our **operation curve**.
5. With the two value pairs (**TP**|**CT**) and the **RCT** we can calculate the two characteristic values **Variability α** and **Capacity Capa** and draw the **operation curve**.



Operation curve

$$CT = \alpha \cdot \frac{\frac{TP}{Capa}}{1 - \frac{TP}{Capa}} \cdot RCT + RCT$$

Little's law

$$TP = \frac{WIP}{CT}$$

Exercise: "Chip-Factory"

1 Measuring the raw cycle time RCT

The RCT enters the OC as a single value
Thus we do three runs to prevent a random value

RCT 1	RCT 2	RCT 3	-> RCT
21 sec	18 sec	18 sec	~19 sec

We enter the RCT in our plot

We may think over, if the mean value is influenced by the outlier of 21 sec. Maybe we do some more measurements.

2. Measuring of the first point with reduced input

The cycle time CT and the throughput TP are measured after appr. 5min with a marked chip. Afterwards the production is stopped to count the current WIP.

CT 1	TP 1	WIP 1	Test Little´s Law
60 sec	0,68 units/sec	47 units	$WIP = CT_1 \cdot TP_1 = 41\,units$

We enter (TP_1|CT_1) to our plot

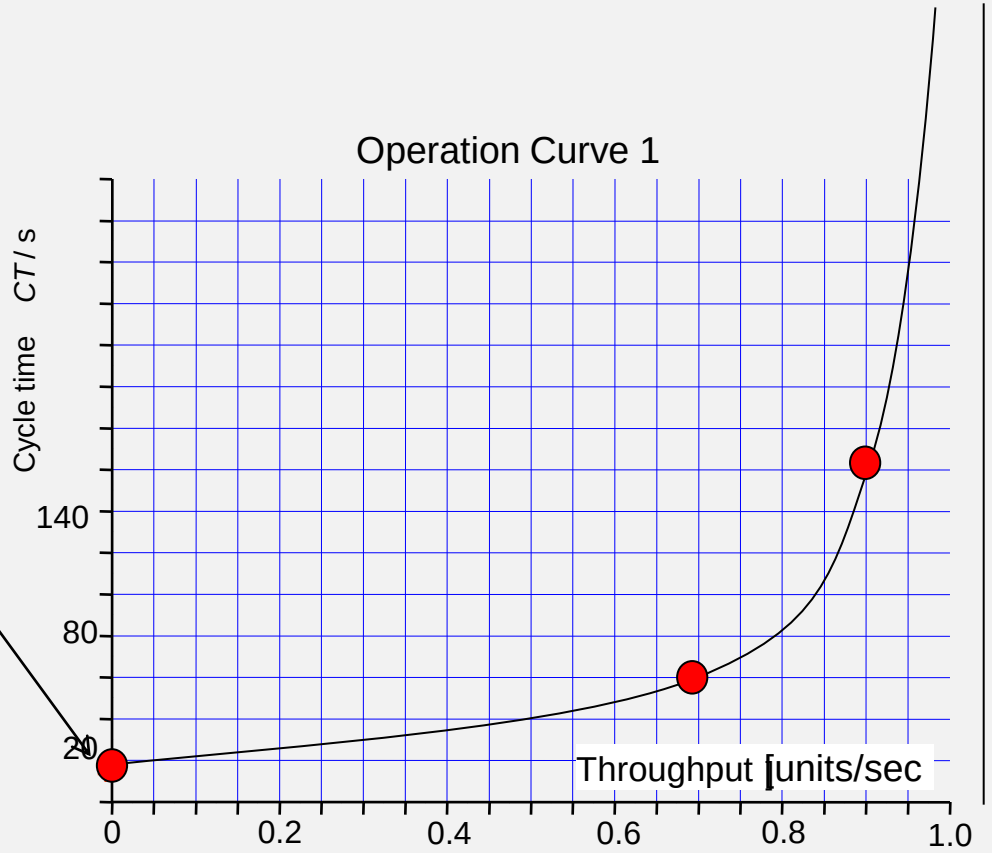
compare !

3. Measuring the second point with full input

CT 2	TP 2	WIP 2	Test Little´s Law
162 sec	0,92 units/sec	250 units	150 units

We enter (GR_2|CT_2) to our plot

compare !



$$\alpha = \left(\frac{CT}{RCT} - 1\right) \cdot \frac{1 - \frac{TP}{Capa}}{\frac{TP}{Capa}} = 1{,}18$$

Capa

(details on next slide)

$$Capa = \frac{TP_1 \cdot TP_2 \cdot (CT_1 - CT_2)}{TP_2 \cdot (CT_1 - RCT) - TP_1 \cdot (CT_2 - RCT)} = 1{,}06$$

4. Calculation of the performance indicators α and Capa

$$CT = \alpha \cdot \frac{\frac{TP}{Capa}}{1-\frac{TP}{Capa}} \cdot RCT + RCT$$

Transformation to calculate α:

$$\alpha = \left(\frac{CT}{RCT} - 1\right) \cdot \frac{1-\frac{TP}{Capa}}{\frac{TP}{Capa}}$$

Both pairs (TP|CT) should result in the same value for α, if the dicing speed (=RCT) is kept constant

$$\alpha_1 = \left(\frac{CT_1}{RCT} - 1\right) \cdot \frac{1-\frac{TP_1}{Capa}}{\frac{TP_1}{Capa}} = \alpha_2 = \left(\frac{CT_2}{RCT} - 1\right) \cdot \frac{1-\frac{TP_2}{Capa}}{\frac{TP_2}{Capa}}$$

Transformation to calculate Capa:

$$Capa = \frac{TP_1 \cdot TP_2 \cdot (CT_1 - CT_2)}{TP_2 \cdot (CT_1 - RCT) - TP_1 \cdot (CT_2 - RCT)}$$

$$Capa = \frac{0{,}68 \cdot 0{,}92 \cdot (60 - 162)}{0{,}92 \cdot (60 - 19) - 0{,}68 \cdot (162 - 19)}\ units/s = 1{,}06\ units/s$$

CAPA
1,06 units/sec

This value is entered as line to our plot

Now we can calculate α:

$$\alpha = \left(\frac{CT}{RCT} - 1\right) \cdot \frac{1-\frac{TP}{Capa}}{\frac{TP}{Capa}}$$

$$\alpha_1 = \left(\frac{60}{19} - 1\right) \cdot \frac{1-\frac{0{,}68}{1{,}06}}{\frac{0{,}68}{1{,}06}} = \alpha_2 = 1{,}18$$

ALPHA
1,18

Exercise: "Chip-Factory"

5. Drawing of the operation curve

$$CT = \alpha \cdot \frac{\frac{TP}{Capa}}{1-\frac{TP}{Capa}} \cdot RCT + RCT$$

$$CT = 1{,}18 \cdot \frac{\frac{TP}{1{,}06\ units/s}}{1-\frac{TP}{1{,}06\ units/s}} \cdot 19\ s + 19\ s$$

Now we can calculate further value pairs to

TP [unit/s]	0	0,21	0,42	0,64	0,85	0,95	1,05
-> CT [s]	19	24,6	33,9	52,6	108,7	220,8	2239

6. Normalized diagram

The value pairs (TP|CT) are transformed with th following equations:

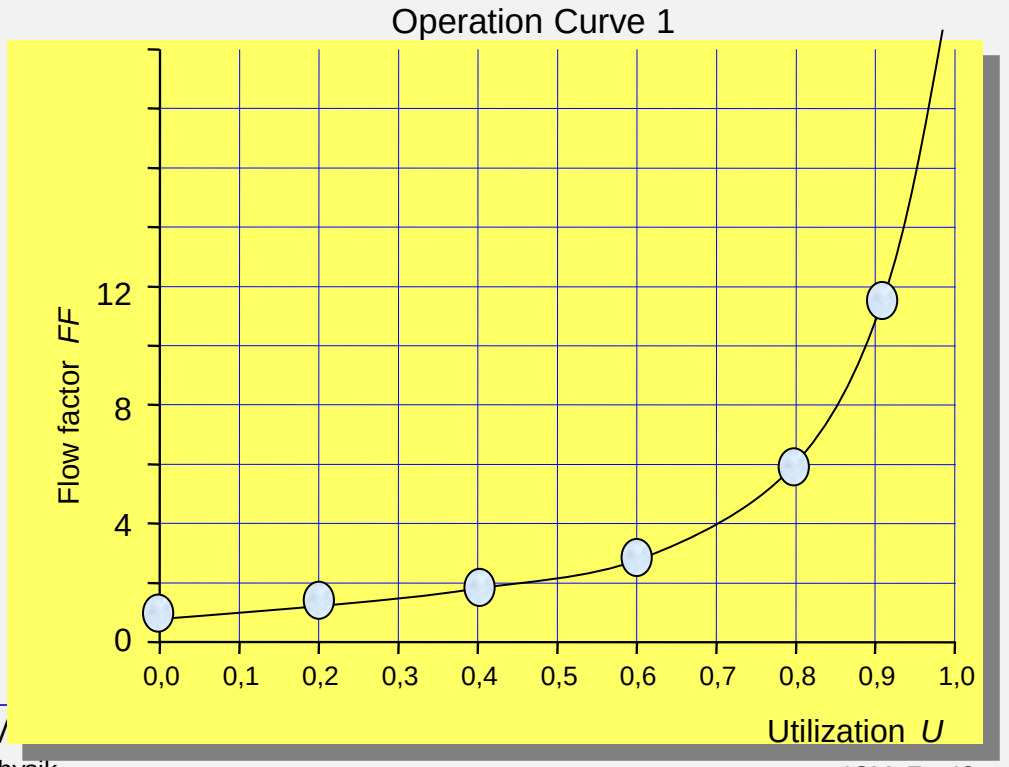
$$FF = \alpha \cdot \left(\frac{U}{1-U}\right) + 1$$

$$FF = \frac{CT}{RCT}$$

$$U = \frac{TP}{Capa}$$

The resulting value pairs (U|FF) are:

TP [units/s]	0	0,21	0,42	0,64	0,85	0,95	1,05
CT [sec]	19	24,6	33,9	52,6	108,7	220,8	2239
U	0	0,2	0,4	0,6	0,8	0,9	0,99
FF	1	1,3	1,79	2,77	5,72	11,6	118



Generating the Operation Curve in Industry

Each product follows an **working plan i,** which processes have to be performed in which sequence

Example:

working plan of product:

HMOS I

1 **LASER MASK**
2 INIT OX
3 **P-TUBE MASK**
4 P-TUBE IM
5 NITRID DEP
6 **FIELD MASK**
7 FIELD DIFF/OX
8 **S/D MASK 1**
9 S/D IM 1
10 S/D OX 1
11 **S/D MASK 2**
12 S/D IM 2
....

working plan of product:

ACMOS V

1 **LASER MASK**
2 EPI DEP
3 INIT OX
4 **P-TUBE MASK**
5 P-TUBE IM
6 **TRENCH MASK**
7 TRENCH ETCH
8 SOD DEP
9 DRIVE-IN
10 TRENCH OX
11 TRENCH POLY
12 **S/D MASK 1**
12 S/D IM 1
....

The number of **mask steps** is used as a count for the complexity of a product

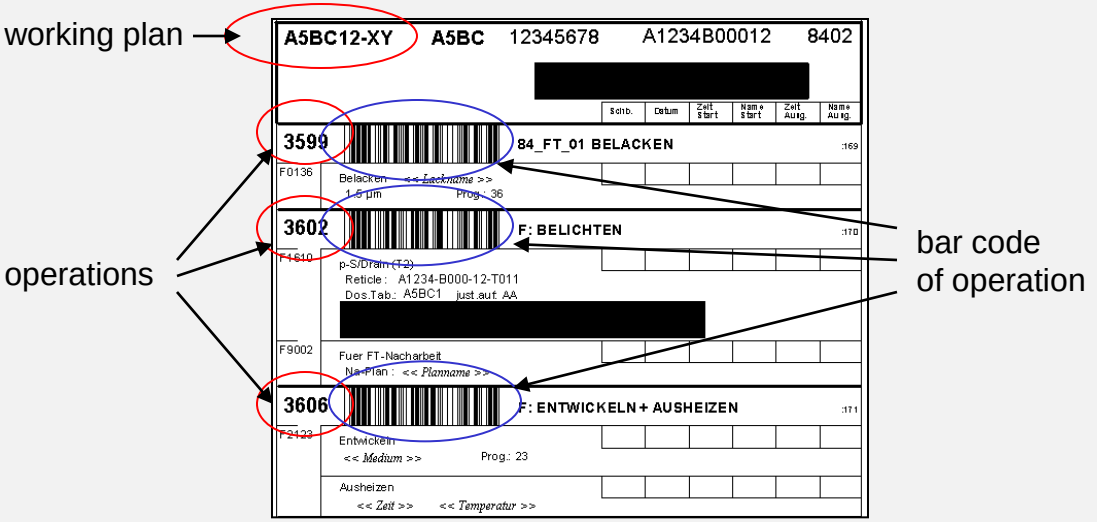
Each of the shown steps consist of 5-20 processes, which are called **operations L** with own numbers, e.g. step MASK may be:

8012 coating: resist DAC22, thickness 1.5µm, spin-on program 15

8040 exposure: mask A1200-B008-T4002, dose plan: A5BC1, ...

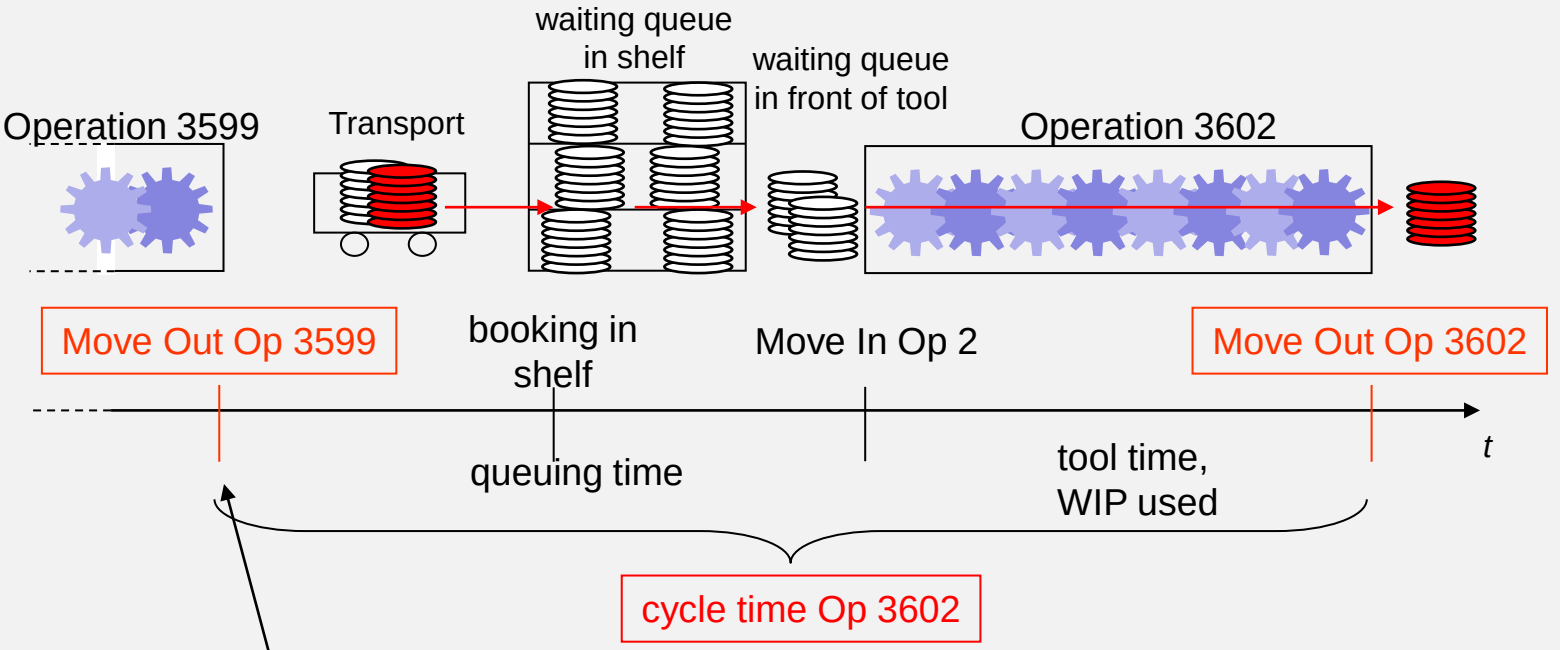
8041 development: development plan: B32

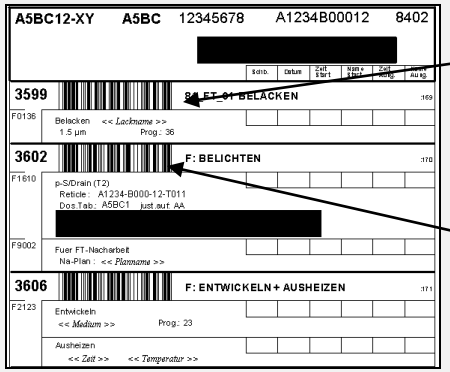
The working plan accompanies each lot (product) in the form of a **run sheet**:



one page of about 200 pages from the whole working plan

Measurement of Cycle Time





After completing operation (e.g. 3599) the operator books out the lot (with a bar code reader).
This time stamp will be stored in the data base: Move-out Op3599.

The operator then has to transport the lot to the waiting shelf before the next operation.
(Here the lot will be booked-in again in another data base to find the actual storage location).

The operator of the following operation 3602 takes the lots out of the shelf following certain serving rules, also booking this lot out of the shelf.
After completing his operation (e.g. 3599) the operator books out the lot (with a bar code reader).
This time stamp will be stored in the data base: Move-out OP3602)

Using the data base, from the **time difference** (Move-out Op3599 - Move-out Op3602) the **actual cycle time** for a special **lot n** from **working plan i** at the **Operation l** at **tool m** can be calculated and from the **amount** of move-out bookings the **actual throughput**.

Calculation of Special Throughputs

! Only using **move-out stamps** the **actual cycle time $CT_{n,lim}$** and the **actual throughput DGR_{lim}** can be calculated

lim = Operation l in work plan i at tool m

For both, cycle time and throughput, summing up can be done:

remember throughput:
(as it is common in semiconductor industry)

DGR = daily going rate = output in units in one day

DGR_{lim} = daily output of **Operation *ℓ*** from **working plan *i*** at **tool *m***

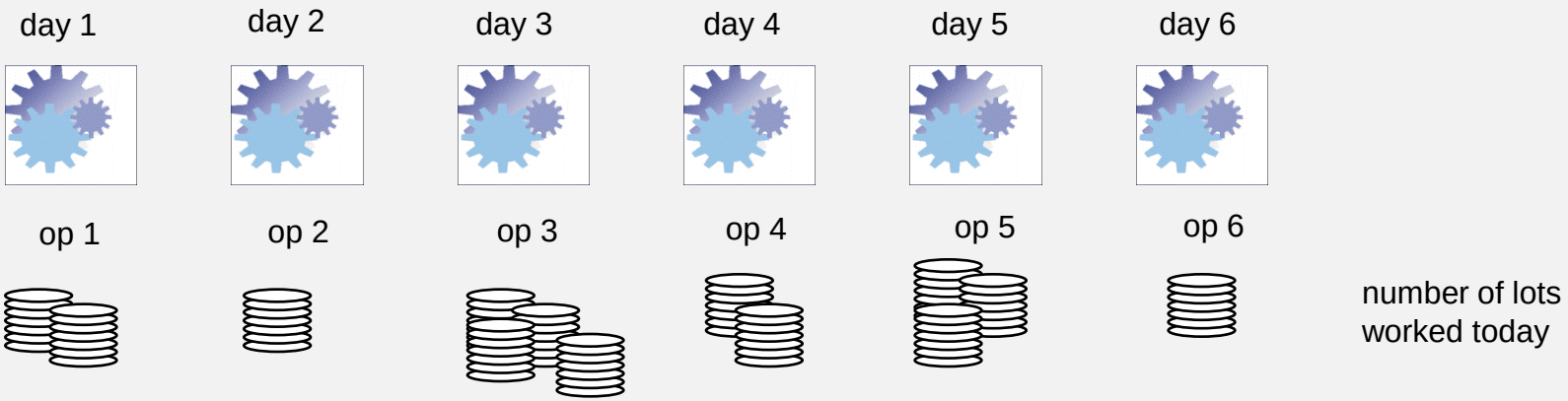
- DGR_{li} = daily output of operation *ℓ* from working plan *i* $= \sum_{k=1}^{m} DGR_{lim}$ (summing over all tools m -> e.g. all exposures for product i)
- DGR_{l} = daily output of operation *ℓ* $= \sum_{k=1}^{k=i} \sum_{k=1}^{m} DGR_{lim}$ → summing for one operation l over all tools m and all products i e.g. all exposures today
- DGR_{i} = daily output of working plan i $= \sum_{k=1}^{k=l} \sum_{k=1}^{m} DGR_{lim}$ → e.g. the total number of moved wafers for product i
- DGR = daily output of complete production $= \sum_{k=1}^{k=l} \sum_{k=1}^{k=i} \sum_{k=1}^{m} DGR_{lim}$ → Sum of all processes and products

Dividing the DGR_i of a special product i by the number of all operations L_i, an "extrapolated" throughput of the product i can be calculated, which is called **dynamical DGR_i** (=$dynDGR_i$).

$$dynDGR_i = \overline{DGR}_i = \frac{1}{L_i} \cdot \sum_{l=1}^{Li} DGR_{li}$$

Summing up all products i, the actual, dynamical throughput of the production can be calculated (=dynDGR). $TP_{line} = \sum_{p=1}^{i} dynTP_i$

Example of Dynamically Throughput Calculation



The situation (in an example):

If a product will be fabricated in 6 operations, each operation consuming 1 day (neglecting interrupts), each tool may be able to produce a maximum of 4 lots per day.
The measured throughputs will be:
1st day: output 1 lot from op 6: TP=1/day
2nd day: output 3 lots moved the day before from op 5 to op 6 : TP = 3 / day
....
-> the measured throughput during the last 6 days will be: 13 lots / 6 days = 2.2 lots/day

These throughputs are measurements of the past 6 days and therefore can not be used to find actual problems

The calculation of the **dynamically throughput** delivers:

Actual (today = now): (2 lots at op1) + (1 lot at op2) + ... + (1 lot at op 6) = 13 lots / 6 ops = 2.2 units/day

This throughput is calculated now, for this moment and can be used to detect actual problems

Calculation of Specific Throughputs

Analogous the **actual cycle times $CT_{n,lim}$** can be calculated from the move-out stamps

$CT_{n,lim}$ = *CT* of **Lot *n*** for **Operation *ℓ*** from **working plan *i*** at tool ***m*** (single-lot CT)

- CT_{lim} = CT of operation *ℓ* from working plan *i* at tool *m*
- CT_{li} = CT of operation *ℓ* from working plan *i* $= \sum_{k=1}^{m} CT_{lim}$
- CT_i = CT from working plan *i* $= \sum_{k=1}^{k=l} \sum_{k=1}^{m} CT_{lim}$
- CT_l = CT of operation *ℓ* $= \sum_{k=1}^{k=i} \sum_{k=1}^{m} CT_{lim}$
- CT = CT of production line $= \sum_{k=1}^{k=l} \sum_{k=1}^{k=i} \sum_{k=1}^{m} CT_{lim}$
- CT...

Analogous to the dynDGR:

Dividing the cycle time CT_i by the number of operations L_i,
an extrapolated cycle time of the product i can be calculated,
which is called **dynamical CT_i** (=$dynCT_i$).

$$dynCT_i = \frac{1}{L_i} \cdot \sum_{l=1}^{Li} CT_{li}$$

But in contrast to the dynDGR, the summing up of all products to calculate a dynCT for the whole fab usually makes no sense, because for various products simultaneously fabricated, the $dynDGR_i$ and the number of wafers may be very different.

But using Little´s Law and summing up the number of all lots in work from dynDGR another form of dynCT can be calculated:

dynamical cycle time: $$dynCT = \frac{WIP}{dynDGR}$$

which is used for consistency in equations and calculations
e.g. to calculate a dynamically flow factor dynFF

Statical and Dynamical Cycle Times

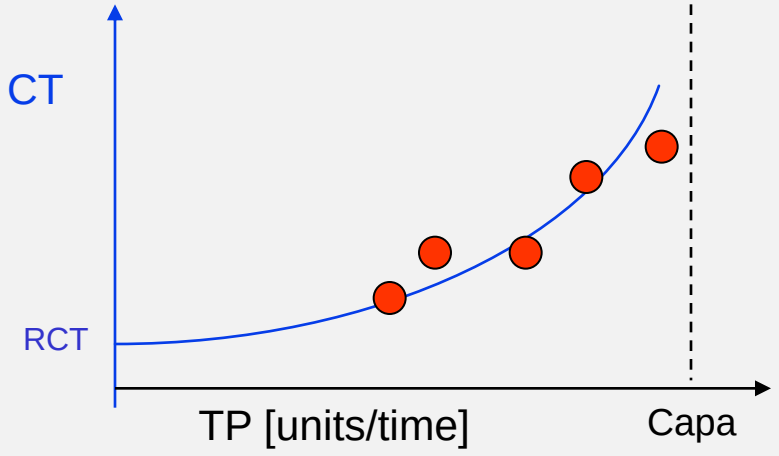
Take care:

1. Placing ourself at the output of a production line and looking to the real cycle time CT (= time between input - output of line) of selected products, we will receive a picture of the production line of the past days. This value of cycle time is relevant for the sales managers and the consumers, but not a usable parameter to react fast on distortions in the production flow. We could call this parameter **statical cycle time**.

2. Using the momentary move-outs and summing ups for every minute (or second or hour), we can calculate the momentary, dynamically CT, which gives us a picture of the momentary status of the production line. These calculated, dynamically values are best-suited to react on distortions in the production flow. The dynamically values are the characterization and control parameters of a semiconductor production line.

- actual, measured CT: consumer relevant parameter
- dynamically CT: Line performance and dynamics

Changing DGRs (TP) and CTs by varying the amount of input units, the operation curve can be constructed



How we can find the long-term performance parameters of variability (alpha) and capacity (capa) ?

$$CT = \alpha \cdot \frac{\frac{TP}{Capa}}{1 - \frac{TP}{Capa}} \cdot RCT + RCT$$

Calculation of Utilization and Variability

We remember:

$$CT = \alpha \cdot \frac{\frac{TP}{Capa}}{1-\frac{TP}{Capa}} \cdot RCT + RCT$$

This form of the operating curve delivers absolute parameters for cycle time CT and throughput TP.
A better (normalized) estimation delivers the normalized form using the flow factor FF.

$$FF = \alpha \frac{U}{1-U} + 1$$

Operation Curve)

with: $FF = \frac{CT}{RCT}$ and $U = \frac{TP}{Capa}$

We rearrange and use dynamically values for the complete production line (because we want to control):

$$\alpha_{line} = \frac{(dynFF_{line} - 1)(1 - U_{line})}{U_{line}}$$

or: $$\alpha_{line} = \frac{(dynFF_{line} - 1)(1 - UUm_{line})}{UUm_{line}}$$

with a little different value UUm (manufacturing utilisation uptime) for utilisation U (see next pages)

the task is now:

We are interested in the variability alpha of the complete production line.
How can we calculate a dynamically flow factor $dynFF_{line}$ and a total utilisation of the complete production line U_{line}, or UUm_{line} ?

Calculation of Utilization and Variability

Calculation of Flow Factor FF:

For each product i, a dynamically flow factor $dynFF_i$ can be defined by using the raw cycle time RCT_i and the calculated dynamically cycle time $dynCT_i$ for this product.
The $dynCT_i$ may be determined independent from the move-out stamps or calculated by finding the $dynDGR_i$ and then using Little´s Law.

$$dynFF_i = \frac{dynCT_i}{RCT_i} = \frac{\frac{WIP_i}{dynTP_i}}{RCT_i}$$

← sum of all lots in the fab, known by the computer
← calculated by the time stamps
← roughly calculated by the tool speeds, taking into account the number of tools used by product i

For the calculation of the total flow factor a summation over all products has to be done

Possibility 1:

$$dynFF_{line} = \frac{1}{P}\sum_{i=1}^{P} dynFF_i$$

The arithmetical mean value of all the $dynFF_i$ of all products i is not a good measurement for the total performance of the production line, because each product may have a complete different volume (WIP and utilisation) in the production. (same argument as before for the calculation of a total dynCT)

Possibility 2:

$$dynFF_{line} = \frac{\sum_{p=1}^{i} dynFF_i \cdot TP_i}{TP_{line}} = \frac{\sum_{p=1}^{i} \frac{dynCT_i}{RCT_i} \cdot \frac{WIP_i}{dynCT_i}}{TP_{line}} = \frac{\sum_{p=1}^{i} \frac{WIP_i}{RCT_i}}{TP_{line}}$$

In consideration of the different volume, the $dynFF_i$ of each product is weighted by its contribution to the whole line output TP_{line}, before we sum up.

using: $TP_{line} = \sum_{p=1}^{i} dynTP_i$ (We use TP instead of DGR to match the units)

Calculation of Utilization and Variability

Calculation of utilisation:

$$U = \frac{TP}{Capa}$$

In the design of a production line a theoretical maximum capacity $Capa_{theo}$ (wafer/day) is planned and according to the process speed (throughput [wafer/hour] for each tool a number of tools can be calculated.
As a consequence for each product i with a special working plan i a specific raw cycle time $RCT_{i,theo}$ can be calculated .
But in practice we have seen (see page 33), that in the raw cycle time always existing failure times (non-productive times) may be included. In this case by back-calculation a different (= lower) Capa is the consequence.
For identification, that a tool-modified, back-calculated Capa is used, the utilization U will be noted as **UUm (manufacturing uptime utilisation).**
In practice Capa or UUm can be calculated in two ways:

time-based manufacturing uptime utilization

contribution of productive time (without engineering time) at the uptime, summed up over all tools.
Not counted are: measurement equipment, transport systems, wafer-handling systems, tools for rework

$$UUm_{line} = \frac{\sum_{m=1}^{M} PRT_m}{\sum_{m=1}^{M} (PRT_m + SBT_m)}$$

PRT = Productive Time
SBT = Standby Time

Empirical values of PRT_m, SBT_m will be introduced in the data base system by hand from the operator or automatically by the tool

$$Capa_{line} = \frac{DGR}{UUm_{line}}$$

Problem: bad tools (which limit throughput) can not be identified by summing up and averaging

throughput-based manufacturing uptime utilization

For the momentary product mix the throughput-limiting tool or work center (bottleneck tool) is searched for. In semiconductor lines this is usually the lithography center (because the number of tools is kept as low as possible, because these tools are very expensive, 30 Mill. € each).

$$Capa_{line} = \begin{cases} maximum\ \ throughput\ of\ \ bottleneck \\ \qquad if\ \ less\ than\ 5\%\ of\ \ lithography \\ maximum\ \ throughput\ in\ lithography\ else \end{cases}$$

$$UUm_{line} = \frac{DGR}{Capa_{line}}$$

The throughput-based calculation of capacity delivers a much lower (but more realistic) value compared to the time-based utilization

Calculation of Utilization and Variability

Now with:

$$UUm_{line} = \frac{DGR}{Capa_{line}}$$

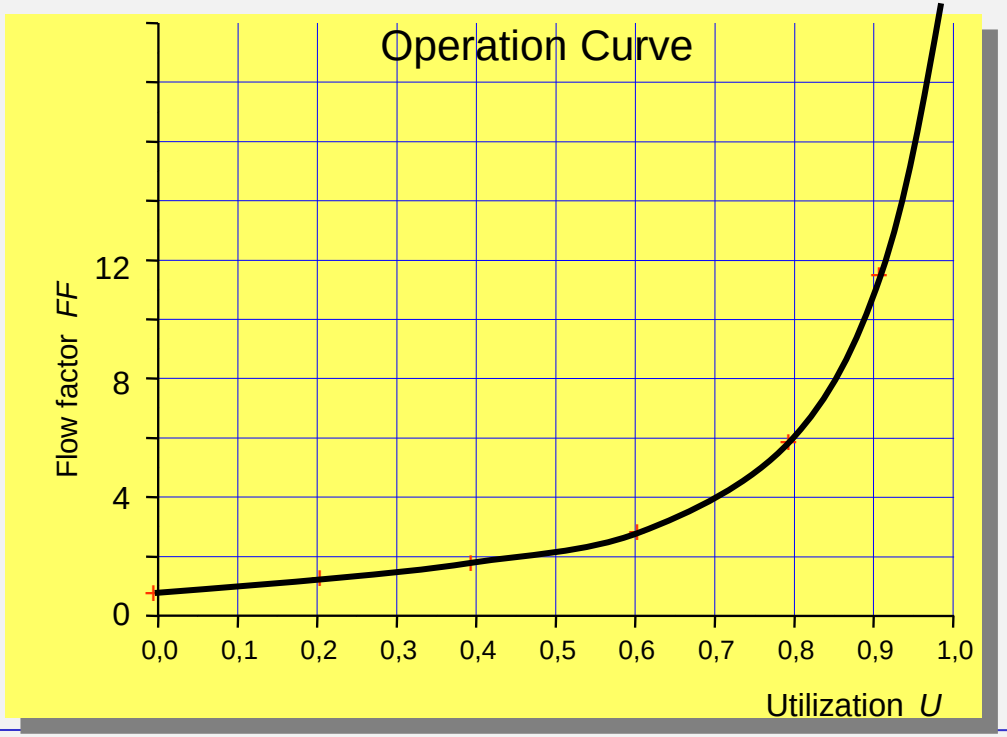
$$dynFF_{line} = \frac{\sum_{p=1}^{i} dynFF_i \cdot TP_i}{TP_{line}} = \frac{\sum_{p=1}^{i} \frac{dynCT_i}{RCT_i} \cdot \frac{WIP_i}{dynCT_i}}{TP_{line}} = \frac{\sum_{p=1}^{i} \frac{WIP_i}{RCT_i}}{TP_{line}}$$

daily measurement !

and calculation of alpha:

$$\alpha_{line} = \frac{(dynFF_{line} - 1)(1 - U_{line})}{U_{line}}$$

the operation curve of the Fab can be constructed:

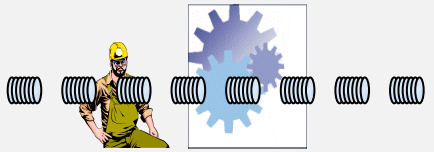


This chapter deals with the derivation of the basic parameters and equations to describe "productivity"

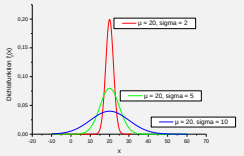
Performance parameters of a production line

short cycle times
high throughput
stable cycle times
high utilization
low inventory

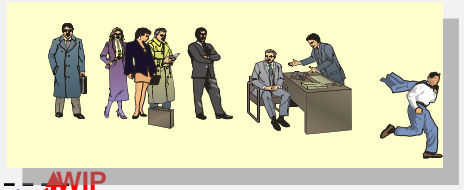
4-Partner Model



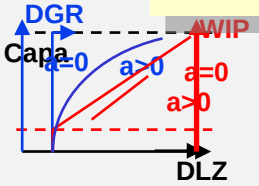
The variability



Queuing theory

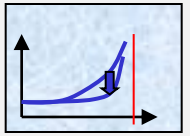


The operation curve

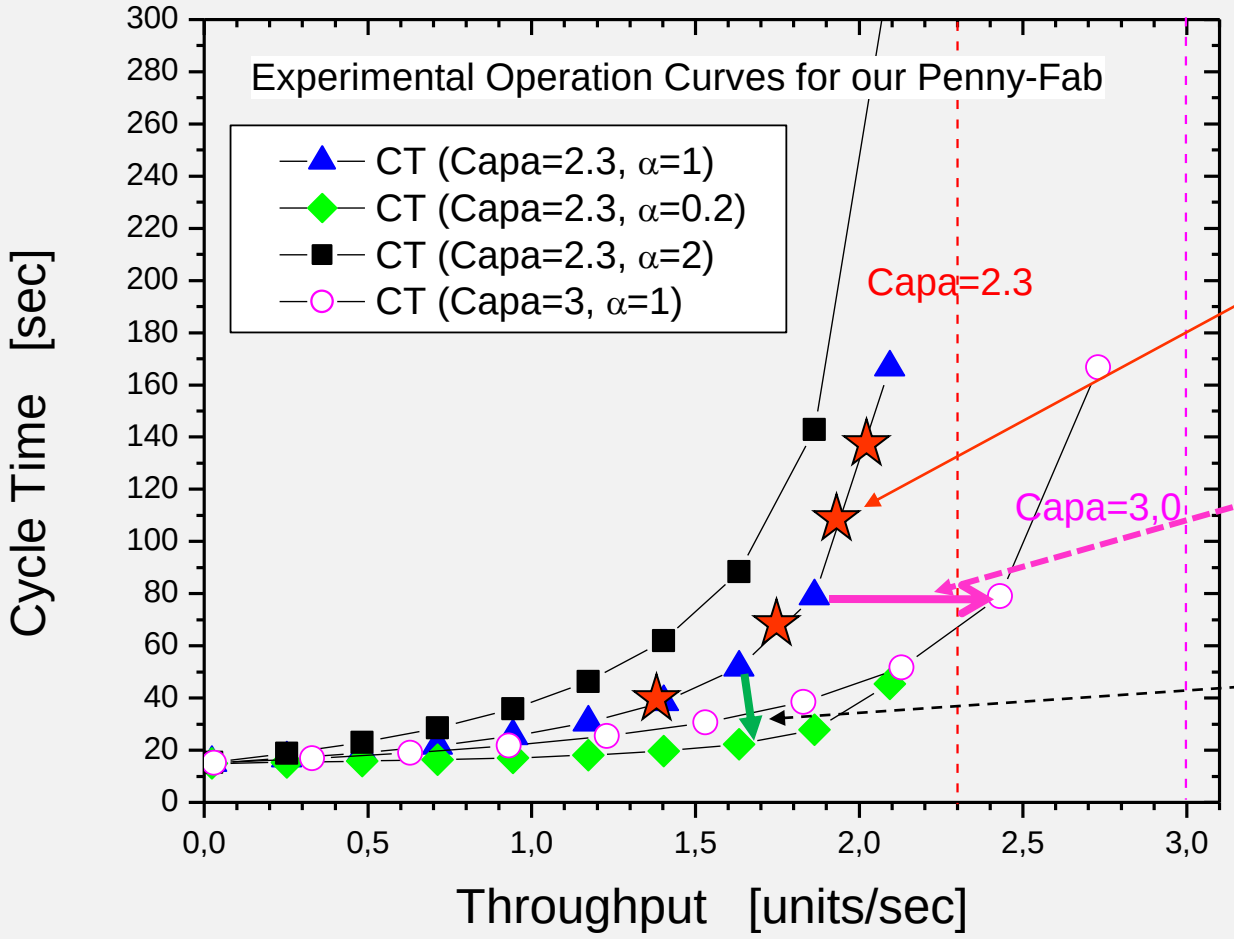


Calculation of the operation curve in a semiconductor fab

Optimization potentials from operation curves



Optimization of Operation Curve



various operation points ★
without improvement of productivity

improved operation curve
by increased capacity capa

improved operation curve
by lower variability α

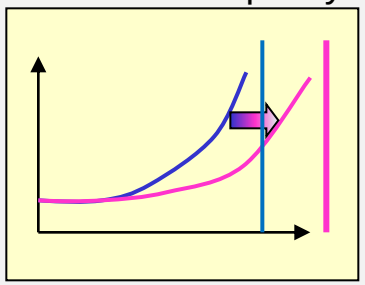
Goal of optimization:

Improved operation curves are achievable by

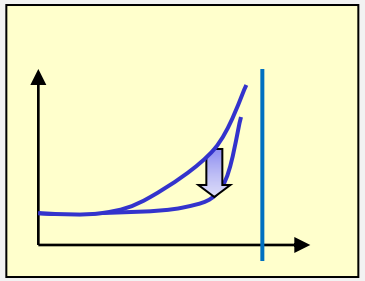
- **Increasing capacity**
- **Reducing variability**

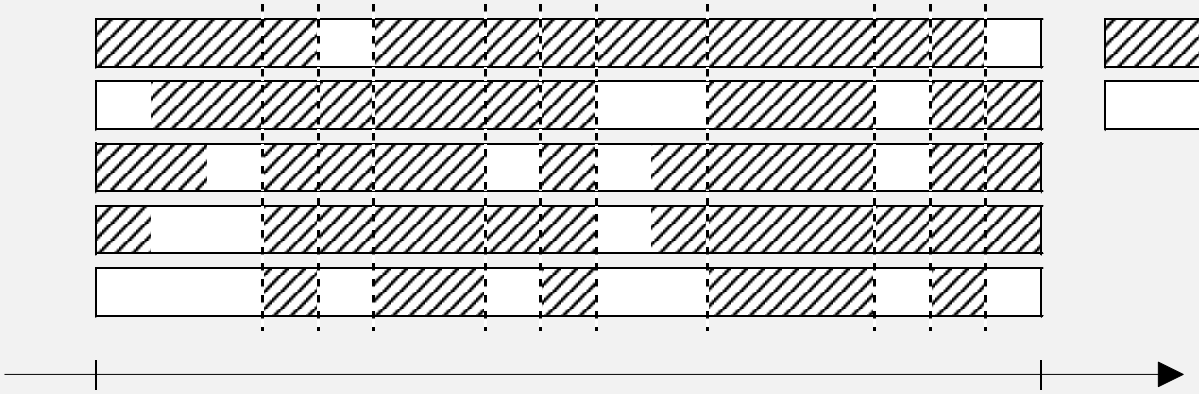
Optimization of Operation Curve

Increase of capacity



Reduction of variability



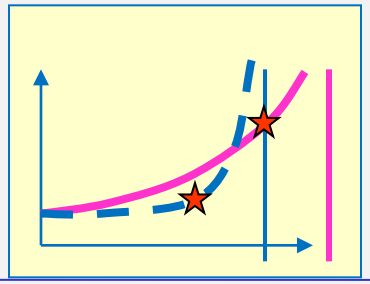


Increase of capacity:
1. increase availability of partners
2. increase the synchronization of partners

Reduction of variability:
1. reduce availability fluctuations of partners
2. reduce fluctuations of synchronisation

Conflict of goals:

In practice, usually changes influence both: capacity and variability.
In these cases investigations have to be done, if the change improves both parameters or if the improvement of one parameter has a corrupted influence on the other parameter (e.g. the addition of an extra tool will increase capacity, but due to additional distortion times the variability will be increased as well).

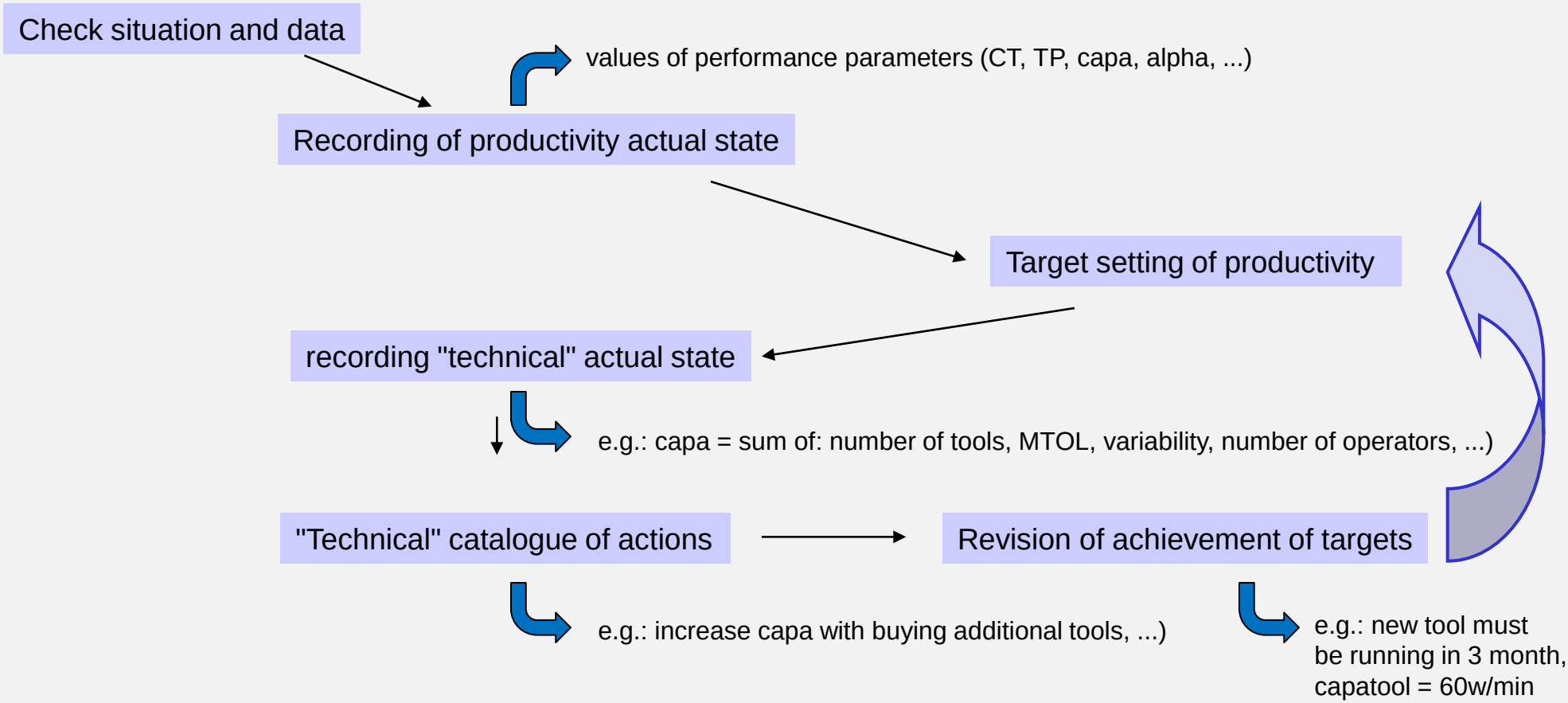


In such a case the old operation curve and the new operation curve may subtend, the best strategy now is utilization-dependent.
But it is always true:

The working point of the respective lower operation curve is always better

Optimization of Operation Curve

Working Plan for Optimization



Working Plan for Optimization

Attention:

Usually the data landscape for measurement of productivity is grown historically: various people may have measured "equivalent data" under different points of view with different measurement tools or different evaluation or the same procedure is done by automation, computer control, data base storage, data analysis.

First step is to prove the definition and consistency of the available data seriously

1. Check actual status:

Raw cycle time RCT ->

* usually from equipment manufacturer specified process speed (throughput)
* but may be measured by login-logout at the tool

momentary cycle time ->

* may be measured at the end of production line from time between input and output of line. For products with long working plans (e.g. semiconductor fabs) only the cycle time of past is delivered.
* Momentary CTs may be calculated by summing up all individual cycle times at all tools

.....

1 Increase of capacity

Capacity is maximum throughput

The achievable throughput may be calculated due to already realized synchronization:

$$24h \bullet process\ speed\left[\frac{wafer}{time}\right] \cdot A_{Op} \cdot A_{WIP} \cdot A_{Proc} \cdot A_{tool} = daily\ going\ rate[units]$$

Daily Going Rate DGR

- A_{Op}, A_{WIP}, A_{Proc}: usually fixed
- A_{tool}: availability of the partners in period of investigation
- $daily\ going\ rate[units]$: **Output** (Ausstoß) [units] in period under observation

$$Pr\,oductive\ time \bullet process\ speed\left[\frac{wafer}{time}\right] \cdot A_{Op} \cdot A_{WIP} \cdot A_{Proc} \cdot A_{tool} = going\ rate[units]$$

Increase of productive time

holydays/celebration days ?
weekend ? -> cultural background
new shift model
flexible working time model

Increase of tool speed

Improvement of tools
Shortening of working procedures

Increase of availability
(especially of bottleneck partner)

Man: stand-by personal
Machine: increase of up-time
Material: better logistics
Method: techn./organ. improvements

for example

Increase of number of one partner
(especially of bottleneck partner)

Employment of people, qualification of people (-> stand-by man)
Buying additional equipment, reduction of dedication, using free capacities in under-worked Fabs

2

Reducing variability

The variability accounts, as before the availability, all partners:

$$\alpha = \left(\frac{c_a^2 + c_e^2}{2} \right)$$

c_a : Variability of inter-arrival times of wip
c_e : Variability from man, machine, method

$$c_e^2 = c_{man}^2 + c_{machine}^2 + c_{method}^2$$

Fluctuation reduction of material

capacity-suited input
homogenous input
reduction of product mix
reduction of c_e

Fluctuation reduction of man

technical actions:
controlled and sufficient recreation
reduction of manual working
-> avoiding sickness times
additional actions:

Fluctuation reduction of maschine

reduction of unexpected downs by scheduled maintenance
adjustment of maintenance strategies

Fluctuation reduction of method

consistent data collection
cultivated data bases

Motivation Willingness

bonus money
carreer perspectives
personel working atmosphere
team spirit
corporate culture

Qualification

trainees
mentor models

Organisation

clearly defined responsibilities
clear information flow
defined decision pathes

selection of examples

Example for Improvement of Operation Curve

Checking actual status:

RCT (due to equipment plan and production plan): 12 days
measured mean cycle time CT: 60 days
measured mean throughput: 4000 wafer/day
measured mean utilisation: 66%

Target setting (to be reached after 1 year)

reduction raw cycle time RCT: 10 days
"cycle time half" : 30 days
twice throughput: 8500 wafer/day
increase utilisation: 76 %

Calculation of missing performance parameters:

Flow Factor FF

$$dynFF_p = \frac{dynCT_p}{RCT_p} = \frac{60\ days}{12\ days} = 5$$

$$dynFF_p = \frac{dynCT_p}{RCT_p} = \frac{30\ days}{10\ days} = 3$$

$$Capa = \frac{throughput}{utilisation} = \frac{4000\ wafer/day}{0{,}66} = 6061\ wafer/day$$

$$Capa = \frac{8500\ wafer/day}{0{,}76} = 11184\ wafer/day$$

$$\alpha_{line} = \frac{(dynFF_{line} - 1)(1-U)}{U} = \frac{(5-1)\cdot(1-0.66)}{0.66} = 2$$

$$\alpha_{line} = \frac{(3-1)\cdot(1-0.76)}{0.76} = 0{,}6$$

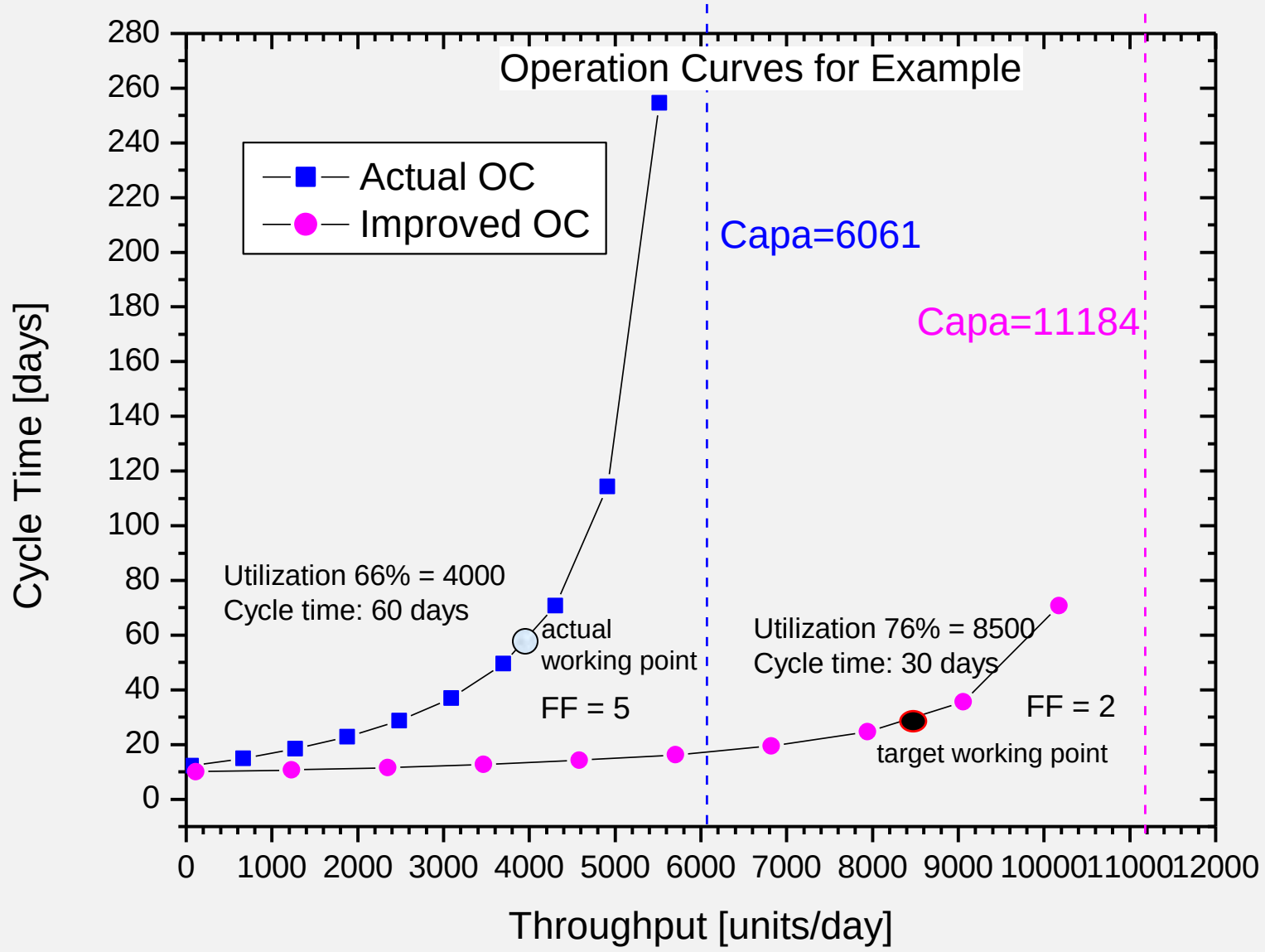
If the target setting is: al parameters must be better, then the long-term parameters capa, α, RCT must be improved

Action catalogue:

short-term: increase of partner availabilities, additional tools
long-term: improvement in logistics, increase of synchronisation

} as we have seen last page

Example for Improvement of Operation Curve



End of Chapter 7

Now we have a good estimation,
what has to be taken care of and to be done
to evaluate our factory performance and increase productivity