Arrival rate $\lambda = \frac{A}{T}$ output rate $X = \frac{C}{T}$ 01B

Utilisation $U = \frac{B}{T}$ mean service time per complete $= \frac{B}{C}$

$U = SX$ $C = A$ or at least $C \approx A$

$V(j)$ = visit ratio of device j $V(j) = \frac{C(j)}{C(0)} = \frac{X(j)}{X(0)}$ forced flow law

$D(j)$ = service demand of a job at device j is the total service time required by that job

$D(j) = V(j)S(j)$ (mean service time of device j)

$D(j) = \frac{U(j)}{X(0)}$ service demand law

$U(j) = S(j)X(j)$

Little's law: X = Throughput of the requests

Rarg = Average response time of the requests

Narg = Average number of requests in the device

$Narg = X * Rarg$

Response time = Depature - Arrival

= waiting + processing

02A
C1)

Utilisation law: $U(j) = X(j)S(j)$

Forced flow law: $X(j) = V(j)X(0)$

Service demand law: $D(j) = V(j)S(j) = \frac{U(j)}{X(0)}$

Little's law: $N = X * R$

Thinking time = processing time of the user

user: waiting time, cpu: response time

N = interactive users Z = mean thinking time

R = mean response time of the computer system.

X_0 = throughput N_{avg} = mean busy users

N_{avg} = average jobs in the computer system

$N_{avg} = R * X_0$, $N = N_{avg} + N_{avg} = X_0 * (Z + R)$

N is the Interactive response time

Bottleneck throughput is limited by the maximum service demand. $X(0) \leq \min\left[\frac{1}{\max D_i}, \frac{N}{\sum_{i=1}^{k} D_i + \text{thinktime}}\right]$

operational analysis allows you to bound the system performance but it does NOT allow you to find the throughput and response time of a system, need to use queueing analysis, need to specify the workload

Performance bound depends on:
- number of users
- service demand

Queue response time depends on:
- job arrival probability distribution
- job service time distribution

δ: small time interval — Exponential inter-arrival time

p: a constant

λ: mean arrival rate of customers

$\text{prob}[\text{no arrival in } [0,t]] = e^{-Npt}$

$\lambda = Np$

An arrival process is Poisson with parameter λ if the probability that n customers arrive in any time interval t is $\frac{(\lambda t)^n e^{-\lambda t}}{n!}$

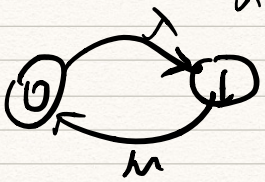
mean number of customers arriving in a time interval of t is

$$\sum_{n=0}^{\infty} n \cdot \frac{(\lambda t)^n e^{-\lambda t}}{n!} = \lambda t$$

$$\text{mean arrival rate} = \frac{1}{\text{mean inter-arrival time}}$$

λ — $\frac{1}{\lambda}$

probability inter-arrival time $\in [x, x+\delta x]$
$= \lambda e^{-\lambda x} \delta x$ mean inter-arrival time $= \frac{1}{\lambda}$
mean number of arrivals in time interval T is λT
mean arrival rate $= \lambda$
An interpretation of poisson arrival:
Prob[no arrival in δ] $= 1 - \lambda\delta$
Prob[1 arrival in δ] $= \lambda\delta$
prob[2 or more ...] $\simeq 0$
call centre with 1 operator and no holding slots
No call at call centre at $t+\Delta t$ can be caused by:
No call at time t and no call arrives in $[t, t+\Delta t]$ or
1 call at time t and call finishes in $[t, t+\Delta t]$
$P_0(t+\Delta t) = P_0(t)(1-\lambda\Delta t) + P_1(t)\mu\Delta t$
1 call during $t+\Delta t$: $P_1(t+\Delta t) = P_0(t)\lambda\Delta t + P_1(t)(1-\mu\Delta t)$
$P_0 = P_0(\infty) = \frac{\mu}{\lambda+\mu}$, $P_1 = P_1(\infty) = \frac{\lambda}{\lambda+\mu}$ $\leftarrow \begin{cases} \lambda P_0 = \mu P_1 \\ P_0 + P_1 = 1 \end{cases}$



rate of transition out of a state
$= \ldots\ldots\ldots\ldots$ into
$\lambda P_0 = \mu P_1$

$$\begin{cases} \lambda P_0 = \mu P_1 \\ \mu P_2 = \lambda P_1 \\ P_0 + P_1 + P_2 = 1 \end{cases} \Rightarrow P_0 = \frac{1}{1 + \frac{\lambda}{\mu} + (\frac{\lambda}{\mu})^2} \quad P_2 = \frac{(\frac{\lambda}{\mu})^2}{1 + \frac{\lambda}{\mu} + (\frac{\lambda}{\mu})^2}$$

$$P_1 = \frac{\frac{\lambda}{\mu}}{1 + \frac{\lambda}{\mu} + (\frac{\lambda}{\mu})^2}$$

Poisson queues with 1 server and (0 or 1) holding slot.

Response time T = waiting time W + service time S 03A (1)

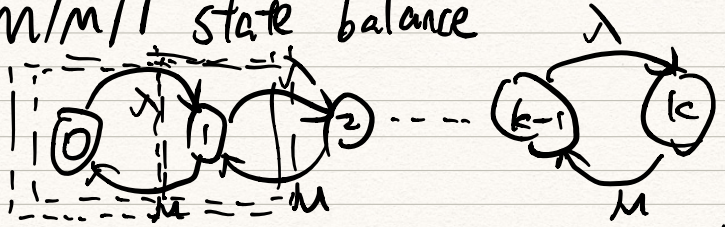
ρ = utilisation

Kendall's notation

$M/M/s/(/B)$ ← ignore when infinite; B = ST? waiting room

- M: inter-arrival exponential
- M: service time exp
- s: server number

$M/M/1$ state balance



$\lambda p_0 = \mu p_1$

$\lambda p_0 + \mu p_2 = (\lambda+\mu) p_1$

$\lambda p_1 = \mu p_2$

$\begin{cases} \lambda p_0 = \mu p_1 \\ \lambda p_1 = \mu p_2 \end{cases} \Rightarrow p_1 = \frac{\lambda}{\mu} p_0$

$p_2 = (\frac{\lambda}{\mu})^2 p_0$

in general $p_k = (\frac{\lambda}{\mu})^k p_0$ means prob k jobs in system

Let $\rho = \frac{\lambda}{\mu}$, $p_k = \rho^k p_0$

$p_0 + p_1 + p_2 + \dots = 1$ $\quad \frac{1}{1-\rho} p_0 = 1$

$(1 + \rho + \rho^2 + \dots) p_0 = 1$ $\quad p_k = (1-\rho)\rho^k$ $\quad \rho < 1 \Rightarrow \lambda < \mu$

The mean number of jobs in the system

$$= \sum_{k=0}^{\infty} k p_k = \sum_{k>0}^{\infty} k(1-\rho)\rho^k = \frac{\rho}{1-\rho}$$

$$\frac{\rho}{1-\rho} = \lambda * T, \quad T = \frac{\rho}{\lambda(1-\rho)} = \frac{1}{\mu(1-\rho)}$$

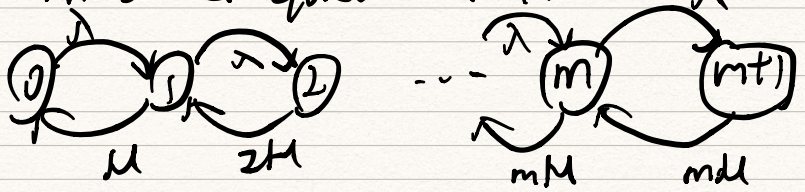
mean response time $T = \frac{1}{\mu - \lambda} = \frac{1}{\mu(1-\rho)}$ 03A (2)

mean service time $S = \frac{1}{\mu}$

mean waiting time $W = \frac{1}{\mu-\lambda} - \frac{1}{\mu} = \frac{\lambda}{\mu(\mu-\lambda)}$

Doubling the service rate can sometimes reduce response time by a factor more than 2

Multi-server queues $M/M/m$



$$T = \frac{C(\rho, m)}{m\mu(1-\rho)} + \frac{1}{\mu} \qquad \rho = \frac{\lambda}{m\mu}$$

$$C(\rho, m) = \frac{\frac{(m\rho)^m}{m!}}{(1-\rho)\sum_{k=0}^{m-1}\frac{(m\rho)^k}{k!} + \frac{(m\rho)^m}{m!}}$$

$M/M/2$: $C(\rho, 2) = \frac{2\rho^2}{(1-\rho)(4\rho^2+2\rho+1)}$

$$T = \frac{\frac{2\rho^2}{(1-\rho)(4\rho^2+2\rho+1)}}{2\mu(1-\rho)} + \frac{1}{\mu}$$

Multi-server queues $M/M/m/m$ with no waiting room

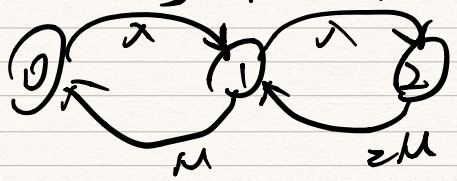
Prob n customers in the system is

$$p_m = \frac{\frac{\rho^m}{m!}}{\sum_{k=0}^{m}\frac{\rho^k}{k!}} \quad \text{where } \rho = \frac{\lambda}{\mu}$$

"Erlang B formula"

continuous-time Markov chain

03B
c1)



X is # users at CPU Y is # users at fast disk

total six states 3 - - - - - - slow ...

(2,0,0), (1,1,0), (1,0,1)

(0,2,0), (0,1,1), (0,0,2)

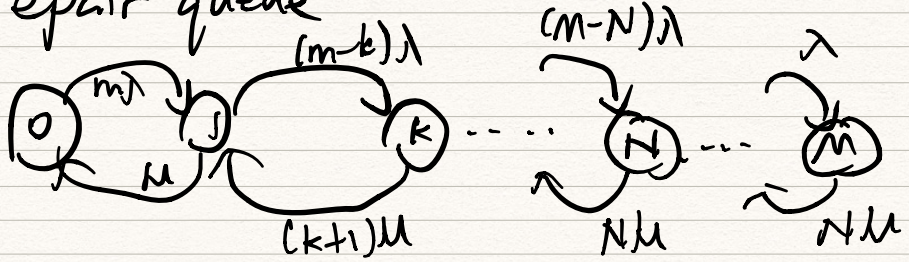
if n users states = $\frac{(n+1)(n+2)}{2}$

CPU utilisation = $P(2,0,0) + P(1,1,0) + P(1,0,1)$

Throughput = utilisation × service rate

Response time $T = \frac{N}{\lambda}$

repair queue



There are (M+1) states

M machines. N staff repair machines

$\lambda = \frac{1}{\text{mean time to failure}}$

$\mu = \frac{1}{\text{mean service time to repair a machine}}$

Machines in operation (max M)

Machines waiting to be repaired (max M-N) (2)

machines being repaired (max N)

P(0) means all machines are working

P(M-k) means k machines are working

$$P(k) = \begin{cases} P(0)\left(\frac{\lambda}{\mu}\right)^k C_k^m & k = 1, \cdots N \\ P(0)\left(\frac{\lambda}{\mu}\right)^k C_k^m \frac{N^{M-k} k!}{N!} & k = N+1, \cdots, M \end{cases}$$

$$P(0) = \left[\sum_{k=0}^{N} \left(\frac{\lambda}{\mu}\right)^k C_k^M + \sum_{k=N+1}^{M} \left(\frac{\lambda}{\mu}\right)^k C_k^M \frac{N^{N-k} k!}{N!} \right]^{-1}$$

Mean machine failure rate

$$\bar{X}_f = \sum_{k=0}^{M-1} (M-k)\lambda P(k)$$

Mean time to failure (MTTF) $= \frac{1}{\lambda}$

Mean time to repair (MTTR)

= queueing time for repair + actual repair time

(use Little's law)

when λ, μ is not exponentially distributed. need to be independence

94A (1)

M/G/1 is not a Markov chain

Arrival is Poisson with rate λ

Service time S has: Mean $= \frac{1}{\mu} = E[S]$ = first moment

second moment $= E[S^2]$

Mean waiting time W of M/G/1:

$$W = \frac{\lambda E[S^2]}{2(1-\rho)} \quad \rho = \frac{\lambda}{\mu}$$ P-K formula

Mean respone time $T = W + E[S]$

By Little's law, Average #messages in the system

= Throughput * mean response time

$= \lambda T$

$$E[S^2] = E[S]^2 + \sigma^2 = E[S]^2(1+C_a^2)$$ σ 标准差

$$W = \frac{\lambda(E[S]^2 + \sigma_s^2)}{2(1-\rho)}$$ smaller variance in S → smaller waiting time

M/D/1 is a special case of M/G/1

D: deterministic, constant service time $E[S]$ and Variance = 0

for the same value of ρ and $E[S]$, deterministic has the smallest mean response time.

$E[X] = \int x f(x) dx$, $E[X^2] = \int x^2 f(x) dx$

if S exponential with rate μ $E[S] = \frac{1}{\mu}$, $E[S^2] = \frac{2}{\mu^2}$

when M/G/1 $E[S] = \frac{1}{\mu}$, $E[S^2] = \frac{2}{\mu^2}$ 04 A

$W = \frac{\rho}{\mu(1-\rho)}$ which is the same as M/M/1 (2)

M/G/1:

W = Mean waiting time N = Mean number of customers in the queue

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

R = Mean residual service time

$W = N * (\frac{1}{\mu}) + R$ Little's law $N = \lambda W$

$W = \frac{R}{1-\rho}$ $\rho = \frac{\lambda}{\mu}$ $R = \frac{1}{2}\lambda E[S^2]$

Assuming M jobs are completed in time T

Mean residual time = $\frac{\sum_{i=1}^{M} \frac{1}{2} S_i^2}{T} = \frac{1}{2} \frac{\sum_{i=1}^{M} S_i^2}{M} \frac{M}{T} =$

G/G/1:

Mean arrival rate = λ

variance of inter-arrival time = σ_a^2

service time S has mean $\frac{1}{\mu} = E[S]$

variance of service time = σ_s^2

$W \approx \frac{\lambda^2(\sigma_a^2 + \sigma_s^2)}{1 + \lambda^2 \sigma_s^2} \frac{\lambda(E[S^2] + \sigma_s^2)}{2(1-\rho)}$ $\rho = \frac{\lambda}{\mu}$

large variance means large waiting time

bound: $W <= \frac{\lambda(\sigma_a^2 + \sigma_s^2)}{2(1-\rho)}$

G/G/m: 04A (3)

$$W_{G/G/m} = W_{M/M/m} \frac{C_a^2 + C_s^2}{2}$$

$W_{M/M/m}$ = waiting time of M/M/m queue

C_a = coeff of variation of inter-arrival time

C_b = ------- service time

coefficient of variation of a random variable X

$$= \frac{\text{standard deviation of } X}{\text{mean of } X} = \frac{\sigma_s}{E[S]}$$

PS sharing time0, time X ....

04B (1)

non-preemptive:

$W_1 = N_1 E[S_1] + R$ High priority

$N_1 = \lambda W_1$ $\boxed{W_1 = \frac{R}{1-\rho_1}}$ $\boxed{\rho_1 = \lambda_1 E[S_1]}$

$\boxed{R = \frac{1}{2} E[S_1^2] \lambda_1 + \frac{1}{2} E[S_2^2] \lambda_2}$

$\boxed{W_2} = R + N_2 E[S_2] + N_1 E[S_1] + \lambda_1 W_2 E[S_1]$

$$= \frac{R + \rho_1 W_1}{1 - \rho_1 - \rho_2} = \boxed{\frac{R}{(1-\rho_1)(1-\rho_1-\rho_2)}}$$

$$W_k = \frac{R}{(1-\rho_1-\rho_2\cdots-\rho_{k-1})(1-\rho_1-\rho_2\cdots-\rho_k)}$$

04B (2)

$$R = \frac{1}{2}\sum_{i=1}^{p} E[S_i^2]\lambda_i \qquad \rho_i = \lambda_i E[S_i]$$

Pre-emptive: T: response time

$$T_1 = E[S_1] + \frac{R_1}{(1-\rho_1)} \quad \text{(highest priority)}$$

$$T_k = E[S_k] + \frac{R_k}{1-\rho_1-\rho_2\cdots-\rho_k} + \left(\sum_{i=1}^{k-1}\rho_i\right)T_k$$

$$R_k = \frac{1}{2}\sum_{i=1}^{k} E[S_i^2]\lambda_i$$

$$\boxed{T_k = T_{k,1} + T_{k,2}}$$

$$T_{k,1} = \frac{E[S_k]}{(1-\rho_1-\rho_2\cdots\rho_{k-1})}$$

$$T_{k,2} = \frac{R_k}{(1-\rho_1-\cdots\rho_{k-1})(1-\rho_1-\cdots\rho_k)}$$

$$R_k = \frac{1}{2}\sum_{i=1}^{k} E[S_i^2]\lambda_i$$

in C: linear congruential generator

$Z_k = a Z_{k-1} + c \pmod{m}$

python, matlab: Mersenne Twister random number generator

A method to generate random number from a particular distribution is the inverse transform method

generate u uniformly distributed in $(0,1)$

compute $F^{-1}(u)$

CPU (10s) Fast Disk (15s) Slow Disk (30s) 05B

$S_1 = 10,\ S_2 = 15,\ S_3 = 30$ MVA only exponential service time

$V_1 = 1,\ V_2 = \frac{1}{2},\ V_3 = \frac{1}{2}$

$R_j(n) = (\bar{n}_j(n-1) + 1) S(j)$

$n = 0,\ \bar{n}_1(0) = \bar{n}_2(0) = \bar{n}_3(0) = 0$

$n = 1\ R_1(1) = (\bar{n}_1(0) + 1) S_1 = 10$

$R_2(1) = 15\quad R_3(1) = 30$

$R_0(1) = V_1 R_1(1) + V_2 R_2(1) + V_3 R_3(1) = 32.5$

$X_0(1) = \frac{1}{R_0(1)} = \frac{1}{32.5}$

$X_1(1) = V_1 X_0(1) = \frac{1}{32.5}$ $\quad X_2(1) = X_3(1) = \frac{1}{2} \cdot \frac{1}{32.5} = \frac{1}{65}$

$\bar{n}_1(1) = R_1(1) \cdot X_1(1) = 10 \cdot \frac{1}{32.5} = \frac{4}{13}$

$\bar{n}_2(1) = R_2(1) \cdot X_2(1) = \frac{3}{13}$ $\quad \bar{n}_3(1) = R_3(1) \cdot X_3(1) = \frac{6}{13}$

$n = 2\ R_1(2) = (\bar{n}_1(1) + 1) S_1 = (\frac{4}{13} + 1) \cdot 10 = \frac{120}{13}$

$R_2(2) = (\bar{n}_2(1) + 1) S_2 = \frac{240}{13}$

$R_3(2) = \ldots$

$X_0(n)$ = Throughput of the system

$\bar{n}_i(n)$ = Mean # of customers in device i

$R_i(n)$ = Mean response time in device i

$R_0(n)$ = Mean response time of the system

$X_i(n)$ = Throughput of device i

transient removal:

$$\frac{X(m+1)+X(m+2)\cdots X(N)}{N-m}$$

mean: $\hat{\bar{T}} = \frac{\sum_{i=1}^{n} T(i)}{n}$

standard deviation: $\hat{S} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{\bar{T}} - T(i))^2}{n-1}}$

there is a probability $(1-\alpha)$ mean response time in the interval

$$\left[\hat{\bar{T}} - t_{n-1,1-\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}},\ \hat{\bar{T}} + t_{n-1,1-\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}}\right]$$

see from student t distribution

95% confidence interval $\alpha = 0.05$