به نام خدا

Performance Evaluation of Computer Systems

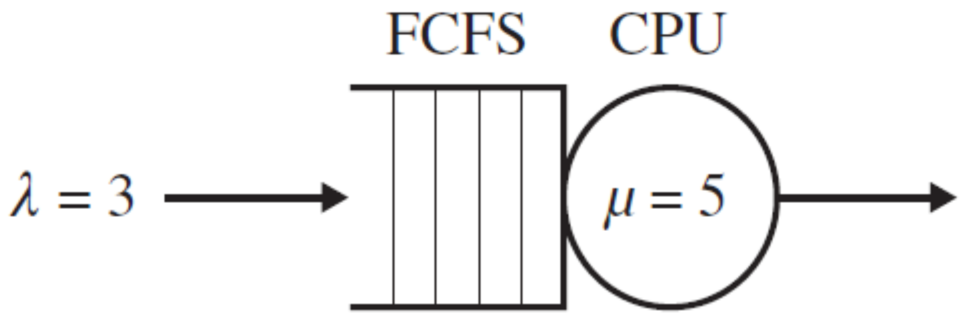
Prof. Ali Movaghar

Fall 2022

Performance Modeling and Design of Computer Systems

1- INTRODUCTION TO QUEUING

1- Examples



CPU

Disk

Router

Databases

Server farms

…

Metrics

- Delay
- Utilization
- Completion rate
- …

2- Terminology
Single Server

- Service order
  - Default is FCFS
- Average arrival rate
- Mean interarrival time
- Service requirement, size
- Mean service time
- Average service time

2- Terminology
Single Server

- Performance metrics
  - Response time, Turnaround time (T)
    - $T = t_{depart} - t_{arrive}$
  - Waiting time, Delay (T_Q)
    - $E(T) = E\left(T_Q\right) + E(S)$
  - Number of jobs in the system
  - Number of jobs in queue

2- Terminology
Single Server

Question: What if $\lambda > \mu$?

Answer: Queue length goes to infinity over time.

- Large time, t
- Number of jobs in the system at time t, N(t)
- Number of arrivals by time t, A(t)
- Number of departures by time t, D(t)

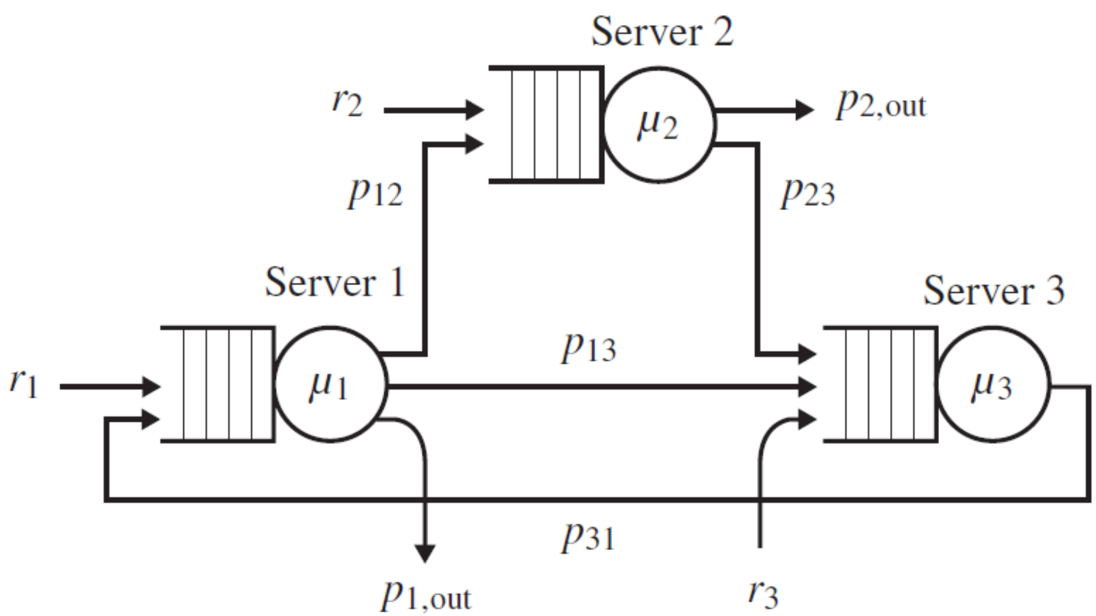
$$E[N(t)] = E[A(t)] - E[D(t)] \geq \lambda t - \mu t = t(\lambda - \mu)$$

$$if\ \lambda > \mu\ then\ t \to \infty \Rightarrow t(\lambda - \mu) \to \infty$$

2- Terminology
Classification

- Open networks
- Closed networks

2- Terminology
Open Networks



2- Terminology
Throughput and Utilization

- In multi-queue, multi-server system
  - E(T) , Mean time a job spends in the whole system
- For addressing the ith queue
  - E(T_i), expected time a job spends queueing and in service at server i

2- Terminology
Throughput and Utilization

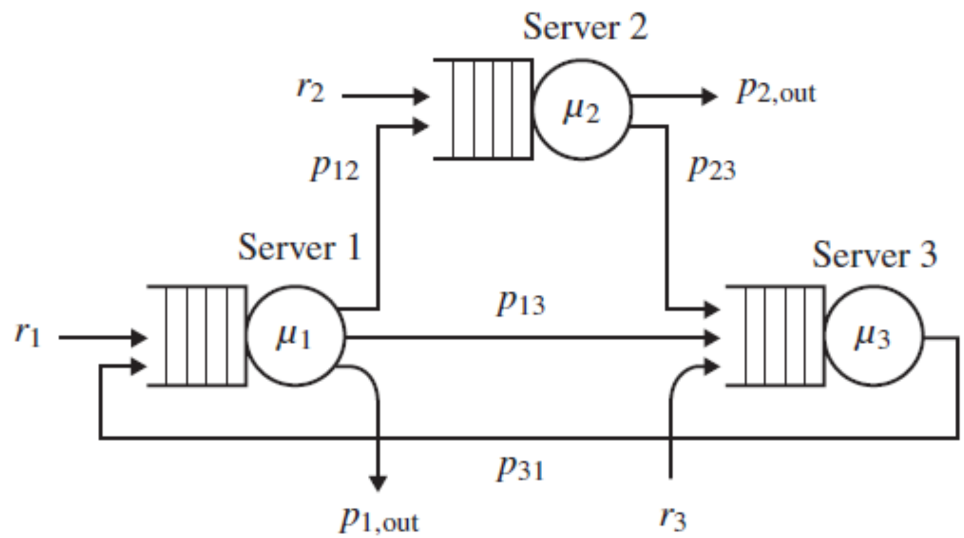
- Utilization, ρ_i
  - Fraction of time device i is busy
    - $\rho_i = \frac{B}{\tau}$
- Throughput, X_i
  - Rate of completions at device i
    - $X_i = \frac{C}{\tau}$

$$\frac{C}{\tau} = \left(\frac{C}{B}\right) . \frac{B}{\tau}$$

$X_i = \mu_i . \rho_i$ or $\rho_i = X_i . E[S]$ (utilization law)

2- Terminology
Throughput and Utilization

- Example:
  - Assume $\lambda_i < \mu_i, \forall i$
  - System throughput
    - $X = \sum_i r_i$
  - Server i throughput
    - $X_i = \lambda_i$
    - $\lambda_i = r_i + \sum_j \lambda_j P_{ij}$

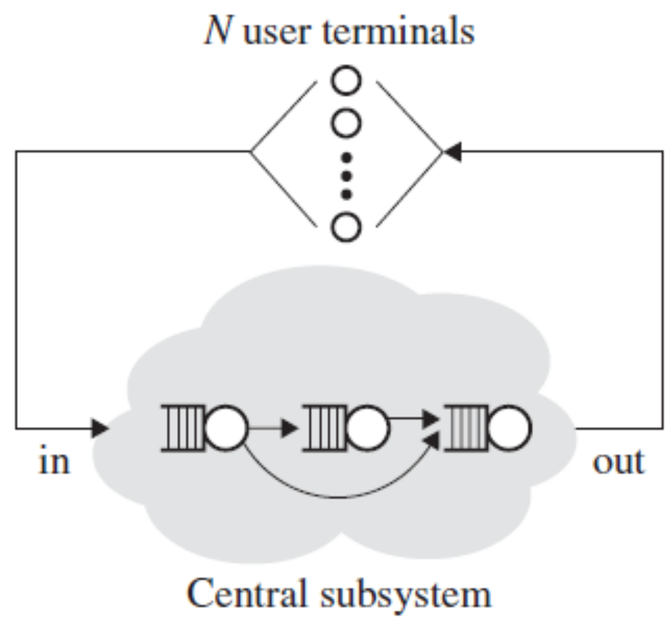


2- Terminology
Closed Networks

- Interactive (terminal-driven)
- Batch system

2- Terminology
Closed Networks

- Interactive (terminal-driven)
  - Think time: Z
  - Response time: time between in and out
  - System time: T
  - $E(T) = E(R) + E(Z)$

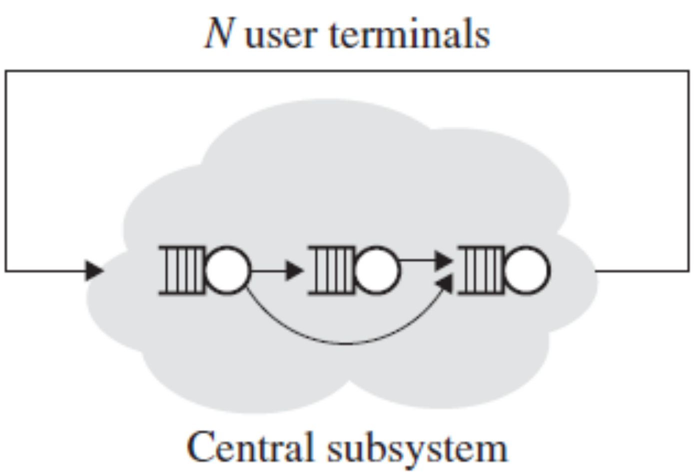


- MPL(Multi-Programming Level)
  - Is at most N

2- Terminology
Closed Networks

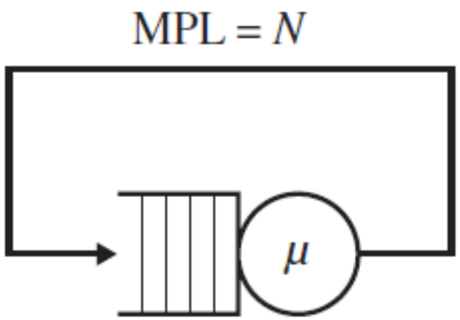
- Batch systems
  - Think time is zero
  - “in” and “out” are equal
    - X is the number of jobs crossing “out” per second

- MPL is exactly N



2- Terminology
Closed Networks

- Throughput in a closed system
  - What is X?
    - X=μ
  - What is E(R)?
    - E(R) = E(T)
    - E(T) = $\frac{N}{\mu}$



2- Terminology
Differences

- Open systems
  - Throughput, X is independent of μ_i's
  - Throughput and Response time are not related
- Closed systems
  - X depends on μ_i's
  - Higher throughput $\Leftrightarrow$ Lower avg. response time