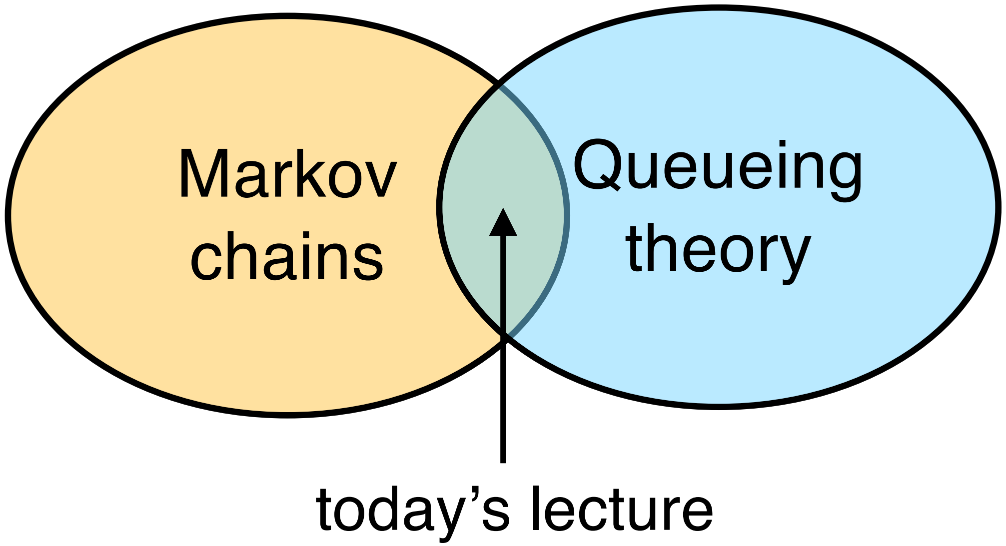
Performance Evaluation of Networks

Sara Alouf

Part II – Queueing Theory

- Part I – Markov chains
  - Irreducible
    - Discrete-time Markov chains (Chapter 1)
    - Continuous-time Markov chains (Chapter 2)
  - Absorbing
    - Discrete-time and continuous-time (Chapter 3)
- Part II – Queueing Theory



Part II – Queueing Theory

- What is a queue?
  - Supermarket, bank, postoffice, administrations, etc.
  - CPU, servers, clusters, etc.
  - Manufacturing, product lines, etc.
  - …
- System with
  - at least one service facility
  - potentially a waiting room (finite or infinite)
  - customers
- Representation

Kendall's Notation

- Describe a queueing system

$$A \,/\, B \,/\, c \,/\, K$$

A ➜ distribution of interarrivals

B ➜ distribution of service times

c ➜ number of servers

K ➜ number of customers in system (omitted if infinite)

- Distributions often used
  - Exponential ➜ M
  - Deterministic ➜ D
  - General ➜ G
  - Erlang ➜ E
  - Phase-type ➜ PH

What is Not in Kendall's Notation

- Service discipline (scheduling)
  - First-in-first-out (First-come-first-served)
  - Last-in-first-out (last-come-first-served)
  - Processor sharing
  - Random
  - Shortest-job-first
  - Shortest-processing-time-first
- Multiple waiting rooms / queues: Join discipline
  - Random
  - Join-shortest-queue
  - Best-out-of-*d*

$M / M / 1$ Queue

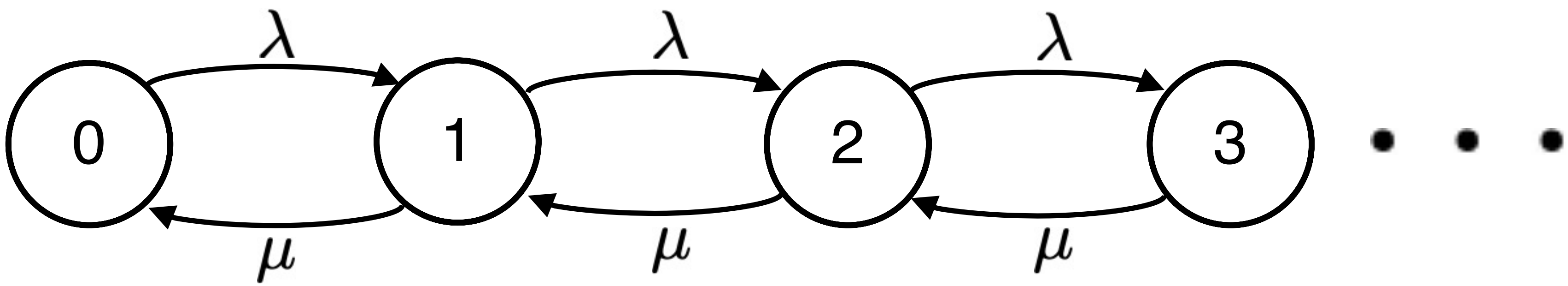
- Arrivals: Poisson rate λ ➜ interarrival time Exp(λ)
- Service time Exp(μ)
- Independence between arrivals and service times
- Service discipline ➜ not relevant (memoryless property)
- $X(t)$ number of customers in system at time t (queue size)
- Proposition 12: $\{X(t), t \geq 0\}$ is a birth-and-death process, birth rate λ, death rate μ
- Proof: construction rule #2

$$i \to i+1 \qquad \mathrm{Exp}(\lambda)$$

$$i \to i-1 \qquad \mathrm{Exp}(\mu) \quad i > 0$$

$M / M / 1$ Queue

- Transition diagram, $\mathcal{E} = \mathbb{N}$



- Global balance equations

$$\lambda \pi_{i-1} = \mu \pi_i, \qquad i \geq 1$$

- System utilization $\rho = \dfrac{\lambda}{\mu}$

- Proposition 13: If $\rho < 1$ limiting/stationary distribution is

$$\pi_i = (1-\rho)\rho^i, \quad i \geq 0$$

$M\,/\,M\,/\,1$ Queue

- Proof
$$\begin{aligned}\pi_i &= \rho\pi_{i-1}\\ &= \rho^i\pi_0\end{aligned}$$

- Normalization $\sum_{i\geq 0}\pi_i = 1$

$$\Leftrightarrow \quad \pi_0\sum_{i\geq 0}\rho^i = 1$$ sum of terms in geometric progression

- If $\boxed{\rho < 1} \Rightarrow \quad \pi_0\dfrac{1}{1-\rho} = 1$

stability condition

$$\Rightarrow \quad \pi_0 = 1-\rho$$ (ρ system utilization)

$$\Rightarrow \quad \pi_i = (1-\rho)\rho^i, \quad i \geq 0$$

$M\,/\,M\,/\,1$ Queue

- Let X stationary version of queue size ➜ $X \sim \mathrm{Geom}(1-\rho)$

  $1-\rho$ probability to find system empty

  X number of « failed trials » before finding system empty

- Expected queue size

$$E[X] = \sum_{i \geq 0} i\,\pi_i = (1-\rho) \sum_{i \geq 0} i\,\rho^i$$

$$= \rho(1-\rho) \sum_{i \geq 1} i\,\rho^{i-1} \qquad = \rho(1-\rho) \left(\sum_{i \geq 1} \rho^i \right)'$$

$$\Rightarrow \boxed{E[X] = \frac{\rho}{1-\rho}}$$

- Throughput (= rate of everything that goes through the system)

$$\mathrm{Thpt} = \mu\,(1-\pi_0) = \mu\,\rho = \lambda$$

Burke's Theorem

- Suppose $M/M/1$ starts in steady-state
  - Departure process is Poisson with rate λ
  - Queue size at t independent of departures before time t

- Proof: use time-reversibility of $M/M/1$

  Forward chain identical to Backwards chain

  ✔ Arrival process is Poisson with rate λ

  ✔ Queue size at t independent of arrivals after time t

$M / M / 1 / K$ Queue

- $M / M / 1$ finite waiting room ➜ at most K customers in system
- Queue size is a finite birth-and-death process

$$\pi_i = \rho^i \pi_0, \quad 0 \le i \le K$$

- Normalization $\displaystyle\sum_{i=0}^{K} \pi_i = 1 \quad \Leftrightarrow \quad \pi_0 \sum_{i=0}^{K} \rho^i = 1$

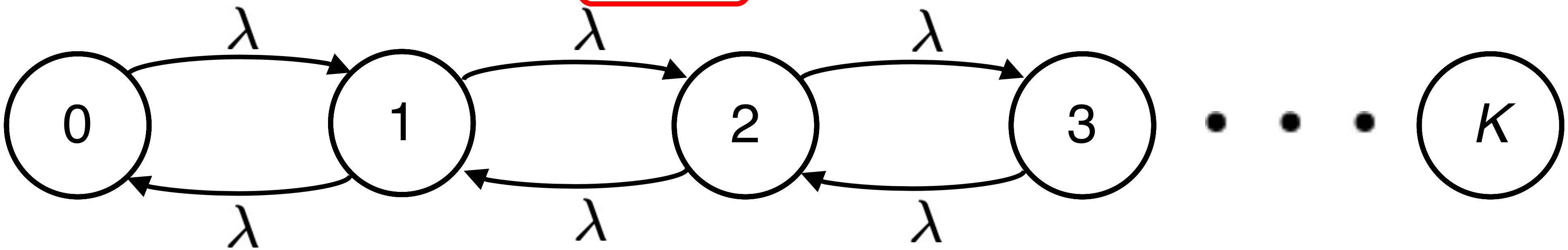
- If $\rho \neq 1 \;\Rightarrow\; \displaystyle\pi_0 \frac{1-\rho^{K+1}}{1-\rho} = 1 \quad \Leftrightarrow \quad \pi_0 = \frac{1-\rho}{1-\rho^{K+1}}$

$$\Rightarrow \quad \boxed{\pi_i = \frac{(1-\rho)\,\rho^i}{1-\rho^{K+1}}, \quad i = 0, 1, \ldots, K}$$

- If $\rho = 1 \;\Rightarrow\; \displaystyle\pi_0 = \boxed{\frac{1}{K+1} = \pi_i, \quad i = 0, 1, \ldots, K}$

$M / M / 1 / K$ Queue

- Transition diagram, $\boxed{\rho = 1}$ $\mathcal{E} = \{0, 1, \ldots, K\}$



- We are equally likely to go left or right
- Stationary process X is Uniform between 0 and K

$$\pi_i = \frac{1}{K+1}, \quad i = 0, 1, \ldots, K$$

- Expected queue size

$$E[X] = \sum_{i=0}^{K} i\,\pi_i = \frac{1}{K+1}\sum_{i=0}^{K} i = \frac{1}{K+1}\frac{K(K+1)}{2} = \frac{K}{2}$$

$M / M / 1 / K$ Queue

- Stationary distribution always exists!

  no stability condition (finite system)

- Finite system ➜ customers may find system full!

  loss probability ?

- PASTA : Poisson Arrivals See Time Averages
- Birth-and-death process is ergodic
  - Time averages = stationary distribution

- Loss probability = prob customer arrives and sees full queue

$$P_{\text{loss}} = \pi_K = \frac{(1-\rho)\,\rho^K}{1-\rho^{K+1}}$$

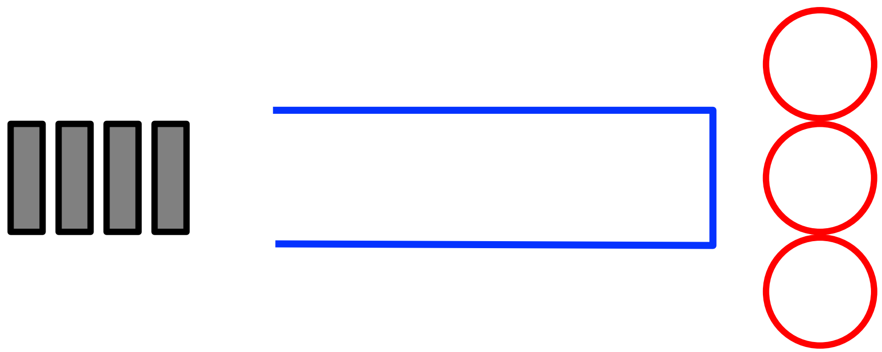
- Throughput $\quad \text{Thpt} = \mu\,(1-\pi_0) = \lambda\,(1-\pi_K)$

$M / M / c$ Queue

- Poisson arrivals rate λ ➜ interarrival time Exp(λ)
- Service time Exp(μ)
- Multiple servers
- Infinite waiting room
- Queue size is a birth-and-death process
  - $i \to i+1$ birth rate λ
  - $i \to i-1$ death rate $\begin{aligned} \mu_i &= i\mu, \quad i = 1, 2, \ldots, c-1 \\ &= c\mu, \quad i \geq c \end{aligned}$
- System utilization $\rho = \dfrac{\lambda}{c\mu}$
- Infinite system ➜ stability condition might be needed!

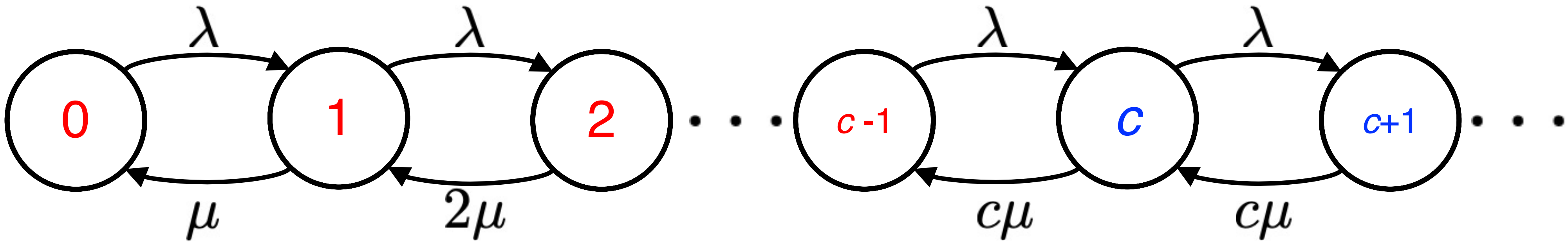
$M\,/\,M\,/\,c$ Queue

- Representation



Example:
call center

- Transition diagram, $\mathcal{E} = \mathbb{N}$



Customers served upon arrival

new customer needs to wait

$M/M/c$ Queue

- Stationary distribution: for $i = 1, 2, \ldots$

$$\pi_i = \frac{\lambda_0\lambda_1\cdots\lambda_{i-1}}{\mu_1\mu_2\cdots\mu_i}\pi_0 = \begin{cases} \left(\frac{\lambda}{\mu}\right)^i \frac{1}{i!}\pi_0 & i = 0, 1, \ldots, c \\ \left(\frac{\lambda}{\mu}\right)^i \frac{1}{c!}\frac{1}{c^{i-c}}\pi_0 & i \geq c \end{cases}$$

- Normalization: if $\boxed{\rho < 1}$ (stability condition)

$$\pi_0 = \left[\sum_{i=0}^{c-1}\left(\frac{\lambda}{\mu}\right)^i \frac{1}{i!} + \left(\frac{\lambda}{\mu}\right)^c \frac{1}{c!}\left(\frac{1}{1-\rho}\right)\right]^{-1}$$

$$\rho = \frac{\lambda}{c\mu}$$

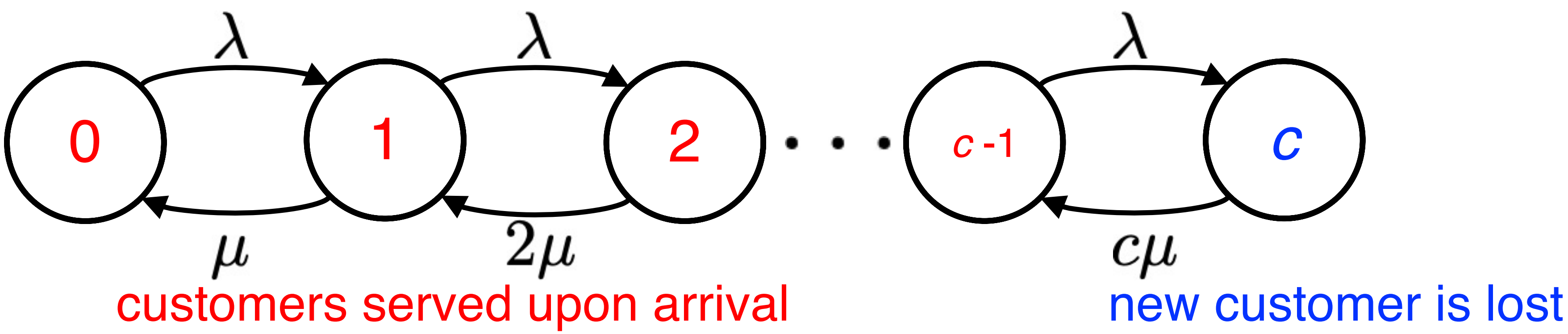
- Probability of waiting (PASTA)

$$P_{\text{wait}} = \sum_{i\geq c}\pi_i = \frac{\pi_0(\lambda/\mu)^c}{c!}\sum_{i\geq 0}\left(\frac{\lambda}{c\mu}\right)^i = \frac{\pi_0(c\rho)^c}{c!(1-\rho)}$$

15 minutes break

$M / M / c / c$ Queue

- Multi-server queue with no waiting room
- Pure loss system (call center without music)
- Finite system ➜ no stability condition
- Queue size is a birth-and-death process
  - $i \to i+1$ birth rate $\lambda_i = \lambda \quad i = 0, 1, \ldots, c-1$
  - $i \to i-1$ death rate $\mu_i = i\mu, \quad i = 1, 2, \ldots, c$
- Transition diagram $\mathcal{E} = \{0, 1, \ldots, c\}$



$M\,/\,M\,/\,c\,/\,c$ Queue

- Stationary distribution

$$\pi_i = \pi_0 \left(\frac{\lambda}{\mu}\right)^i \frac{1}{i!} \quad i = 0, 1, \ldots, c$$

$$\pi_0 = \left[\sum_{i=0}^{c} \left(\frac{\lambda}{\mu}\right)^i \frac{1}{i!}\right]^{-1}$$

- Loss probability

$$\boxed{P_{\text{loss}} = \pi_c = \frac{\left(\frac{\lambda}{\mu}\right)^c \frac{1}{c!}}{\sum_{i=0}^{c} \left(\frac{\lambda}{\mu}\right)^i \frac{1}{i!}}}$$

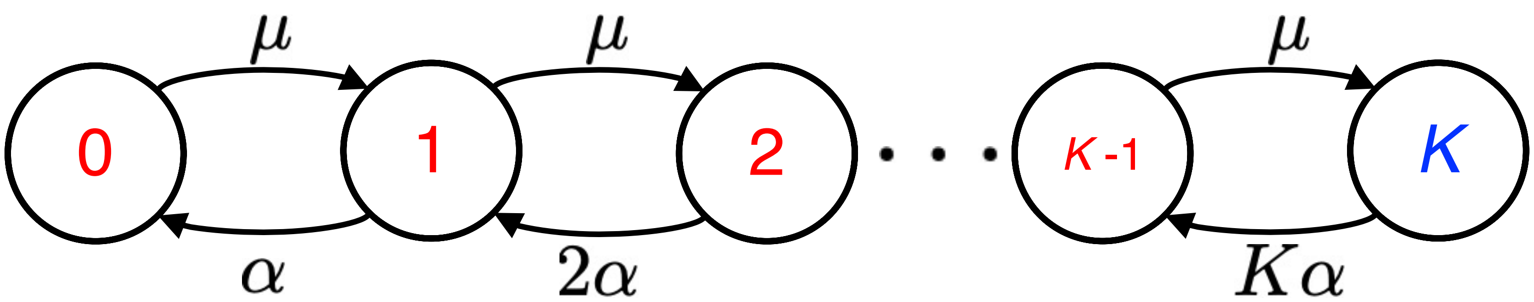
- Erlang's loss formula (1917)
- Major historical role in dimensioning phone systems
- Insensitivity property: holds for any service time distribution

Example: Repair Person Model

- K machines
- One repair person
- Each machine breaks after time Exp(α) (break rate is α)
- Upon a breakdown repair request is sent
- Repair person spends time Exp(μ) to repair one machine
- Questions
  - Probability that i machines are working normally
  - Overall failure rate?
- Define $X(t)$ number of functional machines at time t
- State space $\mathcal{E} = \{0, 1, \ldots, K\}$

Example: Repair Person Model

- Possible transitions?
  - $i \rightarrow i+1$ repair is over, time Exp(μ)
  - $i \rightarrow i-1$ break occurs, time Exp($i\alpha$)
- Using construction rule #2, process is a (homogeneous) CTMC
- It is also a birth-and-death-process
  - Birth rate (of functional machines) $\lambda_i = \mu$
  - Death rate $\mu_i = i\,\alpha$
- Transition diagram



Example: Repair Person Model

- Number of functional machines

  = Queue size of $M / M / K / K$

- Probability that i machines are working normally

$$\pi_i = \frac{\left(\frac{\mu}{\alpha}\right)^i \frac{1}{i!}}{\sum_{j=0}^{K} \left(\frac{\mu}{\alpha}\right)^j \frac{1}{j!}}$$

- Overall failure rate

$$\sum_{i=1}^{K} (i\,\alpha)\,\pi_i = \alpha E[X]$$

- Overall repair rate

$$\sum_{i=0}^{K-1} \mu\,\pi_i = \mu\,(1 - \pi_K)$$

- In steady-state overall failure rate = overall repair rate

Little's Formula

- Relates three quantities in steady-state
  - Occupancy in a system $\overline{N}$
  - Entrance rate to the system λ
  - Sojourn time in system $\overline{T}$

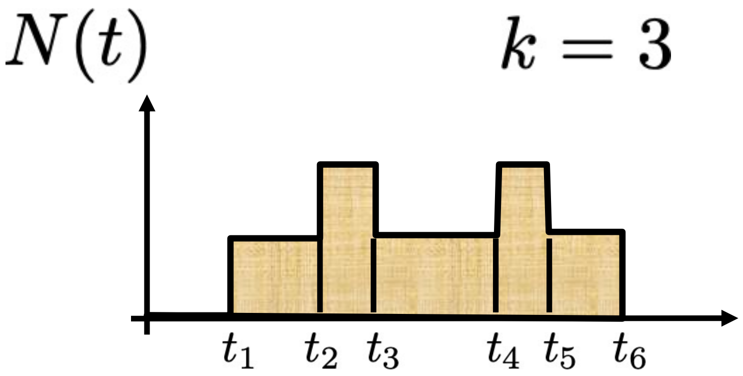
$$\boxed{\overline{N} = \lambda \overline{T}}$$

- Valid for work-conserving systems

  system may not be idle if customers waiting
- No assumption on any distribution
- No assumption on service discipline
- Independence assumption

Proof of Little's Formula

- Steady-state ➜ system empties infinitely often
- Let 0 and C be two times when system is empty
- Let k be number of customers served in $(0, C)$
- $\{a_i\}_{i=1,\ldots,k}$ arrival instants
- $\{d_i\}_{i=1,\ldots,k}$ departure instants
- $\{t_n\}_{n=1,\ldots,2k}$ all instants
- Number of customers over time $N(t)$

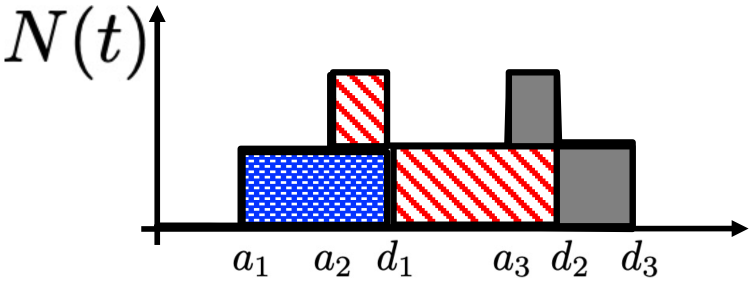


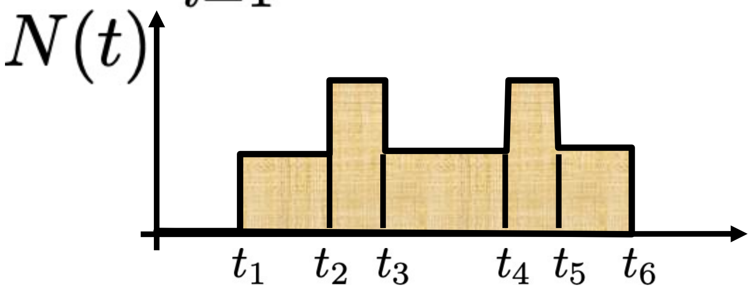
$$\overline{N} = \frac{1}{C} \int_0^C N(t)\, dt$$

$$= \frac{1}{C} \sum_{i=1}^{2k-1} N(t_i^+)\,(t_{i+1} - t_i)$$

Proof of Little's Formula

- Mean sojourn time $$\overline{T} = \frac{1}{k}\sum_{i=1}^{k}(d_i - a_i)$$





sum of horizontal boxes = sum of vertical boxes

$$\sum_{i=1}^{k}(d_i - a_i) = \sum_{i=1}^{2k-1} N(t_i^+)\,(t_{i+1} - t_i)$$

$$\Rightarrow \quad \overline{T}\,k = \overline{N}\,C$$

- When $C \to \infty,\ \frac{k}{C} \to \lambda \quad \Rightarrow \quad \boxed{\overline{T}\lambda = \overline{N}}$

Example 5 Page 39

- Consider $M\,/\,M\,/\,1$, arrival rate λ, service rate μ
- If $\rho = \dfrac{\lambda}{\mu} < 1$ (queue is stable)
- Mean number of customers $\overline{N} = E[X] = \dfrac{\rho}{1-\rho} > 0$
- Entrance rate = arrival rate (no losses)
- Expected sojourn time

  By Little's formula $\overline{T} = \dfrac{\overline{N}}{\lambda} = \dfrac{\rho}{\lambda(1-\rho)} \Rightarrow \boxed{\overline{T} = \dfrac{1}{\mu - \lambda}}$
- Expected waiting time

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

For next week

- Lesson 4 to revise
- Homework 4 to return on Tuesday 8 October before 9 am
- Lesson 5 to read before Lecture 5