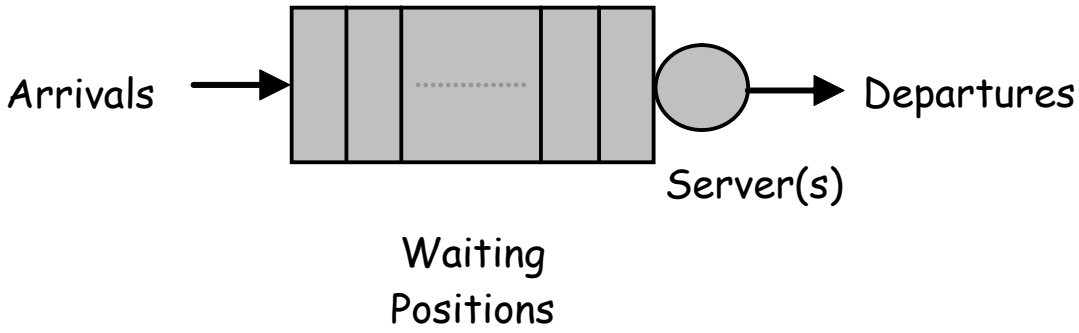
Queues
and
Queueing Networks

Sanjay K. Bose
Dept. of EEE, IITG

Introduction to Queueing Models and Queueing Analysis

Model of a Queue



Input Specifications

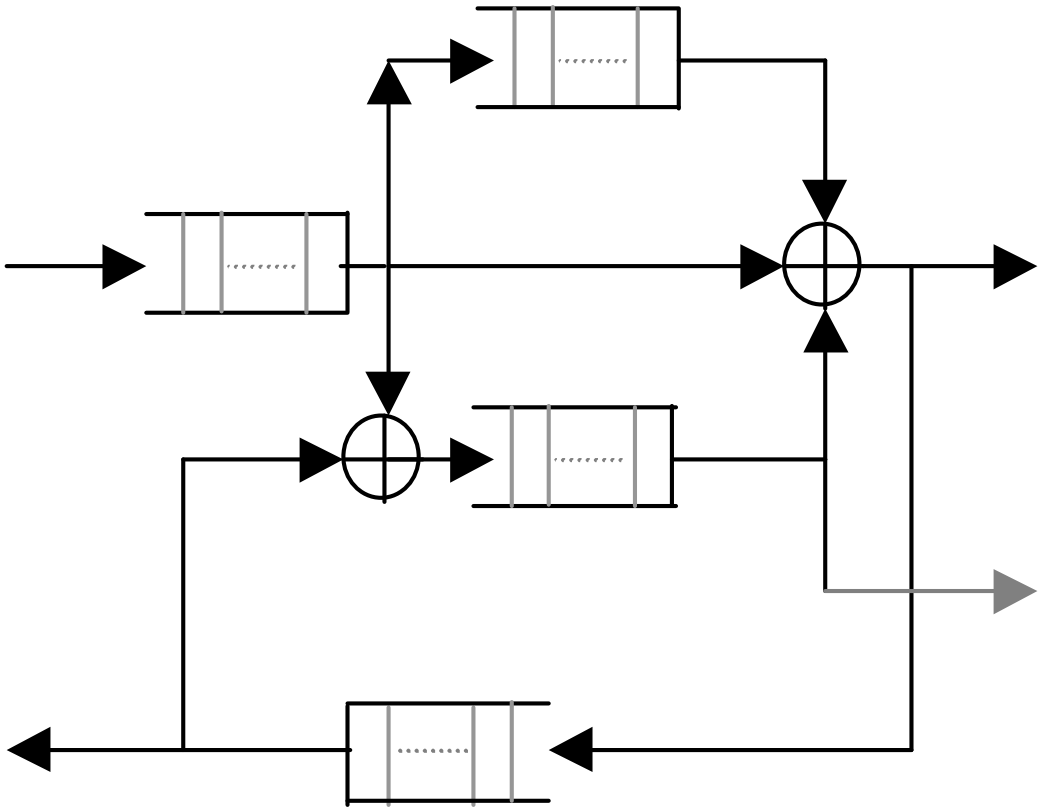
- Arrival Process Description
- Service Process Description
- Number of Servers
- Number of Waiting Positions
- Special Queueing Rules, e.g. -
  - order of service (FCFS, LCFS, SIRO, etc.)
  - baulking, reneging, jockeying for queue position

Input Specifications

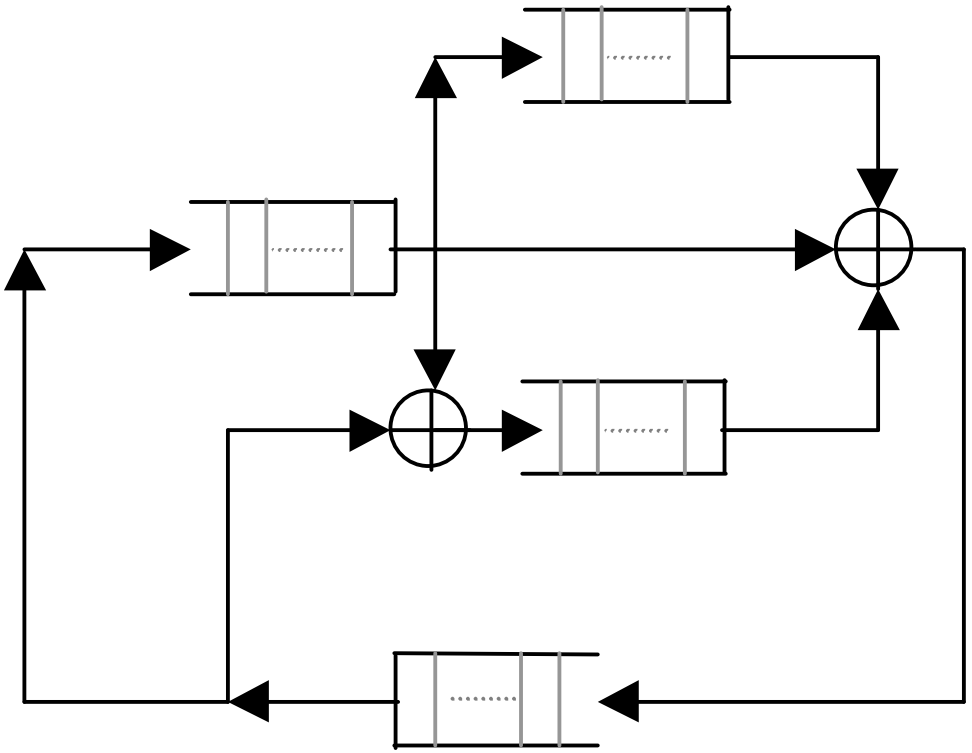
For networks of queues, one must provide additional information, such as -

- 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)

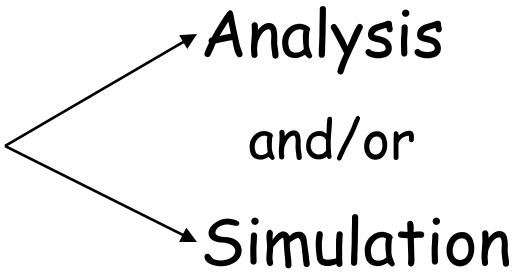
An Open Queueing Network



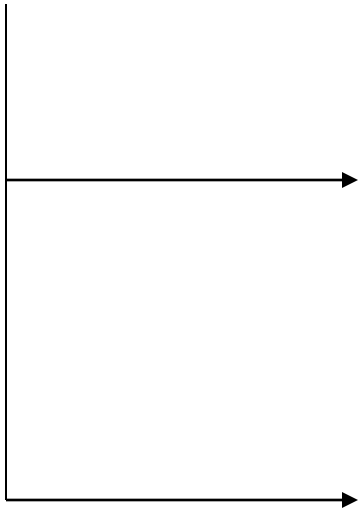
A Closed Queueing Network



A Queue or a Queueing Network may be studied in different ways



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.)

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

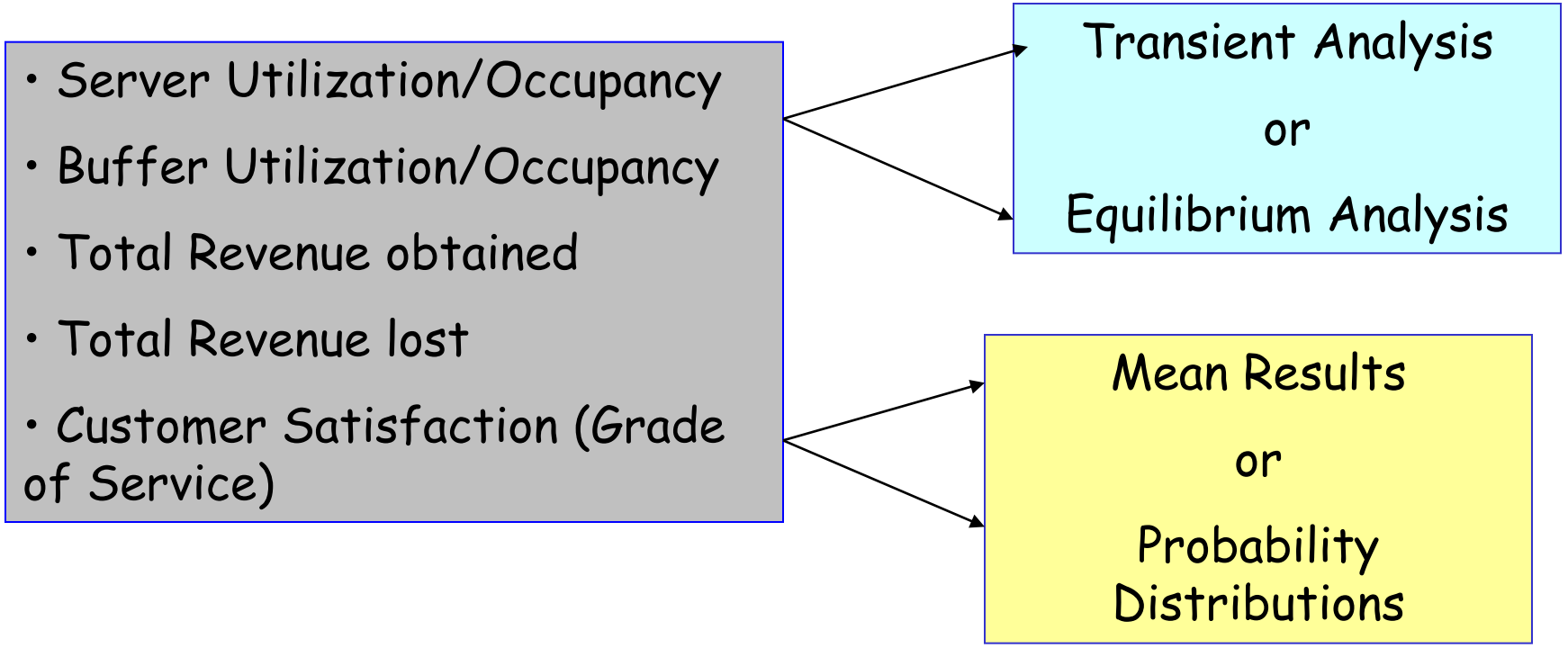
→ Mean Results

or

→ Probability Distributions

Parameters of interest for the Service Provider

(Service Parameters)



Analytical Approach

"Subject to appropriate modelling assumptions, obtain exact analytical results for the mean performance parameters under equilibrium conditions"

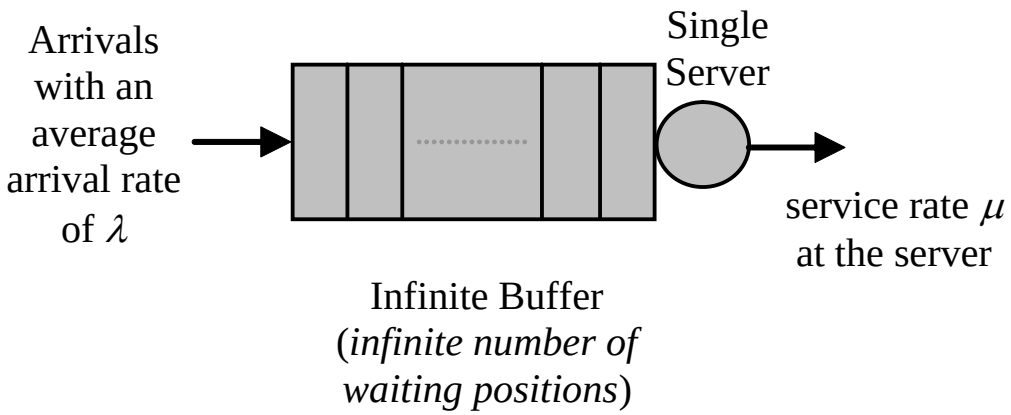
In some special cases, we can also obtain results on higher moments (variance etc.) or probability distributions and/or their transforms.

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*)



Assume that, as $\Delta t \to 0$

Arrival Process

P{one arrival in time Δt} = $\lambda \Delta t$

Mean Inter-arrival time = $\frac{1}{\lambda}$

P{no arrival in time Δt} = $1-\lambda \Delta t$

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

Service Process

P{one departure in time Δt} = $\mu \Delta t$

Mean Service time = $\frac{1}{\mu}$

P{no departure in time Δt} = $1- \mu \Delta t$

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

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

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

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

$$N>0 \qquad (1.2)$$

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

Taking the limits as $\Delta t \to 0$, and subject to the same normalisation, 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

$p_i(t)=p_i$ for $i=0, 1, 2 \ldots . \propto$

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

$$p_1 = \rho p_0$$

$$p_{N+1} = (1+\rho)p_N - \rho p_{N-1} = \rho p_N = \rho^{N+1} p_0 \qquad N \geq 1 \qquad (1.5)$$

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

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

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

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

(b) Mean Number Waiting in Queue, N_q

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

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 *(continued)*

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 *(continued)*

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} = $1-p_0 = \rho$

Mean Performance Parameters of the Queue *(continued)*

(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

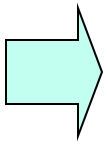
Some other simple queues which may be similarly analyzed, under the same assumptions -

- Queue with Finite Capacity
- Queue with Multiple Servers
- Queue with Variable Arrival Rates
- Queue with "Balking"

Kendall's Notation for Queues

A/B/C/D/E

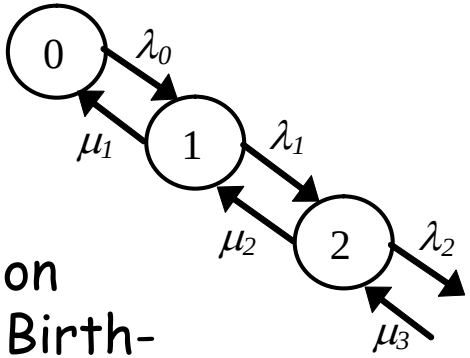
A	Inter-arrival time distribution
B	Service time distribution
C	Number of servers
D	Maximum number of jobs that can be there in the system (waiting and in service)
	Default ∞ for infinite number of waiting positions
E	Queueing Discipline (FCFS, LCFS, SIRO etc.)
	Default is FCFS



M	exponential
D	deterministic
E_k	Erlangian (order k)
G	general

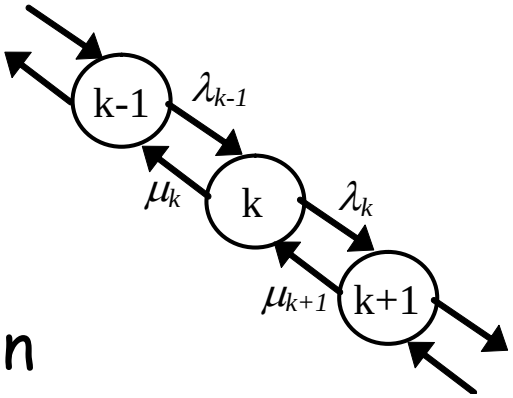
M/M/1 or M/M/1/∞ Single server queue with Poisson arrivals, exponentially distributed service times and infinite number of waiting positions

Equilibrium Analysis of M/M/-/- Type of Queues can be easily done using a State Transition Diagram



State Transition Diagram for a Birth-Death Process

$p_k(t) = P\{X(t)=k\}$
= Probability system in state k at time t.



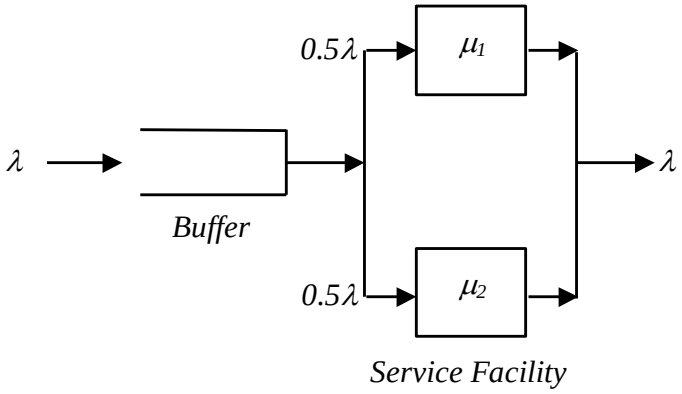
$$\sum_{i \neq j} p_i p_{ij} = p_j \sum_{i \neq j} p_{ji}$$

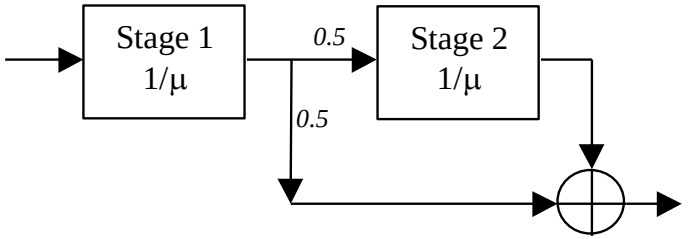
Flow Balance

$$\sum_{\forall i} p_i = 1$$

Normalization Condition

Equilibrium Analysis of M/-/-/- Type of Queues where the service times can be expressed as a combination of exponentially distributed random variables can also be done in a similar way using Method of Stages

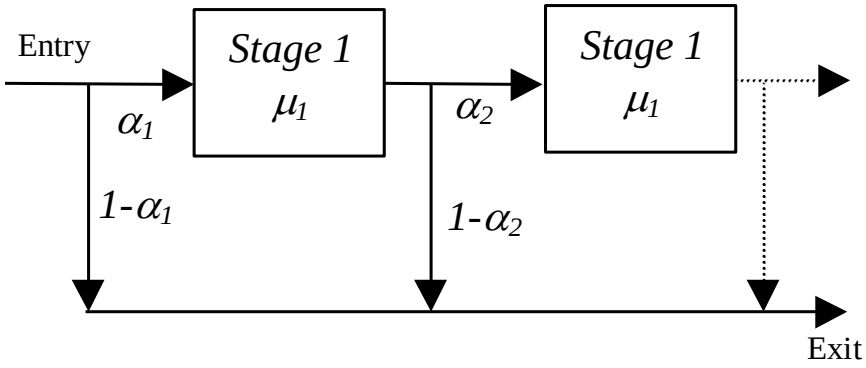




$$L_B(s) = \frac{0.5s(\mu_1 + \mu_2) + \mu_1\mu_2}{(s+\mu_1)(s+\mu_2)}$$

$$L_B(s) = \frac{0.5\mu}{s+\mu} + \frac{0.5\mu^2}{(s+\mu)^2}$$

Method of Stages

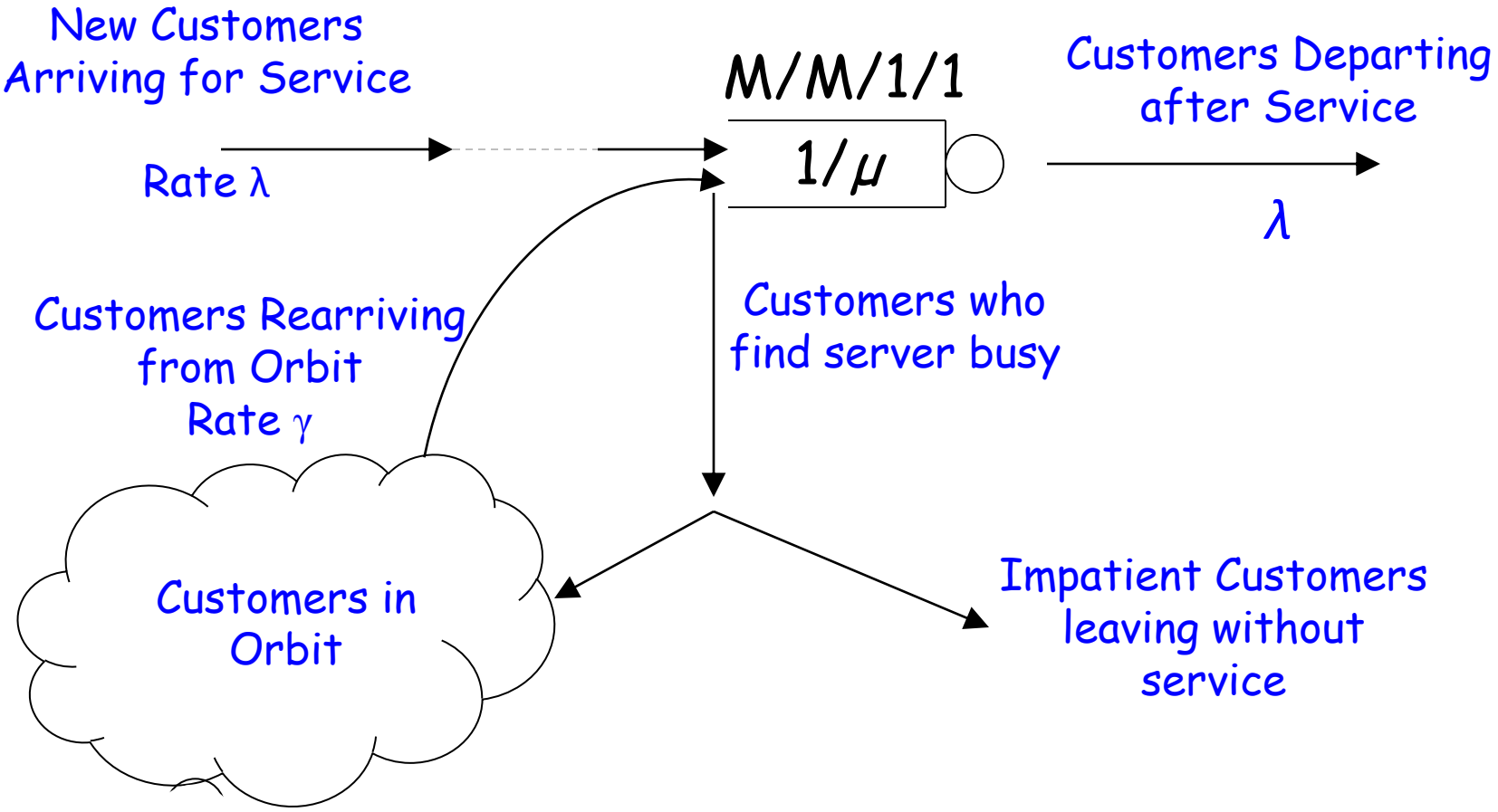


$$L_B(s) = (1-\alpha_1) + \sum_j \alpha_1 \ldots\ldots \alpha_{j-1}(1-\alpha_j)\prod_{i=1}^{j}\frac{\mu_i}{s+\mu_i}$$

$$= \beta_0 + \sum_i \frac{\beta_i}{s+\mu_i}$$

Note: Feedback type connections are also allowed

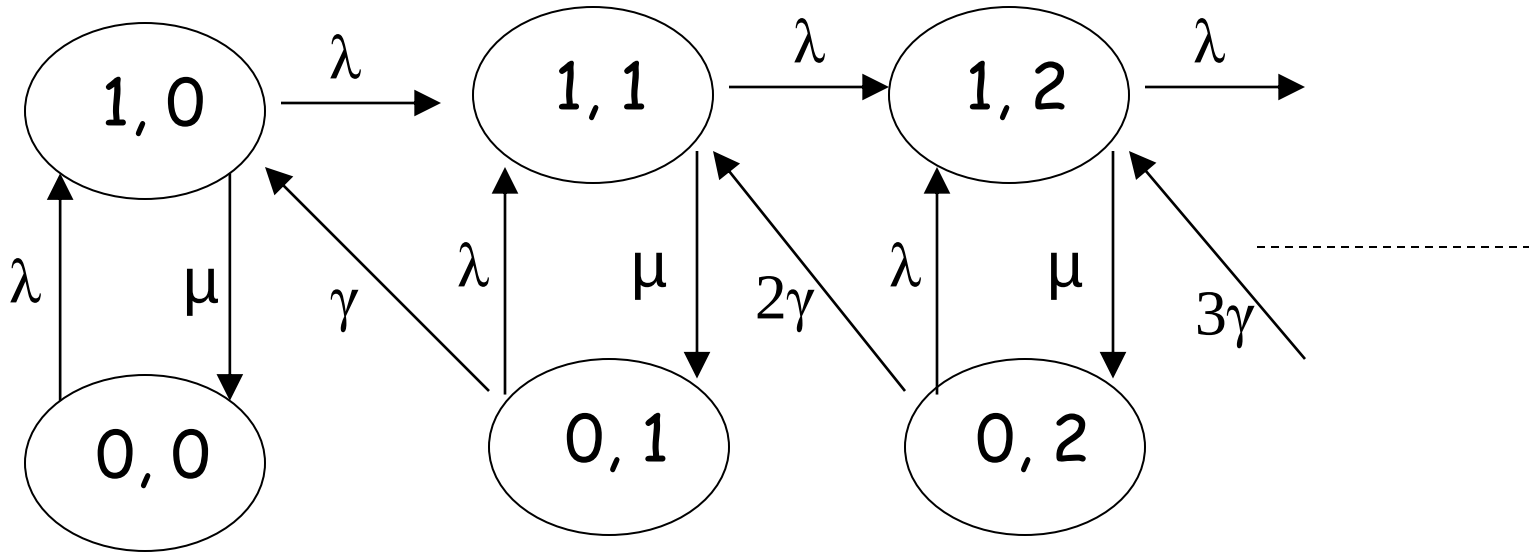
Retrial Queue – A Simple Example



Retrial Queue – A Simple Example

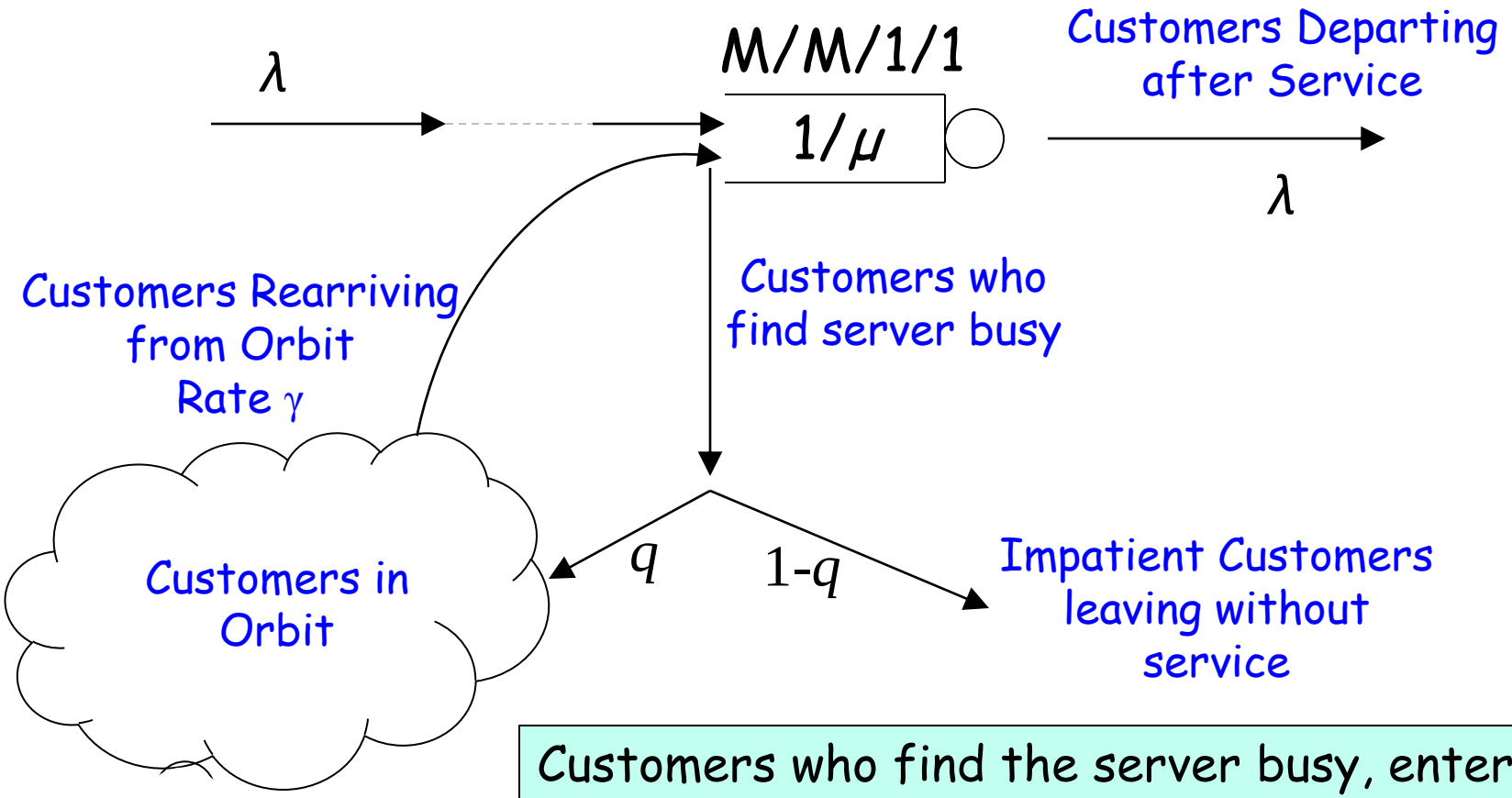
State Descriptor: (i, n) $i = 0,1$ $n = 0,1,2, \ldots\ldots \infty$

i number of customers in service, n number of customers in orbit



State Transition Diagram

M/M/1 Retrial Queue with Impatient Customers



Customers who find the server busy, enter the orbit with probability q and leave the system with probability $(1-q)$

Equilibrium Analysis of the M/G/1 Queue

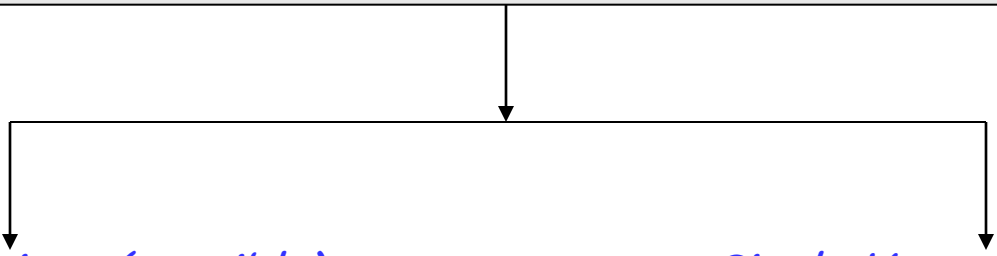
- Mean Analysis using Residual Life Arguments
- Analysis using an Imbedded Markov Chain Approach
- Method of Supplementary Variables

Method of Stages or other exact/approximate analytical methods may also be used

Matrix-Geometric Method may be used to analyze G/M/1, G/E_k/1 and G/PH_k/1 kind of queues

M/G/1 Queue with Vacations

Vacation: After a *busy period*, the server goes on *vacation* of random length. It examines the queue once again when it returns from the vacation



Multiple Vacations (possibly)

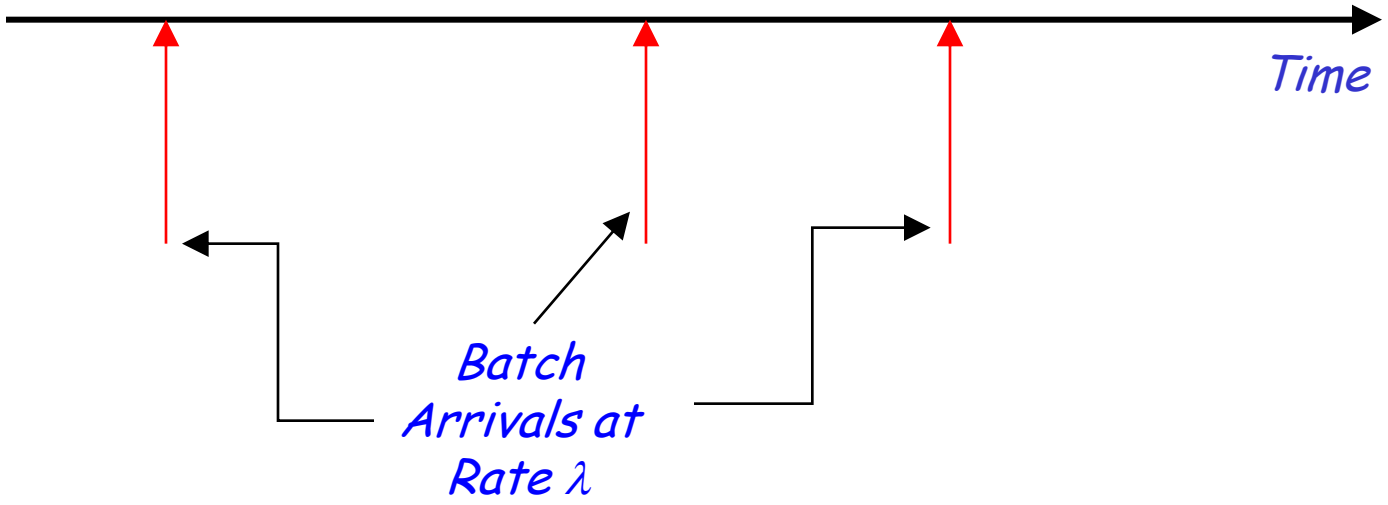
If system still empty when the server returns from a vacation, it goes for another vacation. This continues until it finds system non-empty on return from vacation; it then resumes service normally

Single Vacation (per idle)

After a busy period ends, server goes on only one vacation. If system is still empty when it returns, it stays and waits for a job to arrive.

Other models are also possible, e.g. server goes on (possibly multiple) vacations following the busy period until there are *K* waiting jobs or the server goes for at most *K* vacations if it comes back and finds system empty

M[x]/G/1 Queue : Queue with Batch Arrivals



Number of jobs in a batch = r
(random variable) $1 \le r \le \infty$

β_r = P{r jobs in a batch}

$$\beta(z) = \sum_{r=1}^{\infty} \beta_r z^r$$

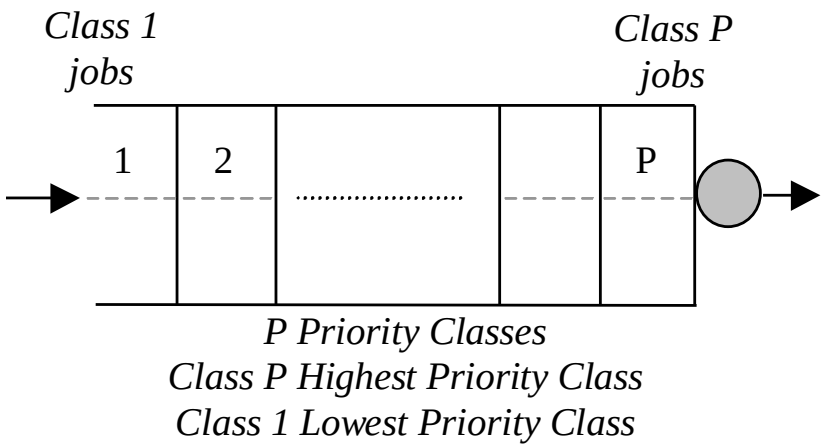
$$E\{r\} = \bar{r} = \beta'(1) = \sum_{r=1}^{\infty} r\beta_r$$

M[x]/G/1 Queue : Queue with Batch Arrivals

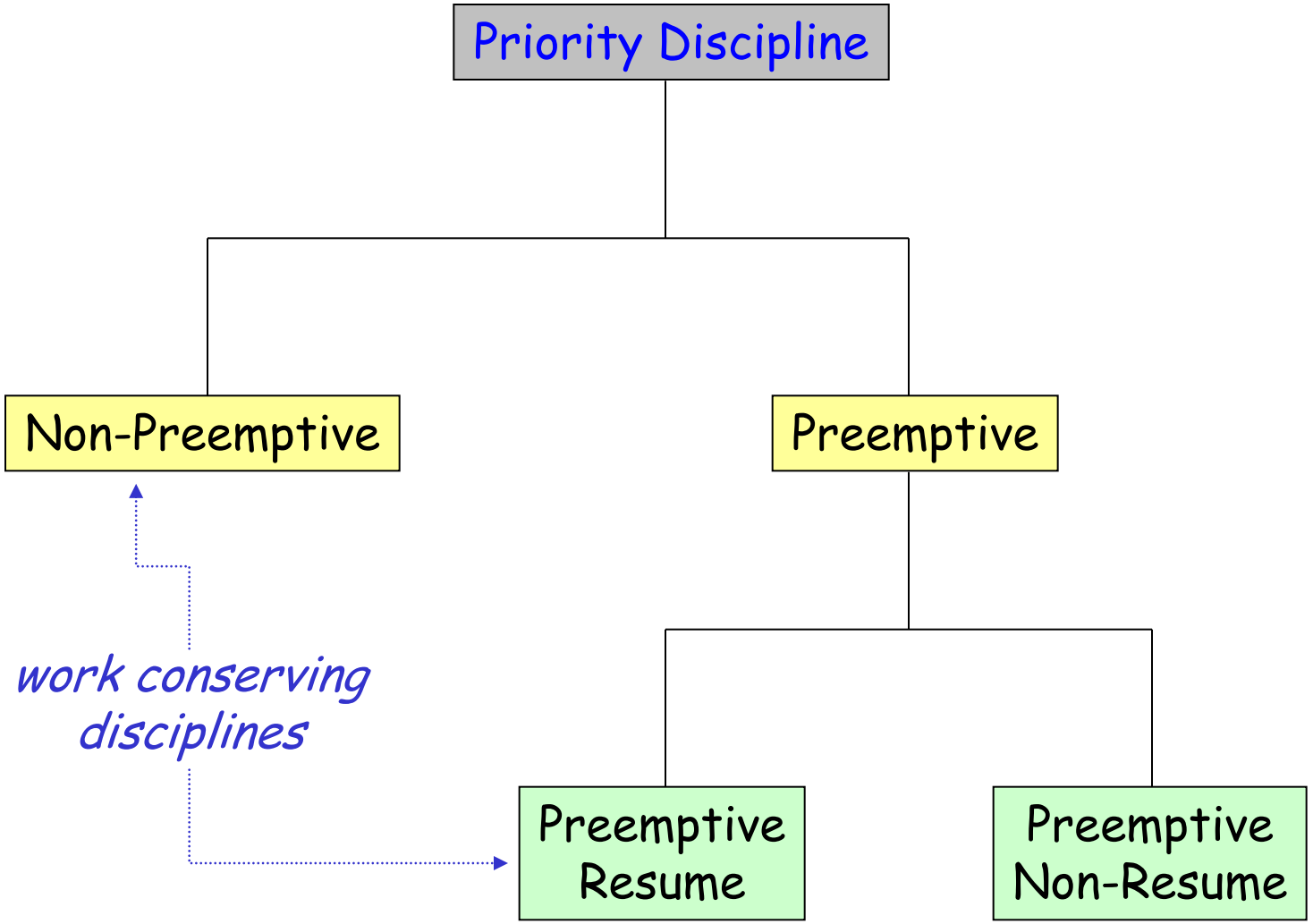
Equilibrium solution for performance measures like mean delay and mean number in system may be done using an Imbedded Markov Chain approach –

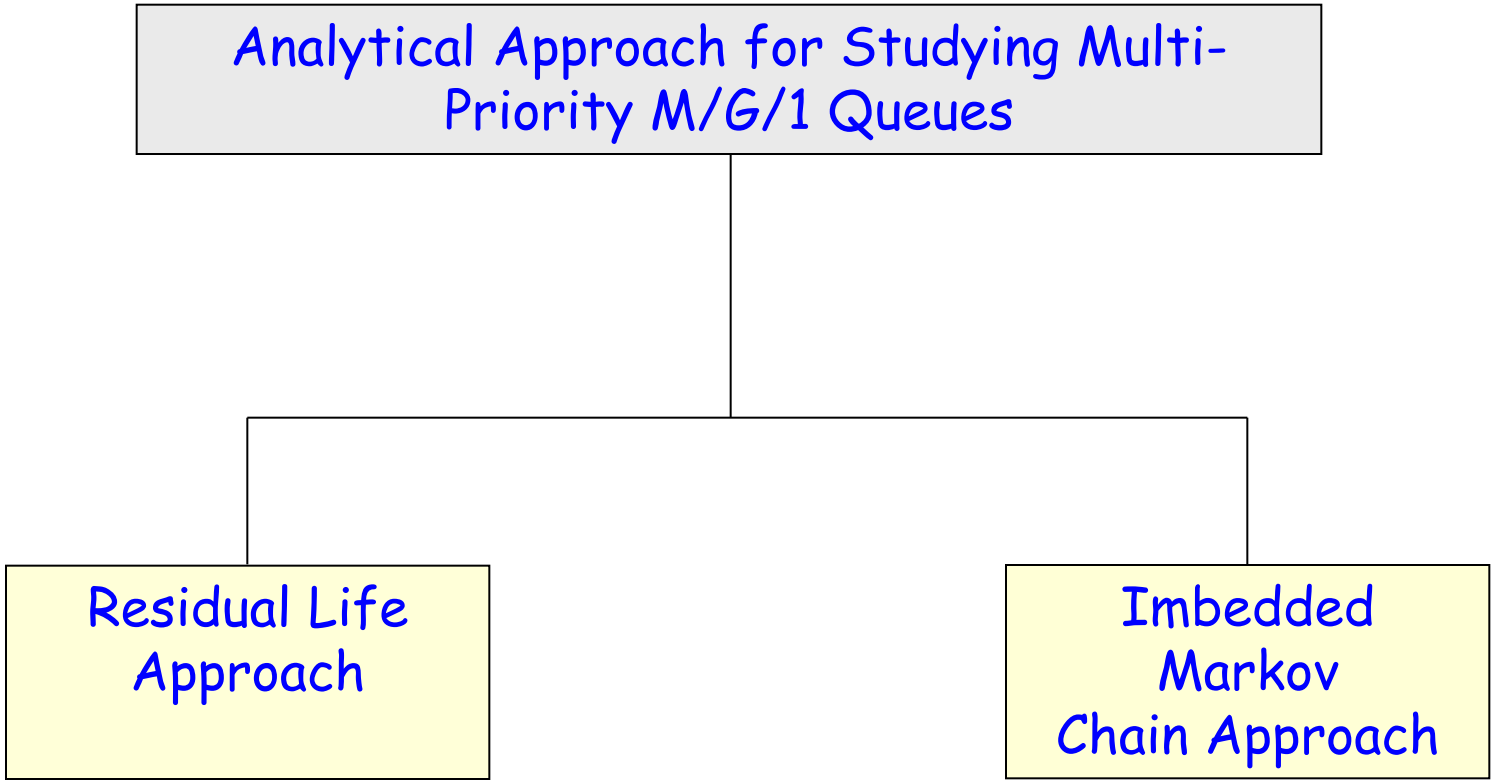
- Equilibrium solutions first obtained consider each batch as a single job
- Subsequently, examine service of a job within a batch

Priority Operation of a M/G/1 Queue



Head of Line (HOL) Priority Operation of M/G/1 Queue





The G/M/1 Queue

- The G/M/1 queue is the dual of the M/G/1 queue where the arrival process is a general one but the service times are exponentially distributed.

- Service time distribution is exponential with parameter $1/\mu$

- General Arrival Process with mean arrival rate λ.

- Total Traffic $\rho = \lambda/\mu$

Stability consideration require that $\rho < 1$ for the queue to be at equilibrium

G/G/1 and G/G/m Queue

Exact analysis cannot be done but various bounds and approximations are available

Special approximations which are quite good available under heavy traffic conditions