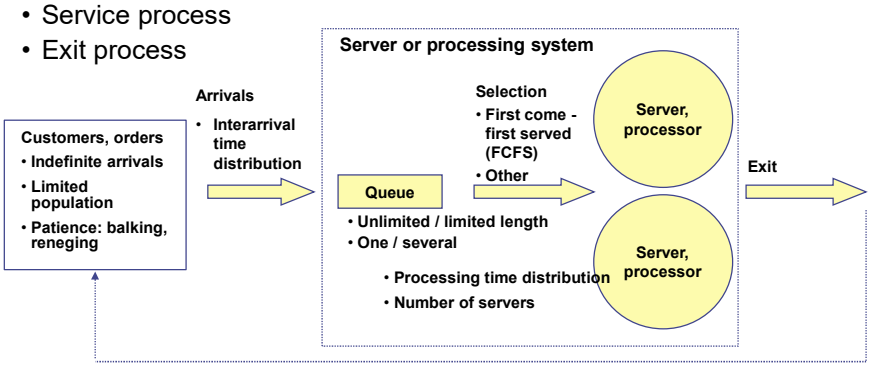
Queuing theory

Main components of a queuing system are:

- Customers
- Arrival process
- Queue
- Queuing discipline
- Service process
- Exit process



esko.niemi@aalto.fi

1

Queuing theory

Queuing theory mathematically examines queuing phenomena. Formulas are developed for various queuing processes and their arrival and processing time distributions. These processes can be further combined to form a network. The formulas can be used to calculate queue length and for example:

- Average number of customers (workpieces) in the process
- Average queuing time
- Average time in the system
- Probability for the system to be empty
- Probability for queuing
- Probability for *n* workpieces in the system

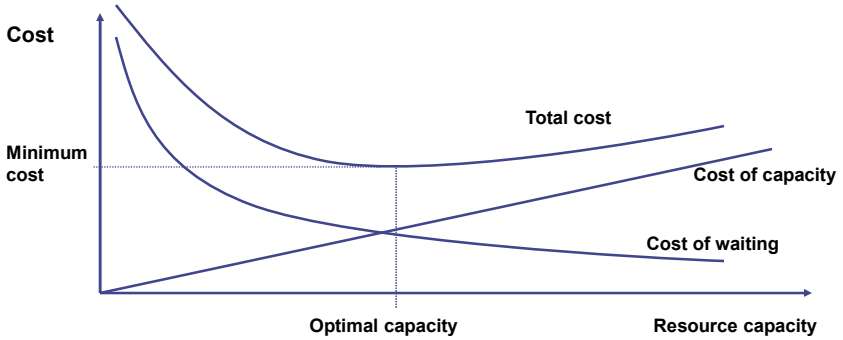
Results are statistical quantities for a stable system. Queuing models can not be used for testing complicated control rules or monitor a single object in the system. Queuing network modelling can be used to model real systems, but its main value in manufacturing is in teaching, because models are simple and they are analytically derived and proven.

esko.niemi@aalto.fi

2

Queuing theory

Final objective is often to reach balance or minimum of service cost and cost of waiting



esko.niemi@aalto.fi

3

Queuing theory

Queuing process is often described with the following notation:

Interarrival time distribution / Service time distribution / Number of parallel servers / Maximum length of queue / Maximum number of customers

where

M = exponential distribution
D = Deterministic (constant) times
G = General distribution (standard deviation given)
E_k = Erlang (Gamma) distribution

esko.niemi@aalto.fi

4

Queuing theory

For example, *M/M/s* denotes a process with exponential interarrival times and service times and that there are *s* parallel servers

Default assumptions:

- Queue length or number of customers are not limited
- Interarrival times and service times are independent of other such events and of each other
- FCFS (First Come First Served) queuing discipline is followed
- The system is stable, i.e. results are static quantities
- Utilization rate of the system is < 100 %

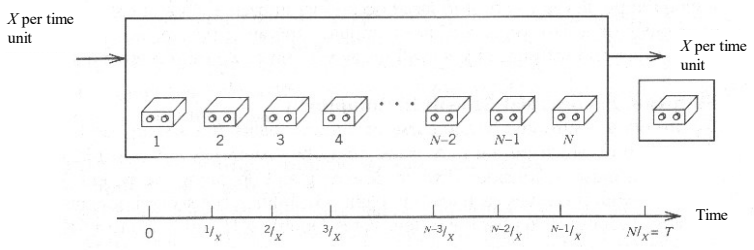
esko.niemi@aalto.fi

5

Little's law[1]

For any production system, machine, workshop etc. in the long run:

*Work in process (N) = Production rate (X) * Throughput time (T)*
$L = \lambda W$



Throughput time (T) = *N* steps * 1/*X* time units per step. That is $N = X\,T$

In other words: "On average *T* days production should be found in the factory"

[1] Little, J.D.C. (1961) "A Proof for the Queuing Formula: $L = \lambda W$", Operations Research, 9(3).

esko.niemi@aalto.fi

6

Little's law

The unit used to measure work in progress (WIP) can be piece, kg, euro etc., as long as the same unit is used. Production rate is correspondingly the same unit per time unit, which must be the same as used for throughput time (min, h, day, year...).

The condition of stability excludes a situation, in which production has started, but no complete products have been finished. Throughput time remains undetermined in such a situation. In the long run the system saturates, and the averages are valid.

Example 1: A machine is processing 12 products per hour and there are on average 20 workpieces waiting on the floor on front of the machine. Average waiting time is = 20 pcs / 12 pcs/h = 1.67 h = 100 min.

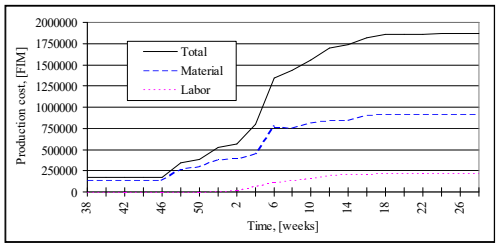
Example 2: A factory produces 2000 products a year and the manufacturing takes about 5 weeks. There should be an average of 2000 pcs/year * 5/52 year = 192 incomplete or complete poducts in the factory.

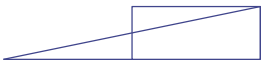
esko.niemi@aalto.fi

7

Little's law

Example 3: The total production cost of an electrical motor factory is 100 MEUR a year, and the value of WIP is 35 MEUR. Costs accumulate over time during production as the figure below shows.. What is the average throughput time?





If all costs would be generated immediately when the production starts, throughput time would be 35/100 = 0.35 years = 18 weeks. If we approximate that the cost accumulation is linear, we can conclude that the throughput time is about 36 weeks in reality.

esko.niemi@aalto.fi

8

Practical implications of Little's law

According to the rule, WIP is in direct relation to throughput time. Therefore:

- If throughput time can be reduced by developing the production process, WIP decreases.
- If WIP is reduced by decreasing order intake, it will reduce as well as output. Actually according to queuing theory, the throughput time decreases, but this is due to lower utilization rate and reduced waiting, but not Little's law.
- Correspondingly, when production rate increases due to improved order intake, WIP increases and cash flow may turn negative for a while. When production approaches full capacity, throughput time (and WIP) may increase sharply due to increased waiting in the process. Under no circumstances can throughput time be decreased by allowing more production on the floor.

esko.niemi@aalto.fi

9

Queuing theory

For queuing systems:

$L = \lambda W$ (WIP)

$L_q = \lambda W_q$ (Queue length)

$L_p = \lambda * 1/\mu = \lambda/\mu$ (In process, machine, being served)

$W = W_q + 1/\mu$ (Throughput time)

$L = L_q + L_p = L_q + \lambda/\mu$ (WIP),

where

λ = Arrival rate (note: $1/\lambda$ is interarrival time)

μ = Processing rate (products per time unit, $1/\mu$ is processing time)

L = Products in system (= WIP, Work in Progress)

W = Throughput time

L_q = Queue length

W_q = Waiting time in queue

esko.niemi@aalto.fi

10

Queuing theory

If queue length or waiting time is known, all other interesting values can be calculated. For given utilization rate, queue length L_q depends mainly on interarrival time and processing time distributions.

For M/M/1

$L_q = \lambda^2 / \mu(\mu - \lambda) = \rho^2/(1 - \rho)$, $L = \rho^2/(1 - \rho) + \rho = \rho/(1 - \rho)$

where

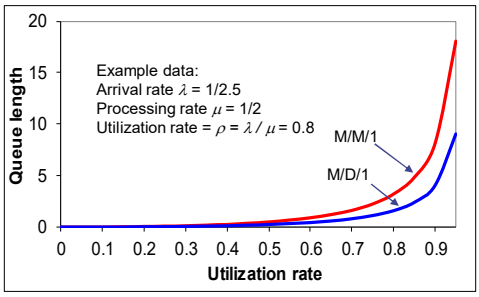
$\rho = \lambda / \mu = L_p$, utilization rate

For example $1/\lambda = 2.5$ ja $1/\mu = 2$:

$L_q = 0.8^2/(1-0.8) = 3.2$

$L = L_q + \lambda/\mu = 3.2 + 0.4/0.5 = 4$

$W = L / \lambda = 4 / (1 / 2.5) = 10$



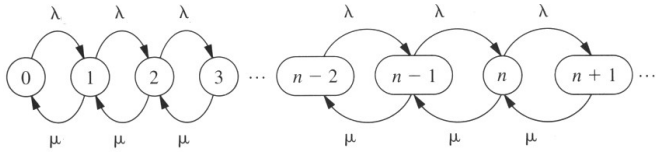
esko.niemi@aalto.fi

11

M/M/1 – queue length

For a Poisson process (like the one described with the negative exponential distribution), the probability of events taking place is the same independent of previous events in the system. State diagram:

State (number of customers, orders in system):



Balance equations (P_n = probability for the system to be in state n):

	In = Out
0	$\mu P_1 = \lambda P_0$
1	$\lambda P_0 + \mu P_2 = (\lambda + \mu) P_1$
2	$\lambda P_1 + \mu P_3 = (\lambda + \mu) P_2$
⋮	⋮
$n - 1$	$\lambda P_{n-2} + \mu P_n = (\lambda + \mu) P_{n-1}$

As sum of P_n must be 1, a numerical solution for this system is possible

esko.niemi@aalto.fi

12

Closed form solution

P_n can be solved for P_0:

State:

$$0: \quad P_1 = \frac{\lambda}{\mu} P_0$$

$$1: \quad P_2 = \frac{\lambda}{\mu} P_1 + \frac{1}{\mu}(\mu P_1 - \lambda P_0) = \frac{\lambda}{\mu} P_1 = \frac{\lambda^2}{\mu^2} P_0$$

$$2: \quad P_3 = \frac{\lambda}{\mu} P_2 + \frac{1}{\mu}(\mu P_2 - \lambda P_1) = \frac{\lambda}{\mu} P_2 = \frac{\lambda^3}{\mu^3} P_0$$

$$\vdots \qquad \vdots$$

$$n-1: \quad P_n = \left(\frac{\lambda}{\mu}\right)^n P_0 = \rho^n P_0, \qquad n = 0, 1, 2, \ldots,$$

In addition, we know that the sum of sate probabilities must be 1:

$$\sum_{n=0}^{\infty} P_n = 1$$

Now P_0 can be solved:

$$P_0 = \left(\sum_{n=0}^{\infty} \rho^n\right)^{-1} = \left(\frac{1}{1-\rho}\right)^{-1} = 1 - \rho.$$

For geometric series: $\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}, \quad \text{if } |x| < 1.$

And further $P_n = (1 - \rho)\rho^n, \qquad n = 0, 1, 2, \ldots.$

esko.niemi@aalto.fi

13

Now L can be solved for ρ, because there must be customers in the system on average:

$$L = \sum_{n=0}^{\infty} nP_n.$$

$$= \sum_{n=0}^{\infty} n(1-\rho)\rho^n$$

$$= (1-\rho)\rho \sum_{n=0}^{\infty} \frac{d}{d\rho}(\rho^n)$$

$$= (1-\rho)\rho \frac{d}{d\rho}\left(\sum_{n=0}^{\infty} \rho^n\right)$$

$$= (1-\rho)\rho \frac{d}{d\rho}\left(\frac{1}{1-\rho}\right)$$

$$= \frac{\rho}{1-\rho} = \frac{\lambda}{\mu - \lambda}.$$

And finally, the queue length:

$L_q = L - \rho = \rho/(1 - \rho) - \rho = (\rho - \rho + \rho^2)/(1 - \rho) = \rho^2/(1 - \rho).$

esko.niemi@aalto.fi

14

Queuing models

For manufacturing systems, some readily available queuing models are:

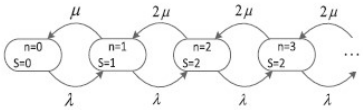
- *M/M/s*
- *M/M/s/-/N*
- *(M/M/s/K)*
- *M/M/s* and state dependent μ
- *M/G/*1
- *M/D/s*
- *M/E$_k$/s*
- Models for other than *M*-arrivals
- Models for other than FCFS priorities

esko.niemi@aalto.fi

15

M/M/2 system

State diagram:



Here:
λ = arrival rate for the whole system
μ = processing rate for one server (machine)
n* = *state
***N* = number of servers**
S = units currently processed

Numerical solution:

		λ	μ	k
		0.4	0.25	2
n	p_n	$p_n\lambda$		
0	0.111845	0.044738	0.044738	$p_{n+1}\mu$
1	0.178952	0.071581	0.071581	$p_{n+1}2\mu$
2	0.143161	0.057264	0.057264	$p_{n+1}2\mu$
3	0.114529	0.045812	0.045812	$p_{n+1}2\mu$
4	0.091623	0.036649	0.036649	$p_{n+1}2\mu$

Balance equations can also be written for flows between nodes; "cut partition method"

$$\lambda p_0 = \mu p_1$$

$$\lambda p_1 = 2\mu p_2$$

$$\vdots$$

$$\lambda p_{N-1} = N\mu p_N$$

$$\lambda p_N = N\mu p_{N+1}$$

$$\lambda p_{N+1} = N\mu p_{N+2}$$

$$\vdots$$

$$\sum_{n=0}^{\infty} p_n = 1.$$

esko.niemi@aalto.fi

16

M/M/s formulas

Here:
λ = arrival rate for the whole system
μ = processing rate for one server (machine)
k = number of servers

Probability for empty system:

$$P_0 = \frac{1}{\sum_{n=0}^{k-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^k}{k!}\left(\frac{k\mu}{k\mu-\lambda}\right)}$$

Queue length:

$$L_q = \frac{(\lambda/\mu)^k \lambda\mu}{(k-1)!(k\mu-\lambda)^2} P_0$$

Waiting time in queue:

$$W_q = \frac{L_q}{\lambda}$$

Probability of n customers (parts, orders) in system:

$$P_n = \frac{(\lambda/\mu)^n}{n!} P_0 \quad \text{for } n \le k$$

$$P_n = \frac{(\lambda/\mu)^n}{k!k^{(n-k)}} P_0 \quad \text{for } n > k$$

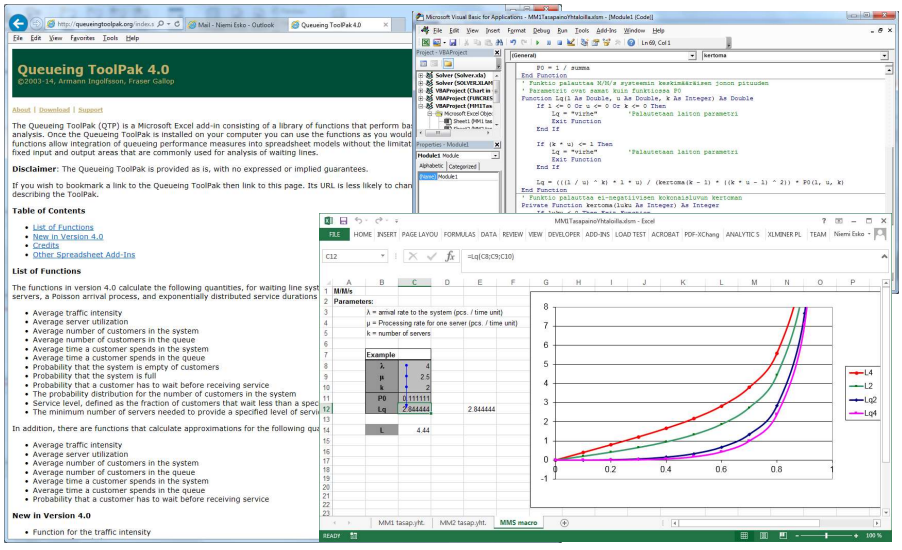
Probability for waiting:

$$P_w = \frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k \left(\frac{k\mu}{k\mu-\lambda}\right) P_0$$

esko.niemi@aalto.fi

17

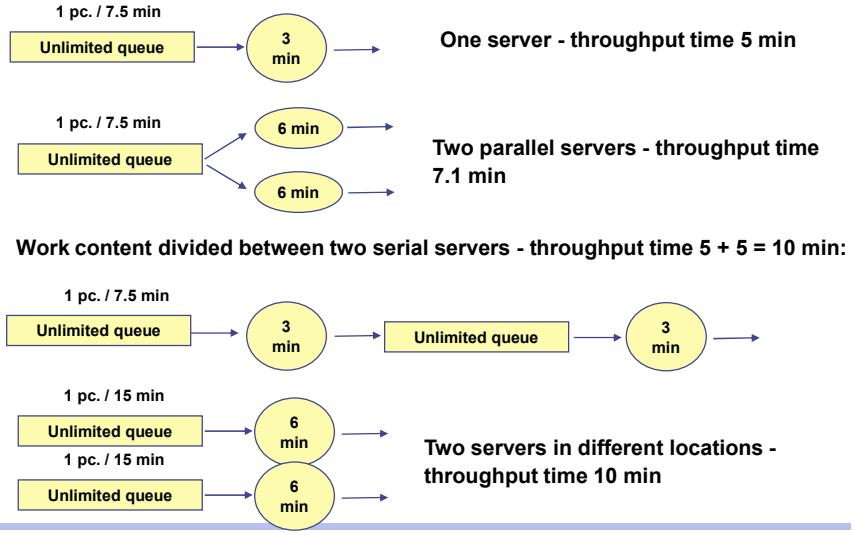
M/M/2 and M/M/4 with Excel functions



esko.niemi@aalto.fi

19

M/M/s server systems with equal capacity



esko.niemi@aalto.fi

20

Cost model for queuing

Final objective is to reach balance or minimum of service cost and cost of waiting

Minimize E(TC) = E(service cost) + E(waiting cost)

Cost of service is typically machine hour, labor hour etc. and easy to determine. Waiting time can be estimated using a suitable formula. Cost of waiting time may be easy to calculated, if the customer is a machine or person the cost of which is the opportunity cost of lost sales. If the customer is external to the service organisation, estimation of costs may be difficult. For example, badwill caused by late deliveries may result in lost sales, but you may not know that you were blacklisted as a supplier and why this happened.

esko.niemi@aalto.fi

21

Cost of waiting – example 1

Tool storage:

Workers fetch tools from a central tool storage 30 times an hour on average. A person in the tool storage can serve 20 customers per hour. Cost of lost production due to workers waiting for service is 48 48 euro/h. How many service persons should be employed, if one costs 20 euro/h?

We apply *M/M/s* model. Workers waiting or being served:

1 storage person: not sufficient; 2 storage persons: 3.43 kpl; 3 storage persons: 1.74 kpl; 4 storage persons: 1.54 kpl; 5 storage persons: 1.51 kpl

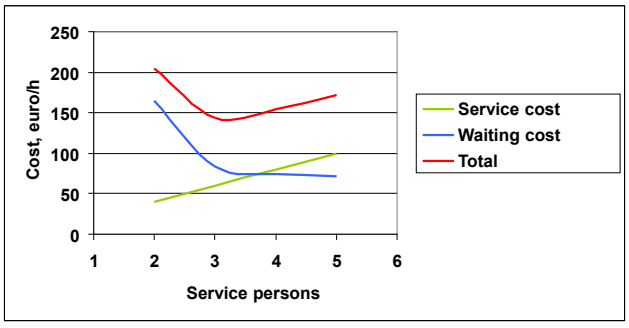
esko.niemi@aalto.fi

22

Cost of waiting – example 1

Costs:

2 storage persons: TC = 20 x 2 + 3.4 x 48 = 205 euro/h
3 storage persons: TC = 20 x 3 + 1.74 x 48 = 144 euro/h
4 storage persons: TC = 20 x 4 + 1.54 x 48 = 154 euro/h
5 storage persons: TC = 20 x 5 + 1.51 x 48 = 172 euro/h



esko.niemi@aalto.fi

23

M/M/1/-/N model formulas

Here:
λ = arrival rate for the whole system
μ = processing rate for one server (machine)
***N* = number of customers**

Probability for empty system:

$$P_0 = \frac{1}{\sum_{n=0}^{N} \frac{N!}{(N-n)!}\left(\frac{\lambda}{\mu}\right)^n}$$

Queue length:

$$L_q = N - \frac{\lambda + \mu}{\lambda}(1 - P_0)$$

WIP:

$$L = L_q + (1 - P_0)$$

Waiting time in queue:

$$W_q = \frac{L_q}{(N - L)\lambda}$$

Probability of *n* customers (parts, orders) in system:

$$P_n = \frac{N!}{(N-n)!}\left(\frac{\lambda}{\mu}\right)^n P_0 \qquad n = 0, 1, \ldots, N$$

esko.niemi@aalto.fi

24

Cost of waiting – example 2

Machine tools break down:

Our machine shop has 10 machine tools, which break down every 10 hours on average. Cost of machine not in production is 48 euro/h. Repairing a machine takes one hour on average for one repair person. We allocate all repair persons to machines according to FIFO priority. How many maintenance persons do we employ, if one costs 20 euro/h?

We assume that Time Between Failures and Time To Repair are exponentially distributed. We apply *M/M/1/-/10* model. We have machines down as follows:

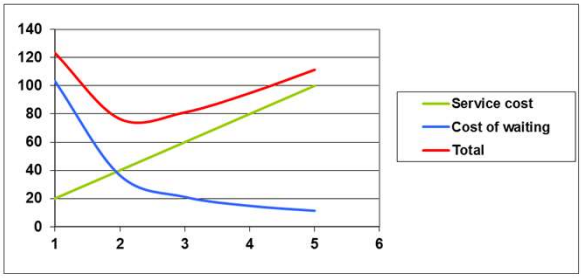
1 repair person: 2.15 machines; 2 repair persons: 0.759 machines; 3 repair persons: 0.44 machines; 4 repair persons: 0.308 machines; 5 repair persons: 0.236 machines

esko.niemi@aalto.fi

25

Cost of waiting – example 2

Costs:
1 repair person: TC = 20 x 1 + 2.15 x 48 = 123.2 euro/h
2 repair persons: TC = 20 x 2 + 0.759 x 48 = 76.4 euro/h
3 repair persons: TC = 20 x 3 + 0.440 x 48 = 81.1 euro/h
4 repair persons: TC = 20 x 4 + 0.308 x 48 = 94.8 euro/h
5 repair persons: TC = 20 x 5 + 0.236 x 48 = 111.3 euro/h



esko.niemi@aalto.fi

26

Waiting and batch sizes – an example

A turning facility uses an automatic bar lathe to manufacture 48 000 hydraulic connectors a year, 40 different types, 3 minutes cutting time / pc. and the set-up time is 100 min. The production is made to order so, that an optimal batch is manufactured each time the stock level is lower than the ordered amount. The demand is uniformly distributed between the different connector types. Material cost is 1 euro/connector and machining cost is 0.5 euro/min in two shifts. Set-up cost is 1 euro/min because it requires the operator in addition to the machine.

We assume that *M*/*M*/1 can be applied and examine the effect of batch size on throughput time and cost.

esko.niemi@aalto.fi

27

Automatic bar turning machines

esko.niemi@aalto.fi

28

Waiting and batch sizes – an example

BatchesAndWaiting.xls [Compatibility Mode] - Excel

FILE | HOME | INSERT | PAGE LAY | FORMULA | DATA | REVIEW | VIEW | DEVELOPE | ADD-INS | LOAD TES | ACROBAT | PDF-XCha | ANALYTIC | XLMINER | TEAM | Niemi Esko

C9 | =B9

	A	B	C	D	E	F	G	H	I	J	K
1	**Throughput time study**										
2	Annual working hours, 2-shift	198000	198000	198000	198000	198000	198000	198000	198000	198000	1980
3	Batch size	100	150	200	250	300	350	400	450	500	5
4	Set-up time	100	100	100	100	100	100	100	100	100	1
5	Processing time	3	3	3	3	3	3	3	3	3	
6	Processing cost	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2
7	Batch time	400	550	700	850	1000	1150	1300	1450	1600	17
8	Batch time, days	0.4	0.6	0.7	0.9	1.0	1.2	1.4	1.5	1.7	
9	Volume/year	48000	48000	48000	48000	48000	48000	48000	48000	48000	480
10	Batch interarrival time	413	619	825	1031	1238	1444	1650	1856	2063	22
11	Utilization rate	0.970	0.889	0.848	0.824	0.808	0.797	0.788	0.781	0.776	0.7
12	Lq	31.03	7.11	4.75	3.87	3.40	3.12	2.93	2.79	2.68	2
13	Lamda	0.0024	0.0016	0.0012	0.0010	0.0008	0.0007	0.0006	0.0005	0.0005	0.00
14	Wq	12800	4400	3920	3986	4211	4502	4829	5175	5535	59
15	Waiting time in queue	13.3	4.6	4.1	4.2	4.4	4.7	5.0	5.4	5.8	6
16	W	13200	4950	4620	4836	5211	5652	6129	6625	7135	76
17	Throughput time, days	13.8	5.2	4.8	5.0	5.4	5.9	6.4	6.9	7.4	8
18											
19	Set-up cost	100	100	100	100	100	100	100	100	100	1
20	Annual volume/type	1200	1200	1200	1200	1200	1200	1200	1200	1200	12
21	Holding cost-%	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.
22	Set-up cost/pc.	1.00	0.67	0.50	0.40	0.33	0.29	0.25	0.22	0.20	0.
23	Holding cost	0.04	0.05	0.06	0.08	0.09	0.10	0.11	0.13	0.14	0.
24	Cost/piece	3.54	3.22	3.06	2.98	2.92	2.89	2.86	2.85	2.84	2.
25											
26											

Throughput time

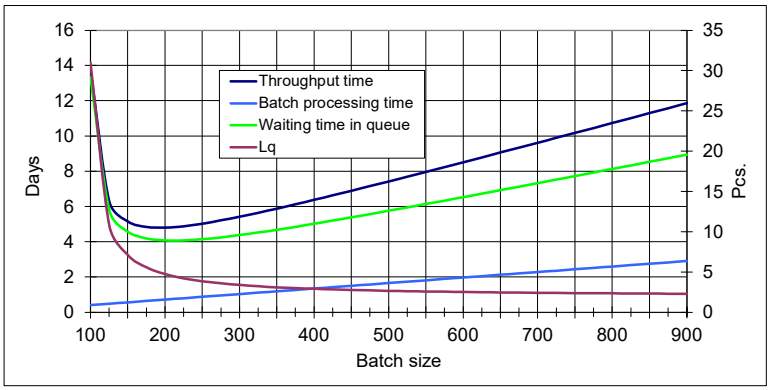
READY | CALCULATE | 100 %

esko.niemi@aalto.fi

29

Waiting and batch sizes – throughput times

When batch size is increased, utilization rate decreases, because the number of set-ups reduces. This shortens queues when measured in number of batches, but the larger batches increase waiting time.



esko.niemi@aalto.fi

30

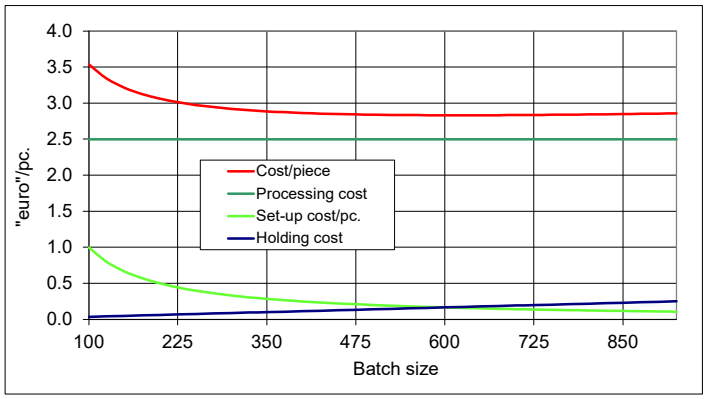
Waiting and batch sizes – cost

Economic Order Quantity can be calculated using the classical formula $Q_{eoq} = \sqrt{2DS/H}$,

which gives the minimum total cost batch size: $C_{tot} = DC + (D/Q)S + (Q/2)H$

Here D = Demand, C = Cost/piece, Q = Batch size, S = Cost/batch (set-up), H = Holding cost

This is mainly intended for production or purchasing to stock, and it does not consider throughput time or limited capacity

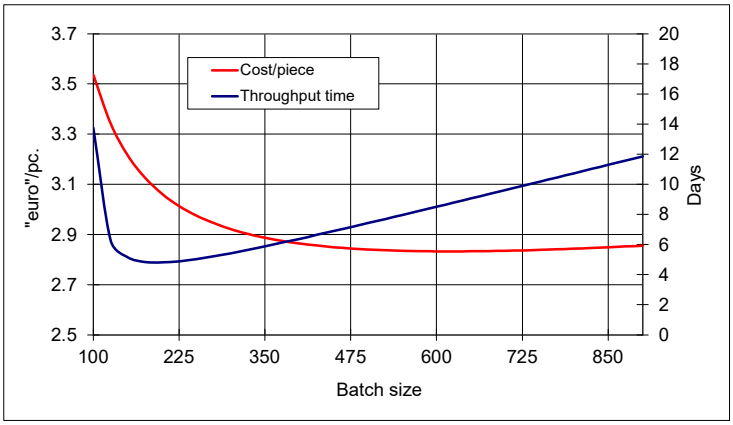


esko.niemi@aalto.fi

31

Waiting and batch sizes

Typically, *EOQ* formula gives such a large ”optimal” batch size, that the corresponding throughput times are too long for MTO production

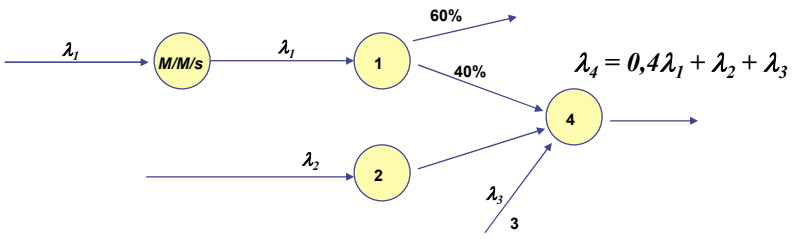


esko.niemi@aalto.fi

32

Queuing networks

If the arrivals are a Poisson process with parameter λ and the μ in for *M/M/s* servers is equal, exit process is also a Poisson process with parameter λ. In these conditions queuing networks can be analysed as a collection of single queuing systems with unlimited queue lengths. The arrival rate from other nodes and from outside has to be determined, for example:

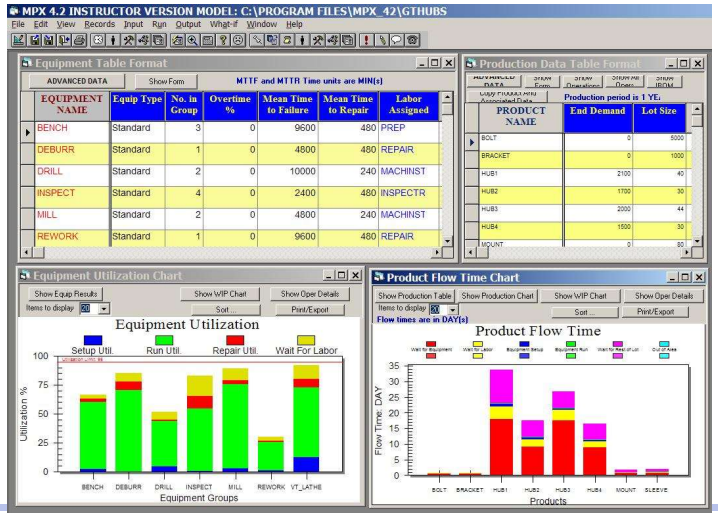


esko.niemi@aalto.fi

33

Queuing networks

A screenshot of MPX queuing network modeling software:



esko.niemi@aalto.fi

34

Summary - throughput time

- In a stochastic process throughput time depends on
  - Utilization rate
  - Variation
  - Processing time (and in case of batch production also of the batch size and set-up time)
- G/G/1 approximation[1]

  Cycle Time = *VUT* + *T*,

  where

  $V = (CV^2_{processing\ time} + CV^2_{interarrival\ time})/2$

  $U = u / (1 - u)$, where u = utilization rate

  T = processing time.
- Because high utilization rate increases *U* strongly, it is beneficial to focus improvement efforts on bottle necks

[1] also known as *Kingman's formula*

esko.niemi@aalto.fi

35

Production control and queuing models

Queuing models provide estimates of averages, variation and probabilities based on assumptions of system and input characteristics. They can be used for rough analyses of real production systems. Practical production is controlled by rough cut capacity and workload matching and prioritizing of orders and tasks in the short run. Under these circumstances the assumptions about applicability of typical probability distributions may be quite valid, especially for arrival processes. For service processes, the main sources of variability are product variation and batch size variation. In manual work, also differences between workers and random human behavior increase variability.

Real (optimized) production control usually requires formulating and solving of binary mathematical programming problems. This can be very hard…

esko.niemi@aalto.fi