Service Engineering April 15, 1997

Last Revised December 15, 2003

REVIEW: MARKOV JUMP-PROCESS (MJP)

MJP $X = \{X_t,\ t \ge 0\}$ on $\mathcal{S} = \{i, j, \ldots\}$ countable.
Markov property: $P_r\{X_t = j | X_r,\ r < s;\ X_s = i\} = P_{ij}(s,t), \quad \forall\, s < t,\ \forall\, i, j \in \mathcal{S}$.
Time homogeneity: $P_r\{X_{s+t} = j | X_s = i\} = P_{ij}(t), \quad \forall\, s, t,\ i, j,$ transition probabilities.

Characterization: π^0 = initial distribution and $P(t) = [P_{ij}(t)],\ t \ge 0$, stochastic.
Finite-dimensional distributions:
$P_r\{X_0 = i_0,\ X_{t_1} = i_1, \ldots, X_{t_n} = i_n\} = \pi^0(i_0) P_{i_0,i_1}(t_1) \ldots P_{i_{n-1},i_n}(t_n - t_{n-1})$.

$P(t)$: stochastic ; $P(s+t) = P(s)P(t), \quad \forall\, s, t$ (Chapman Kolmogorov);
$\exists\, P(0) = I$; $\exists\, \dot{P}(0) = Q = [q_{ij}]$, infinitesimal generator $\left(\sum_{j \in \mathcal{S}} q_{ij} = 0\right)$.

Micro to Macro : $\dot{P}(t) = P(t)Q \quad (= QP(t))$ and $P(0) = I$
Forward (Backward) equations.

Solution : $P(t) = \exp[tQ] = \sum_{n=0}^{\infty} \frac{t^n}{n!} Q^n ,\ t \ge 0$.

Animation: $i \xrightarrow{q_{ij}} j; \quad \forall\, i, j \in \mathcal{S}\ \exists$ exponential clock at rate q_{ij}, call it (i,j).
Given i, consider clocks $(i,j),\ j \in \mathcal{S}$; move to the "winner" when rings.
Thus: stay at $i \sim \exp(q_i = \sum_{j \ne i} q_{ij})$ and switch to j with probability $P_{ij} = q_{ij}/q_i$
$(q_{ij} = q_i P_{ij}, i \ne j; q_{ii} = -q_i)$.

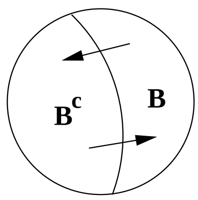
Transient analysis	vs. long-run/limit	stability/steady-state
	$\exists\, \lim_{t \uparrow \infty} P_{ij}(t) = \pi_j,\ \forall\, i;$	$\pi = \pi P(t),\ \forall\, t.$

Calculation via **steady-state equations**: $\dot{P}(\infty) = P(\infty)Q \Rightarrow \left\{ \begin{array}{l} 0 = \pi Q \\ \sum_i \pi_i = 1,\ \pi_i \ge 0 \end{array} \right\}$

or balance equations: $\sum_{i \ne j} \pi_i q_{ij} = -\pi_j q_{jj} = \sum_{i \ne j} \pi_j q_{ji},\ \forall\, j$.

Transition rates: $\pi_i q_{ij}$ = long-run average number of switches from i to j.

Cuts: $\sum_{i \in B} \sum_{j \in B^c} \pi_i q_{ij} = \sum_{i \in B^c} \sum_{j \in B} \pi_i q_{ij},\ \forall\, B \subset \mathcal{S}$.

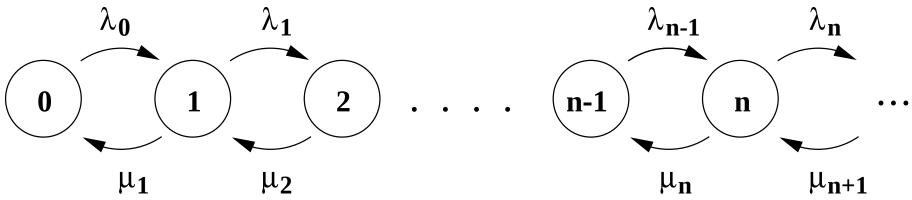


Ergodic Theorem: Let X be *irreducible* ($i \leftrightarrow j$). Assume that there exists a solution π to its steady-state equations. Then, X must be "unexplosive" and π must be its stationary distribution, its limit distribution and

SLLN $\bullet \lim_{T\uparrow\infty} \frac{1}{T} \int_0^T f(X_t)dt = \sum_i \pi_i f(i)$ ("$=$"$Ef(X_\infty)$) ; eg. $f(x) = 1_B(x)$.

$\bullet \lim_{T\uparrow\infty} \frac{1}{T} \sum_{t\le T} g(X_{t-}, X_t) = \sum_i \pi_i \sum_j q_{ij} g(i,j)$, for $g(x,x) = 0$, $\forall x$; e.g. $g(x,y) = 1_C(x,y)$.

Birth & Death Model of a Service Station (*Hall, §5.4*)



Cuts at $n \leftrightarrow n+1$ yield: $\pi_n \lambda_n = \pi_{n+1}\mu_{n+1}$, $n \ge 0$;

$$\pi_{n+1} = \frac{\lambda_n}{\mu_{n+1}} \pi_n = \frac{\lambda_n \lambda_{n-1}}{\mu_{n+1}\mu_n} \pi_{n-1} = \cdots = \frac{\lambda_0 \lambda_1 \ldots \lambda_n}{\mu_1 \mu_2 \ldots \mu_{n+1}} \pi_0 \ .$$

The required solution exists if and only if

$$\sum_{n=0}^{\infty} \frac{\lambda_0 \ldots \lambda_n}{\mu_1 \ldots \mu_{n+1}} < \infty \ .$$

The Ergodic Theorem then yields

$$\begin{cases} \pi_n = \frac{\lambda_0 \ldots \lambda_{n-1}}{\mu_1 \ldots \mu_n} \pi_0 \ , \ n \ge 0 \\ \pi_0 = \left[\sum_{n\ge 0} \frac{\lambda_0 \ldots \lambda_n}{\mu_1 \ldots \mu_{n+1}}\right]^{-1} \end{cases}$$

Measures of Performance (MOP's):

L = number of customers at the service station;
L_q = number of customers in the queue;
W = sojourn time of a customer at the service station;
W_q = waiting time of a customer in the queue;

$$E(L) = \sum_{n\ge 0} n\pi_n = \lim_{T\uparrow\infty} \frac{1}{T} \int_0^T L(t)dt.$$

Let $m(n)$ = number of active servers at state n, $0 \le m(n) \le n$; the servers are **statistically identical**.

$$E(L_q) = \sum_{n \ge 0} [n - m(n)]\pi_n \quad = \text{ also long-run average, as above.}$$

Service rate per server is $\mu(n)/m(n),\ n \ge 1$.

Average (actual) **service rate**: $\sum_{n \ge 1} \frac{\mu(n)}{m(n)} \pi_n = E[\mu(L)/m(L)]$.

Potential service rate of each server: $E[\mu(L)/m(L) | L > 0] = \frac{E[\mu(L)/m(L)]}{1-\pi_0}$.

Inflow rate: $\lambda = \sum_{n \ge 0} \pi_n q_{n,n+1} = \sum_{n \ge 0} \pi_n \lambda(n) = E\lambda(L)$.

↑ arrival

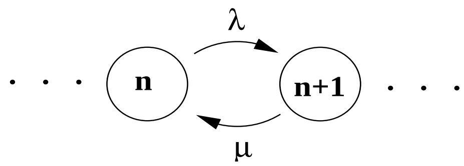
↓ departure

Outflow rate: $\delta = \sum_{n \ge 0} \pi_n q_{n,n-1} = \sum_{n \ge 0} \pi_n \mu(n) = E\mu(L)$. (Assume $\mu(0) = 0$.)

Note: in steady state, $\pi_n \lambda_n = \pi_{n+1} \mu_{n+1}$, $\forall\ n \ge 0 \Rightarrow$ inflow rate = outflow rate.

Throughput rate: $E\lambda(L) = E\mu(L)$ (the common quantity).

Example. **M/M/1**



$\lambda_j = \lambda,\ j \ge 0;\quad \mu_j = \mu \cdot 1_{j \ge 1}$.

$\rho = \frac{\lambda}{\mu} < 1$ assumed for steady state (traffic intensity).

$\pi_n = (1-\rho)\rho^n,\ n \ge 0.$ Geometric distribution!

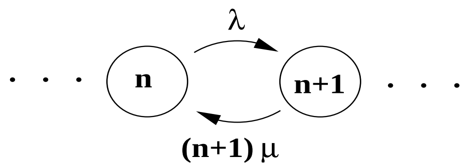
Actual service rate = $\sum_{n \ge 1} \pi_n \cdot \mu = \mu(1-\pi_0) = \mu \cdot \rho = \mu \cdot \frac{\lambda}{\mu} = \lambda$, contrasted with

Potential service rate = $\frac{\lambda}{1-\pi_0} = \frac{\lambda}{\rho} = \mu$, as anticipated.

Additional properties: $W \sim \exp\left(\text{mean} = \frac{1}{\mu(1-\rho)} = \frac{1}{\mu}\left[1 + \frac{\rho}{1-\rho}\right]\right)$, geometric mixture of exp's.
Departure process is Poisson (λ) (Burke's Theorem).

$$\frac{W_q}{1/\mu} \stackrel{d}{=} \begin{cases} 0 & \text{wp } 1-\rho \\ \exp\left(\text{mean} = \frac{1}{1-\rho}\right) & \text{wp } \rho \end{cases}$$

Example M/M/∞



Always stable.

$\pi_n = e^{-\rho}\frac{\rho^n}{n!}$, $n \geq 0$, Poisson distribution!

E (# busy servers) $= \lambda \cdot \frac{1}{\mu} = \frac{\lambda}{\mu} = \rho.$

Very *useful*: ∞-server models provide upper bound (e.g., Israel Electric Company).

Example M/M/S

$$\mu_j = (j \wedge s)\mu\ , \quad \lambda_j \equiv \lambda\,,$$

$$\rho = \frac{\lambda}{s\mu} < 1 \qquad \text{assumed, as before, to ensure stability.}$$

$$\begin{aligned} \pi_k &= \frac{a^k}{k!}\,\pi_0, \qquad k \leq S, \\ &= \frac{s^s\rho^k}{s!}\,\pi_0, \qquad k \geq S, \\ \pi_0 &= \left[\sum_{j=0}^{s-1}\frac{a^j}{j!} + \frac{a^s}{s!(1-\rho)}\right]^{-1}, \text{ where } a = \frac{\lambda}{\mu}\ , \text{ offered load.} \end{aligned}$$

Note: "Wait | Wait > 0" is exponential, having the same distribution as that in an M/M/1 queue with arrival rate λ and service rate $S \cdot \mu$.

Erlang-C Formula (1917):

$$E_{2,S} = \sum_{k \geq s}\pi_k = \frac{a^S}{S!}\,\frac{1}{1-\rho}\cdot\pi_0, \qquad \text{delay probability (PASTA).}$$

Example M/M/S/S

$$\lambda_j \equiv \lambda, j = 0, \ldots, S-1, \quad \mu_j = j \cdot \mu \ \text{ for } j = 1, 2, \ldots, S\,.$$

Always reaches steady state.

$$\pi_k = \frac{a^k}{k!} \Big/ \sum_{j=0}^{S}\frac{a^j}{j!}\ , \qquad k = 0, 1, \ldots, S.$$

Erlang-B Formula:

$$E_{1,S} = \pi_S = \frac{a^s}{s!} \Bigg/ \sum_{j=0}^{S} \frac{a^j}{j!} \,, \qquad \text{loss probability (PASTA).}$$

$\lambda\pi_s$ – rate of lost customers,
$\lambda(1-\pi_s)$ – effective throughput.

Note: Useful relations between the Erlang-B and Erlang-C formulae are

$$E_{1,S} = \frac{(S-a)E_{2,S}}{S-aE_{2,S}} \;;\; E_{2,S} = \frac{E_{1,S}}{(1-\rho)+\rho E_{1,S}} \;;$$

$$E_{2,S} > E_{1,S}, \text{ as expected: why?}$$

The expression of $E_{2,S}$ in terms of $E_{1,S}$ will become especially useful later on.

Example **M/M/S/N** $(S \le N)$

$$\lambda_j = \lambda, \qquad 0 \le j \le N-1, \qquad (\lambda_N = 0)$$

$$\mu_j = (j \wedge S)\mu, \qquad 1 \le j \le N. \qquad (\mu_0 = 0)$$

Formulae straightforward but cumbersome (simply truncate M/M/S).

Always reaches steady state.

Note: Mainly M/M/S (Erlang-C) and sometimes M/M/S/S (Erlang-B) are the prevalent models used in the world of call centers. However, M/M/S/N is more appropriate, and even more so M/M/S/N + Abandonment: Erlang-A.

But the following question then arises: How to model Abandonment?

Erlang's Formulae

(Exact Results for M/M/m = Erlang-C, and M/M/m/m = Erlang-B)

R = offered load $\left(= \lambda/\mu = m \cdot \rho \; ; \; \rho = \frac{R}{m}\right)$

Erlang B: $E_{1,m} = \frac{\frac{R^m}{m!}}{\sum_{k=0}^{m} \frac{R^k}{k!}}$ Probability of *blocking*/loss

Erlang C: $E_{2,m} = \frac{\frac{R^m}{m!} \frac{1}{1-\rho}}{\sum_{k=0}^{m-1} \frac{R^k}{k!} + \frac{R^m}{m!} \frac{1}{1-\rho}}$ Probability of *delay*

Relations (Palm, 1943?)

- Some observations on the Erlang formulae.........pg. 18
- Contributions to the Theory of Delay Systems ......pg. 37

1. $$E_{2,n} = \frac{nE_{1,n}}{(n-R) + RE_{1,n}} = \frac{E_{1,n}}{(1-\rho) + \rho E_{1,n}} \quad \text{for} \quad \rho < 1 \; \left(\rho = \frac{R}{n}\right)$$

   $E_{2,n} > E_{1,n} \quad ; \quad \frac{d}{dR} E_{2,n}(n) = \frac{1}{nE_{1,n}(n)}$

2. $$E_{2,n} = \frac{R(n-1-R)E_{2,n-1}}{(n-1)(n-R) - RE_{2,n-1}} \quad \text{for} \quad R < n-1.$$

   (Must have $R < 1$ to start with $E_{2,1} = \rho$)

3. $$E_{1,n} = \frac{RE_{1,n-1}}{n + RE_{1,n-1}} = \frac{\rho E_{1,n-1}}{1 + \rho E_{1,n-1}} \quad ; \quad E_{1,0} = 1.$$

Recursions are useful for calculations.
For example, to calculate $E_{2,n}$, it is convenient to calculate recursively $E_{1,n}$ via 3. and then calculate $E_{2,n}$ via 1.
They will also be useful for us in asymptotic analysis of systems with many servers.
For example, to analyze the behavior of $E_{2,n}$, as $n \uparrow \infty$, it is convenient to analyze first $E_{1,n}$, and then use 1.

Recall: Erlang B/C/A formulae, and much more, are implemented in *4CallCenters* that you have been using.