Introduction to Queues and Queueing Theory

Lecture 1

History

- Queueing Theory was developed by Agner Krarup Erlang, an employee of the Danish Telephone Company.

Jan. 1, 1878 Born at Lonber in Jutland, Denmark

1892 Passed the Preliminary Examinations to the University of Copenhagen at 14 years old.

1896 Passed Entrance Exams to the University of Copenhagen

1991 Graduated with an MA in Mathematics

1908 Joined the Danish Telephone Company

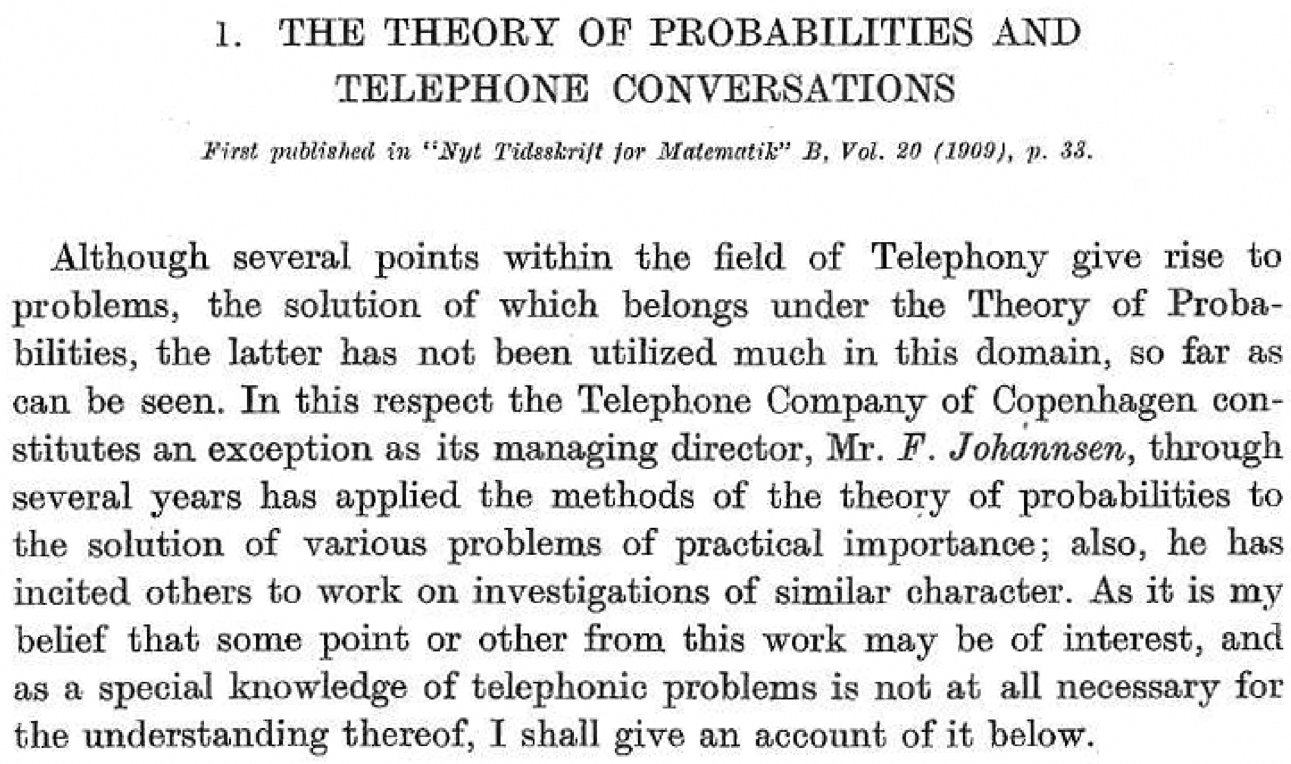
1909 Publish The Theory of Probabilities and Telephone Conversations

1917 Published Solutions of Some Problems in the Theory of Probabilities of Significance in Automatic Telephone Exchanges

Feb. 3, 1929 Died following an abdominal operation.

Applications (I)

- 1. Telephone Networks
  - How many external telephone lines would a Danish village need if the probability of a busy signal with a long-distance call is to be at most 5%?



1. THE THEORY OF PROBABILITIES AND TELEPHONE CONVERSATIONS

First published in "Nyt Tidsskrift for Matematik" B, Vol. 20 (1909), p. 33.

Although several points within the field of Telephony give rise to problems, the solution of which belongs under the Theory of Probabilities, the latter has not been utilized much in this domain, so far as can be seen. In this respect the Telephone Company of Copenhagen constitutes an exception as its managing director, Mr. *F. Johannsen*, through several years has applied the methods of the theory of probabilities to the solution of various problems of practical importance; also, he has incited others to work on investigations of similar character. As it is my belief that some point or other from this work may be of interest, and as a special knowledge of telephonic problems is not at all necessary for the understanding thereof, I shall give an account of it below.

Applications (II)

- 2. Air Traffic Control

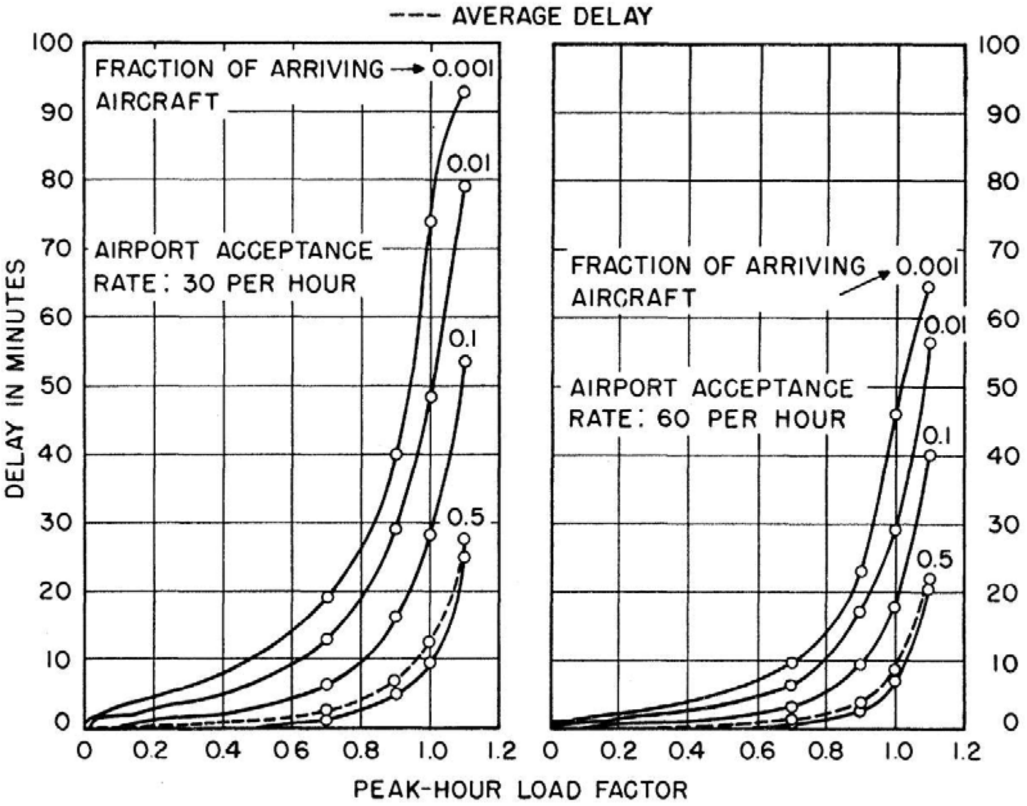
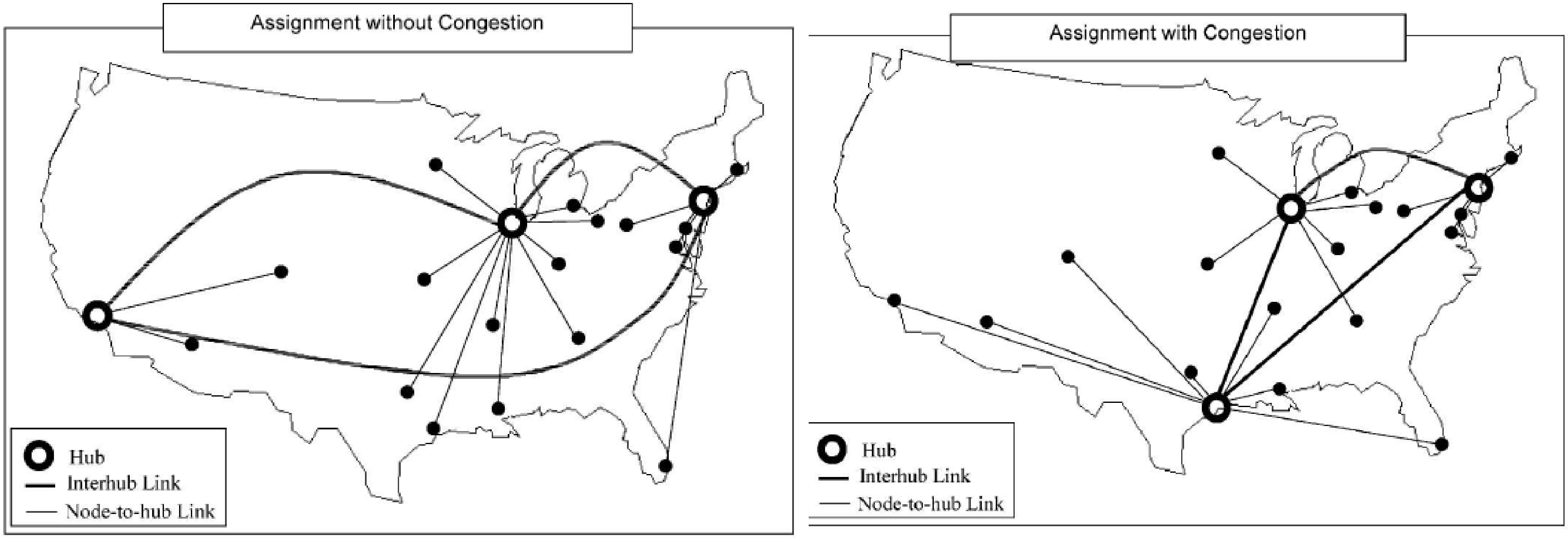


Fig. 7. Delay probabilities, as function of peak-hour load factor, at time of highest average delay.

H. P. Galliher and R. C. Wheeler, Nonstationary Queueing Probabilities for Landing Congestion of Aircraft, Operations Research, 6,264{275, (1958).

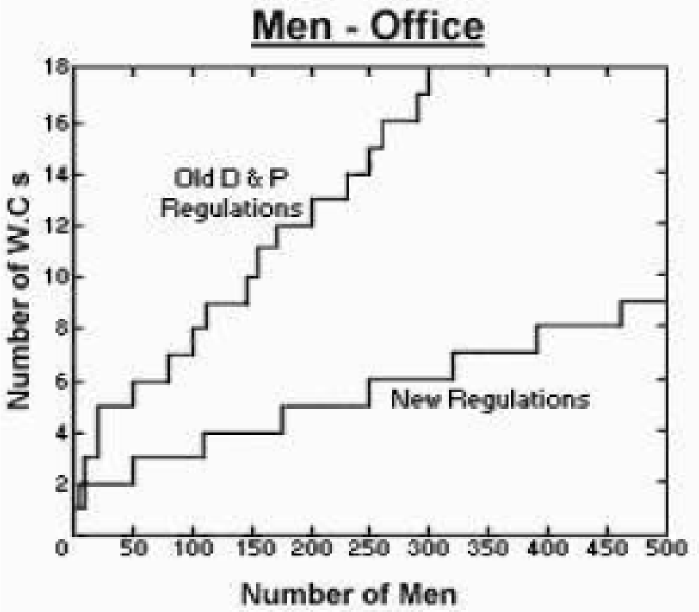


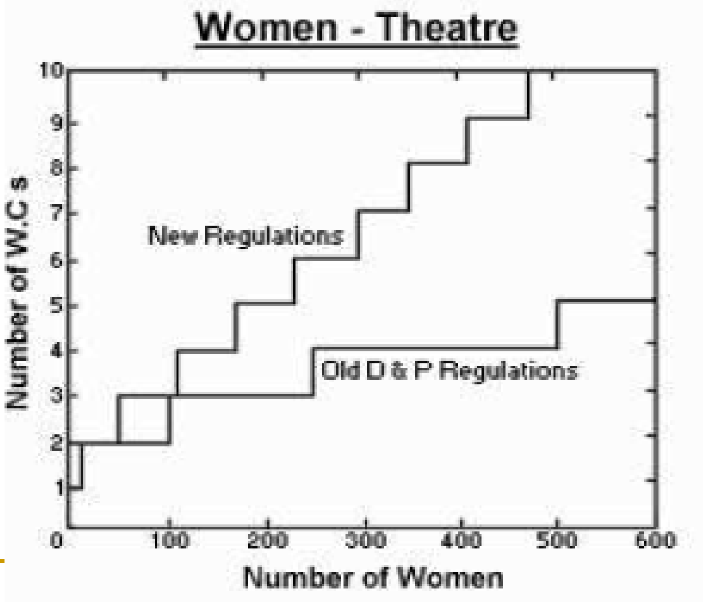
Marianova and D. Serrab, Location models for airline hubs behaving as M/D/c queues, Computers & Operations Research, 30, 983-1003 (2003).

Applications (III)

- 3. Queueing for Toilets
  - Current New Zealand Building Code stipulates that 90% of toilet users in a public building should wait less than 1 minute.
  - The number of toilets is determined by an M/M/c queue model.

D. McNickle, Queueing for Toilets, OR Insight, April-June (1998).





Applications (VI)

- 4. Highway Toll Booths

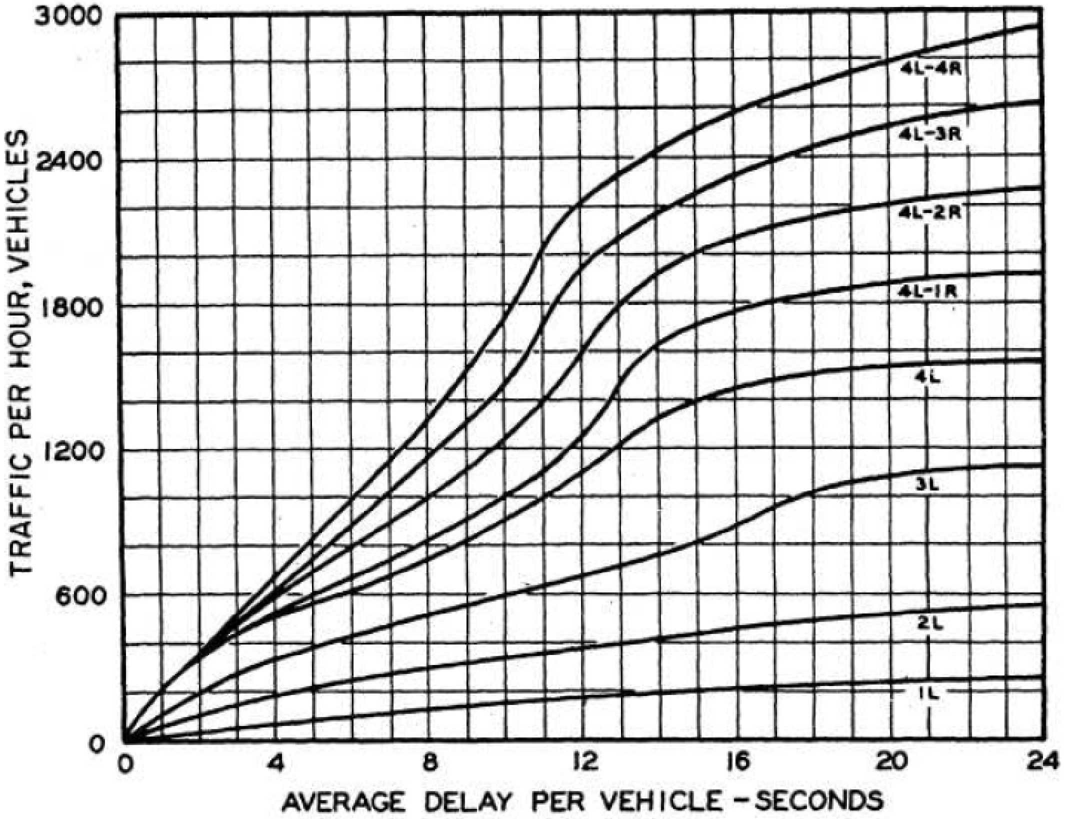
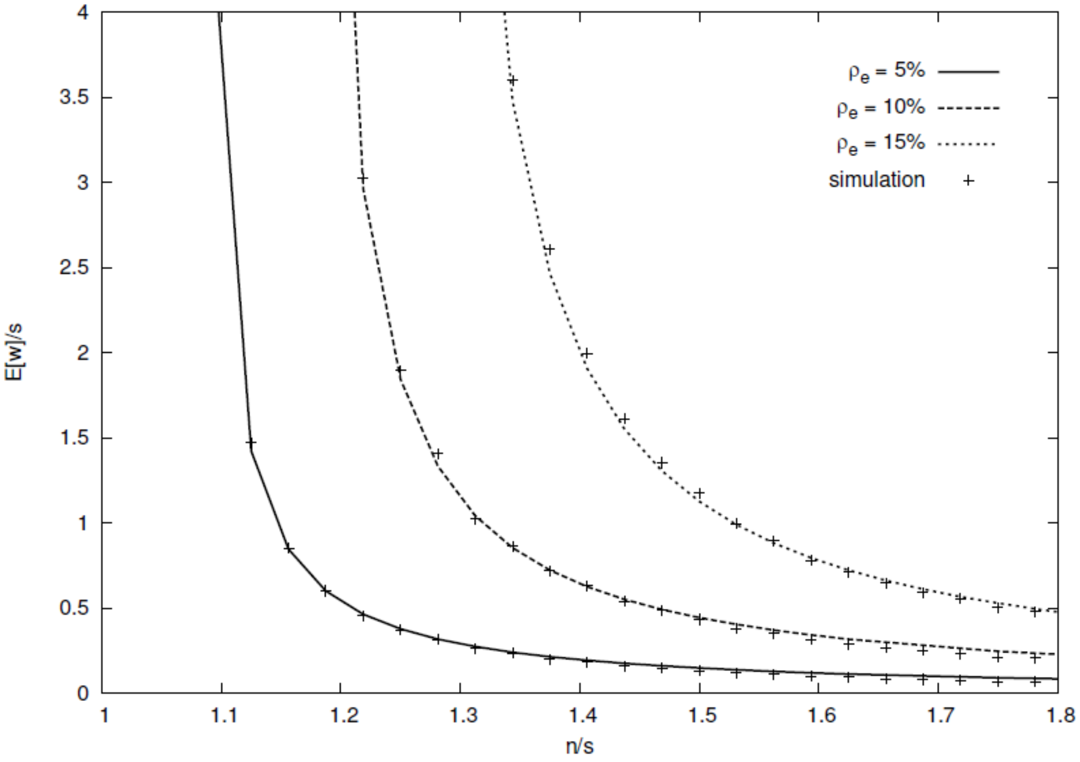


FIG. 8. Average delay for various volumes of traffic at George Washington Bridge.

L. C. Edie, Traffic Delays at Toll Booths, Journal of the Operations Research Society of America, 2, 107{138 (1954).

Applications (V)

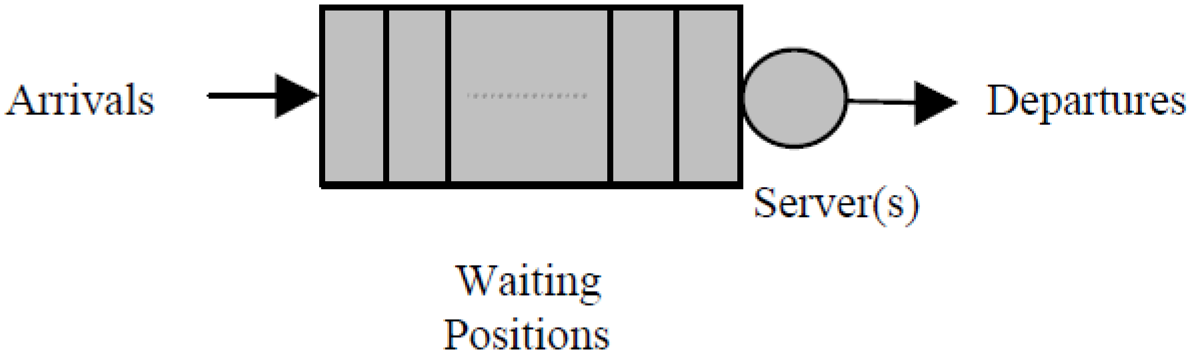
- 5. Waiting times of patients



Average waiting times of the elective patients are depicted versus the Inter-arrival times for various values of the emergency load

Fiems, D., Koole, G. and Nain, P. Waiting times of scheduled patients in the presence of emergency requests. August 6, 2007.

Model of a Queue



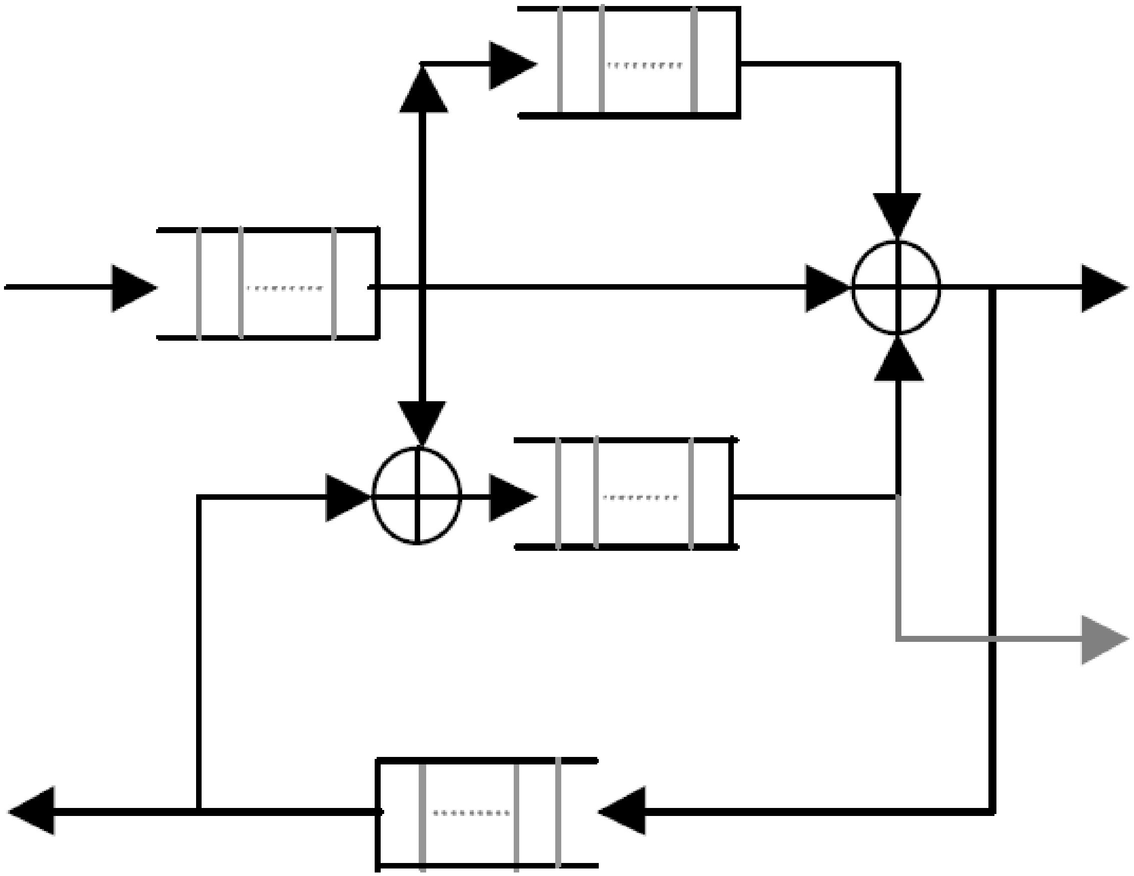
Input Specifications

- Arrival Process Description
- Service Process Description
- Number of Servers
- Number of Waiting Positions
- Special Queueing Rules, e.g. -
  - order of service (FCFS, LCFS, SIRO, etc.)
  - baulking, reneging, jockeying for queue position

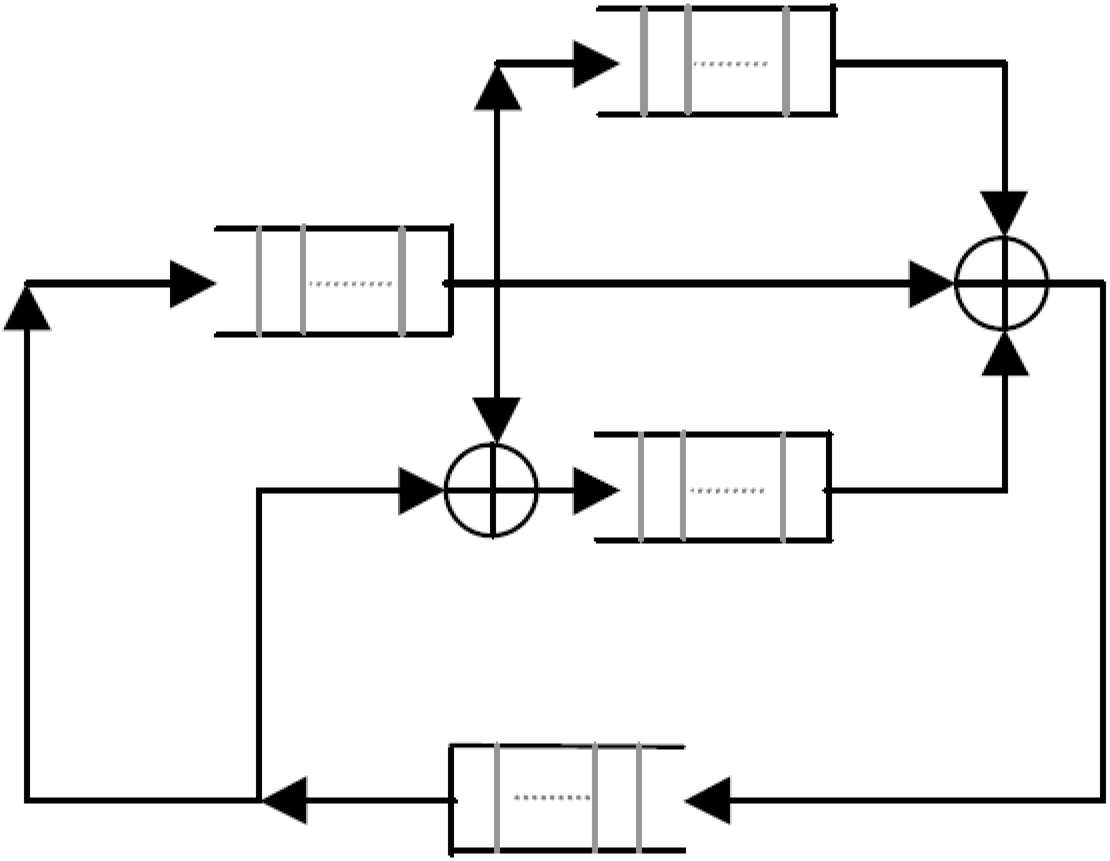
Input Specifications

- For networks of queues, one must provide additional information, such as -
    - Interconnections between the queues
    - Routing Strategy - deterministic, class based or probabilistic with given routing probabilities
    - Strategy followed to handle blocking if the destination queue is one of finite capacity (i.e. with finite number of waiting positions)

An Open Queueing Network



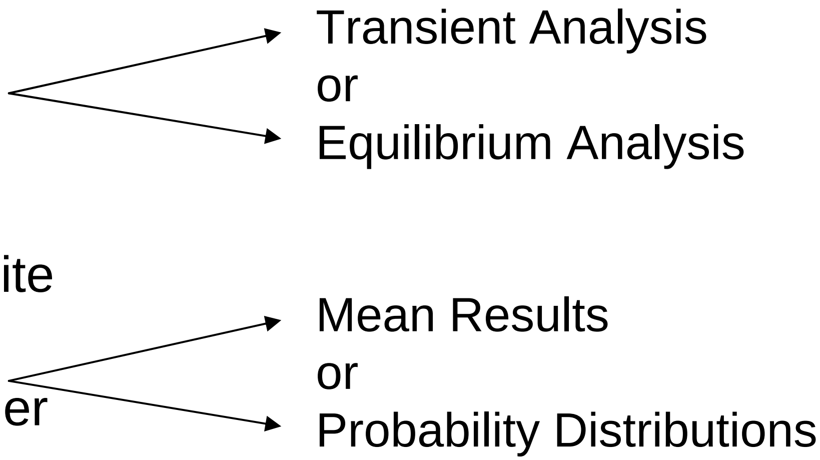
A Closed Queueing Network



- A Queue or a Queueing Network may be studied in different ways
  - Analysis OR/AND
  - Simulation
- The results may be provided from different points of view
  - That of a customer entering the system for service
  - That of a service provider who provides the resources (servers, buffers etc.)

Service Parameters

- Parameters of interest for a customer arriving to the queue for service (Service Parameters)
  - Queueing delay
  - Total delay
  - Number waiting in queue
  - Number in the system
  - Blocking probability (for finite capacity queues)
  - Probability that the customer has to wait for service



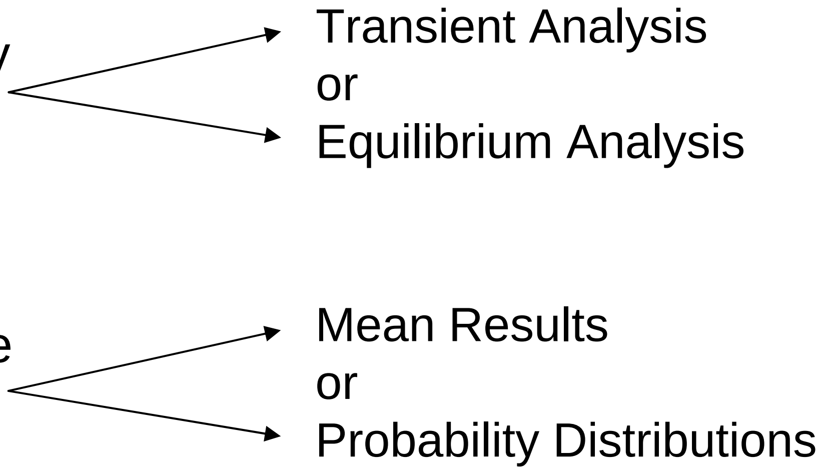
Transient Analysis
or
Equilibrium Analysis

Mean Results
or
Probability Distributions

Service Parameters

- Parameters of interest for the Service Provider (Service Parameters)

  - Server Utilization/ Occupancy
  - Buffer Utilization/ Occupancy
  - Total Revenue obtained
  - Total Revenue lost
  - Customer Satisfaction (Grade of Service)



Transient Analysis
or
Equilibrium Analysis

Mean Results
or
Probability Distributions

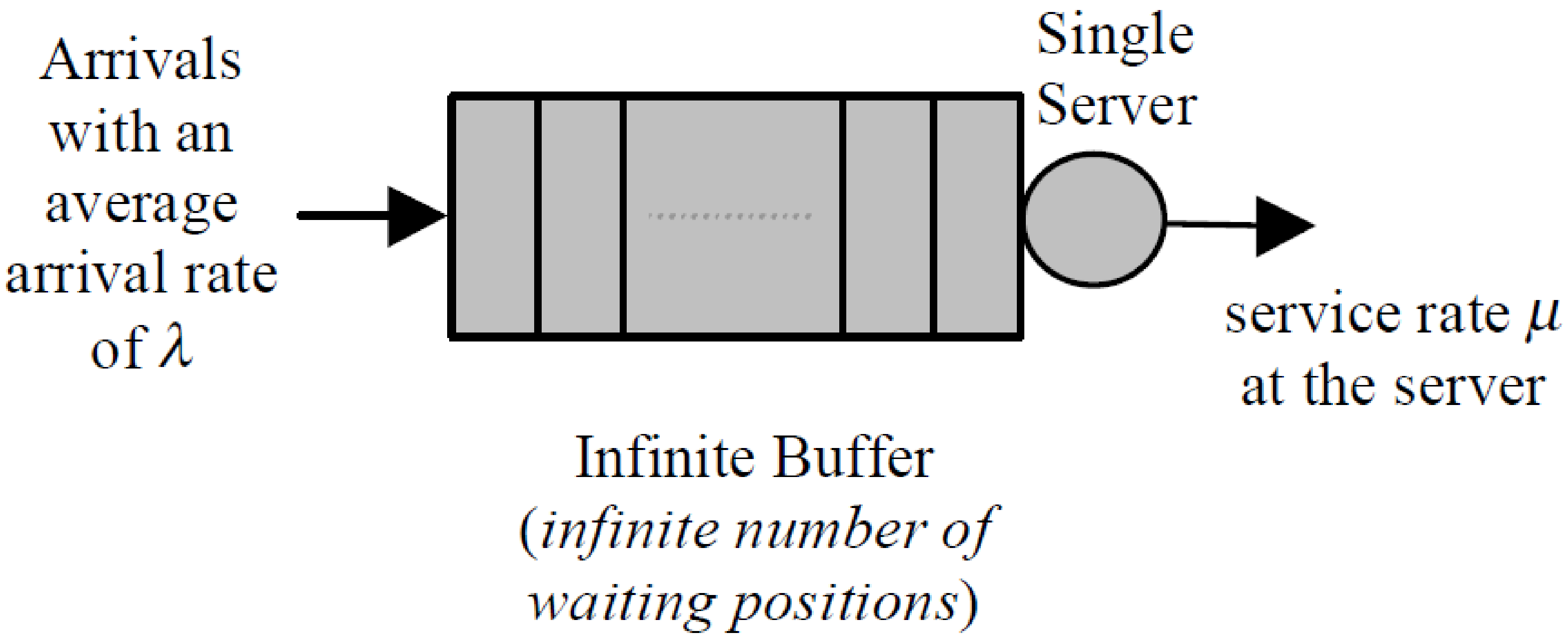
- Our approach to the study of queues and queueing networks

  “Subject to appropriate modeling assumptions, obtain exact analytical results for the mean performance parameters under equilibrium conditions”

- In some special cases, we can also obtain results on higher moments (variance etc.) or probability distributions and/or their transforms.
- Transient analysis is not generally feasible, except for some very simple cases. For this, simulation methods are preferred.
- In some case, especially for queueing networks, exact analysis is not feasible but good approximate analytical methods are available.

Analysis of a Simple Queue

(with some simplifying assumptions)



Arrivals with an average arrival rate of λ

Single Server

service rate μ at the server

Infinite Buffer (*infinite number of waiting positions*)

- Assume that, as $\Delta t \rightarrow 0$
- P{one arrival in time Δt } $= \lambda \Delta t$
- P{no arrival in time Δt } $= 1 - \lambda \Delta t$
- P{more than one arrival in time Δt } $= O((\Delta t)^2) = 0$

Arrival Process
Mean Inter-arrival time
$= 1/\lambda$

- P{one departure in time Δt } $= \mu \Delta t$
- P{no departure in time Δt } $= 1 - \mu \Delta t$
- P{more than one departure in time Δt } $= O((\Delta t)^2) = 0$

Service Process
Mean Service time
$= 1/\mu$

- P{one or more arrival and one or more departure in time Δt } $= O((\Delta t)^2) = 0$

- We have not really explicitly said it, but the implications of our earlier description for the arrivals and departures as $\Delta t \rightarrow 0$ is that –
    - The arrival process is a Poisson process with exponentially distributed random inter-arrival times
    - The service time is an exponentially distributed random variable
    - The arrival process and the service process are independent of each other

- The *state of the queue* is defined by defining an appropriate *system state* variable

- System State at time t = $N(t)$
  = Number in the system at t (waiting and in service)

- Let $p_N(t)$ =P{system in state N at time t }

Note that, given the initial system state at t=0 (which is typically assumed to be zero), if we can find $p_N(t)$ then we can actually describe probabilistically how the system will evolve with time.

- By ignoring terms with (Δt)^2 and higher order terms, the probability of the system state at time $t+\Delta t$ may then be found as -

$$p_0(t+\Delta t) = p_0(t)[1-\lambda\Delta t] + p_1(t)\mu\Delta t \qquad N=0 \qquad (1.1)$$

$$p_N(t+\Delta t) = p_N(t)[1-\lambda\Delta t-\mu\Delta t] + p_{N-1}(t)\lambda\Delta t + p_{N+1}(t)\mu\Delta t \qquad N>0 \qquad (1.2)$$

subject to the normalization condition that $\sum_{\forall i} p_i(t) = 1 \text{ for all } t \geq 0$

- Taking the limits as $\Delta t \rightarrow 0$, and subject to the same normalization, we get

$$\frac{dp_0(t)}{dt} = -\lambda p_0(t) + \mu p_1(t) \qquad N=0 \qquad (1.3)$$

$$\frac{dp_N(t)}{dt} = -(\lambda + \mu) p_N(t) + \lambda p_{N-1}(t) + \mu p_{N+1}(t) \qquad N>0 \qquad (1.4)$$

These equations may be solved with the proper initial conditions to get the ***Transient Solution***.

If the queue starts with N in the system, then the corresponding initial condition will be

$p_i(0)=0$ *for* $i \neq N$

$p_N(0)=1$ *for* $i = N$

- For the *equilibrium solution*, the conditions invoked are -

$$\frac{dp_i(t)}{dt} = 0$$

and

$$p_i(t) = p_i \qquad \text{for } i = 0, 1, 2 \ldots\ldots \infty$$

- or this, defining $\rho = \lambda/\mu$ *erlangs*, with $\rho<1$ for stability, we get

$$\left.\begin{aligned} &p_1 = \rho p_0 \\ &p_{N+1} = (1+\rho)p_N - \rho p_{N-1} = \rho p_N = \rho^{N+1} p_0 \qquad N \geq 1 \end{aligned}\right\} \quad (1.5)$$

- Applying the Normalization Condition $\sum_{i=0}^{\infty} p_i = 1$ we get

$$p_i = \rho^i (1-\rho) \qquad i = 0,\ 1,\ \ldots\ldots, \propto \qquad (1.6)$$

- as the equilibrium solution for the state distribution when the arrival and service rates are such that $\rho=\lambda/\mu<1$

Note that the equilibrium solution does not depend on the initial condition but requires that the average arrival rate must be less than the average service rate

Mean Performance Parameters of the Queue

- (a) Mean Number in System, *N*

$$N = \sum_{i=0}^{\infty} i p_i = \sum_{i=0}^{\infty} i \rho^i (1-\rho) = \frac{\rho}{1-\rho} \qquad (1.7)$$

- (b) Mean Number Waiting in Queue, N_q

$$N_q = \sum_{i=1}^{\infty} (i-1) p_i = \frac{\rho}{1-\rho} - (1-p_0) = \frac{\rho}{1-\rho} - \rho = \frac{\rho^2}{1-\rho} \qquad (1.8)$$

Mean Performance Parameters of the Queue

- (c) Mean Time Spent in System *W*
  - This would require the following additional assumptions
  - FCFS system though the mean results will hold for any queue where the server does not idle while there are customers in the system
  - The equilibrium state probability p_k will also be the same as the probability distribution for the number in the system as seen by an arriving customer
  - The mean residual service time for the customer currently in service when an arrival occurs will still be $1/\mu$ Memory-less Property satisfied only by the exponential distribution

Mean Performance Parameters of the Queue

- Using these assumptions, we can write

$$W = \sum_{k=0}^{\infty} \frac{(k+1)}{\mu} p_k = \frac{1}{\mu(1-\rho)} \qquad (1.9)$$

- (d) Mean Time Spent Waiting in Queue W_q

This will obviously be one mean service time less than *W*

$$W_q = W - \frac{1}{\mu} = \frac{\rho}{\mu(1-\rho)} \qquad (1.10)$$

Mean Performance Parameters of the Queue

- Alternatively, W_q may be obtained using the same kind of arguments as those used to obtain W earlier. This will give

$$W_q = \sum_{k=0}^{\infty} \frac{k}{\mu} p_k = \frac{\rho}{\mu(1-\rho)}$$

which is the same result as obtained earlier.

- (e) P{ Arriving customer has to wait for service} = $1\text{-}p_0 = \rho$

Mean Performance Parameters of the Queue

- (f) Server Utilization “*Fraction of time the server is busy*”

= P{server is not idle}

$= 1 - p_0 = \rho$

The queue we have analyzed is the single server M/M/1/∞ queue with Poisson arrivals, exponentially distributed service times and infinite number of buffer positions

- The analytical approach given here may actually be applied for simple queueing situations where
  - The arrival process is Poisson, i.e. the inter-arrival times are exponentially distributed
  - The service times are exponentially distributed
  - The arrival process and the service process are independent of each other

- Some other simple queues which may be similarly analyzed, under the same assumptions –
    - Queue with Finite Capacity
    - Queue with Multiple Servers
    - Queue with Variable Arrival Rates
    - Queue with “Balking”