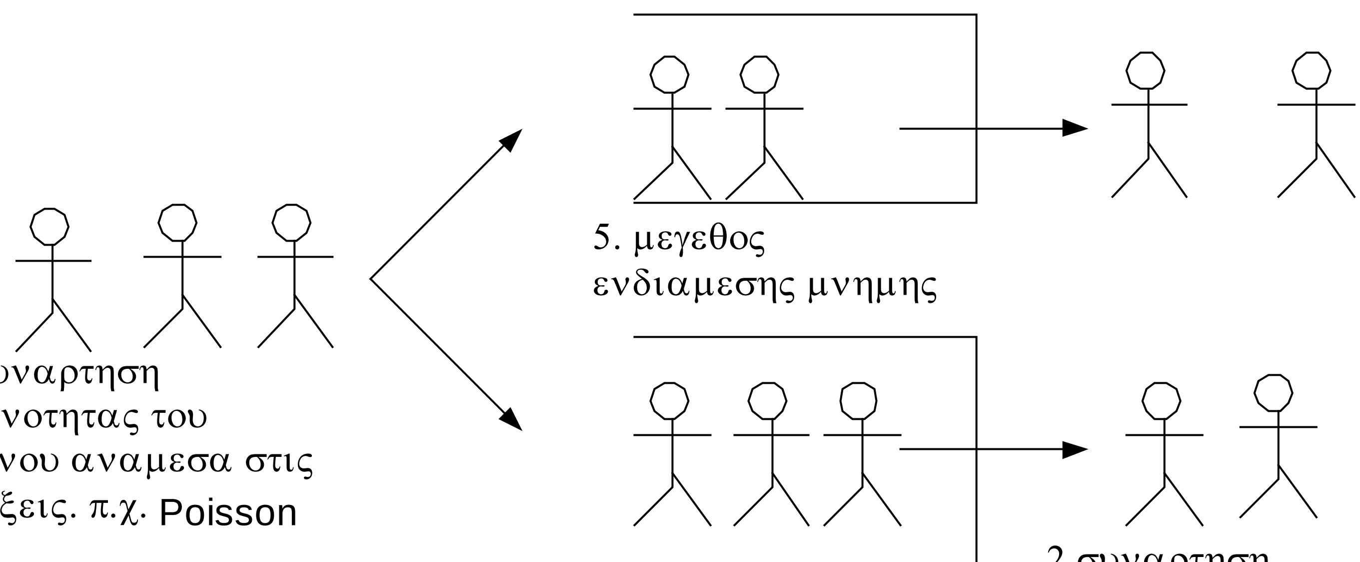
Εισαγωγή στην θεωρία ουρών.

Εισαγωγή στη Θεωρία Ουρών/Queuing Theory.

- Από τα πιο ισχυρά μαθηματικά εργαλεία για την εκτέλεση ποσοτικών αναλύσεων.
- Αρχικά αναπτύχθηκε για ανάλυση της στατιστικής συμπεριφοράς των συστημάτων μεταγωγής τηλεφώνου/telephone switching systems αλλά έχει εφαρμογές σε πολλά προβλήματα της δικτύωσης υπολογιστών.

Συστήματα Ουρών

Μπορουν να χρησιμποποιηθουν για την μοντελλοποιηση διεργασιων, στις οποιες οι πελατες γτανουν, περιμενουν την σειρα τους για εξυπηρετηση, εξυπηρετουνται και αναχωρουν.



Για να αναλυθεί ένα σύστημα πρέπει να ειναί γνωστά:

- η συνάρτηση πυκνότητας πιθανότητας (probability density function) άφιξης και η συνάρτηση πυκνότητας πιθανότητας εξυπηρέτησης (1,2).
- ο αριθμός των μονάδων εξυπηρέτησης (3).
- ο τρόπος εξυπηρέτησης (4).
- μέγεθος ενδιάμεσης μνήμης (5).

Θα συγκεντρωθούμε στα συστήματα με άπειρο χώρο μνήμης, μια μονάδα εξυπηρέτησης, FIFO τρόπο εξυπηρέτησης.

Συμβολισμός A/B/m/K/M

- A-πυκνότητα πιθανότητας των χρηστών μεταξύ των αφίξεων.
- B-πυκνότητα πιθανότητας του χρόνου εξυπηρέτησης .
- m-αριθμός των μονάδων εξυπηρέτησης.
- K- χωριτικότητα   capacity
- M- Πληθυσμός   population

Arrival Process / Service Time / Servers / Max Occupancy

Interarrival times τ

M = exponential

D = deterministic

G = general

Arrival Rate:

$\lambda = 1/ E[\tau]$

Service times X

M = exponential

D = deterministic

G = general

Service Rate:

$\mu = 1/ E[X]$

1 server

c servers

infinite

K customers

unspecified if unlimited

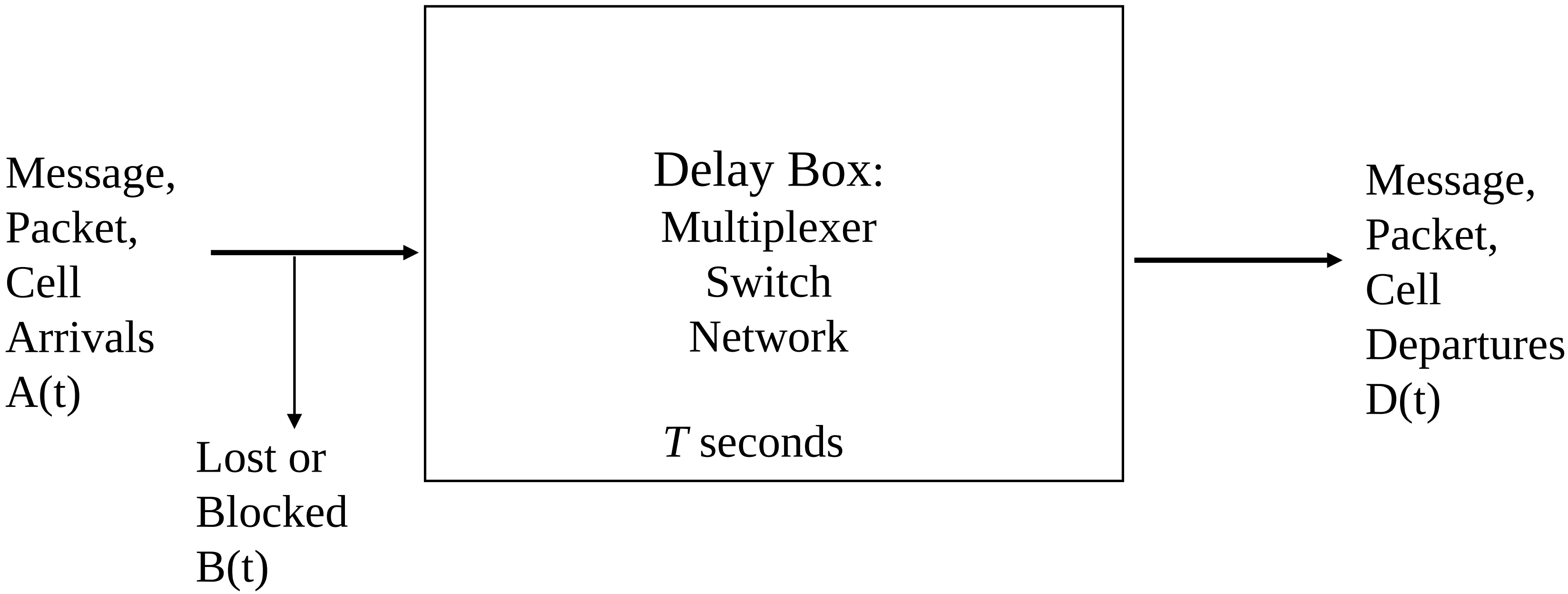
Multiplexer Models:	M/M/1/K, M/M/1, M/G/1, M/D/1
Trunking Models:	M/M/c/c, M/G/c/c
User Activity:	M/M/∞, M/G/ ∞

Figure A.7

Είδη Ουρών

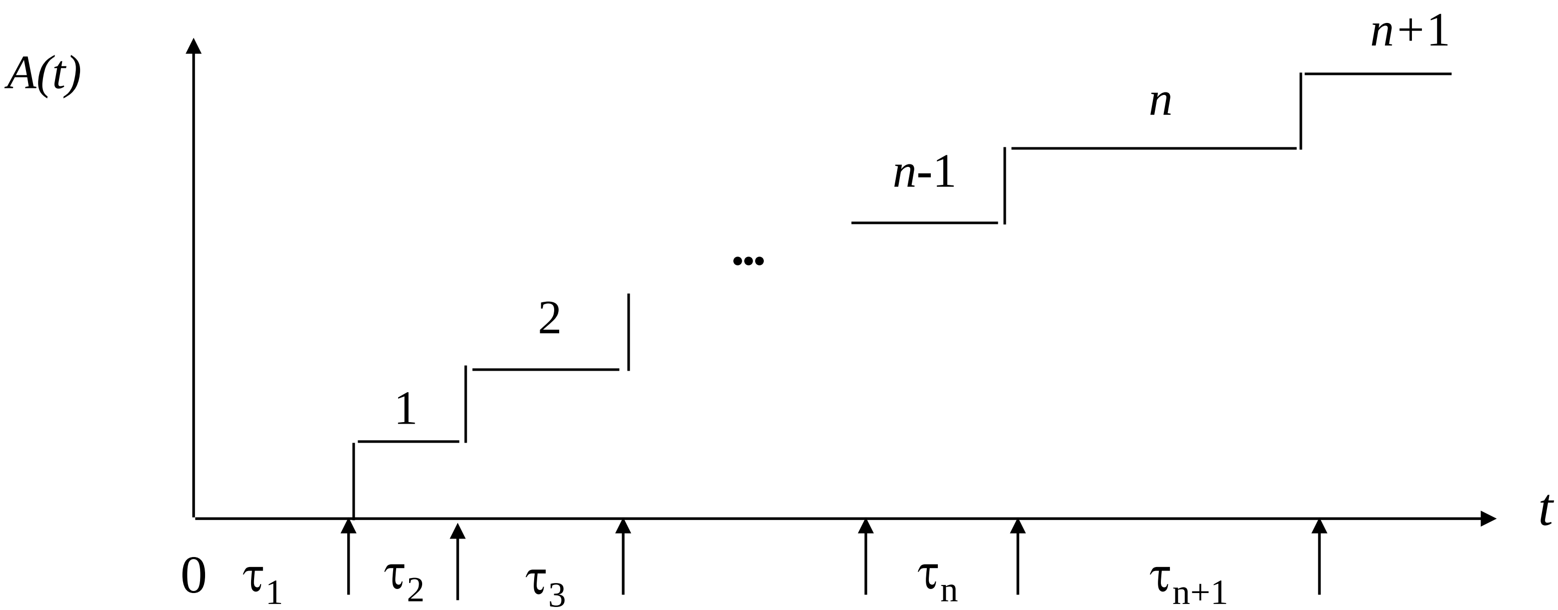
- M/M/1- για μοντελλοποίηση συστημάτων με μεγάλο αριθμό από ανεξάρτητους πελάτες (π.χ. Το τηλεφωνικό σύστημα). Τα πάντα είναι γνωστά (π.χ. Ο αριθμός πελατών στην ουρά, η μέση καθυστέρηση, κ.ο.κ) και οι λύσεις προσφέρωνται σε ακριβή αναλυτική μορφή (closed form).
- G/G/1-για μοντελλοποίηση πιο γενικών συστημάτων. Ακριβές αναλυτικές λύσεις δεν είναι γνωστες.
- M/D/1
- G/D/1

Arrival Rates and Traffic Load



Number of users in system $N(t) = A(t) - D(t) - B(t)$

Figure A.1



Time of nth arrival = $\tau_1 + \tau_2 + \ldots + \tau_n$

$$\text{Arrival Rate} = \frac{n \text{ arrivals}}{\tau_1 + \tau_2 + \ldots + \tau_n \text{ seconds}} = \frac{1}{(\tau_1+\tau_2+\ldots+\tau_n)/n} \longrightarrow \frac{1}{E[\tau]}$$

Arrival Rate = 1 / mean interarrival time

Figure A.2

Little’s Law

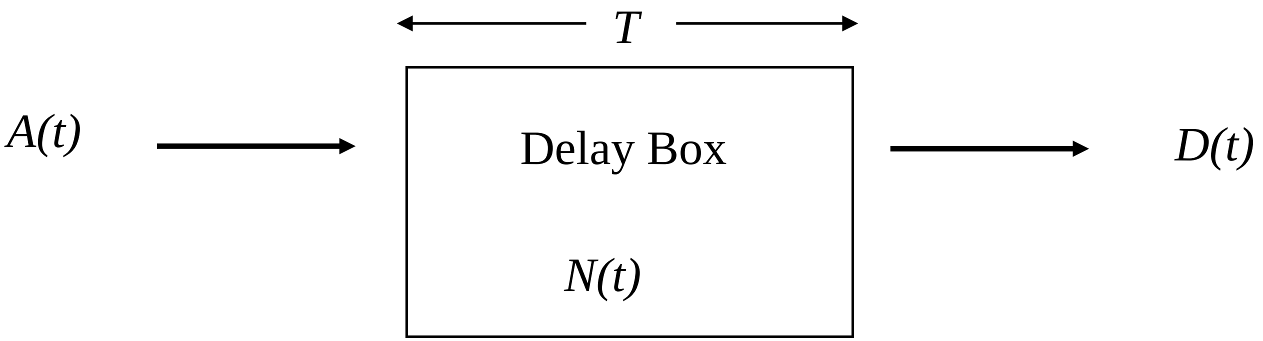


Figure A.3

Little’s Law

Assumes
first-in
first-out

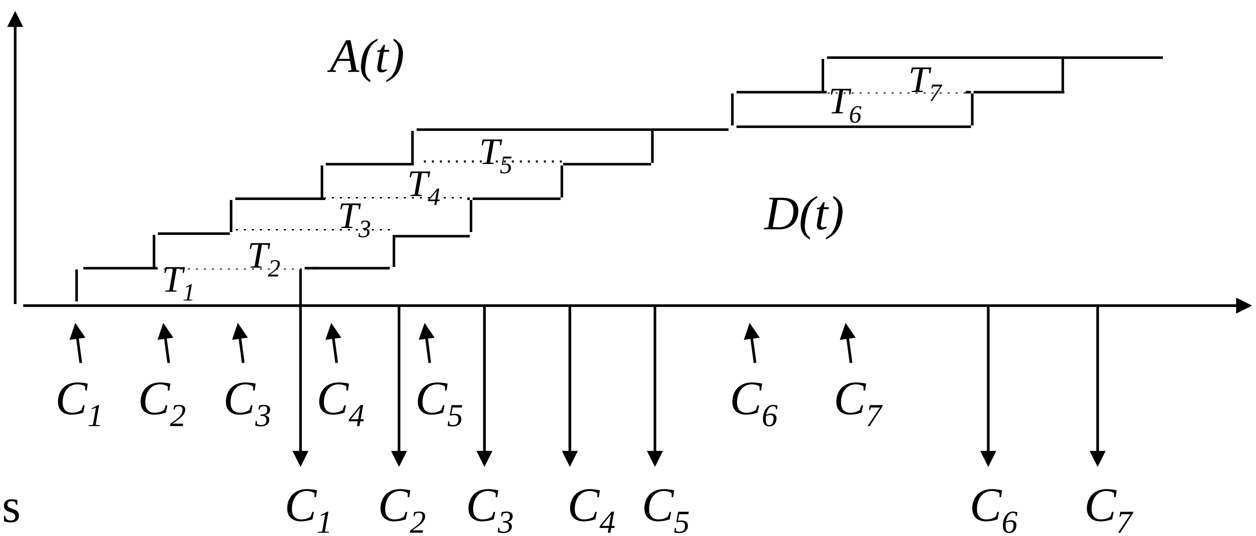


Figure A.4

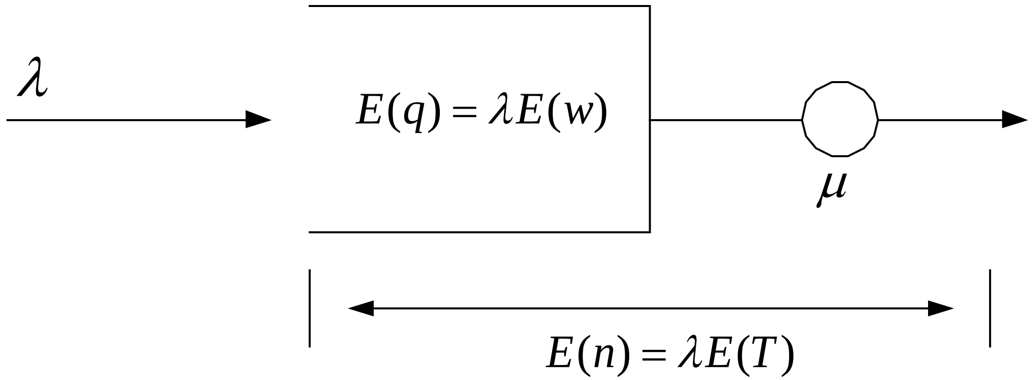
Little's Formula

A queuing system with arrival rate λ, mean delay E(T) through the system and an average queue length E(n) is governed by Little's Formula:

$$E(n) = \lambda E(T)$$

If we consider a system where customers will be blocked then

$$E(n) = \lambda(1-P_b)E(T)$$



E(T)	=	E(w)	+	1/μ
Average time delay		Average wait time		Average service time

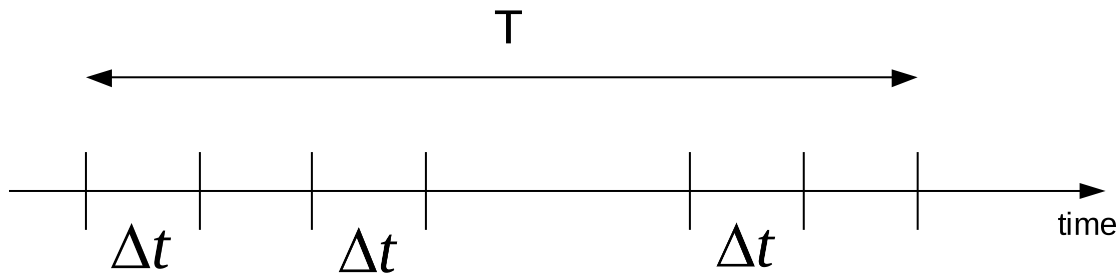
The average number of customers E(q) waiting in the queue is:

$$E(q) = \lambda E(w) = \lambda E(T) - \frac{\lambda}{\mu} = E(n) - \rho$$

Arrival Processes

- *Deterministic* – when interarrival times are all equal to the same constant
- *Exponential* – when the interarrival times are exponential random variables with mean $E[\tau] = 1/\lambda$
- $P[\tau > t] = e^{-t/E[\tau]} = e^{-\lambda t}$ for $t > 0$

Poisson Process



Consider a small interval $\Delta t (\Delta t \rightarrow 0)$:

1. The probability of one arrival in the interval Δt is defined to be $\lambda\Delta t + o(\Delta t)$, $\lambda \Delta t \ll 1$ and λ is a specified proportionality constant.
2. The probability of zero arrivals in Δt is $1-\Delta t + o(\Delta t)$.
3. Arrivals are memoryless: An arrival (event) in one time interval of length Δt is independent of events in previous or future intervals.

Poisson Distribution

Taking a larger finite time interval T one can find the probability of k arrivals in T:

$$p(k) = (\lambda T)^k \cdot \frac{e^{-\lambda T}}{k!}$$

The mean or expected value of k arrivals:

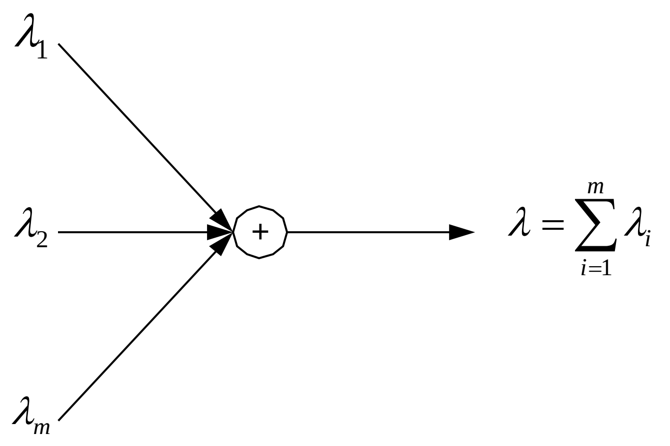
$$E(k) = \lambda T$$

The variance is:

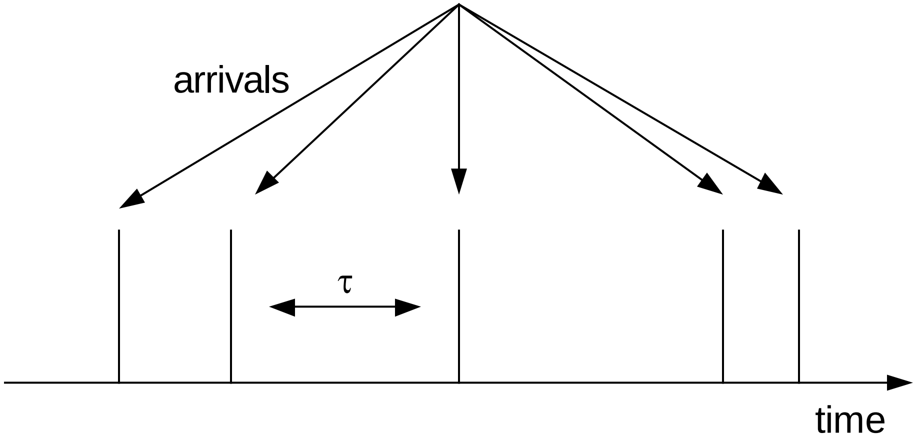
$$\sigma^2_{(k)} = E(k) = \lambda T$$

Distribution Conservation

- If there are m independent Poisson process streams of arbitrary arrival rates, $\lambda_1, \lambda_2, \ldots \lambda_m$, and these are merged , the composite stream , is itself a Poisson process with parameter $\lambda = \sum \lambda_i$.
- Sums of Poisson processes are distribution conserving. They retain the Poisson property.



Time between successive arrivals, τ

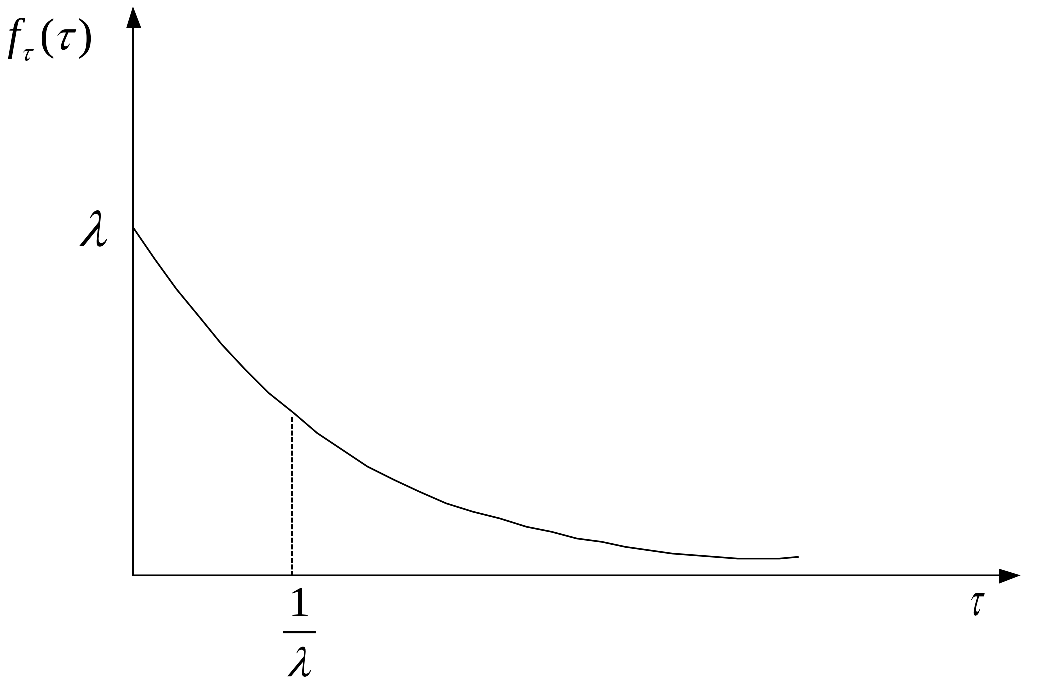


The time between successive arrivals, τ, is an exponentially distributed random variable i.e. its probability density function is as follows:

$$f_{\tau}(\tau) = \lambda e^{-\lambda\tau} \qquad \tau \geq 0$$

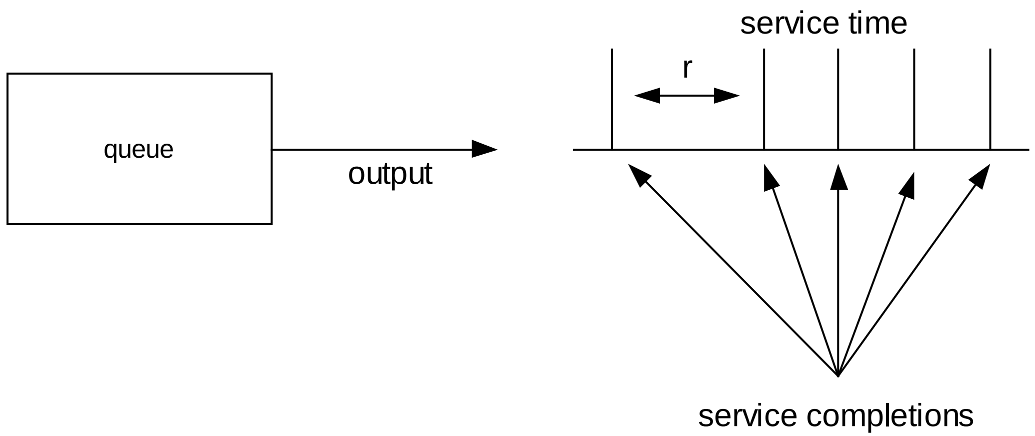
$$E(\tau) = 1/\lambda \qquad Var(\tau) = 1/\lambda^2$$

Time between successive arrivals



For Poisson arrivals, the time between arrivals is more likely to be small than large. The probability between 2 successive events decreases exponentially with the time τ between them.

Service Process



Following similar arguments as for the arrival process, it can be observed that the service process is the complete analogue of the arrival process. For the case where r, the time between completions, is exponentially distributed with mean value $1/\mu$, the completion times themselves must represent a Poisson Process.

M/M/1 Queue

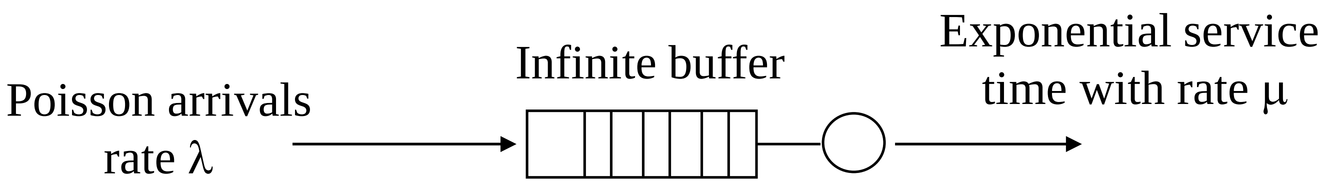
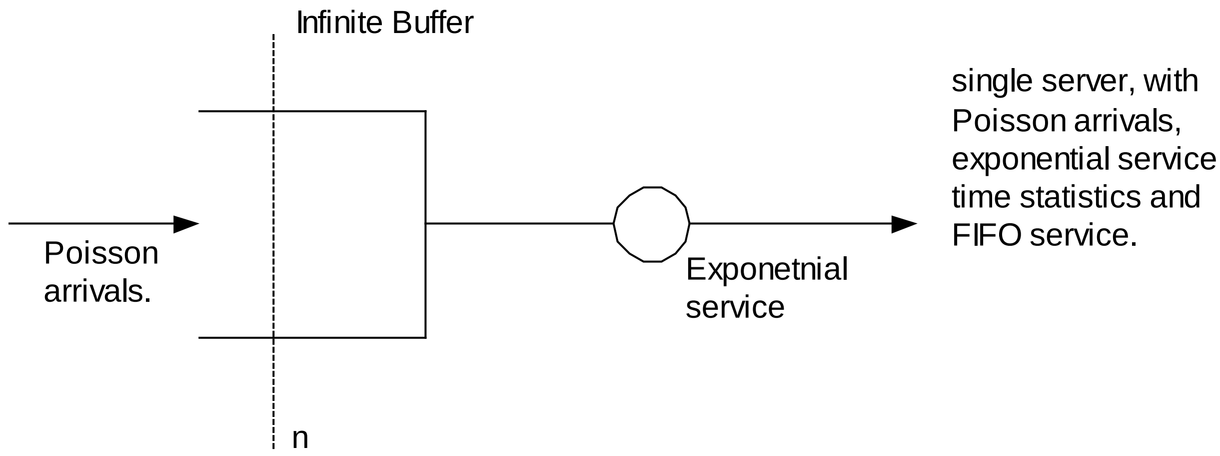
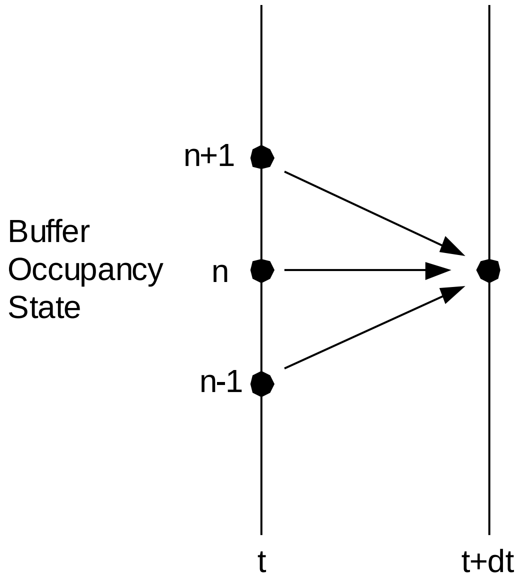


Figure A.9

The M/M/1 Queue.



The aim is to find the probability of state n at the queue as a function of time (Pn(t)). The probability Pn ($t+\Delta t$) that the queue is in state n at time $t+\Delta t$ must be the sum of the mutually exclusive probabilities that the queue was in states n-1, n, n+1 at time t, each multiplied by the independent probability of arriving at state n in the intervening Δt units of time.



$$P_n(t+\Delta t) = P_n\ (t)[(1-\lambda\Delta t)(1-\mu\Delta t) + \mu\Delta t\lambda\Delta t + o(\Delta t)$$
$$+ P_{n-1}(t)[\lambda\Delta t(1-\mu\Delta t) + o(\Delta t)] + P_{n+1}(t)[(1-\lambda\Delta t)\mu\Delta t + o(\Delta t)]$$

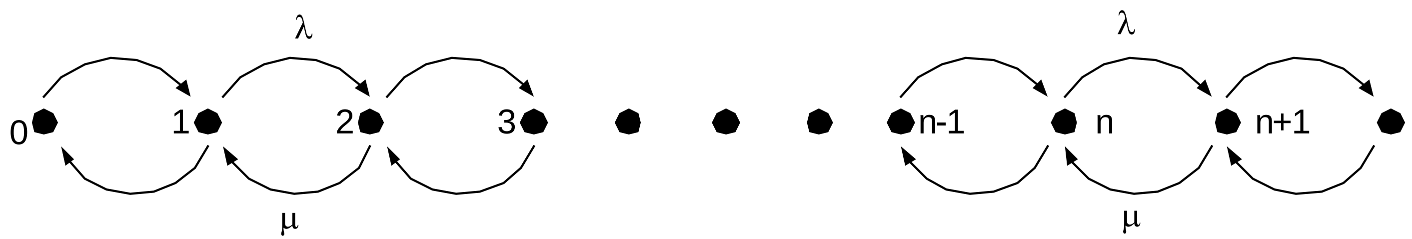
Simplifying, dropping o(Δt) and expanding as a Taylor series about t a Differential-Difference equation can be derived:

$$\frac{dP_n(t)}{dt} = -(\lambda+\mu)P_n(t) + \lambda P_{n-1}(t) + \mu P_{n+1}(t)$$

In steady state:

$$(\lambda+\mu)P_n = \lambda P_{n-1} + \mu P_{n+1}$$

Deriving the Equation using Balance Equations



$$(\lambda+\mu)P_n \quad = \quad \lambda P_{n-1} \quad + \quad \mu P_{n+1}$$

rate of leaving state n given the systems was in state n with probability Pn	rate of entering state n from state n-1	rate of entering state n from state n+1

M/M/1 Queue
State diagrams

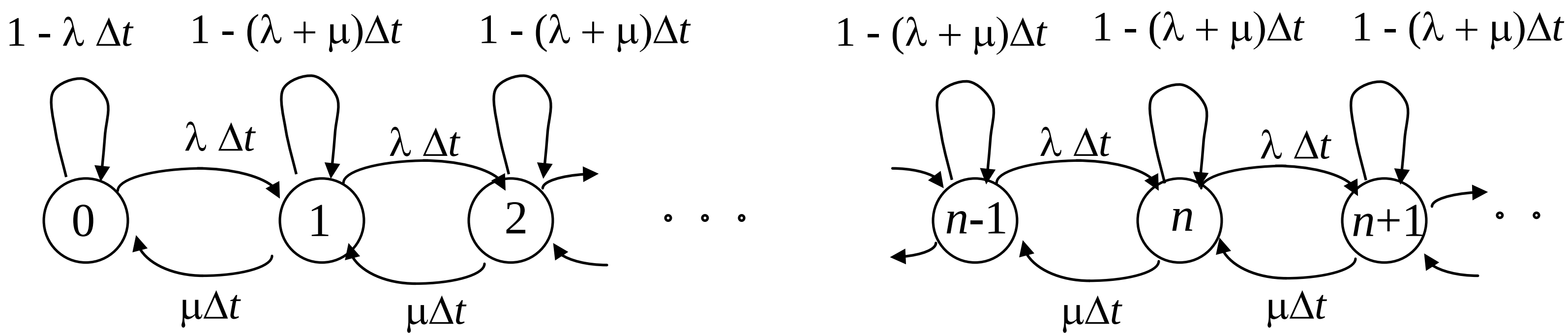
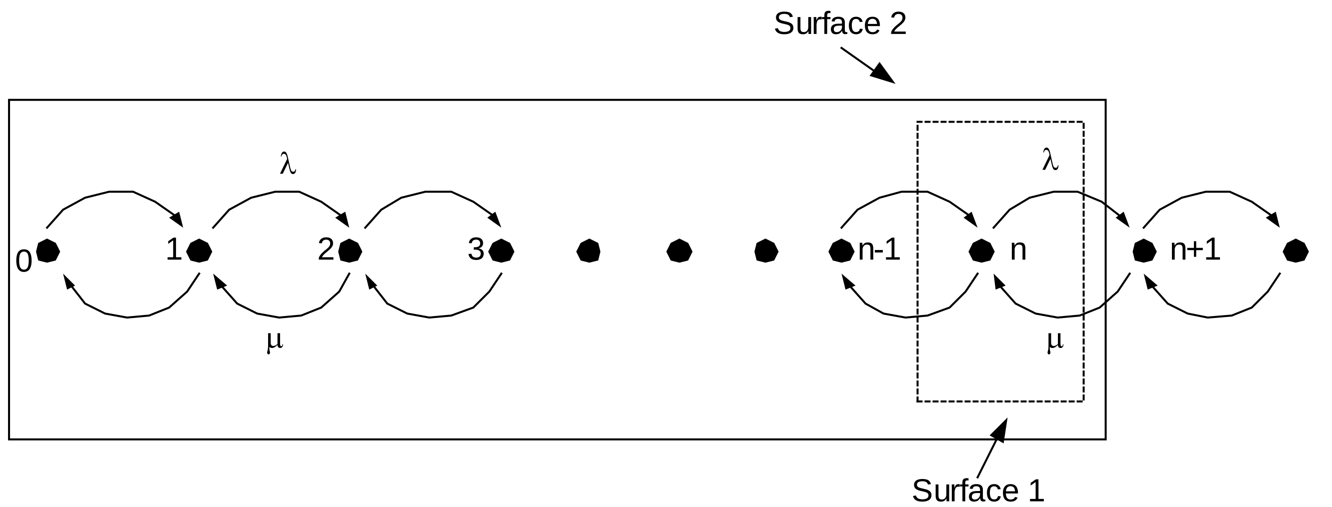


Figure A.10

Solution using the Flow Balance Diagram



Equating input and output flux around:

- Surface 1:

$$(\lambda + \mu)P_n = \lambda P_{n-1} + \mu P_{n+1}$$

- Surface 2:

$$\mu P_{n+1} = \lambda P_n$$

Solving recursively:

$$P_1 = \frac{\lambda}{\mu} P_0 = \rho P_0$$

$$P_2 = \rho . \rho . P_0$$

$$P_n = \rho^n P_0$$

where $\frac{\lambda}{\mu} = \rho$ is the line utilization or traffic intensity.

By utilizing the probability normalization condition $\sum_n P_n = 1$:

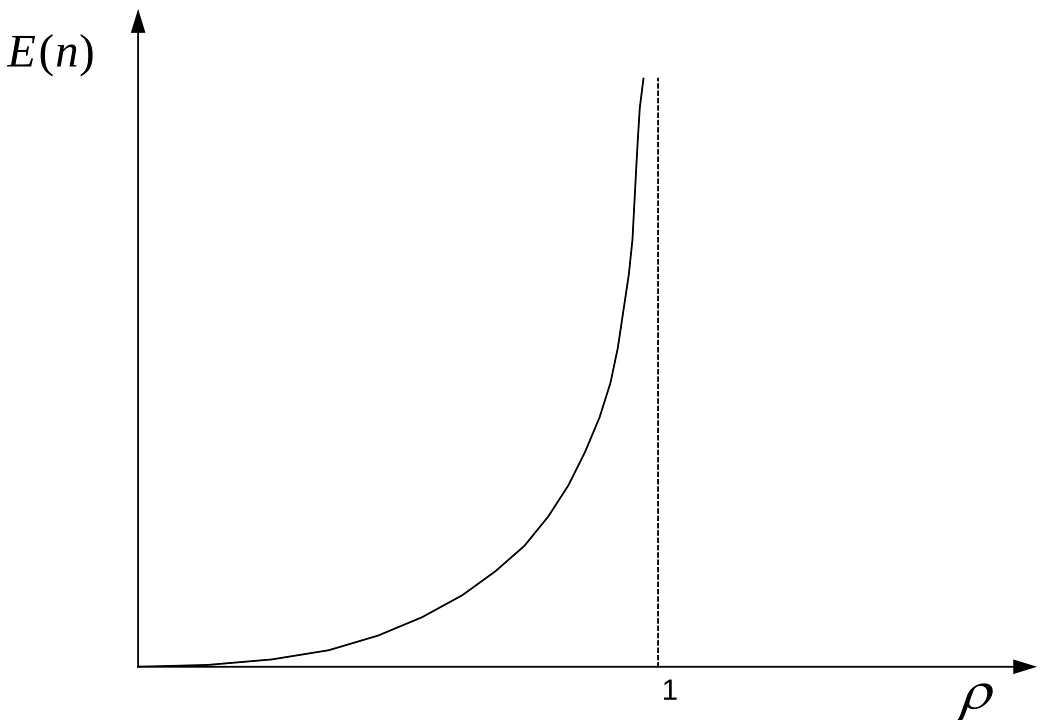
$$\Rightarrow P_0 = 1 - \rho$$

$$\Rightarrow P_n = (1 - \rho)\rho^n$$

The above distribution is called a geometric distribution and it can only be derived if ρ<1.

Expected number of customers in M/M/1 queue with infinite buffer space:

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



Extension to Finite Queues.

The queue has a finite maximum queue length N:

$$P_n = \frac{\rho^n (1-\rho)}{1-\rho^{N+1}} \qquad \rho \neq 1$$

The probability that the queue is full, which is equal to the Blocking probability is equal to:

$$P_N = \frac{\rho^N (1-\rho)}{1-\rho^{N+1}}$$

The probability that the queue is empty is equal to:

$$P_0 = \frac{1-\rho}{1-\rho^{N+1}}$$

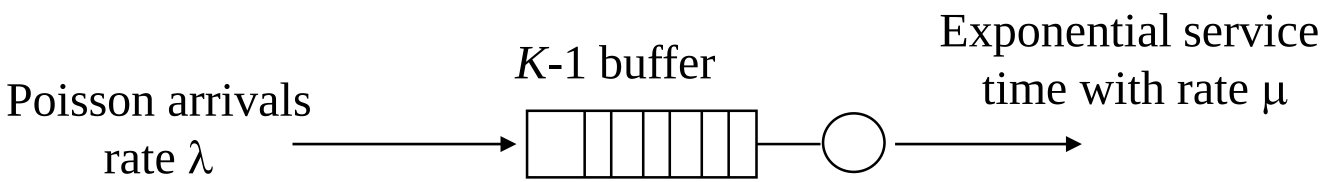
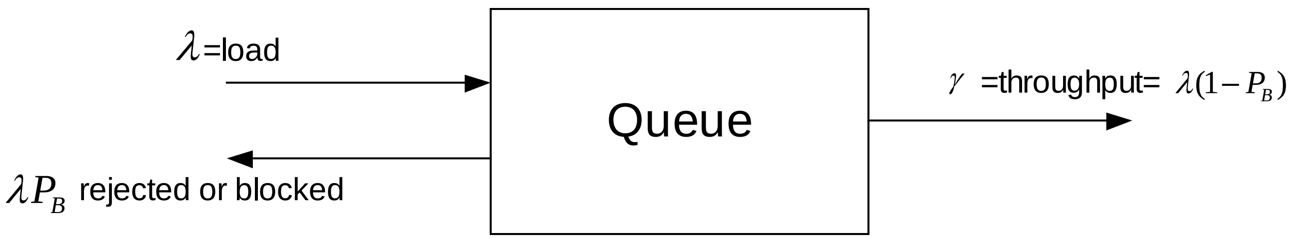

Poisson arrivals
rate λ
K-1 buffer
Exponential service
time with rate μ

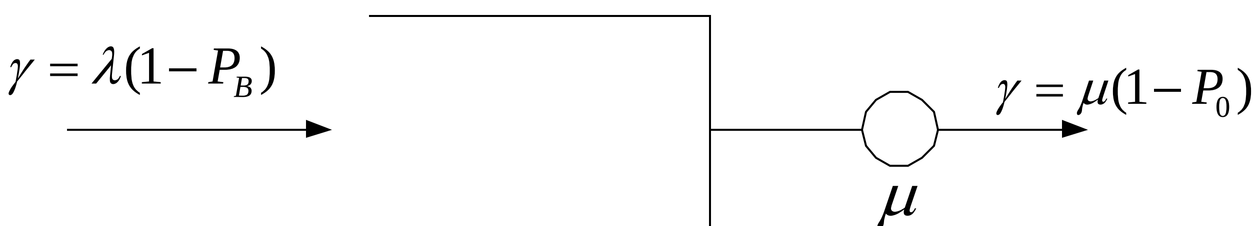
Figure A.9

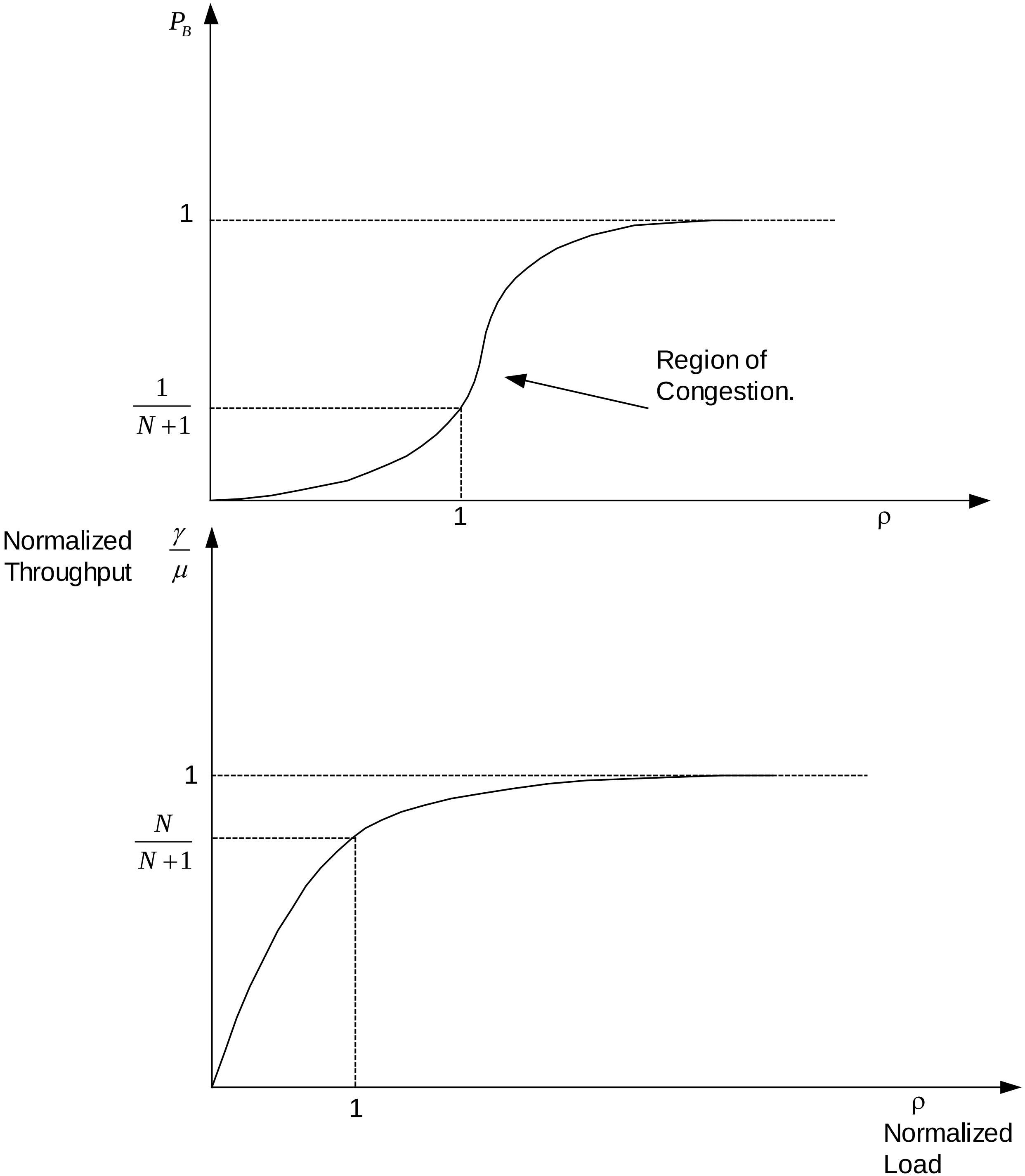
Relation between Throughput and Load



$$\gamma = \lambda(1-P_B) = \mu(1-P_0)$$

throughput | net arrival rate | net departure rate





P_B
1
$\frac{1}{N+1}$
Region of Congestion.
1
ρ
Normalized Throughput
$\frac{\gamma}{\mu}$
1
$\frac{N}{N+1}$
1
ρ
Normalized Load

Queue Performance

- As the load of the system increases the throughput increases as well.
- More customers are blocked.
- The average number of customers in the queue and thus the average wait time increases as well .
- At high loads queuing deadlocks can occur and throughput may drop to zero.
- There is a trade-off in performance.

Nonpreemptive Priority Queuing Systems

Need to provide priority in many systems:

- Computer systems
- Computer control of telephone digital switching exchanges
- Deadlock prevention in packet switching

Nonpreemptive Priority: Higher priority customers move ahead of lower priority ones in the queue but do not preempt lower priority customers already in service.

Preemptive Priority: Interrupt lower priority customers in service until all higher priority customers are served.

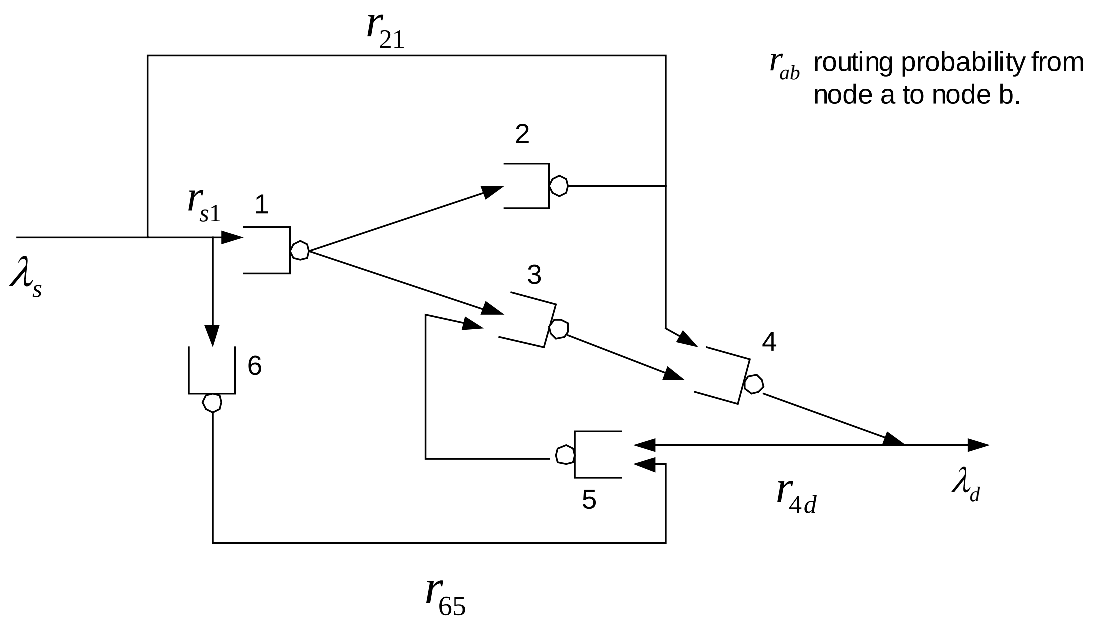
Queuing Networks

- For M/M/1 queues, models handling network of queues are relatively easy. They make use of the so called product form solution (Jackson Network). Much of the research since 1970s is devoted to these two problem areas:

  -finding conditions for which the product form solution applies.

  -developing improved and efficient algorithms for reducing the computational complexity.
- Two generic classes can be considered: open and closed queuing networks.

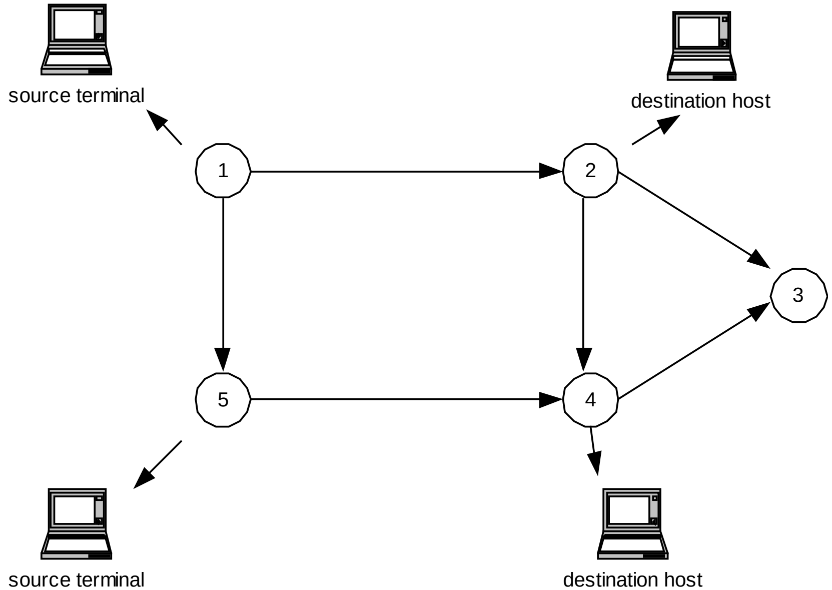
Open Queuing Networks

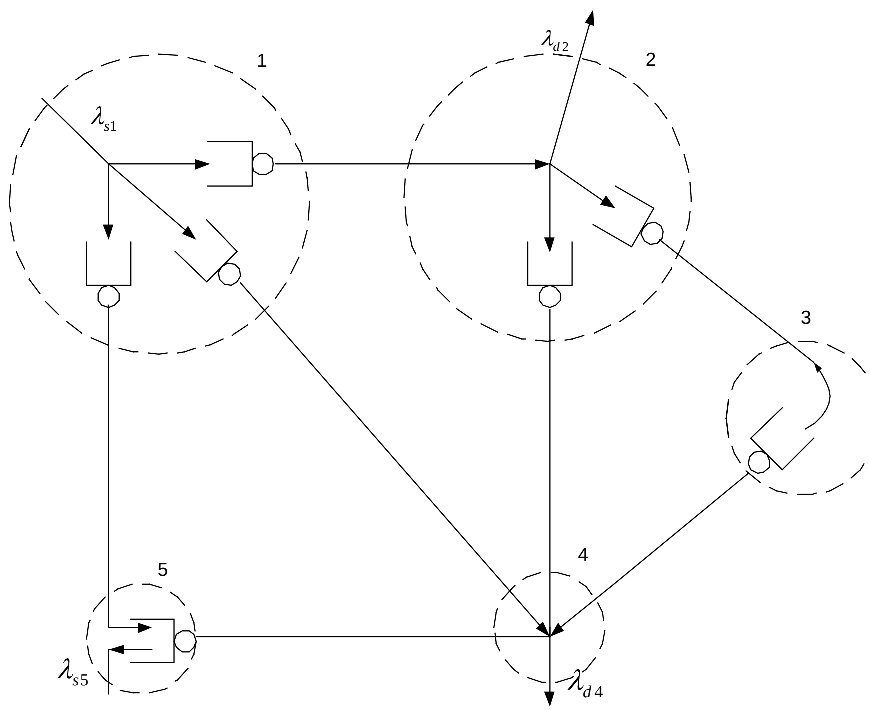


- Packets enter and leave the network without losses.
- From flow conservation principles

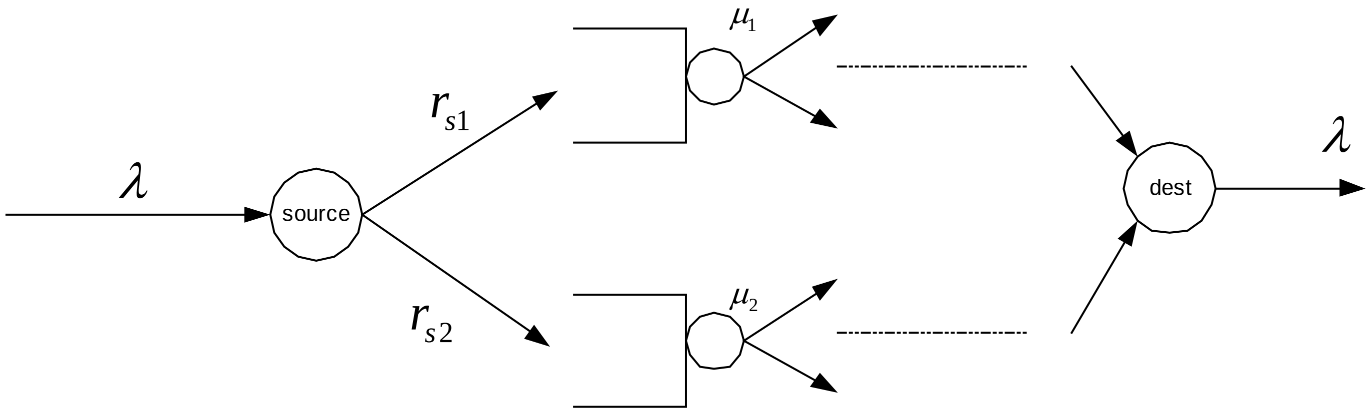
Net arrival rate= Net departure rate

$$\lambda s \quad = \quad \lambda d$$


source terminal
destination host
1
2
3
4
5
source terminal
destination host


λ_{d2}
1
2
λ_{s1}
3
4
5
λ_{s5}
λ_{d4}

Consider a portion of the network with M queues:



- The Poisson arrival rate at a source is labelled λ.
- The symbol r_{ij} represents the probability that a packet (customer) completing service at queue i is routed to queue j.
- The queue service rate at a node i is labelled μ_i .

- Normalization condition:

$$r_{id} + \sum_{j=1}^{M} r_{ij} = 1$$

- Continuity of flow:

$$\lambda_i = r_{is}\lambda + \sum_{K=1}^{M} r_{ki}\lambda_k$$

- Product form solution:

$$P(n) = \prod_{i=1}^{M} P_i(n_i) \qquad P_i(n_i) = (1 - \rho_i)\rho_i^n$$

- The various queues even though interconnected though the continuity expression behave as if they are independent. More remarkably they appear as M/M/1 queues with the familiar state probability distribution.

	M/D/1	M/ Er/1	M/M/1	M/H/1
Service Time	Constant	Erlang	Exponential	Hyperexponential
Coefficient of Variation	0	<1	1	>1
$E[W]/E[W_{\mathrm{M/M/1}}]$	1/2	1/2< , <1	1	>1

Figure A.13