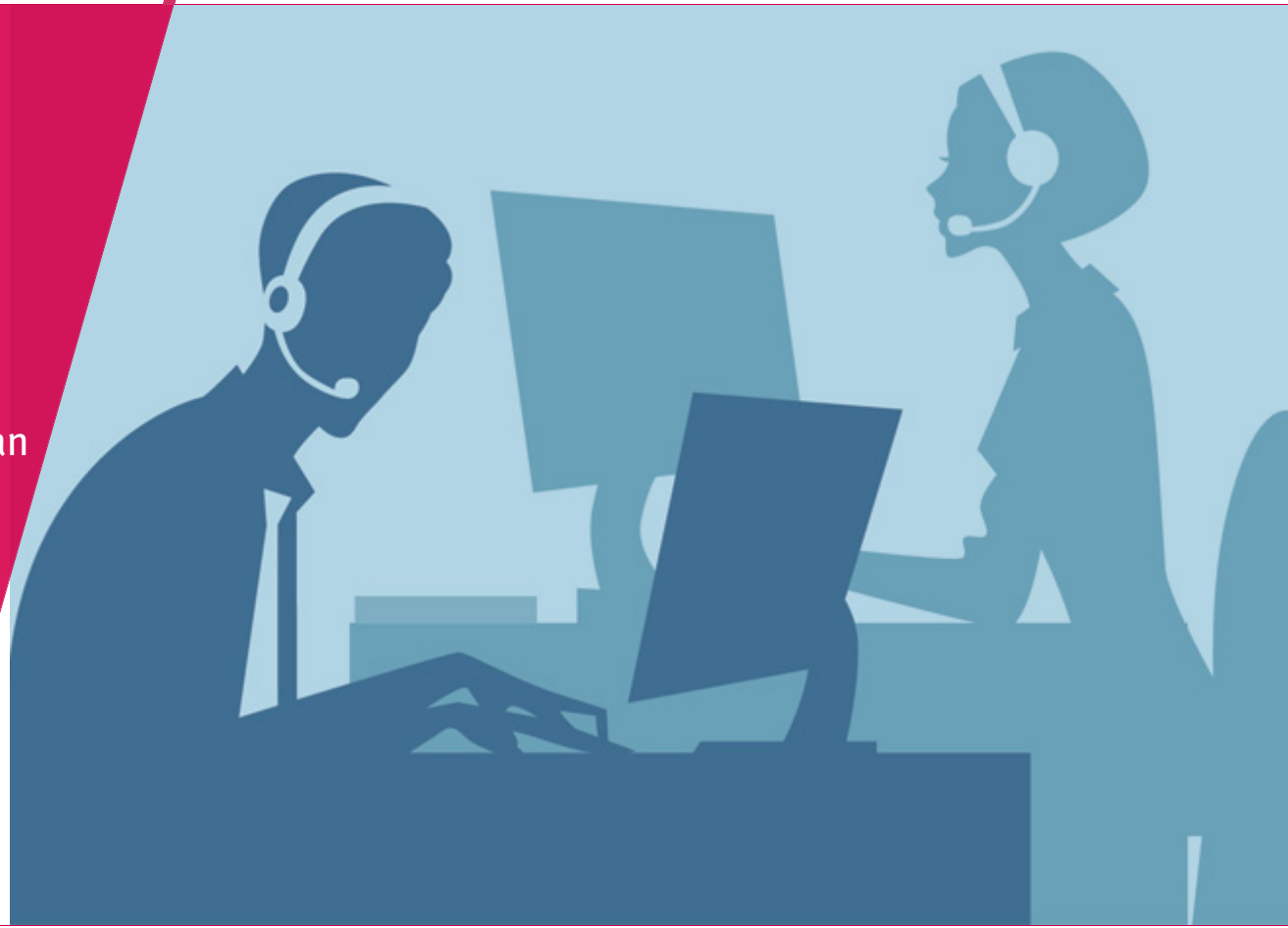
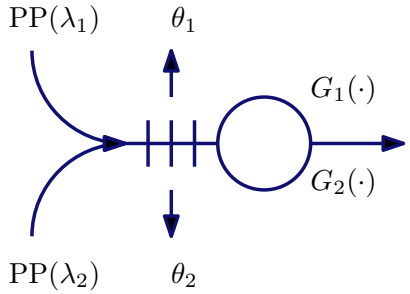
Technische Universiteit
Eindhoven
University of Technology

Queueing model with heterogeneous reneging customers

Vidyadhar Kulkarni, Brett Hathaway, Ivo Adan

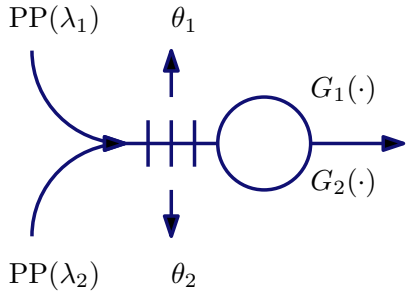


Single server system



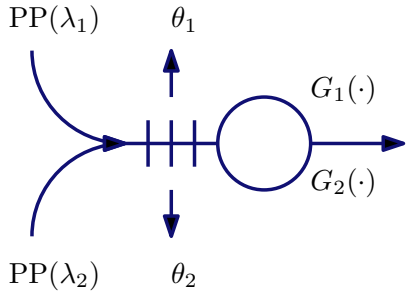
- Two types of customers, type 1 and 2
- Customers of type i
  - arrive according to PP(λ_i)
  - need service times with cdf G_i, LST $\tilde{G}_i$ and mean τ_i
  - exponential patience times with rate θ_i:
    customer leaves the queue if service does not start before his patience time expires

Single server system



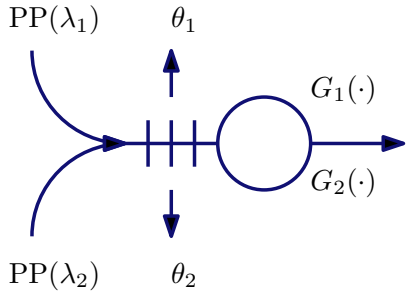
- Questions:
  - What is the fraction of customers leaving without service?
  - What is the (mean) waiting time?

Single server system



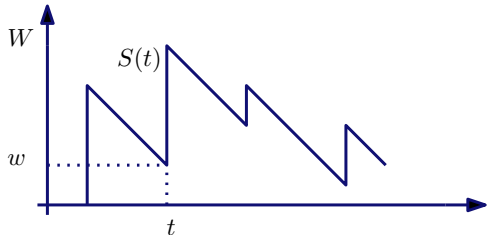
- Queue length process:
  - total number in queue is not sufficient for Markovian description
  - keep track of detailed composition of queue $(1, 1, 2, 1, 2)$ and residual service time $R(t)$ at time t
- Workload or virtual queueing time process $W(t)$:
  - workload decreases at rate 1 (while positive)
  - if arrival of type i occurs at time t and $W(t) = w$, then arrival
    - *immediately leaves* with probability $1 - e^{-\theta_i w}$, or
    - enters service with probability $e^{-\theta_i w}$, and needs an amount of service with cdf G_i and mean τ_i

Single server system



- Workload process of $M/G/1$ with workload dependent service times
- Distribution of the size $S(t)$ of upward jump at time t, when $W(t) = w$, is given by (where $\lambda = \lambda_1 + \lambda_2$)

$$P(S(t) \leq y | W(t) = w) = \sum_{i=1}^{2} \frac{\lambda_i}{\lambda}(1 - e^{-\theta_i w} + e^{-\theta_i w} G_i(y)) = 1 - \sum_{i=1}^{2} \frac{\lambda_i}{\lambda} e^{-\theta_i w}(1 - G_i(y))$$



Analysis of workload process

- Let $\psi(s,t) = E(e^{-sW(t)})$
- Conditioning on whether $W(t) > 0$ or $W(t) = 0$, we get for small $h > 0$

$$\psi(s,t+h) = sh(\psi(s,t) - P(W(t)=0)) + (1-\lambda h)\psi(s,t) + \lambda h\left(\psi(s,t) - \sum_{i=1}^{2}\frac{\lambda_i}{\lambda}\psi(s+\theta_i,t)(1-\tilde{G}_i(s))\right) + O(h^2)$$

- Dividing by h and letting $h \to 0$

$$\frac{d}{dt}\psi(s,t) = s\psi(s,t) - sP(W(t)=0) - s\sum_{i=1}^{2}\lambda_i\psi(s+\theta_i,t)\frac{1-\tilde{G}_i(s)}{s}$$

- Take $t \to \infty$, then $\psi(s,t) \to \psi(s)$, $\frac{d}{dt}\psi(s,t) \to 0$, and $P(W(t)=0) \to p_0$, so

$$\psi(s) = p_0 + \sum_{i=1}^{2}\psi(s+\theta_i)H_i(s)$$

where $H_i(s) = \lambda_i\frac{1-\tilde{G}_i(s)}{s}$ (note that $H_i(s)/(\lambda_i\tau_i)$ is LST of residual service time)

Analysis of workload process

- Using the normalization $\psi(0) = 1$ yields

$$p_0 = 1 - \sum_{i=1}^{2} \psi(\theta_i)\lambda_i\tau_i$$

Analysis of workload process: Remarks

- Alternative derivation by level crossing:
  Equating the rate of down crossings through work load level $W = w$ and the rate of up crossings

$$f(w) = p_0 \sum_{i=1}^{2} \lambda_i (1 - G_i(w)) + \int_{y=0}^{w} \sum_{i=1}^{2} \lambda_i e^{-\theta_i y} (1 - G_i(w - y)) f(y) dy.$$

  where f is density of W. Multiplying by e^{-sw} and integrating from 0 to ∞

$$\psi(s) - p_0 \;=\; p_0 \sum_{i=1}^{2} H_i(s) + \sum_{i=1}^{2} (\psi(s + \theta_i) - p_0)\, H_i(s) \;=\; \sum_{i=1}^{2} \psi(s + \theta_i) H_i(s)$$

- Hyper-exp(θ_{ij}, p_{ij}) patience of type i customers:

$$\psi(s) - p_0 = \sum_{i=1}^{2} \sum_{j} p_{ij} \psi(s + \theta_{ij}) H_i(s)$$

  where $p_0 = 1 - \sum_{i=1}^{2} \sum_j p_{ij} \psi(\theta_{ij}) \lambda_i \tau_i$

Performance measures

- Type i customer enters service if his patience time T_i is longer than W, so probability of entering service

$$P(T_i > W) = E(e^{-\theta_i W}) = \psi(\theta_i)$$

- Probability that server is busy serving a type i customer

$$\rho_i = \lambda_i \tau_i \psi(\theta_i)$$

- Probability that the server is busy

$$\rho = \psi(\theta_1)\lambda_1\tau_1 + \psi(\theta_2)\lambda_2\tau_2$$

- Steady state throughput

$$\lambda_1\psi(\theta_1) + \lambda_2\psi(\theta_2)$$

- Reneging rate

$$\lambda_1(1 - \psi(\theta_1)) + \lambda_2(1 - \psi(\theta_2))$$

Performance measures

- Expected time waiting for service

$$E(\min(W, T_i)) = E(E(\min(W, T_i)|W)) = E\left(\frac{1 - e^{-\theta_i W}}{\theta_i}\right) = \frac{1 - \psi(\theta_i)}{\theta_i}$$

- Expected number of type i customers waiting for service

$$E(L_i^q) = \lambda_i E(\min(W, T_i)) = \frac{\lambda_i}{\theta_i}(1 - \psi(\theta_i))$$

- Conclusion: We need the quantities $\psi(\theta_i)$ to compute performance measures

Calculation of $\psi(s)$

- Solve $\psi(s)$ recursively from

$$
\begin{aligned}
\psi(s) &= p_0 + \psi(s+\theta_1)H_1(s) + \psi(s+\theta_2)H_2(s) \\
&= p_0(1 + H_1(s) + H_2(s)) \\
&\quad +\psi(s+2\theta_1)H_1(s+\theta_1)H_1(s) \\
&\quad +\psi(s+\theta_1+\theta_2)(H_2(s+\theta_1)H_1(s) + H_1(s+\theta_2)H_2(s)) \\
&\quad +\psi(s+2\theta_2)H_2(s+\theta_2)H_2(s)
\end{aligned}
$$

Calculation of $\psi(s)$

- Recursion yields the following expression for $\psi(s)$:

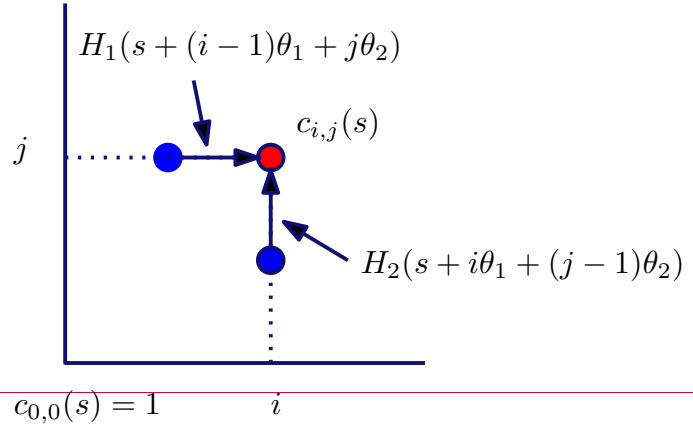
$$\psi(s) = p_0 c(s)$$

where $c(s) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} c_{i,j}(s)$

- Terms $c_{i,j}(s)$ are determined from recursion

$$c_{i,j}(s) = H_1(s + (i-1)\theta_1 + j\theta_2) c_{i-1,j}(s) + H_2(s + i\theta_1 + (j-1)\theta_2) c_{i,j-1}(s)$$

with initially $c_{0,0}(s) = 1$ and $c_{i,j}(s) = 0$ if $i < 0$ or $j < 0$



Calculation of $\psi(s)$

- Probability $p_0 = P(W = 0)$ follows from normalization

$$p_0 = 1 - \sum_{i=1}^{2} \psi(\theta_i)\lambda_i\tau_i = 1 - p_0 \sum_{i=1}^{2} c(\theta_i)\lambda_i\tau_i$$

Special case: Single Class with exponential service rate μ

- Set $\lambda_2 = 0$ and $\lambda_1 = \lambda$, $\theta_1 = \theta$, $H_1(s) = \frac{\lambda}{s+\mu}$
- Equation for $\psi(s)$ simplifies to

$$\psi(s) = p_0 + \psi(s+\theta)\frac{\lambda}{s+\mu}$$

- Recursively solving yields

$$\psi(s) = p_0\left[1 + \sum_{n=1}^{\infty}\prod_{m=0}^{n-1}\frac{\lambda}{s+m\theta+\mu}\right]$$

- Setting $s = \theta$ and substituting $p_0 = 1 - \psi(\theta)\lambda/\mu$

$$\psi(\theta) = \frac{1 + \sum_{n=1}^{\infty}\prod_{m=0}^{n-1}\frac{\lambda}{m\theta+\mu}}{1 + \sum_{n=0}^{\infty}\prod_{m=0}^{n}\frac{\lambda}{m\theta+\mu}}$$

- Throughput rate

$$\psi(\theta)\lambda = \frac{1 + \sum_{n=1}^{\infty}\prod_{m=0}^{n-1}\frac{\lambda}{m\theta+\mu}}{1 + \sum_{n=0}^{\infty}\prod_{m=0}^{n}\frac{\lambda}{m\theta+\mu}}\ \lambda$$

CTMC analysis

- $X(t)$ number in system at time t
- $\{X(t), t \geq 0\}$ is BD on $\{0, 1, 2, \cdots\}$ with birth rates $\lambda_n = \lambda \ (n \geq 0)$ and death rates $\mu_n = \mu + (n-1)\theta \ (n \geq 1)$
- From standard analysis

$$p_0 = \lim_{t\to\infty} P(X(t) = 0) = \left[1 + \sum_{n=0}^{\infty} \prod_{m=0}^{n} \frac{\lambda}{\mu + m\theta}\right]^{-1}$$

$$p_n = \lim_{t\to\infty} P(X(t) = 0) = p_0 \prod_{m=0}^{n} \frac{\lambda}{\mu + m\theta}$$

- Throughput rate

$$(1 - p_0)\mu = \frac{\sum_{n=0}^{\infty} \prod_{m=0}^{n} \frac{\lambda}{\mu+m\theta}}{1 + \sum_{n=0}^{\infty} \prod_{m=0}^{n} \frac{\lambda}{\mu+m\theta}} \mu = \frac{1 + \sum_{n=1}^{\infty} \prod_{m=0}^{n-1} \frac{\lambda}{m\theta+\mu}}{1 + \sum_{n=0}^{\infty} \prod_{m=0}^{n} \frac{\lambda}{\mu+m\theta}} \lambda = \psi(\theta)\lambda$$

Multi server system with k servers and identical exponential service rates $\mu_i = \mu$

- $W(t)$ is virtual queueing time at time t
- $K(t)$ is number of busy servers at time t
- $W(t) > 0$ if and only if $K(t) = k$, and $W(t) = 0$ if and only if $K(t) \leq k - 1$
- $S(t)$ is the size of the upward jump at time t if an arrival occurs at time t:
  - jump size is minimum of
    - service time of the arriving customer
    - residual service times of the customers in service *at the moment he enters service* at time $t + W(t)$
  - if $W(t) = w > 0$ the jump size is exponential with parameter $k\mu$

$$P(S(t) > y | W(t) = w) = \sum_{i=1}^{2} \frac{\lambda_i}{\lambda} e^{-\theta_i w} e^{-k\mu y}$$

  - if $W(t) = 0$ there is an upward jump if and only if $K(t) = k - 1$

Multi server system with k servers and identical exponential service rates $\mu_i = \mu$

- Let $f(w)$ be density of steady state W and $\psi(s) = E(e^{-sW}; K \geq k-1)$
- Let $p_n = P(W = 0, K = n)$ for $n = 0, \ldots, k-1$
- Equating the rate of down crossings through work load level $W = w$ and the rate of up crossings

$$f(w) = p_{k-1}\lambda e^{-k\mu w} + \int_{y=0}^{w} \sum_{i=1}^{2} \lambda_i e^{-\theta_i y} e^{-k\mu(w-y)} f(y)\, dy$$

Multiplying by e^{-sw} and integrating from 0 to ∞

$$\begin{aligned} \psi(s) - p_{k-1} &= p_{k-1}\frac{\lambda}{k\mu + s} + \sum_{i=1}^{2} (\psi(s+\theta_i) - p_{k-1})\frac{\lambda_i}{k\mu + s} \\ &= \sum_{i=1}^{2} \psi(s+\theta_i)\frac{\lambda_i}{k\mu + s} \end{aligned}$$

Calculation of $\psi(s)$

- Recursion yields the following expression for $\psi(s)$:

$$\psi(s) = p_{k-1}c(s)$$

  where $c(s) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} c_{i,j}(s)$

- Terms $c_{i,j}(s)$ are determined from recursion

$$c_{i,j}(s) = \frac{\lambda_1}{k\mu + s + (i-1)\theta_1 + j\theta_2} c_{i-1,j}(s) + \frac{\lambda_2}{k\mu + s + i\theta_1 + (j-1)\theta_2} c_{i,j-1}(s)$$

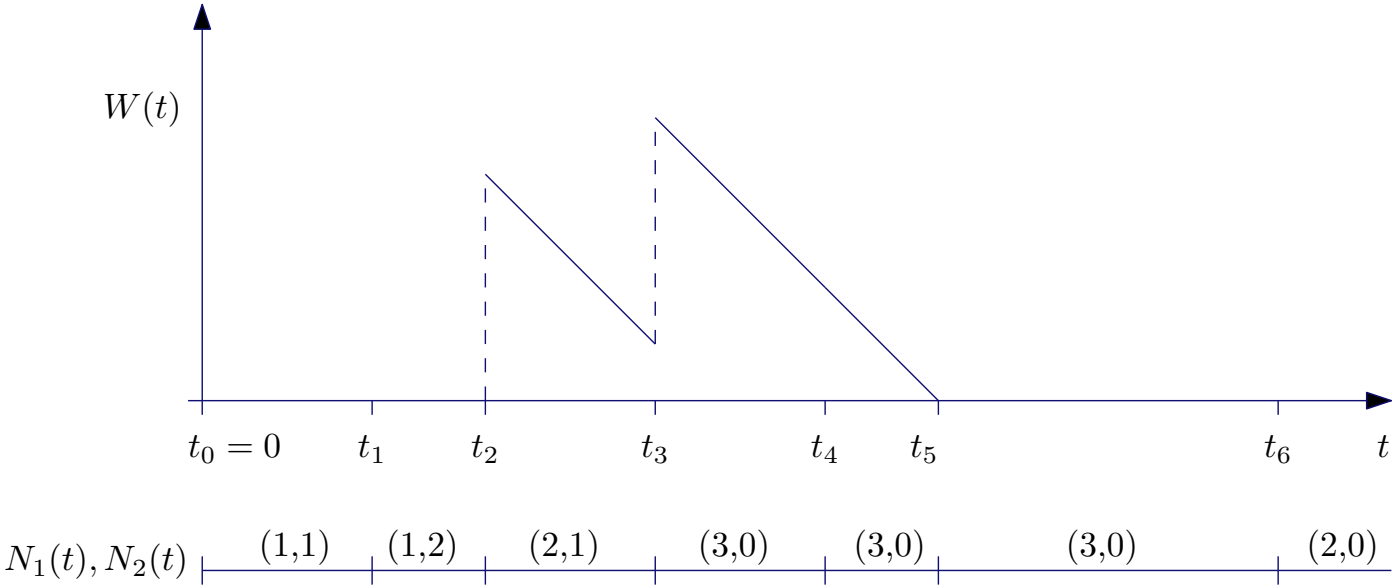
  with initially $c_{0,0}(s) = 1$ and $c_{i,j}(s) = 0$ if $i < 0$ or $j < 0$

- $p_0, p_1, \ldots, p_{k-1}$ follow from normalization $\psi(0) = 1$ and balance equations $p_{n-1}\lambda = p_n n\mu$ for $n = 1, \ldots, k-1$
- **Remarks:**
  - For $k = 1$ solution reduces to single exponential server case
  - Throughput is $\lambda_1\psi(\theta_1) + \lambda_2\psi(\theta_2)$ and mean number of type i customers in service is $\psi(\theta_i)\frac{\lambda_i}{\mu}$

Multi server system with k servers and non-identical exponential service rates μ_i

- $W(t)$ is virtual queueing time at time t
- $K(t)$ is number of busy servers at time t
- $W(t) > 0$ if and only if $K(t) = k$, and $W(t) = 0$ if and only if $K(t) \leq k - 1$
- $S(t)$ is the size of the upward jump at time t if an arrival occurs at time t:
  - jump size is minimum of
    - service time of the arriving customer
    - residual service times of the customers in service *at the moment he enters service at time* $t + W(t)$
    - What are these $k - 1$ residual service times?
  - $N_i(t)$ is number of servers busy with type i customer *at time* $t + W(t)$
- System can be described by Markov process $\{(W(t), N_1(t), N_2(t)), t \geq 0\}$ with workload dependent jumps

Multi server system with k servers and non-identical exponential service rates μ_i



Sample path of virtual queueing time process of system with $k = 4$ servers

Multi server system with k servers and non-identical exponential service rates μ_i

- Let

$$\begin{aligned} \psi_i(s) &= E(s^{-sW}; N_1 = i, N_2 = k-i-1), & 0 \le i \le k-1 \\ p_{i,j} &= P(W = 0, N_1 = i, N_2 = j), & 0 \le i+j \le k-1 \end{aligned}$$

- Let

$$\psi(s) = [\psi_0(s) \;\; \psi_1(s) \;\; \cdots \;\; \psi_{k-1}(s)], \quad p_{k-1} = [p_{0,k-1} \;\; p_{1,k-2} \;\; \cdots \;\; p_{k-1,0}]$$

- $\psi(s)$ and p_{k-1} satisfy

$$\psi(s) = p_{k-1} + \psi(s+\theta_1)H_1(s) + \psi(s+\theta_2)H_2(s)$$

  where $H_1(s)$ and $H_2(s)$ are $k \times k$ matrices

- Solution

$$\psi(s) = p_{k-1}C(s)$$

  where $C(s) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} C_{i,j}(s)$

- Terms $C_{i,j}(s)$ are recursively calculated and p_{k-1} follows form boundary conditions

Numerical examples

- $k = 5, \mu_1 = 1, \mu_2 = 2, \lambda_1 = \lambda_2, \lambda = \lambda_1 + \lambda_2$