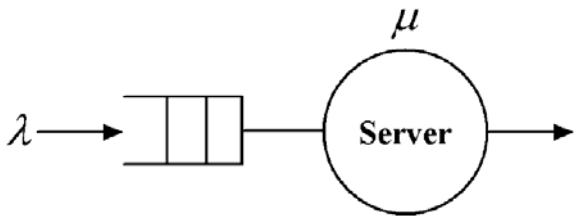
Analytic Modeling

Stochastic Model

Infinite Population Model

- Single server case



Notation

L = length of observation period

n = number of customers arrived and served in L

$\lambda = n/L$ = arrival rate

x_j = service time of j^{th} customer

$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$ = mean service time

Performance Bounds

Model parameters

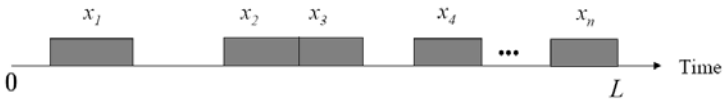
- mean interarrival time, $1/\lambda$
- mean service time, $\bar{x}$

Performance measures

- utilization factor, U
- mean response time, $\bar{T}$
- throughput, R

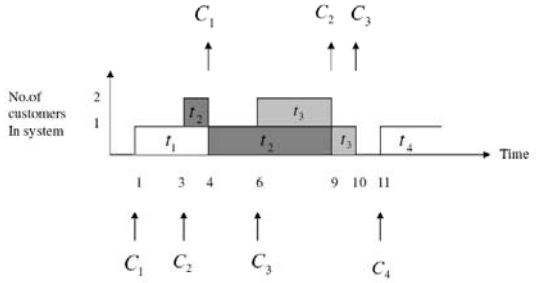
Utilization Factor –Single Server Case

Consider the following case:



- total busy time $= \sum_{j=1}^{n} x_j$
- U = % of time server is busy $= \frac{1}{L}\sum_{j=1}^{n} x_j = \lambda \overline{x}$
- note: $U \leq 1$
- results can be extended to multiple servers

Mean Response Time (Little's Formula)



- Sum of response time = $\sum_{j=1}^{n} t_j$ = total area under curve = $L\overline{N}$
- Thus, $\frac{n}{L}\frac{1}{n}\sum_{j=1}^{n} t_j = \overline{N}$ or
- $\lambda \overline{T} = \overline{N}$ - Little's Formula

Throughput

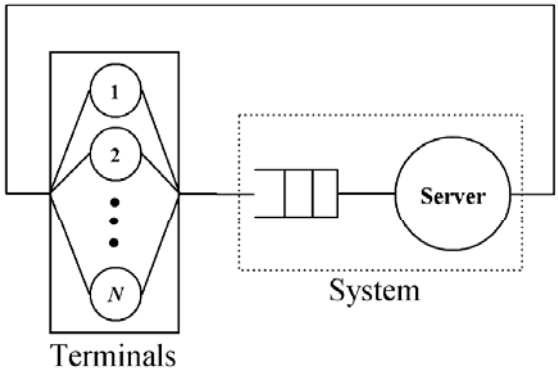
- $R = n/L = \lambda$ where n is the number of customers served in time L
- Since $\lambda = \frac{U}{\overline{x}}$, then $R = \frac{U}{\overline{x}}$

Some Properties

- Utilization factor, throughput, and Little's formula are not affected by scheduling discipline
- Little's formula can also be applied to an arbitrary queuing system (including queuing networks)
- Note: the fundamental results are based on the system returning to the idle state at time *L*, an indication that the system is stable

Finite Population Model

Single server case



Performance Bounds

Model parameters

- mean think time, $\overline{h}$
- mean service time, $\overline{x}$

Performance measures

- utilization factor, U
- mean response time, $\overline{T}$

Utilization Factor (Upper Bound)

When $N = 1,\ U = \dfrac{\overline{x}}{\overline{x} + \overline{h}}$

Let $N^* = \dfrac{(\overline{x} + \overline{h})}{\overline{x}}$

$$U \le \begin{cases} N\dfrac{\overline{x}}{\overline{x} + \overline{h}} & 1 \le N \le N^* \\ 1 & N \ge N^* \end{cases}$$

Mean Response Time (Lower Bound)

- Arrival rate to system (queue and server) = arrival rate to user terminals = λ
- Apply Little's formula to system

  $\lambda\overline{T} = \overline{N}(s) =$ mean number in system
- Apply Little's formula to user terminals

  $\lambda\overline{h} = \overline{N}(t) =$ mean number in thinking state
- Since $\overline{N}(s) + \overline{N}(t) = N,\ \lambda(\overline{T} + \overline{h}) = N$ or $\overline{T} = \dfrac{N}{\lambda} - \overline{h}$
- Since $U = \lambda\overline{x} \le 1,\ \overline{T} \ge N\overline{x} - \overline{h}$ (lower bound)

Summary

- Little's formula
  - mathematical relationship between λ, $\overline{N}$, and $\overline{T}$, two of them are performance measures
  - derivation is based on infinite population model, but result is application to other models, e.g., finite population model, queueing networks
- Infinite population model
  - exact analytic result for utilization factor is available and is a function of input parameters

Summary

- Finite population model
  - exact analytic results are generally not available
  - performance bounds are available -upper bound for utilization and lower bound for mean response time
  - performance bounds are functions of input parameters
- In general, exact analytic results are available if certain assumptions are made on input parameters such as interarrival time, service time and think time
  - for the case of exponential distribution, results are available for infinite population model, finite population model and queuing networks
  - for non-exponential distributions, selected cases where analytic results are available