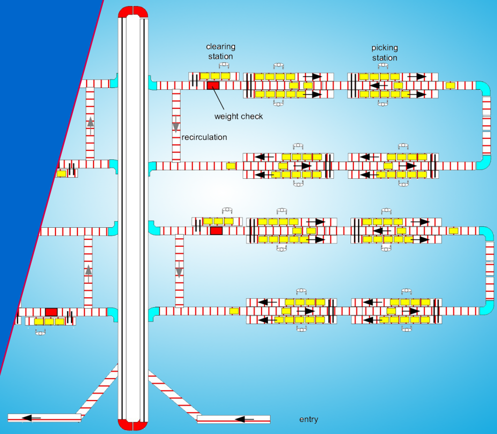
Stochastic Models of Manufacturing Systems

Ivo Adan



Technische Universiteit **Eindhoven** University of Technology

Tuesday June 2

Single server: exponential

- Poisson arrivals with rate λ
- Exponential service times with mean $1/\mu$
- Stability: $\lambda < \mu$ or $\rho = \lambda/\mu < 1$
- Single server and FCFS service

Then:

$$p_k = P(k \text{ jobs in system}) = (1-\rho)\rho^k, \quad k = 0, 1, \ldots$$

and

$$E(W) = \frac{\rho}{1-\rho}\frac{1}{\mu}$$

Single server: general service

- Poisson arrivals with rate λ
- General service times B with distribution $F_B(\cdot)$
- Stability: $\rho = \lambda E(B) < 1$
- Single server and FCFS service

Then:

$$E(W) = \frac{\rho}{1-\rho}\, E(R)$$

where R is residual service time,

$$\begin{aligned} E(R) &= \frac{E(B^2)}{2E(B)} \\ &= \frac{1}{2}\, E(B)\,(1 + c_B^2) \end{aligned}$$

Multi server: exponential

- Poisson arrivals with rate λ
- Exponential service times with mean $1/\mu$
- Stability: $\lambda < c\mu$ or $\rho = \lambda/(c\mu) < 1$
- c parallel servers and FCFS service

Then:

$$E(W) = \frac{\Pi_W}{1-\rho} \frac{1}{c\mu}$$

where Π_W is probability of waiting,

$$\begin{aligned} \Pi_W &= P(W > 0) \\ &= p_c + p_{c+1} + \cdots \\ &= \frac{(c\rho)^c}{c!} \left(\frac{(c\rho)^c}{c!} + (1-\rho) \sum_{n=0}^{c-1} \frac{(c\rho)^n}{n!} \right)^{-1} \end{aligned}$$

Multi server: general service

- Poisson arrivals with rate λ
- General service times B with distribution $F_B(\cdot)$
- Stability: $\rho = \lambda E(B)/c < 1$
- c parallel servers and FCFS service

Then:

$$E(W) \approx \frac{\Pi_W}{1-\rho}\frac{E(R)}{c}$$

where Π_W is probability of waiting in corresponding exponential system, so

$$E(W) \approx \frac{\Pi_W}{1-\rho}\frac{1}{2}\frac{E(B)}{c}(1 + c_B^2)$$

Single server: NP priority

- Poisson arrivals of type i with rate λ_i, $i = 1, \ldots, r$
- General service times B_i for type i, residual service time is R_i
- Stability: $\rho = \sum_{i=1}^{r} \rho_i < 1$ where $\rho_i = \lambda_i E(B_i)$
- Single server
- Non-Preemptive priority service (type 1 highest priority)

Then:

$$E(W_i) = \frac{\sum_{j=1}^{r} \rho_j E(R_j)}{(1 - \rho_{<i})\left(1 - \rho_{\le i}\right)}$$

where

$$\rho_{<i} = \sum_{j=1}^{i-1} \rho_j, \quad \rho_{\le i} = \sum_{j=1}^{i} \rho_j.$$

Single server: SPTF priority

- Poisson arrivals with rate λ
- General service times B with distribution $F_B(\cdot)$ en density $f_B(x)$
- Stability: $\rho = \lambda E(B) < 1$
- Single server
- Shortest Processing Time First service discipline

Then, for waiting time $W(x)$ of a job with service time x:

$$E(W(x)) = \frac{\rho E(R)}{(1 - \rho(x))^2}$$

where $\rho(x)$ is utilization due to jobs with service time $\leq x$,

$$\rho(x) = \int_0^x y \lambda f_B(y) dy$$

Multi server: NP priority

- Poisson arrivals of type i with rate λ_i
- General service times B for all types i with distribution $F_B(\cdot)$
- Stability: $\rho = \sum_{i=1}^{r} \rho_i < 1$ where $\rho_i = \lambda_i E(B)/c$
- c parallel servers
- Non-Preemptive priority service (type 1 highest priority)

Then:

$$E(W_i) \approx \frac{\Pi_W}{(1 - \rho_{<i})\left(1 - \rho_{\leq i}\right)} \frac{E(R)}{c}$$

where Π_W is probability of waiting in corresponding exponential system

Production lines

- Target throughput TH
- Total work content of job W
- Number of machines m

Minimum number of required machines: $m \geq TH \cdot W$

Typically more than minimum required because of, for example,

- unbalance
- variability in processing times
- machine failures

Issues in design and operation of production lines:

- Degree of paralleling of workstations
- Location and size of buffers
- Choice of material handling system
- Allocation of tasks and operators to workstations

Production lines

Synchronous lines:

- Coordinated (simultaneous) movement of jobs
- WIP (number of jobs) is constant
- No buffers needed
- Paced (maximum limit for processing time) or unpaced

Asynchronous lines:

- No coordination of movement of jobs
- WIP fluctuates
- Blocking and starvation
- Buffers needed

Unpaced synchronous lines

- m machines in series
- B_i processing time of machine i with distribution $F_{B_i}(\cdot)$
- C is the cycle time, so $C = \max\{B_1, \ldots, B_m\}$ and thus

$$F_C(t) = P(C \le t) = P(\max\{B_1, \ldots, B_m\} \le t) = F_{B_1}(t) \cdots F_{B_m}(t)$$

Then throughput TH of the line

$$TH = \frac{1}{E(C)}$$

where

$$E(C) = \int_0^\infty t f_C(t) dt = \int_0^\infty (1 - F_C(t))\, dt = \int_0^\infty \left(1 - F_{B_1}(t) \cdots F_{B_m}(t)\right) dt$$

Unpaced synchronous lines

What can variability of processing times do to the throughput?

Examples:

- B_i are uniform on $(0, 1)$, then

$$E(C) = 1 - \frac{1}{m+1}$$

- B_i are exponential with rate 2, then

$$E(C) = \frac{1}{2}\left(\frac{1}{m} + \frac{1}{m-1} + \cdots + \frac{1}{2} + 1\right) \approx \frac{1}{2}\log m$$

Paced synchronous lines

- m machines in series
- B_i processing time of machine i with distribution $F_{B_i}(\cdot)$
- Fixed cycle time c

Then throughput TH of jobs with no defects

$$TH = \frac{Q(c)}{c}$$

where

$$Q(c) = P(B_1 \leq c) \cdots P(B_m \leq c) = F_{B_1}(c) \cdots F_{B_m}(c)$$

So trade-off between volume TH of output and quality $Q(c)$ of output

Paced synchronous lines

Cycle time c^* maximizing TH is the solution of

$$Q'(c) = \frac{Q(c)}{c}$$

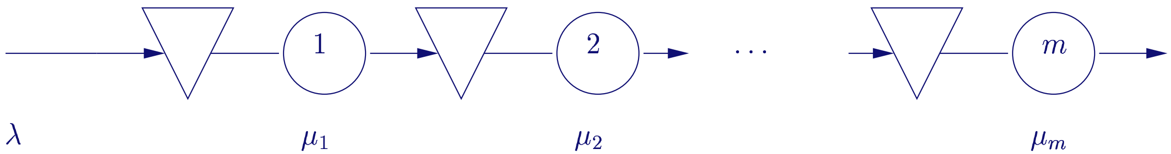
Then cycle time c should be set greater or equal to c^*!

Examples: B_i are exponential with rate 1

m	c	c^*	$Q(c^*)$	TH
5	5.51	2.55	0.66	0.26
10	6.21	3.60	0.76	0.21
20	6.90	4.50	0.80	0.18

c is minimal cycle time to meet $Q = 0.98$

Asynchronous exponential lines



- Machines $1, \ldots, m$
- Jobs arrive according to Poisson process with rate λ
- Processing times at machine i are exponential with rate μ_i
- Buffers have infinite (unlimited) capacity
- Stability: $\rho_i = \frac{\lambda}{\mu_i} < 1$ for all i

Output of $M/M/1$ is again Poisson, so every workstation is $M/M/1$!

$$E(L_i) = \frac{\rho_i}{1-\rho_i}, \quad i = 1, \ldots, m$$

$$E(S) = \frac{E(L)}{\lambda} = \frac{\sum_{i=1}^{m} E(L_i)}{\lambda}$$

Workload allocation

- Machines $1, \ldots, m$
- Jobs arrive according to Poisson process with rate λ
- Mean total work content of jobs is W
- Processing times at machine i are exponential with mean w_i
- Stability: $\rho_i = \lambda w_i < 1$ for all i

Question: How to allocate w_i so as to minimize $E(L)$?

$$\min \sum_{i=1}^{m} \frac{\lambda w_i}{1 - \lambda w_i}$$

subject to

$$\sum_{i=1}^{m} w_i = W,$$

$$0 \leq \lambda w_i < 1, \quad i = 1, 2, \ldots, m.$$

Workload allocation

- Machines $1, \ldots, m$
- Jobs arrive according to Poisson process with rate λ
- Mean total work content of jobs is W
- Processing times at machine i are exponential with mean w_i
- Stability: $\rho_i = \lambda w_i < 1$ for all i

Question: How to allocate w_i so as to minimize $E(L)$?

$$\min \sum_{i=1}^{m} \frac{\lambda w_i}{1 - \lambda w_i}$$

subject to

$$\sum_{i=1}^{m} w_i = W,$$

$$0 \leq \lambda w_i < 1, \quad i = 1, 2, \ldots, m.$$

Solution: $w_i = \frac{W}{m}$ for all i, so balance the line!

Impact of unbalance

- Line with 4 machines, $1, 2, 3, 4$
- Arrival rate $\lambda = 1$
- Mean processing time at machine i is w_i
- Average work load per machine, $\rho = \frac{1}{4}(\rho_1 + \rho_2 + \rho_3 + \rho_4)$

ρ	w_1	w_2	w_3	w_4	$E(L_1)$	$E(L_2)$	$E(L_3)$	$E(L_4)$	$E(L)$
0.80	0.85	0.65	0.90	0.80	5.7	1.9	9.0	4.0	20.5
0.80	0.80	0.80	0.80	0.80	4.0	4.0	4.0	4.0	16.0
0.90	0.95	0.83	0.97	0.85	19.0	4.9	32.3	5.7	61.9
0.90	0.90	0.90	0.90	0.90	9.0	9.0	9.0	9.0	36.0
0.95	0.96	0.93	0.97	0.94	24.0	13.3	32.3	15.7	85.3
0.95	0.95	0.95	0.95	0.95	19.0	19.0	19.0	19.0	76.0

Asynchronous general lines

- Machines $1, \ldots, m$
- Arrival process with general inter-arrival times A, mean $\frac{1}{\lambda}$ and scv c_A^2
- General processing times B_i at machine i with mean $E(B_i)$ and scv $c_{B_i}^2$
- Buffers have infinite (unlimited) capacity
- Stability: $\rho_i = \lambda E(B_i) < 1$ for all i

Arrival process of workstation i is output of workstation $i-1$, so workstation i can be approximated by $G/G/1$ with arrival rate λ (conservation of flow) and

$$c_{A_1}^2 = c_A^2, \quad c_{A_i}^2 \approx (1-\rho_{i-1}^2)c_{A_{i-1}}^2 + \rho_{i-1}^2 c_{B_{i-1}}^2, \quad i = 2, \ldots, m$$

Approximation of the mean flow time at workstation i

$$E(S_i) \approx \frac{\rho_i}{1-\rho_i} \frac{1}{2} E(B_i)(c_{A_i}^2 + c_{B_i}^2) + E(B_i)$$

Optimal ordering

- Balanced line $E(B_1) = \cdots = E(B_m)$
- Any ordering of machine is feasible

Question: What ordering minimizes total flow time $E(S)$?

Optimal ordering

- Balanced line $E(B_1) = \cdots = E(B_m)$
- Any ordering of machine is feasible

Question: What ordering minimizes total flow time $E(S)$?

Answer: Machines with best processing reliability should be first!

Let $\pi_1, \ldots, \pi_m$ be permutation of $1, \ldots, m$

Then optimal ordering should satisfy

$$c^2_{B_{\pi_1}} \leq c^2_{B_{\pi_2}} \leq \cdots \leq c^2_{B_{\pi_m}}$$

Optimal ordering

- Machines $0, 1, 2$
- Poisson inflow with rate λ
- $E(B_i) = 1$ (balanced line)
- $c^2_{B_i} = i$

Machine order	λ	$E(S)$
2 1 0	0.80	12.8
0 1 2	0.80	10
2 1 0	0.90	30
0 1 2	0.90	22.3
2 1 0	0.95	64.8
0 1 2	0.95	47.2

Asynchronous general lines

- Use $G/G/c$ approximation for workstation with c parallel machines
- Use for diverging and converging production lines:
  - Merging of streams: If two streams with rates λ_1 and λ_2 and scv $c^2_{A_1}$ and $c^2_{A_2}$ are merged, then the resulting stream has rate $\lambda_1 + \lambda_2$ and its scv can be approximated by

$$c^2_A \approx \frac{\lambda_1}{\lambda_1 + \lambda_2} c^2_{A_1} + \frac{\lambda_2}{\lambda_1 + \lambda_2} c^2_{A_2}$$

  - Random splitting of stream: If arriving jobs are randomly split with probability p from a stream with rate λ and scv c^2, then the resulting stream has rate $p\lambda$ and

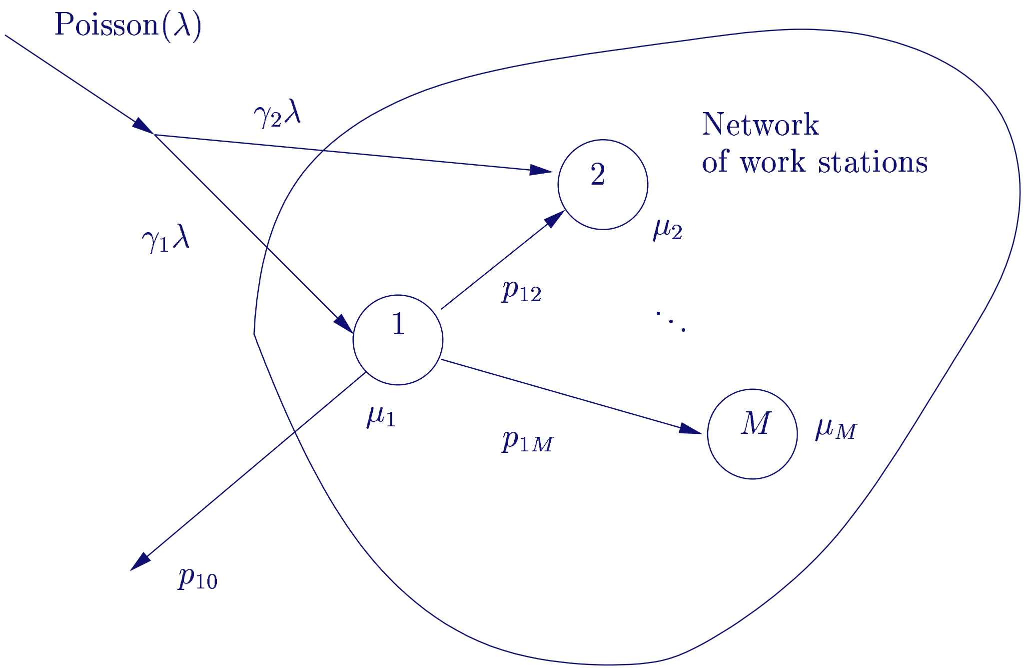
$$c^2_A = pc^2 + 1 - p$$

Exponential open job shops

- Workstations $1, \dots, M$
- Workstation m has c_m parallel identical machines
- Jobs arrive according to Poisson stream with rate λ
- Arriving job joins workstation m with probabilty γ_m
- Processing times in workstation m are exponential with rate μ_m
- Processing order is FCFS
- Buffers are unlimited
- Markovian routing:
  job moves from workstation m to n with probability p_{mn} and leaves system with probability p_{m0}

This network is also called Open Jackson network

Exponential open job shops



Network capacity

Let v_m be average number of visits (of a job) to work station m

$$v_m = \gamma_m + \sum_{n=1}^{M} v_n p_{nm}, \quad m = 1, \ldots, M.$$

Equations have unique solution for $v_1, \ldots, v_M$.

Then $\lambda_m = \lambda v_m$ is total number of visits per time unit to work station m.

Bottleneck station: station with the highest load

$$\max_{1 \le m \le M} \frac{\lambda_m}{c_m \mu_m}$$

Stability: For all m,

$$\rho_m = \frac{\lambda_m}{c_m \mu_m} < 1.$$

Exponential single server network

States of network $\underline{k} = (k_1, \ldots, k_M)$ where k_m is number of jobs in station m

State probabilities $p(k_1, k_2, \ldots, k_M)$ satisfy balance equations ($c_m = 1$)

$$\begin{aligned}
\text{Flow out of } \underline{k} \quad &= \quad \text{Flow into } \underline{k} \\
p(\underline{k})\left(\lambda + \sum_{m=1}^{M} \mu_m \epsilon(k_m)\right) &= \sum_{m=1}^{M} p(\underline{k} + \underline{e}_m)\mu_m p_{m0} \\
&+ \sum_{n=1}^{M}\sum_{m=1}^{M} p(\underline{k} + \underline{e}_n - \underline{e}_m)\mu_n p_{nm}\epsilon(k_m) \\
&+ \sum_{m=1}^{M} p(\underline{k} - \underline{e}_m)\lambda\gamma_m\epsilon(k_m).
\end{aligned}$$

where $\underline{e}_m = (0, \ldots, 1, \ldots, 0)$ with 1 at place m and $\epsilon(k) = \begin{cases} 1 & \text{if } k > 0 \\ 0 & \text{else} \end{cases}$

Exponential single server network

Product form solution Jackson's miracle

$$p(\underline{k}) = p_1(k_1)p_2(k_2)\cdots p_M(k_M),$$

where

$$p_m(k_m) = (1-\rho_m)\rho_m^{k_m}, \quad k_m = 0, 1, \ldots$$

and

$$\rho_m = \frac{\lambda_m}{\mu_m}$$

with $\lambda_m = v_m \lambda$ total arrival rate to workstation m

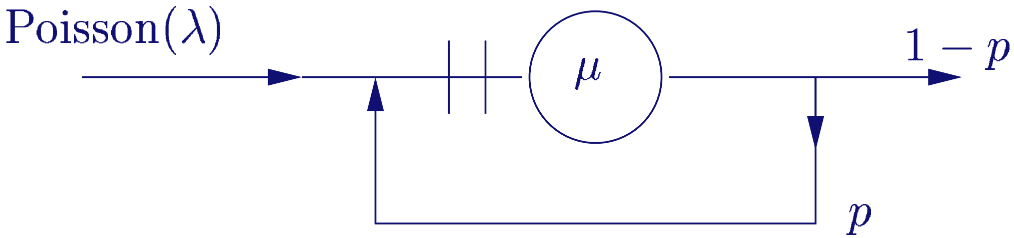
This is just the product of $M/M/1$ solutions!

Exponential single server network

Surprising result:

- Marginal distribution $p_m(\cdot)$ is exactly the same as distribution of $M/M/1$ with arrival rate λ_m and service rate μ_m
- Inflow to workstation m is in general not Poisson!
- Queue lengths at workstations are independent (if you take a snapshot)!

Example: $\mu = 1/\epsilon$, $p = 1 - \epsilon$ (so $\mu(1-p) = 1$), $\lambda \ll 1$



Arrival pattern:  (So not Poisson at all!)

Now you can guess the solution for an exponential multi-server network