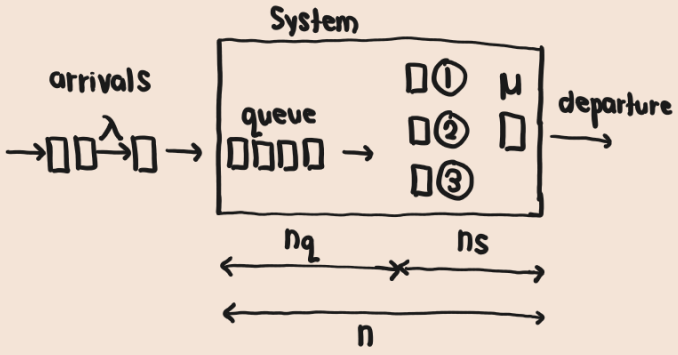
Queuing Theory

1. λ : arrival rate of job per unit time
2. μ : service rate of job per unit time
3. m : # servers



Stability: $\lambda < m\mu$

$E[n] = E[nq] + E[ns]$

$E[r] = E[w] + E[s]$

$E[n] = \lambda E[r]$

Little Law $E[nq] = \lambda E[w]$ ↔ Relation btwn jobs and time

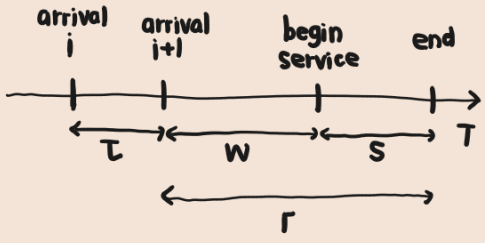
$E[ns] = \lambda E[s]$

Traffic intens $\rho = \lambda E[s] = E[ns]$ (one server)

P_0 = Prob that server is idle

$= 1 - \rho$

$= 1 - \lambda E[s]$



**??

$$E[r] = \frac{E[s]}{1-\rho} \qquad E[w] = \frac{\rho E[s]}{1-\rho}$$

$$E[n] = \frac{\rho}{1-\rho}$$

↳ service time distribution

M/M/1 ← # server

↓

inter-arrival time distribution

$$E[nq] = \frac{\rho^2}{1-\rho} \quad (M/M/1)$$

Operational Laws

1. Utilization Law

$$U_i = X_i S_i = \lambda_i S_i = X D_i \quad (D_i = V_i S_i)$$

2. Forced Flow

$$X_i = X V_i$$

3. Little's Law

$$Q = XR$$

$$Q_i = \lambda_i R_i = X_i R_i$$

4. General Response Time

$$R = \sum_{i=1}^{M} V_i R_i \qquad R_i = \frac{S_i}{1-U_i}$$

5. Interactive Response Time

$$R = \frac{N}{X} - Z$$

Asymptotic Bound

Thruput : $\frac{N}{ND+Z} \leq X(N) \leq \min\left\{\frac{1}{D_{max}}, \frac{N}{D+Z}\right\}$

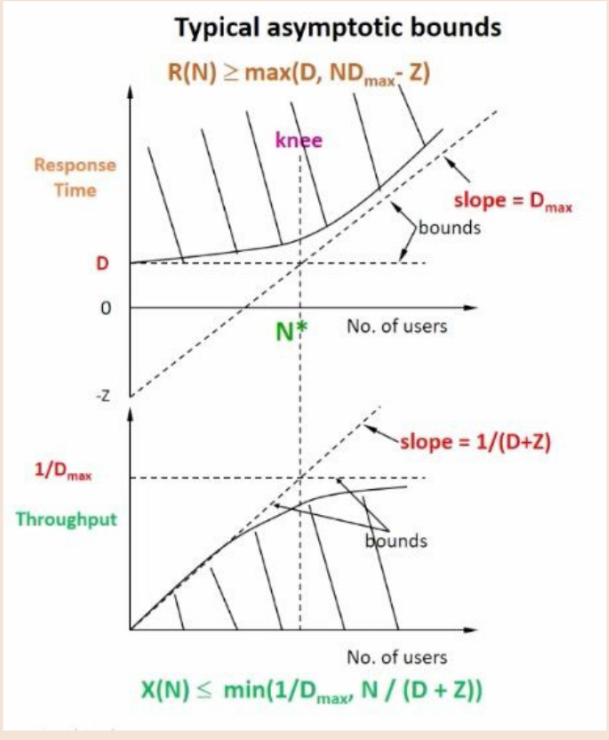
Response time:

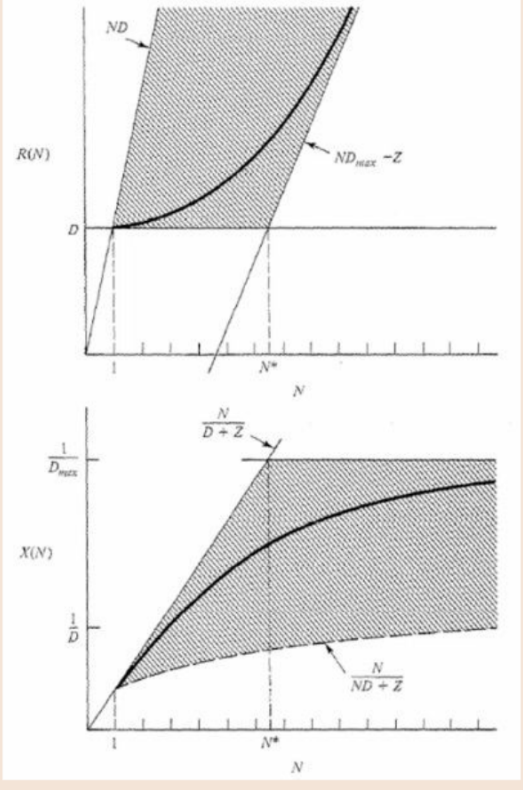
$$\max\{D, ND_{max} - Z\} \leq R(N) \leq ND$$

↑ bottlenecked device

users at knee

$$N^* = \frac{D+Z}{D_{max}}$$





Poisson λ

λ = mean # events in duration unit time

pmf: $f(x) = P(X=x) = \lambda^x \frac{e^{-\lambda}}{x!}$

mean = variance = λ

M: Exponential / Memoryless

$$Pr[M(t)=n] = \frac{e^{-\lambda t}(\lambda t)^n}{n!} \quad (t \geq 0,\ n = 0,1,2\ldots)$$

$Pr[A_1 \leq t] = 1 - e^{-\lambda t}$ (CDF)

$Pr[A_1 > t] = Pr[M(t)=0] = e^{-\lambda t}$

pdf: $f(x) = \lambda e^{-\lambda x}$

cdf: $F(x) = 1 - e^{-x\lambda}$

$0 \leq x \leq \infty$

Mean: $\frac{1}{\lambda}$

Var: $\frac{1}{\lambda^2}$

$$E(x) = \begin{cases} \sum_{i \in I} x_i p(x_i) \\ \int_{-\infty}^{\infty} x f(x)\,dx \end{cases}$$

$$V(x) = \sigma^2 = E[(x - E[x])^2]$$
$$= E[x^2] - E[x]^2$$

Operational Analysis

- Metrics that can be measured
- mean values measured during a finite duration (T or observation period)
- no assumption on distribution

1. # job arrivals (A_i)
2. # busy time of device i (B_i) over T
3. # job departures (C_i)

$\Downarrow$

4. Arrival rate $\lambda_i = \frac{A_i}{T}$
5. Thruput $X_i = \frac{C_i}{T}$
6. Utilization $U_i = \frac{B_i}{T}$

$i=0 \Rightarrow$ system
$i>0 \Rightarrow$ devices

7. Avg service time per job visit

$$S_i = \frac{B_i}{C_i}$$

8. Avg # visits to device i during a job

V_i

9. Utilization law

$$U_i = \frac{B_i}{T} \times 1 = \frac{B_i}{T} \times \frac{C_i}{C_i} = \frac{C_i}{T} \times \frac{B_i}{C_i} = X_i S_i$$

Device with the highest utilization is the bottleneck device

a. Job Flow Balance Assumption

$$A_i = C_i \Leftrightarrow \frac{A_i}{T} = \frac{C_i}{T} \Leftrightarrow U_i = \lambda_i S_i$$

10. Visit Ratio

Each job makes V_i requests for the i^{th} device

Balance $\Rightarrow A_0 = C_0$

$\Rightarrow C_i = C_0 V_i$

11. Forced Flow Law

Relate sys thruput to individual device thruput

Assume each device has job flow balance $(A_i = C_i)$

X is sys thruput

$$X = \frac{C_o}{T}$$

$$X_i = \frac{C_i}{T} \cdot \frac{C_o}{C_o} = \frac{C_o}{T} \cdot \frac{C_i}{C_o} = X V_i$$

let D_i be total service demand by a job at device i

$$D_i = V_i S_i$$

combine forced flow and utilization law

$$U_i = X_i S_i = X V_i S_i \quad \text{(Forced Flow)}$$
$$= X D_i$$

P_{ij} : probability that job moves from i to j

$$C_j = \sum_{i=0}^{M} C_i P_{ij} \quad \Big) \div C_o$$
$$V_j = \sum_{i=0}^{M} V_i P_{ij} \quad \text{visit ratio eqns}$$
$$V_o = 1$$

12. Little's Law

Assume job flow balance

Q_i := avg q size

R_i := avg response time

$$Q_i = \lambda_i R_i = X_i R_i$$

13. General Response Time Law

Apply Little's Law to each component

$$Q = XR = Q_1 + Q_2 + \ldots + Q_M$$

$$XR = \sum_{i=1}^{M} X_i R_i \quad \Big) \div X \text{ and forced flow (which imply Little)}$$
$$R = \sum_{i=1}^{M} V_i R_i$$

14. Interactive Response Time Law

R := sys response

Z := avg think time

Each user generates $\frac{T}{R+Z}$ jobs in T time

$$\#\text{ jobs completed} = N * \frac{T}{R+Z}$$

$$X = \left(\frac{NT}{R+Z}\right) \cdot \frac{1}{T}$$

$$= \frac{N}{R+Z}$$

$$R = \frac{N}{X} - Z$$

15. Assume N_i jobs at device i

Then $R_i = N_i S_i + S_i$

$= R_i X_i S_i + S_i$

$= R_i * U_i + S_i$

✡ $R_i = \frac{S_i}{1-U_i}$ (M/M/1)