MIEEC

Computer Networks

Lecture note 5

Delay models for queuing

Faculdade de Engenharia da Universidade do Porto

FEUP DEEC > DEPARTAMENTO DE ENGENHARIA ELECTROTÉCNICA E DE COMPUTADORES

Delay models

Types of delay

- Processing delay
- **Queuing delay**
- Transmission delay
- Propagation delay

- Connection establishment delay
- End-to-end delay

TO THINK

Where do you apply queuing delay model?

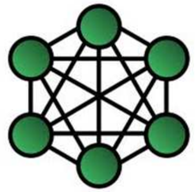
- Statistical multiplexing

What do you need multiplexing for?

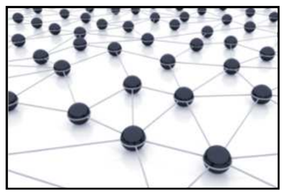
What is multiplexing?

Multiplexing

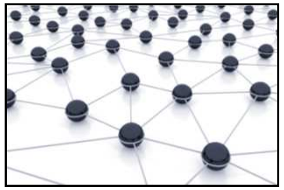
- N^2 connectivty
  - No need for multiplexing
  - Direct channel between all nodes



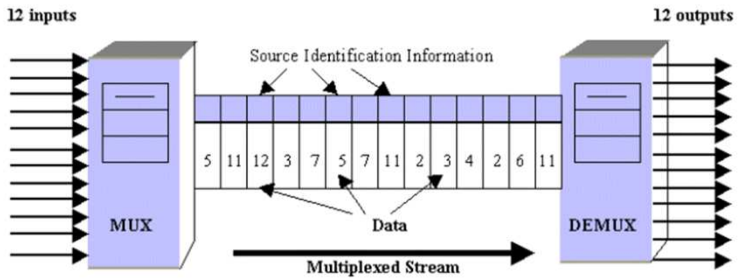
- Network
  - Shared channels
  - Multiplex data into/out of channels



Time Division Multiplexing

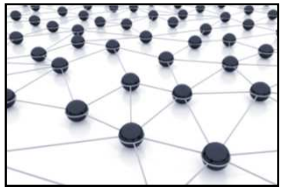


- Time slots
- Assign slots to separate channels
  - Route time slots through the network
- Transmission on a link

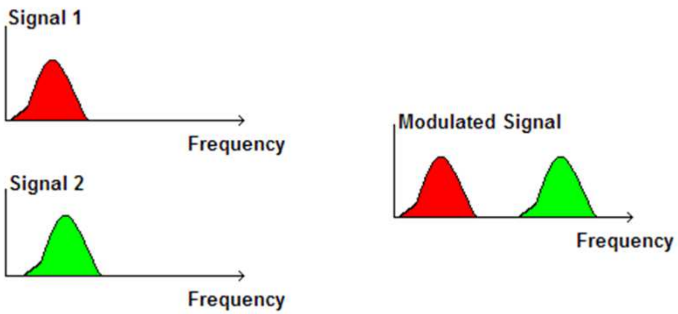


  - m slots
  - Link capacity C
  - Channel line capacity C/m
  - T_{trans}= Lm/C , L >> slot size

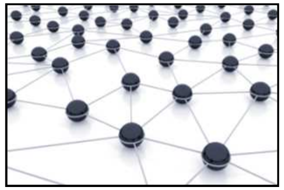
Frequency Division Multiplexing



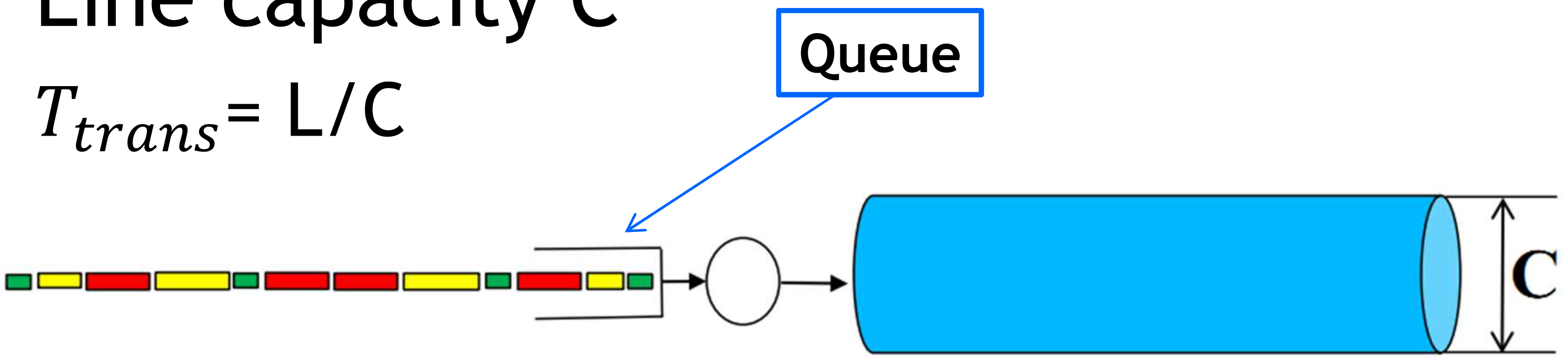
- Frequency slots
- Assign slots to separate channels
  - Route time slots through the network
- Transmission on a link
  - m slots
  - Link capacity C
  - Channel line capacity C/m
  - T_{trans}= Lm/C , L >> slot size



Statistical Multiplexing



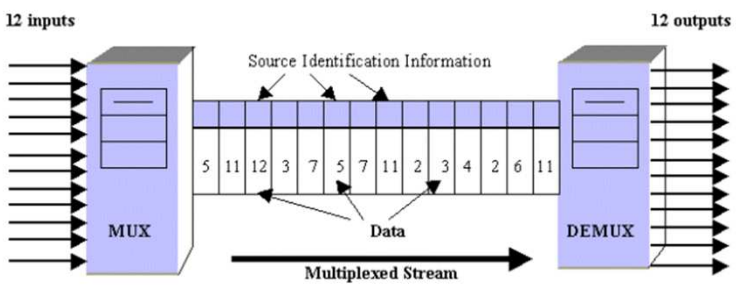
- No slots
- First-come first-served
- Don't drop:
  - Put in the **queue** if not served immediately
- Line capacity C
- T_{trans}= L/C

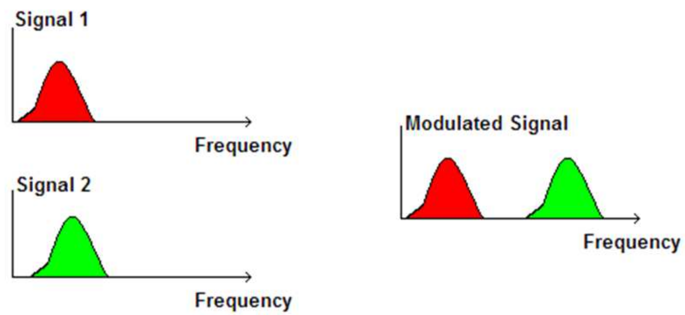


TO THINK

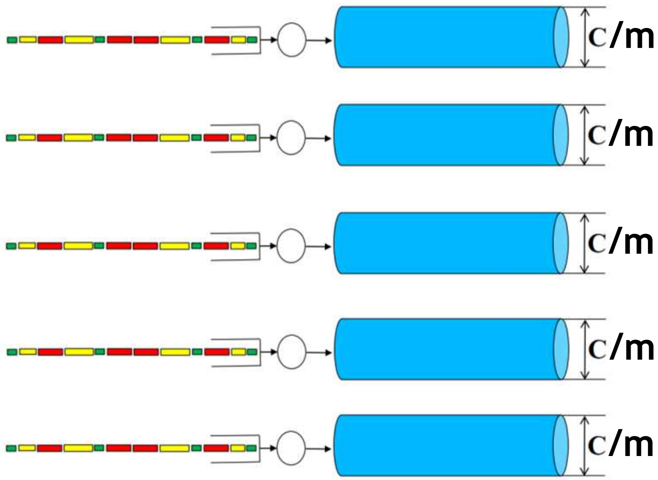
Why are there no queues in the TDM/FDM pictures?

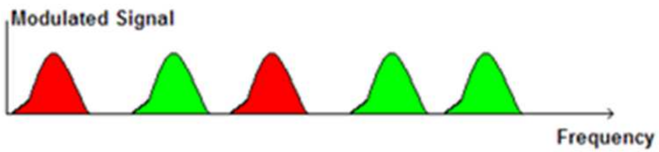
- What’s the nature of the traffic?

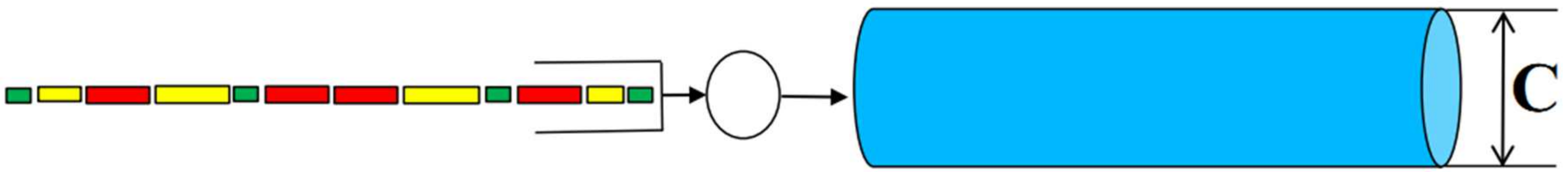




Single or multiple queues?



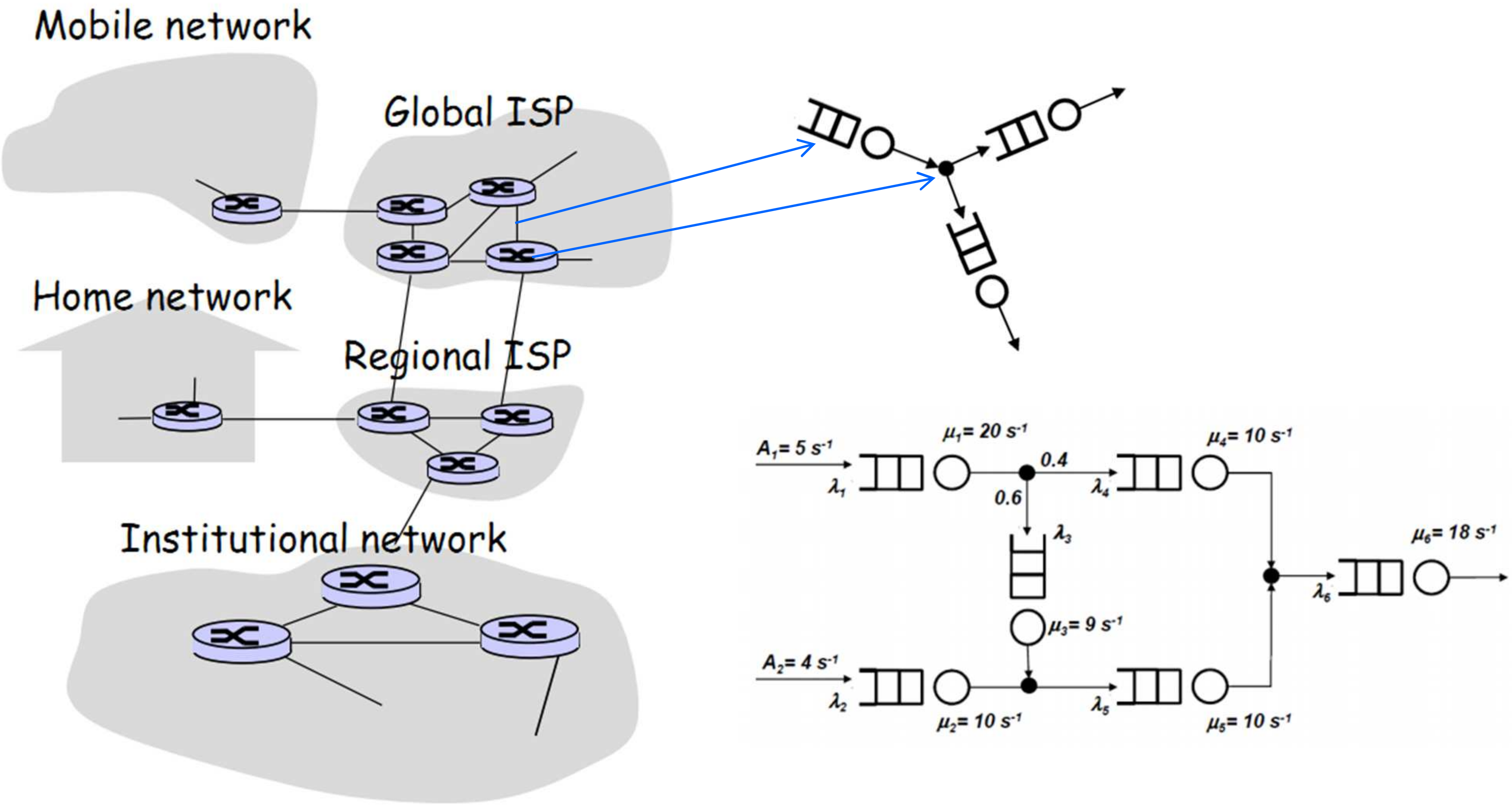




Queue model for delay

- Where can it be used?
  - Any circumstance where packets have to wait on a queue before being sent/received/processed/…
- Model:
  - Packets arrive at random time
  - Need to be served, i.e. sent through link
  - Service time :: packet transmission time L/C
  - Waiting time in queue :: queuing delay
- Average delay per packet (waiting+service)
- Average number of packets in the network

Queue models for computer networks



TO THINK

"Packets arrive at random time"

- If deterministic arrivals then no point in statistical multiplexing

How random is random?

$$\sum_x p(x)\, log_2 p(x)$$

Which randomness?

Poisson Distribution and Process

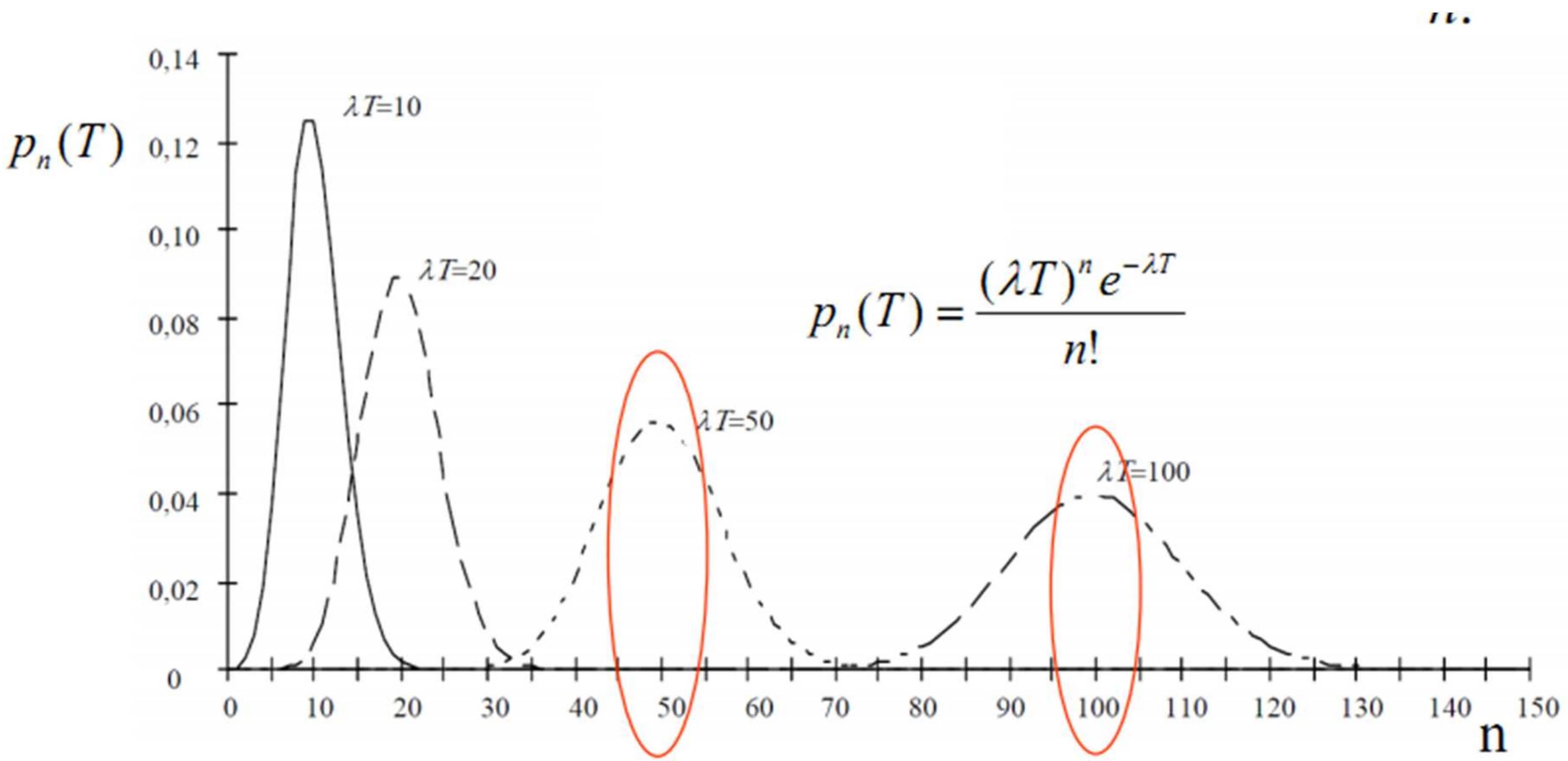
- Poisson distribution

$$P[N = n] = p_n = \frac{m^n e^{-m}}{n!}, n \in \mathbb{N}_0$$

$$E_P[N] = Var_P[N] = m$$

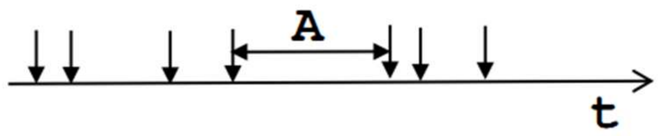
- Poisson process
  - $\lambda\, T = m$ (e.g. λ arrivals/s)
  - $P[n\ arrivals\ in\ T] = p_n(T) = \frac{(\lambda T)^n e^{-\lambda T}}{n!}$
  - $E_P[N] = Var_P[N] = \lambda T$

Poisson Distribution and Process



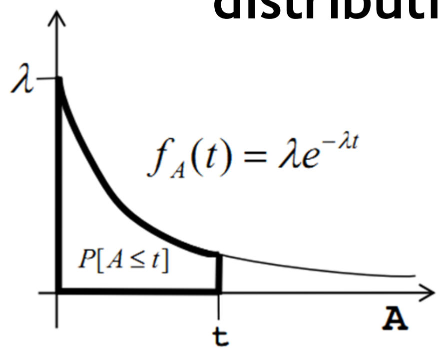
The other side of Poisson: Packet inter-arrival time

- Packet arrival



- $F_A(t) = P[A \leq t] =$
  $= 1 - P[A > t] =$
  $= 1 - p_0(t) = 1 - e^{-\lambda t}$

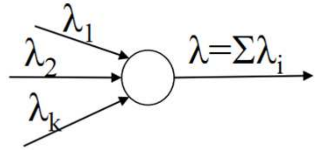
Exponential distribution



- $E_P[A] = \frac{1}{\lambda}$; $Var_P[A] = \frac{1}{\lambda^2}$

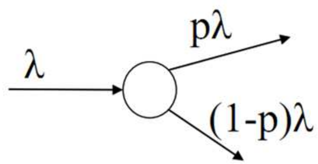
Poisson Process Properties

- Merging property



  - If $A_1, A_2, \ldots, A_k$ are independent Poisson processes with rates $\lambda_1, \lambda_2, \ldots, \lambda_k$
  - Then $A = \sum_{i=1}^{k} A_i$ is still Poisson
  - With rate $\lambda = \sum_{i=1}^{k} \lambda_i$

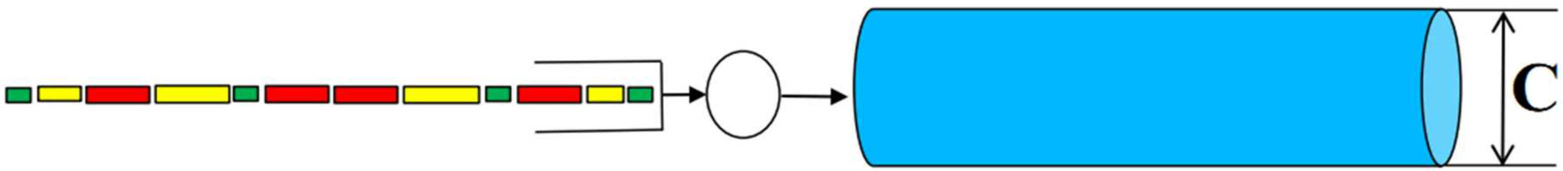
- Splitting property



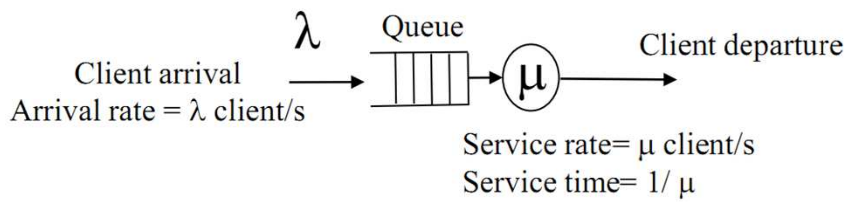
  - Incoming Poisson packets
  - Split randomly with p and (1-p) to output
  - Both outputs are still random processes
  - With rates $p\lambda$ and$(1-p)\lambda$

TO THINK

How many processes are there in a queue?



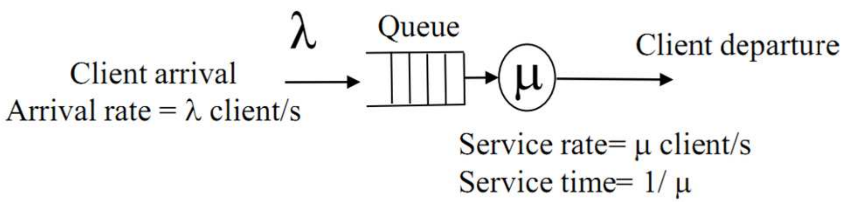
Queue model



- Model for:
  - Customers waiting in queue
  - Packets in network

- Used to determine:
  - N, average number of clients (packets) in the system (network, node)
  - T, average delay experienced by client (packet)

Queue model



- Queue characterized by:
  - λ, arrival rate of customers (placed in the queue)
  - μ, service rate (taken out of the queue and sent)
  - $\rho = \lambda/\mu$, traffic intensity, occupation of the server

- Kendall notation **A/S/s/K**
  - A : statistical process for the arrivals
  - S : statistical process for the service
  - s : number of servers
  - K : system buffer capacity / queue size

Warning

- Queue analysis is always:
  - On the long run
  - In steady state
  - In equilibrium
  - Not instantaneous
- λ
- μ
- ρ

TO THINK

How's the waiting time (**delay**) in the queue related to the number of clients in the queue?

Little's Theorem

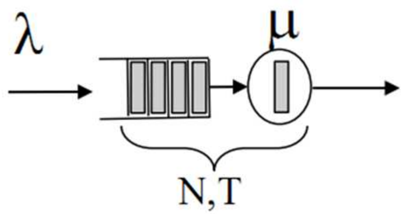
- $N = \lambda T$
  - N: average number of clients in the system
  - T: average time client spends in the system
  - λ: arrival rate of clients in the system

- Queue (w) and service (s) time / number of customers
  - $T = T_w + T_s$
  - $N = N_w + N_s$

- $\boldsymbol{N_w = \lambda T_w}$

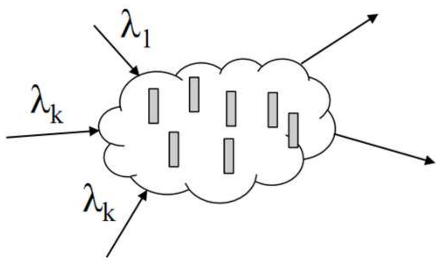
Little's Theorem

- Can be applied to single queue



$$N = \lambda T$$

- Or to a complex system



$$N_i = \lambda_i T_i$$

$$\lambda = \sum \lambda_i ; N = \sum N_i$$

$$N = \lambda T$$

TO THINK

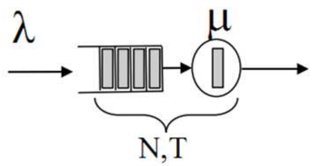
$$N_w = \lambda T_w$$

The mean time a customer has to wait depends only on:

- the number of customers in the queue
- the arrival rate

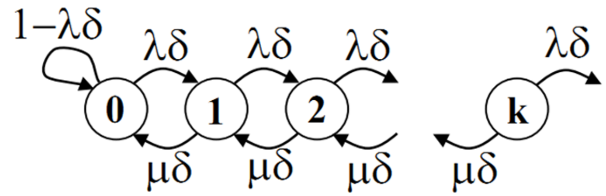
What about service time?

M/M/1 Queue



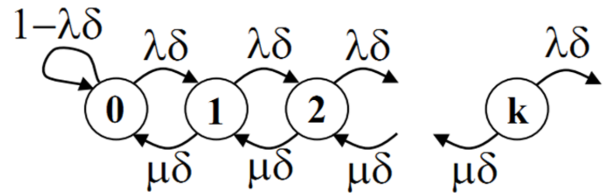
- M/M/1
  - A/S/s/K
  - Poisson arrival, Exponential service time
  - 1 server, infinite queue length
- Goals
  - What’s the average queue size?
  - What’s the average waiting time?
- Underlying model
  - Markov chain

M/M/1
Markov chain



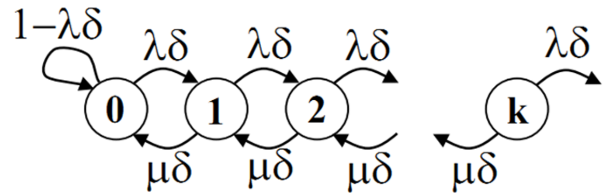
- Markov chain
  - state sequence and transition probability
- State k
  - k clients in the queue
- p(i,j)
  - transition probability from state i to j
- In the limit, when $\delta \rightarrow 0$:
  - $p(i, i+1) = \delta\lambda$: new customer in the queue
  - $p(i, i-1) = \delta\mu$: customer served from queue
  - $p(i, i) = 1 - \delta\lambda - \delta\mu$: nothing happens
  - $p(0,0) = 1 - \delta\lambda$: nothing happens at state 0
  - $p(i, j) = 0, other\ (i, j)$: only allow +1/-1 change in k

M/M/1
Equilibrium Analysis



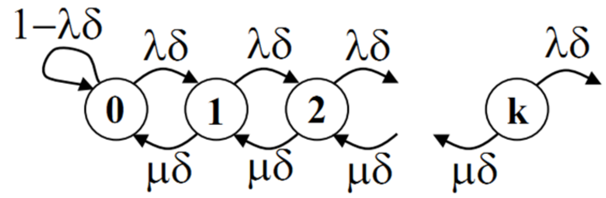
- Transition probabilities known
  - p(i,i+1), etc.
- What's the probability queue in state j?
  - P(j)
- Equilibrium condition
  - Global balance equations for the Markov Chain
  - $P(j)\sum_{\substack{i=0 \\ j\neq i}}^{+\infty} p(j,i) = \sum_{\substack{i=0 \\ j\neq i}}^{+\infty} P(i)\, p(j,i)$

M/M/1
Equilibrium Analysis



- In the case of M/M/1 this leads to:
  - $P(0)\lambda\delta = P(1)\mu\delta \Rightarrow P(1) = \rho P(0)$
  - $P(2) = \rho P(1) = \rho^2 P(0)$
  - $P(n) = \rho^n P(0);$
  - $\sum P_i = 1 \Rightarrow \sum_{i=0}^{+\infty} \rho^i P(0) = 1 \Rightarrow P(0) = 1 - \rho$
  - $P(n) = \rho^n (1 - \rho)$

M/M/1
Equilibrium Analysis

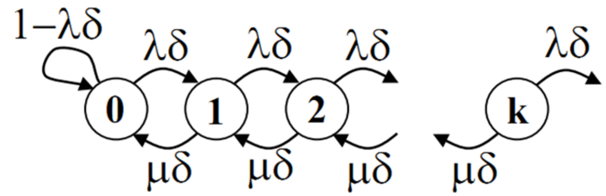


- Goals
  - **What's the average queue size?**
  - What's the average waiting time?

- $N = \sum_{n=0}^{\infty} nP(n) = \sum_{n=0}^{\infty} n\rho^n(1-\rho) = \rho/(1-\rho)$
- $N = \frac{\rho}{1-\rho} = \frac{\frac{\lambda}{\mu}}{1-\frac{\lambda}{\mu}} = \lambda/(\mu-\lambda)$

  Is N really the average queue size?

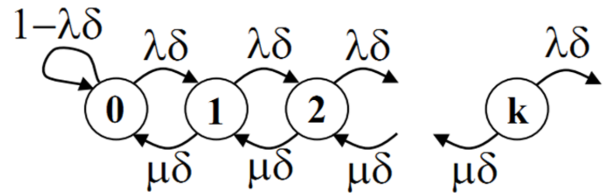
M/M/1
Equilibrium Analysis



- Goals
  - What's the average queue size?
  - **What's the average waiting time?**

- Get T from N using Little's formula
- $N = \frac{\lambda}{\mu-\lambda} = \lambda T \Rightarrow T = 1/(\mu - \lambda)$
- Average waiting time
  - $T_w = T - T_s = \frac{1}{\mu-\lambda} - \frac{1}{\mu} = \frac{\rho}{\mu(1-\rho)}$

M/M/1 Equilibrium Analysis



- Goals
  - **What's the average queue size?**
  - What's the average waiting time?

- Average number of clients waiting in queue
  - $N_w = \lambda T_w = \lambda(T - T_s) = \lambda\left(\frac{1}{\mu-\lambda} - \frac{1}{\mu}\right) = N - \rho$

M/M/1
$N = f(\rho)$

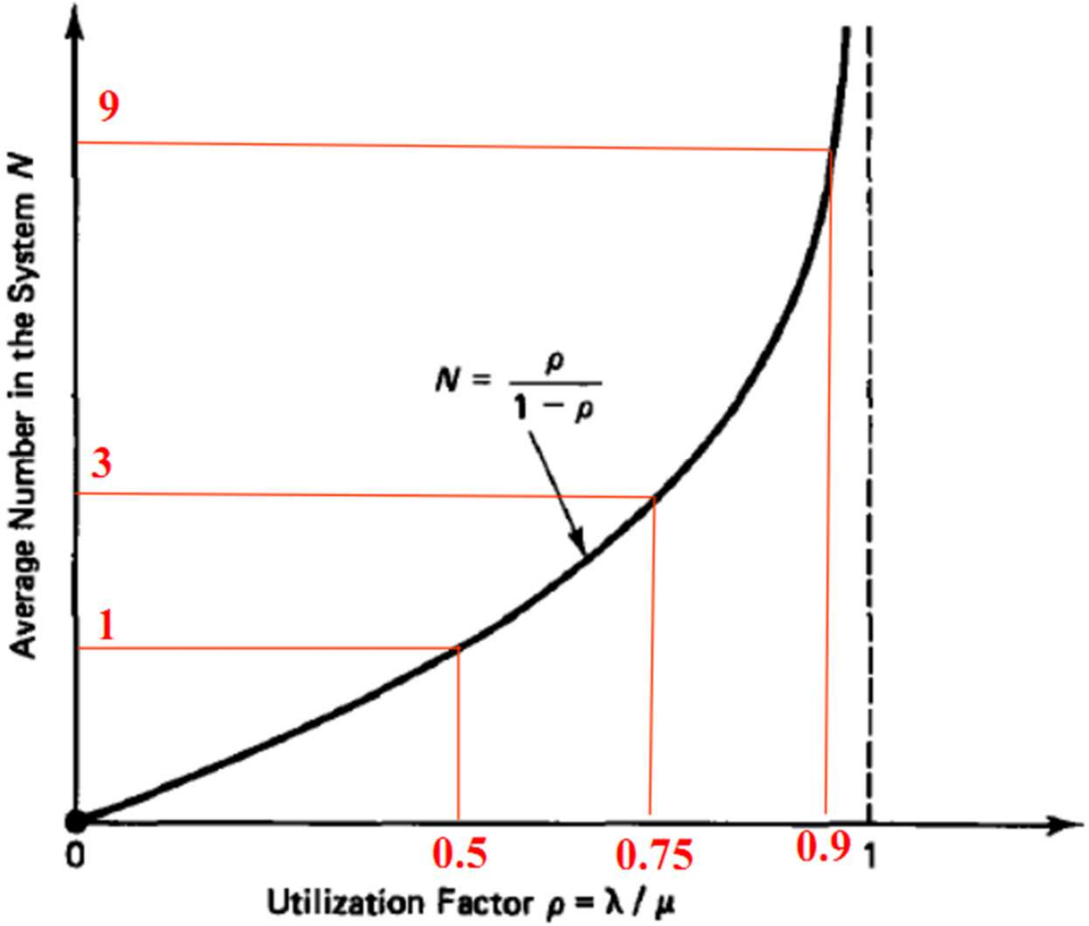


Figure 3.6 The average number in the system versus the utilization factor in the $M/M/1$ system. As $\rho \to 1$, $N \to \infty$.

M/M/1
$N = f(\rho)$

- M/M/1 with $\rho = 0.9 \Rightarrow N = 9$

- Why should customers have to wait if server is busy only 90% of the time (ρ)?

- What would happen for D/D/1, $\rho = 0.9$?
  (D::deterministic)

M/M/1 for computer networks: the more the merrier

- Example
  - 100 packets/s transmitted through link
  - Poisson process arrival
  - Packet lengths exponentially distributed
    - E[L] = 10^4 bit / packet
  - Link capacity C=10 Mbit/s
  - Then
    - $\lambda = 100\, packets/s;\ \mu = \frac{C}{E[L]} = \frac{10^7}{10^4} = 10^3\, packets/s$
    - $\rho = \frac{\lambda}{\mu} = 0.1\ ;\ N = \frac{\rho}{1-\rho} = 1/9\ ;\ T = \frac{N}{\lambda} = 1/900s$
- Now assume increased packets and link capacity
  - $\lambda' = 10\lambda,\ C' = 10C \Rightarrow \mu' = \frac{10C}{E[L]} = 10\mu$
  - Then $\rho' = \rho, N' = N$ but $T' = \frac{N'}{\lambda'} = T/10$
  - The average delay decreases => statistical multiplexing

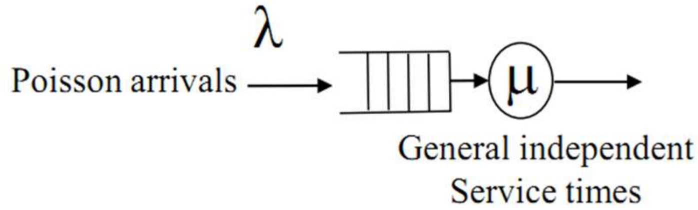
Other queue models

M/M/1/B
Queue size

- This is an M/M/1 queue with limited capacity B (queue size)
  - Packets can be lost
  - Packet lost probability, queue full :: P(B)

- Similar equilibrium analysis to M/M/1
  - $P(n) = \rho^n P(0)\,; P(0) = \frac{1-\rho}{1-\rho^{B+1}} \Rightarrow P(B) = \frac{(1-\rho)\rho^B}{1-\rho^{B+1}}$
- Particular cases
  - $\rho = 1 \Rightarrow P(B) = \frac{1}{B+1}$
  - $\rho \gg 1 \Rightarrow P(B) \sim \frac{\rho-1}{\rho} = (\lambda - \mu)/\mu$

M/G/1
Generic service

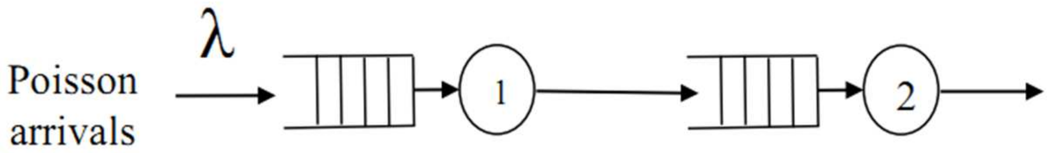


- Poisson arrivals
- Service time has arbitrary distribution
  - With given E[X] and E[X^2]
  - Assume IID (independent, identically distributed)
  - E[service time] = E[X] = $1/\mu$
  - Single server queue

M/G/1
P-K Formula

- P-K :: Pollaczek-Khinchin
- $T_w = \frac{\lambda E[X^2]}{2(1-\rho)}$
  - $\rho = \frac{\lambda}{\mu} = \lambda E[X]$ is the line utilization
- From Little's Theorem
  - $N_w = \lambda T_w;\ T = T_w + E[X] = T_w + 1/\mu$
  - $N = \lambda T = \lambda \left(T_w + \frac{1}{\mu}\right) = N_w + \rho$

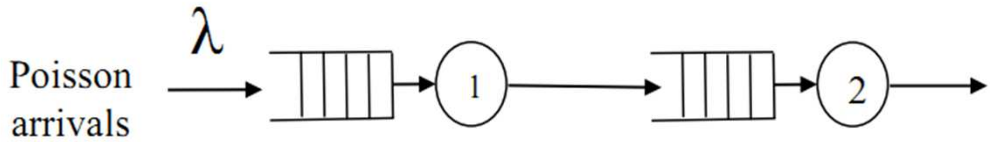
TO THINK



Assume queue 1 is M/M/1.

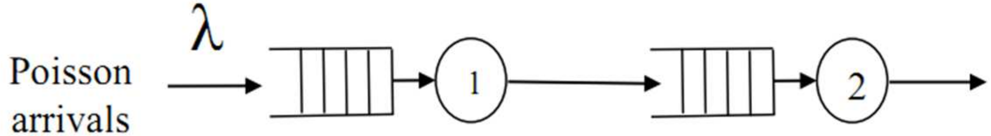
Is the arrival at queue 2 Poisson?

Networks of transmission lines

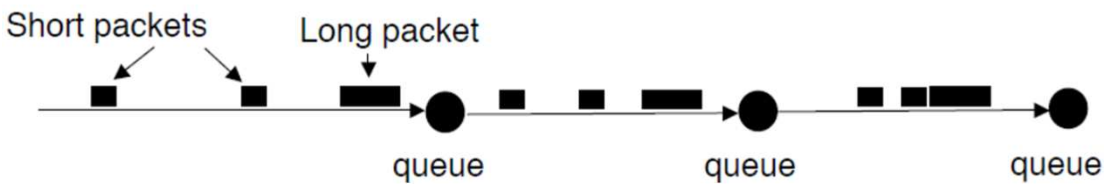


- Case 1
  - Arrival to Q1 is Poisson λ
  - Assume constant packet length
  - Q1 is M/D/1
  - Arrival rate is **not** Poisson
  - If $\lambda_2 < \mu_2 \Rightarrow$ no waiting at Q2

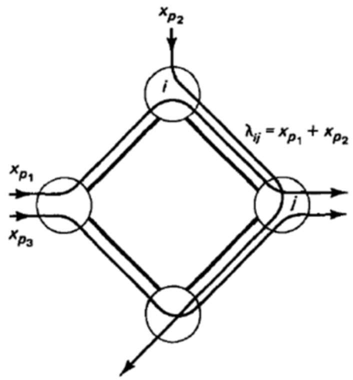
Networks of transmission lines



- Case 2
  - Q1 is M/M/1
  - Arrival at Q2 strongly related to packet length
  - longer packets require longer service
  - shorter packets catch up longer packets
  - inter-arrival time distribution changes
  - Cannot be modeled as M/M/1 (not independent!)



Kleinrock Independence Approximation

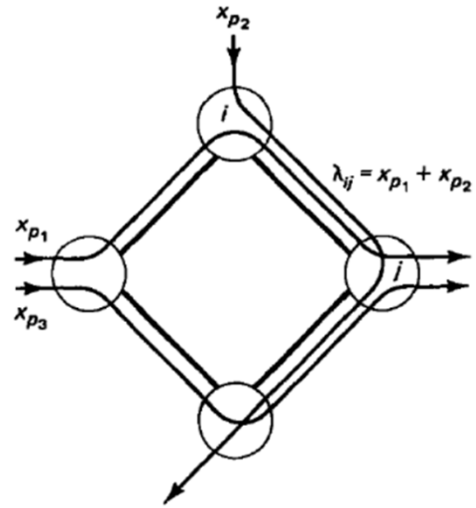


- Queuing destroys inter-arrival time (ITA) independence

- Kleinrock:
  - Merge several packet streams into same line
  - Streams are independent =>restore ITA independence
  - M/M/1 can be used to model each link

- Good approximation for:
  - Poisson streams at entry points
  - Packets length ~ exponential distribution
  - Densely connected networks
  - Moderate to heavy traffic

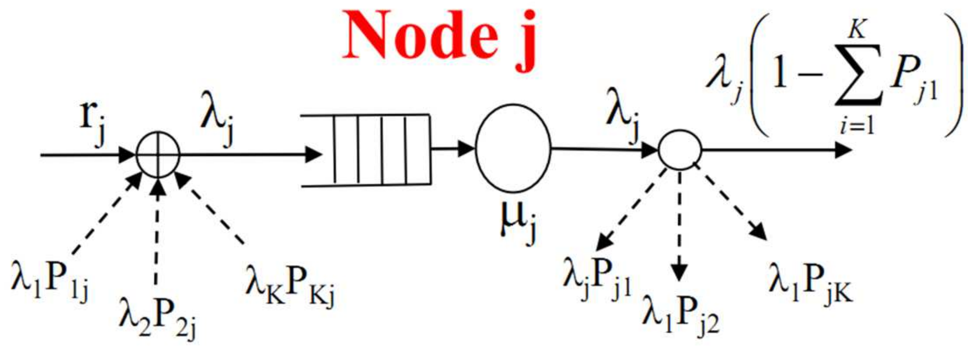
Kleinrock Independence Approximation

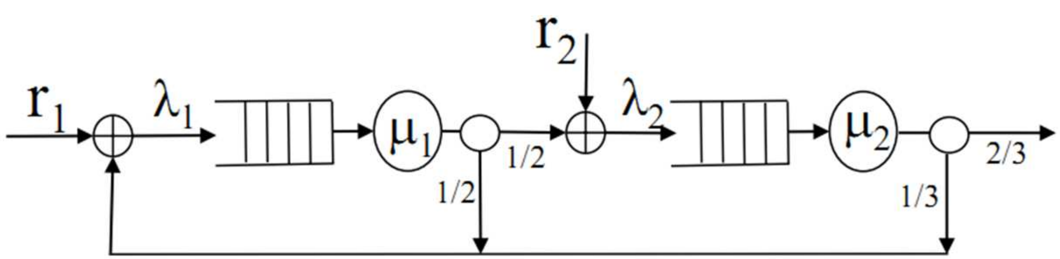


- Paths and links
- Definitions
  - x_p: arrival rate along path p
  - λ_{ij}: arrival rate to link (i,j)
  - μ_{ij}: service rate in link (i,j)
- What's independent? Link queues, M/M/1
  - $\lambda_{ij} = \sum_{all\ p\ in\ link\ (i,j)} x_p$; $\rho_{ij} = \lambda_{ij}/\mu_{ij}$; $N_{ij} = \frac{\rho_{ij}}{1-\rho_{ij}}$
  - $N = \sum_{i,j} N_{ij}$
  - $\lambda = \sum_{all\ paths\ p} x_p = total\ external\ arrival\ rate$; $\mathrm{T} = N/\lambda$

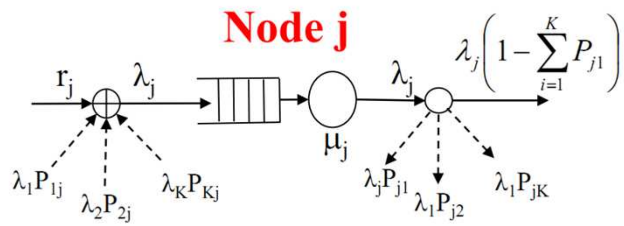
Jackson networks

- Arrivals at j
  - $\lambda_j = r_j + \sum_{i=1}^{K} \lambda_i P_{ij}$
- Ingress rate r_j
- Routing P_{ij}
  - Independent routing
  - Packet leaves i to go to j with probability P_{ij}
  - Loops are possible
  - Packet egress probability at j $P = 1 - \sum_{i=1}^{K} P_{j1}$



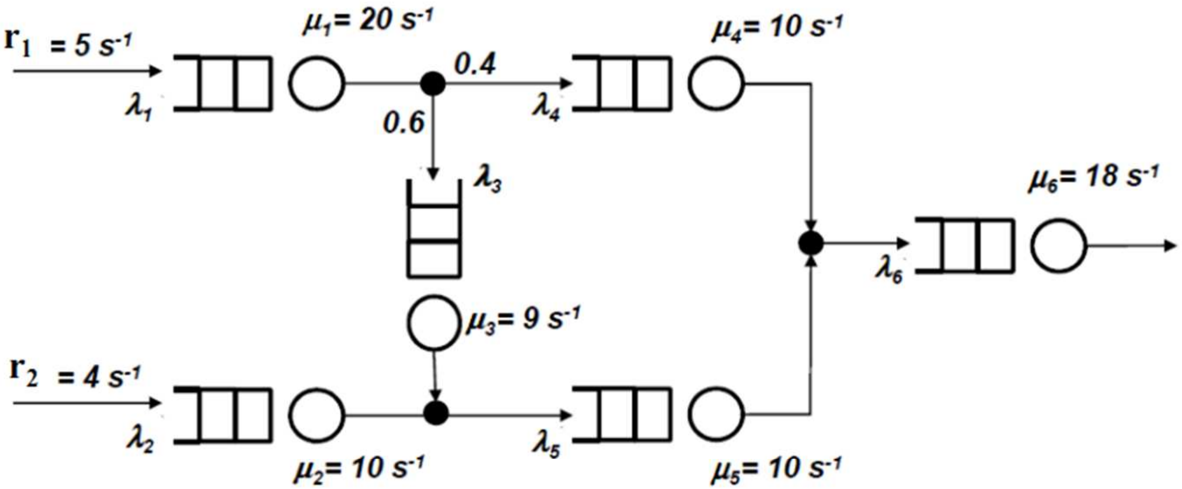


Jackson networks



- System state is defined by $n = (n_1, n_2, \dots, n_k)$
  - n_j is the number of clients in Q_j
- Jackson's theorem
  - $P(n) = \prod_{j=1}^{K} P_j(n_j) = \prod_{j=1}^{K} \rho_j^{n_j}(1-\rho_j)$ with $\rho_j = \frac{\lambda_j}{\mu_j}$
  - State of Q_j is independent of state of other queues
  - Similar to M/M/1 queues and Kleinroch's independence
- By Little's theorem:
  - $N_j = \frac{\rho_j}{1-\rho_j}$; $N = \sum_{j=1}^{K} N_j$; $\lambda = \sum_{j=1}^{K} r_j$; $T = N/\lambda$

Jackson Networks - example



$$\lambda = \sum_{i=1}^{6} r_i = 9\ \mathrm{s}^{-1}$$

$$N = \sum_{i=1}^{6} N_i = 5.08$$

Queue i	$\mathbf{r_i}$ (s^{-1})	λ_i (s^{-1})	μ_i (s^{-1})	$\rho_i = \lambda_i/\mu_i$	$\mathbf{N_i} = \rho_i/(1-\rho_i)$
1	5	5	20	0.25	0.33
2	4	4	10	0.40	0.67
3	-	3	9	0.33	0.50
4	-	2	10	0.20	0.25
5	-	7	10	0.70	2.33
6	-	9	18	0.50	1

$$T = \frac{N}{\lambda} = \frac{5.08}{9} = 0.56\ \mathrm{s}$$

HOMEWORK

- Review slides
- Read:
  - Bertsekas – 3.1, 3.2, 3.3, 3.5, 3.6, 3.8
- Do your Moodle homework