Lecture No 16

Processor Memory Modeling Using Queuing Theory

Processor Memory Modeling Using Queuing Theory

- Most real life processors make buffered requests to memory.
- Whenever requests are buffered the effect of contention and resulting delays are reduced.
- More powerful tools like Queuing Theory are needed to accurately model processor –memory relationships which can incorporate buffered requests.

Queuing Theory

- A statistical tool applicable to general environments where some requestors desire service from a common server.
- The requestors are assumed to be independent from each other and they make requests based on certain ***request probability distribution function.***
- Server is able to process requests one at a time , each independently of others, except that service time is distributed according to ***server probability distribution function.***

Queuing Theory

- The mean of the arrival or request rate (measured in items per unit of time) is called **λ.**
- The mean of service rate distribution is called **μ**.(Mean service time Ts = 1/μ)
- The ratio of arrival rate (λ) and service rate (μ) is called the utilization or occupancy of the system and is denoted by **ρ**.(λ/μ)
- Standard deviation of service time (Ts) distribution is called **σ**.

Queuing Theory

- Queue models are categorized by the triple.

Arrival Distribution / Service Distribution / Number of servers.

- Terminology used to indicate particular probability distribution.
  - M: Poisson / Exponential c=1
  - M_B: Binomial c=1
  - D : Constant c=0
  - G: General c= arbitrary

Queuing Theory

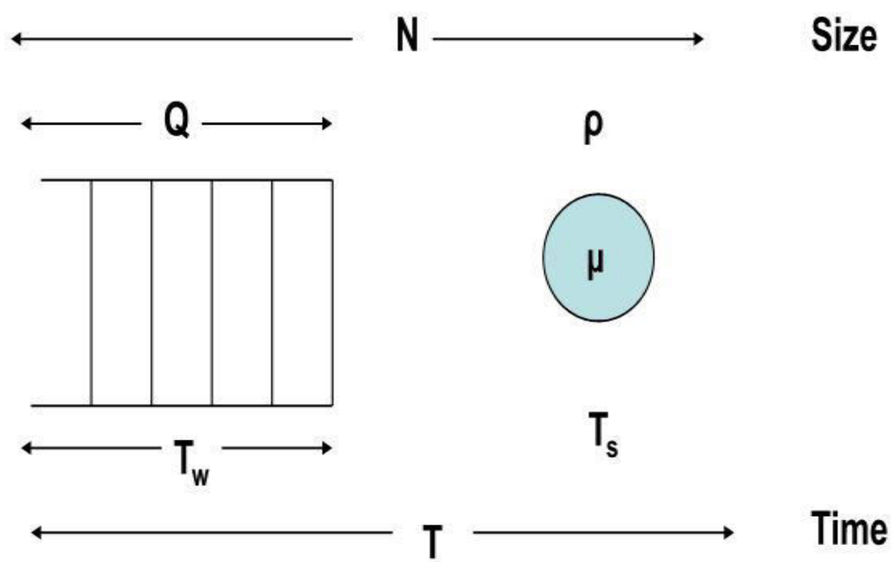
- C is coefficient of variance.

C = variance of service time / mean service time.

$= \sigma / (1/\mu) = \sigma\mu$.

Thus **M/M/1** is a single server queue with poisson arrival and exponential service distribution.

Queue Properties



Queue Properties

- Average time spent in the system (T) consists of average service time(Ts) plus waiting time (Tw).

  $T = T_s + T_w$

  Average Q length (including requests being serviced)

  $N = \lambda T$ (Little's formula).

  Since N consists of items in the queue and an item in service

  $N = Q + \rho$ (ρ is system occupancy or average no of items in service)

Queue Properties

Since $N = \lambda T$

$Q+\rho = \lambda (Ts+Tw)$

$= \lambda (1/\mu + Tw)$

$= \lambda/\mu + \lambda Tw$

$= \rho + \lambda Tw$

Or **$Q = \lambda Tw$**

The **Tw** (Waiting Time) and **Q** (No of items waiting in Queue) are calculated using standard queue formulae for various type of Queue Combinations.

Queue Properties

For M/G/1 Queue Model:

- Mean waiting time $T_w = (1/\lambda)[\ \rho^2(1+c^2)/2(1-\rho)]$
  Mean items in queue $Q = \lambda\ T_w = \rho^2(1+c^2)/2(1-\rho)$

For M/M/1 Queue Model: $C^2 = 1$;

$T_w = (1/\lambda)[\ \rho^2/\ (1-\rho)]$

$Q\ = \rho^2/(1-\rho)$

For M/D/1 Queue Model: $C^2 = 0$;

$T_w = (1/\lambda)[\ \rho^2/\ 2(1-\rho)]$

$Q\ = \rho^2/2(1-\rho)$

Queue Properties

For $M_B/D/1$ Queue Model: $C^2 = 0$;

$T_w = (1/\lambda)[(\rho^2 - p\rho)/2(1-\rho)]$

$Q = (\rho^2 - p\rho)/2(1-\rho)$

For simple binomial **p = 1/m** (Prob of processor making request each Tc is 1)

For δ (Delta) binomial model **p = δ /m** where δ is the probability of processor making request)