Lecture 3: Introduction to Queuing Theory

PAMS'18

Zsolt István

zsolt.istvan@imdea.org

Slide contents heavily influenced by G. Alonso's Advanced Systems Lab lecture slides.

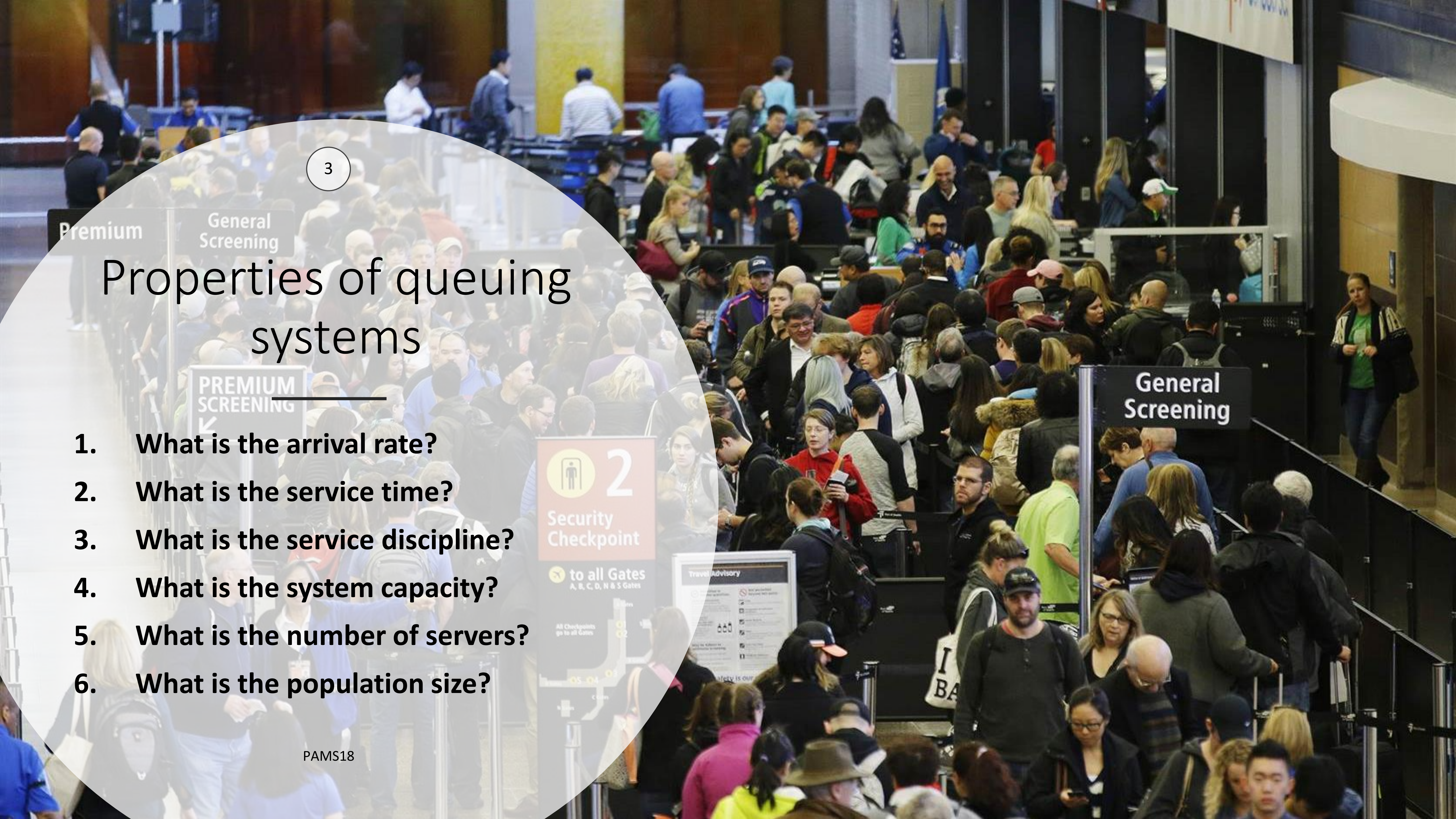
What is a queuing system?

- Jobs arriving to a queue
- One/multiple servers dealing with the jobs from the queue

- When modeling queuing systems, it is important to talk about their properties
  - Some rules apply to all systems

Properties of queuing systems

1. **What is the arrival rate?**
2. **What is the service time?**
3. **What is the service discipline?**
4. **What is the system capacity?**
5. **What is the number of servers?**
6. **What is the population size?**



Interarrival time

- $\tau_{1..N}$: Independent and Identically Distributed random variables

- Events come from a “process”, Often assumed to be Poisson:
  - (1) the event is something that can be counted in whole numbers
  - (2) occurrences are independent, so that one occurrence neither diminishes nor increases the chance of another
  - (3) the average frequency of occurrence for the time period in question is known
  - (4) it is possible to count how many events have occurred, but we are not interested in how many have not occurred

Mean Arrival rate

- Interarrival time **τ** is a random variable
  - Mean value: $E[\tau]$
- Mean arrival rate: **$\lambda = 1/E[\tau]$**
  - Does it look similar to something we discussed previously?
- Useful to assume fixed λ for modeling
  - Do real systems have fixed λ?
- Examples:
  - A new pizza order is received on average every 3 minutes. The arrival rate is 20/hour.
  - The printer receives a new job to print on average every 100ms, the arrival rate is 10jobs/second.

Service time

- Time to process a job ("useful work", no queueing)
  - Random variable: ***s***
- Mean service rate **μ = 1 / E[s]**
  - What if we have *m* servers?
  - Not a random variable
- Example: pizza oven bakes pizza on average in 6 minutes. μ = 10/hour

A word on throughput

- The service rate of a system is rarely the measured throughput!
    - Throughput is client and workload dependent
    - Throughput only counts successful operations

- Is arrival rate the same as throughput?
    - In open systems?
    - In closed systems?
        - If there are no failed jobs?

Default assumptions for other properties

- What is the service discipline?
  - First come, first served (FIFO)

- What is the system capacity?
  - Large enough buffers → Infinite buffers

- What is the population size?
  - Very large → Infinite size

Equations valid for all queueing systems

- Load on system (traffic intensity): $\rho = \lambda/(m\mu)$
- Stability condition: $\rho < 1$ because this meant that $\lambda < (m\mu)$
  - What if $\rho = 1$? Can the system still be considered stable?
  - Remember arrival time is a random variable!
  - Once queueing starts, it never empties…

Traffic intensity example

- A USB thumb drive is serving 5k I/O ops/s
- The average time spent on the I/O operation is 0.1ms
- What is its utilization?

- $\rho = \lambda/(m\mu)$
- $\rho = 5k / (\mu)$
- $\mu = 1/0.1 = 10k$
- $\rho = 5k / 10k = 0.5$ (50%)

More detailed metrics

- Number of jobs in the system is the sum of the jobs in the queue and the ones in service
  - $\mathbf{n = n_s + n_q}$
- Total time spent in system (<u>response time</u>) is the sum of time spent queuing and that in service
  - $\mathbf{r = w_q + s}$

- Remember these are random variables, we'll speak of their expected value.

Little's law

- Remember: $n = n_s + n_q$
- The system as a whole: $E[n] = \lambda * E[r]$
- Only the queue part: $E[n_q] = \lambda * E[w_q]$

- $E[n_s] = \lambda * E[s]$ – looks familiar?
  - If $m=1$ and $E[n_s]=1$, the system is unstable!

- $\rho = \lambda/(m\mu) \rightarrow \rho = E[n_s] /m$

Quick overview of M/M/1 queues

- Interarrival times and service times Poisson
- Single server
- FIFO processing

- Parameters:
  - Mean arrival rate
  - Mean service rate

- Please look at the book for more detail and explanations. Have a look at the list of formulas.

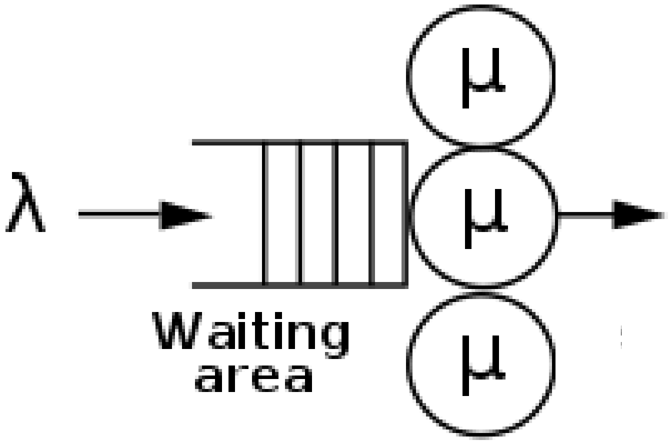
Response time in M/M/1

- The mean number of jobs in the system is computed using the probabilities of having 0..infinity jobs in the system.

$$E[n] = \sum_{n=1}^{\infty} n \cdot p^n = \sum_{n=1}^{\infty} n(1-\rho)\rho^n = \frac{\rho}{1-\rho}$$

- Using Little's law ($E[n] = \lambda * E[r]$), we get
- **E[r]** = $\rho/(\lambda *(1-\rho)) = \boldsymbol{(\mu * (1-\rho))^{-1}}$

Quick overview of M/M/m queues



- Interarrival times and service times Poisson
- Single server
- FIFO processing
- m parallel servers (no queueing if the number of jobs <= m)

- Parameters:
  - Mean arrival rate
  - Mean service rate
  - Server parallelism

- Please look at the book for more detail and explanations. Have a look at the list of formulas.

All formulas for M/M/1 and M/M/m

Box 31.1 M/M/1 Queue

1. Parameters:
   λ = arrival rate in jobs per unit time
   μ = service rate in jobs per unit time
2. Traffic intensity: $\rho = \lambda/\mu$
3. Stability condition: Traffic intensity ρ must be less than 1.
4. Probability of zero jobs in the system: $p_0 = 1-\rho$
5. Probability of n jobs in the system: $p_n = (1-\rho)\rho^n$, $n = 0, 1, \ldots, \infty$
6. Mean number of jobs in the system: $E[n] = \rho/(1-\rho)$
7. Variance of number of jobs in the system: $\text{Var}[n] = \rho/(1-\rho)^2$
8. Probability of k jobs in the queue:
$$P(n_q = k) = \begin{cases} 1-\rho^2, & k = 0 \\ (1-\rho)\rho^{k+1}, & k > 0 \end{cases}$$
9. Mean number of jobs in the queue: $E[n_q] = \rho^2/(1-\rho)$
10. Variance of number of jobs in the queue:
   $\text{Var}[n_q] = \rho^2(1+\rho-\rho^2)/(1-\rho)^2$
11. Cumulative distribution function of the response time:
   $F(r) = 1 - e^{-r\mu(1-\rho)}$
12. Mean response time: $E[r] = (1/\mu)/(1-\rho)$
13. Variance of the response time: $\text{Var}[r] = \dfrac{1/\mu^2}{(1-\rho)^2}$
14. q-Percentile of the response time: $E[r]\ln[100/(100-q)]$
15. 90-Percentile of the response time: $2.3E[r]$
16. Cumulative distribution function of waiting time:
   $F(w) = 1 - \rho e^{-\mu w(1-\rho)}$
17. Mean waiting time: $E[w] = \rho\dfrac{1/\mu}{1-\rho}$
18. Variance of the waiting time: $\text{Var}[w] = (2-\rho)\rho/[\mu^2(1-\rho)^2]$
19. q-Percentile of the waiting time: $\max\left(0, \dfrac{E[w]}{\rho}\ln[100\rho/(100-q)]\right)$
20. 90-Percentile of the waiting time: $\max\left(0, \dfrac{E[w]}{\rho}\ln[10\rho]\right)$
21. Probability of finding n *or more* jobs in the system: ρ^n
22. Probability of serving n jobs in one busy period:
$$\frac{1}{n}\binom{2n-2}{n-1}\frac{\rho^{n-1}}{(1+\rho)^{2n-1}}$$

Box 31.2 M/M/m Queue

1. Parameters:
   λ = arrival rate in jobs per unit time
   μ = service rate in jobs per unit time
   m = number of servers
2. Traffic intensity: $\rho = \lambda/(m\mu)$
3. The system is stable if the traffic intensity ρ is less than 1.
4. Probability of zero jobs in the system:
$$p_0 = \left[1 + \frac{(m\rho)^m}{m!(1-\rho)} + \sum_{n=1}^{m-1}\frac{(m\rho)^n}{n!}\right]^{-1}$$
5. Probability of n jobs in the system:
$$p_n = \begin{cases} p_0\dfrac{(m\rho)^n}{n!}, & n < m \\ p_0\dfrac{\rho^n m^m}{m!}, & n \ge m \end{cases}$$
6. Probability of queueing:
$$\varrho = P(\ge m \text{ jobs}) = \frac{(m\rho)^m}{m!(1-\rho)}p_0$$
   In the remaining formulas below we will use ϱ as defined here.
7. Mean number of jobs in the system: $E[n] = m\rho + \rho\varrho/(1-\rho)$
8. Variance of number of jobs in the system:
$$\text{Var}[n] = m\rho + \rho\varrho\left[\frac{1+\rho-\rho\varrho}{(1-\rho)^2} + m\right]$$
9. Mean number of jobs in the queue: $E[n_q] = \rho\varrho/(1-\rho)$
10. Variance of number of jobs in the queue:
   $\text{Var}[n_q] = \varrho\rho(1+\rho-\varrho\rho)/(1-\rho)^2$
11. Average utilization of each server: $U = \lambda/(m\mu) = \rho$
12. Cumulative distribution function of response time:
$$F(r) = \begin{cases} 1 - e^{-\mu r} - \dfrac{\varrho}{1-m+m\rho}e^{-m\mu(1-\rho)r} - e^{-\mu r}, & \rho \ne (m-1)/m \\ 1 - e^{-\mu r} - \varrho\mu r e^{-\mu r}, & \rho = (m-1)/m \end{cases} \quad r > 0$$

Box 31.2 Continued

13. Mean response time:
$$E[r] = \frac{1}{\mu}\left(1 + \frac{\varrho}{m(1-\rho)}\right)$$
14. Variance of the response time:
$$\text{Var}[r] = \frac{1}{\mu^2}\left[1 + \frac{\varrho(2-\varrho)}{m^2(1-\rho)^2}\right]$$
15. Cumulative distribution function of waiting time:
   $F(w) = 1 - \varrho e^{-m\mu(1-\rho)w}$
16. Mean waiting time: $E[w] = E[n_q]/\lambda = \varrho/[m\mu(1-\rho)]$
17. Variance of the waiting time: $\text{Var}[w] = \varrho(2-\varrho)/[m^2\mu^2(1-\rho)^2]$
18. q-Percentile of the waiting time: $\max\left(0, \dfrac{E[w]}{\varrho}\ln\dfrac{100\varrho}{100-q}\right)$.
19. 90-Percentile of the waiting time: $\dfrac{E[w]}{\varrho}\ln(10\varrho)$

Once again, ϱ in these formulas is the probability of m or more jobs in the system: $\varrho = [(m\rho)^m/\{m!(1-\rho)\}]p_0$. For $m = 1$, ϱ is equal to ρ and all of the formulas become identical to those for M/M/1 queues.

$n=m$

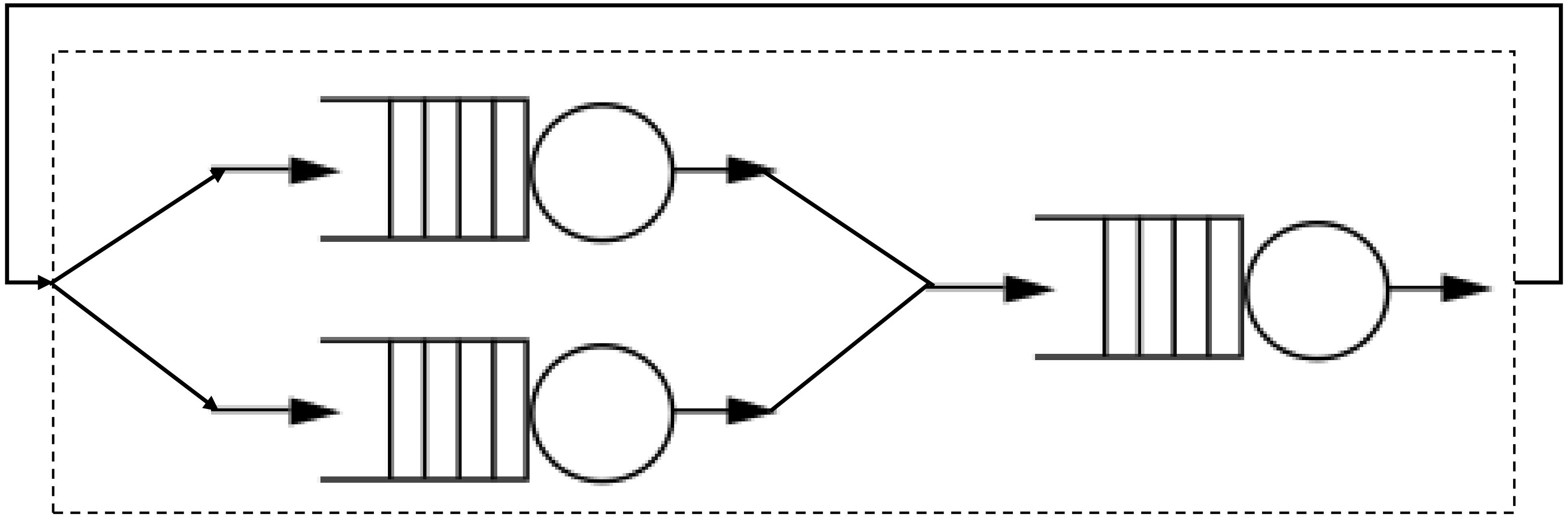
Exercise

- **μ = 250/s**
  - E[s] = ?
- **λ = 1200/s**
- What is better for clients? **5x M/M/1** or **1x M/M/m**?

- For M/M/1: E[r] = 0.1s
- For M/M/m: E[r] = 0.022s
  - The jobs wait in a single queue and can go to any available server. In the other case they need to wait for their pre-chosen server to become available…

Hint: there are many tools/websites that help with computing the outputs of the models (e.g., https://www.supositorio.com/rcalc/rcalclite.htm)

Network of queues



- A collection of queue/server pairs
- Jobs “flow” through the network
- Can represent arbitrarily complex systems
- Open and closed variants

Properties of NoQ devices

- Service discipline
  - FIFO (e.g., M/M/1 and M/M/m)
  - Delay center (imagine M/M/∞)
  - …
- Job classes
  - All jobs are equal

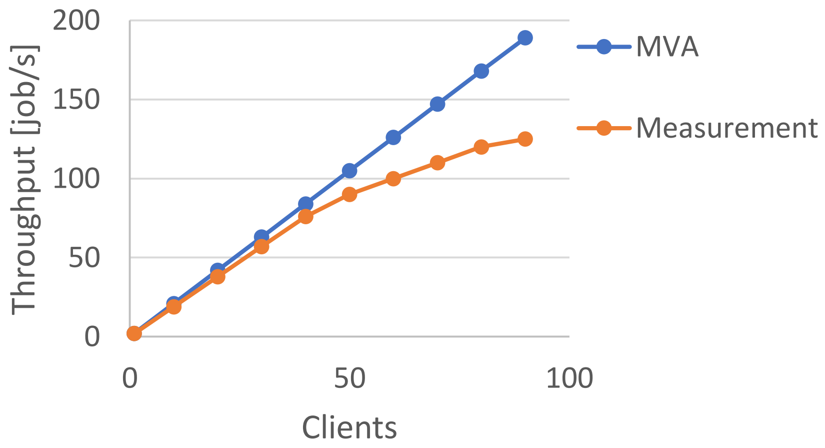
- <u>Job flow balance</u>
  - Number of arrivals at each device equals number of leaving jobs
- <u>One-step behavior</u>
  - The state of the network changes only as a result of a job entering the system (a device)

Operational laws

- Valid for all devices
  - Arrival rate λ_i = (number of arrivals)/time = A_i/T
  - Throughput X_i = (number of completions)/time = C_i/T
  - Utilization U_i = (busy time)/time = B_i/T
  - Mean service time = (busy time)/(number of completions) = B_i/C_i
- Utilization Law
  - $U_i=B_i/T = C_i/T * B_i/C_i$
  - $U_i=X_i * S_i$ (Device with highest utilization is the bottleneck device)
- Forced flow law
  - $A_i = C_i$
- System throughput X = (jobs completed)/time
  - Device throughput $X_i = X * V_i$
  - V_i is the visit ratio; how many times a job is handled by the device *i*

Mean Value Analysis (MVA)

- Algorithm to compute the behavior of a NoQ with increasing clients
  - Might have to map throughput levels to number of clients!



Box 34.2 MVA Algorithm

Inputs:
N = number of users
Z = think time
M = number of devices (not including terminals)
S_i = service time per visit to the ith device
V_i = number of visits to the ith device

Outputs:
X = system throughput
Q_i = average number of jobs at the ith device
R_i = response time of the ith device
R = system response time
U_i = utilization of the ith device

Initialization: FOR $i = 1$ TO M DO $Q_i = 0$

Iterations:
FOR $n = 1$ TO N DO
BEGIN

FOR $i = 1$ TO M DO $R_i = \begin{cases} S_i(1 + Q_i) & \text{Fixed capacity} \\ S_i & \text{Delay centers} \end{cases}$

$$R = \sum_{i=1}^{M} R_i V_i$$

$$X = \frac{N}{Z + R}$$

FOR $i = 1$ TO M DO $Q_i = X V_i R_i$

END

Device throughputs: $X_i = X V_i$
Device utilizations: $U_i = X S_i V_i$