Multi-type queues with general customer impatience

Benny Van Houdt

PATS Research Group
Dept. Mathematics and Computer Science
University of Antwerp - IBBT

YEQT-VI

Universiteit Antwerpen

Aim of this lecture

Queueing system

- Single server, infinite waiting room, FCFS
- Multi-type Markovian arrivals (correlated types)
- Type dependent service times (PH)
- Type dependent customer impatience (general, e.g., Weibull)

$\Rightarrow$ In CONTINUOUS time

Universiteit Antwerpen

Stepwise approach

Step 1

- Single server, infinite waiting room, FCFS
- Multi-type Markovian arrivals (correlated types)
- Type dependent service times (PH)
- ~~Type dependent customer impatience (general, e.g., Weibull)~~

$\Rightarrow$ In DISCRETE time

Universiteit Antwerpen

Stepwise approach

Step 2

- Single server, infinite waiting room, FCFS
- Multi-type Markovian arrivals (correlated types)
- Type dependent service times (PH)
- ~~Type dependent customer impatience (general, e.g., Weibull)~~

$\Rightarrow$ In CONTINUOUS time

Universiteit Antwerpen

Stepwise approach

Step 3

- Single server, infinite waiting room, FCFS
- Multi-type Markovian arrivals (correlated types)
- Type dependent service times (PH)
- Type dependent customer impatience (general, e.g., Weibull)

$\Rightarrow$ In CONTINUOUS time

Universiteit Antwerpen

Step 1: Arrival Process (discrete-time)

MMAP[K]: Markovian arrival process with marked transitions

- Characterized by a set of $m_a \times m_a$ matrices $\{D_k | k = 0, \ldots, K\}$ such that
  - D_k is substochastic matrix
  - $D = \sum_{k=0}^{K} D_k$ is a transition matrix

Interpretation

- The (j, j')-th entry $(D_k)_{j,j'}$ of D_k holds the probability that the underlying discrete time MC changes its phase from j to j', while generating a type k arrival
- Like a D-BMAP, but a size k batch arrival is now a type k arrival

Universiteit Antwerpen

Step 1: Service Times (discrete-time)

Type dependent service

- Type k customers: discrete phase-type (DPH) distributed amount of service with representation (α_k, S_k) of order $m_{s,k}$.

DPH definition

- Order n DPH is the time to absorption in an $n+1$ state discrete time Markov chain with transition matrix

$$P = \begin{bmatrix} S & s \\ 0 & 1 \end{bmatrix}$$

(note, $s = e - Se$) and initial probability vector

$$(\alpha, 1 - \alpha e) = (\alpha, \alpha_0),$$

such that states $\{1, \ldots, n\}$ are transient.

Universiteit Antwerpen

Step 1: Solution method

Step 1a

- Construct a GI/M/1-type MC that allows us to compute
  - Waiting/sojourn time distributions
  - Queue length distributions

  from its steady state distribution.

Step 1b

- Reduce the GI/M/1-type MC to a Quasi-Birth-Death (QBD) MC to compute the steady state distribution more efficiently (in terms of time and memory usage).

GI/M/1-type transition matrix P

- QBD is skip-free in both directions, GI/M/1-type is skip-free to the right

$$P = \begin{bmatrix} B_1 & A_0 & & & 0 \\ B_2 & A_1 & A_0 & & \\ B_3 & A_2 & A_1 & A_0 & \\ B_4 & A_3 & A_2 & A_1 & \ddots \\ \vdots & \vdots & \ddots & \ddots & \ddots \end{bmatrix}$$

Universiteit Antwerpen

Discrete time GI/M/1-type MC: Main results

Positive recurrence and stationary vector π

- Markov chain is positive recurrent if and only if

$$\theta \sum_{i \geq 1} i A_i e > 1,$$

with $\theta A = \theta$ and $A = \sum_{i \geq 0} A_i$, which is equivalent to $sp(R) < 1$

- Stationary vector $\pi = (\pi_0, \pi_1, \ldots)$ obeys, for $i > 0$

$$\pi_i = \pi_{i-1} R = \pi_0 R^i$$

and $\pi_0 = \pi_0 \sum_{i=1}^{\infty} R^{i-1} B_i$ and $\pi_0 (I - R)^{-1} e = 1$

Discrete time GI/M/1-type MC: Main results

Key Equation

- Smallest nonnegative solution to nonlinear matrix equation

$$R = \sum_{i=0}^{\infty} R^i A_i$$

  R has the same probabilistic interpretation as for the QBD

- To compute R we make use of the (Ramaswami/Bright) dual (in SMCSolver) and compute G via:
  - Functional iterations (FI), (Neuts, Latouche)
  - Newton Iteration (NI), (Perez, Telek, Van Houdt)
  - Cyclic Reduction (CR), (Bini, Meini)
  - Invariant Subspace (IS), (Akar, Sohraby)
  - Ramaswami Reduction (RR), (Bini, Meini, Ramaswami)

Universiteit Antwerpen

Step 1a: the GI/M/1-type MC

Main idea

- Observe the system when the server is busy.
- Define MC $\{(A_t, (T_t, S_t, M_{t-A_t}))\}_{t\geq 0}$ with
  - A_t: age of the customer in service ($\in \{1, 2, \ldots\}$) at time t,
  - T_t: represents the type of the customer in service at time t,
  - S_t: the phase of the server at time t,
  - M_t: state of the MMAP[K] at time t.

$\Rightarrow$ Keep track of MMAP[K] state at arrival time.

Step 1a: the GI/M/1-type MC

Transitions (1/2)

- No service completion:
  - $A_{t+1} = A_t + 1$, age increases by one.
  - $T_{t+1} = T_t = k$, type remains the same.
  - $P[S_{t+1} = j | S_t = i] = (S_k)_{i,j}$, due to (α_k, S_k) PH service.
  - $M_{t+1-A_{t+1}} = M_{t-A_t}$, MMAP[K] state remains the same.

$\Rightarrow$ Level increases by one: $A_0 = S_{ser} \otimes I_{m_a}$, with

$$S_{ser} = \begin{bmatrix} S_1 & & \\ & \ddots & \\ & & S_K \end{bmatrix}.$$

Universiteit Antwerpen

Transitions (2/2)

- Service completion (of customer n):
  - Let i be the inter-arrival time between customer n and $n+1$, then $A_{t+1} = \max(1, A_t + 1 - i)$.
  - Age decreases by $i-1 \rightarrow$ covered by matrix A_i.

$\Rightarrow$ Level reduces by $i-1$: $A_i = s_{ser} \otimes (D_0)^{i-1} L$, with

$$s_{ser} = e - S_{ser} e,$$

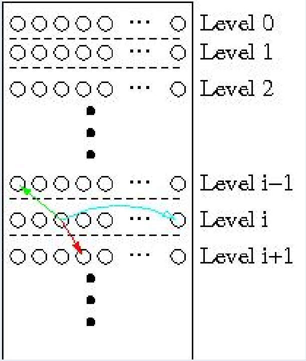
and

$$L = \left[(\alpha_1 \otimes D_1) \quad \ldots \quad (\alpha_K \otimes D_K) \right].$$

Universiteit Antwerpen

Discrete-time Quasi-Birth-Death (QBD) type Markov chain:

Infinite Quasi-Birth-Death Markov chain



- infinite state space $\mathcal{S}$
- $\mathcal{S}$ partitioned into levels of size m (except for level 0)
- level de- or increases by at most one
- characterized by $m \times m$ matices A_0, A_1 and A_2
- plus some boundary matrices

Universiteit Antwerpen

Discrete time QBD

QBD transition matrix P

$$P = \begin{bmatrix} B_1 & A_2 & & & 0 \\ B_0 & A_1 & A_2 & & \\ & A_0 & A_1 & A_2 & \\ & & A_0 & A_1 & \ddots \\ 0 & & & \ddots & \ddots \end{bmatrix}$$

Universiteit Antwerpen

Discrete time QBD: Main results

Positive recurrence and stationary vector π

- Markov chain is positive recurrent if and only if

$$(\text{Downward drift} \approx)\ \theta A_0 e > \theta A_2 e\ (\approx \text{Upward drift})\ ,$$

  with $\theta A = \theta$ and $A = A_0 + A_1 + A_2$, which is equivalent to $sp(R) < 1$
- Stationary vector $\pi = (\pi_0, \pi_1, \ldots)$ obeys, for $i > 0$

$$\pi_i = \pi_{i-1} R = \pi_0 R^i$$

  and $\pi_0 = \pi_0(B_1 + RB_0)$ and $\pi_0(I - R)^{-1} e = 1$

Universiteit Antwerpen

Discrete time QBD: Main results

Key Equation

- Smallest nonnegative solution to nonlinear matrix equation

$$R = A_2 + RA_1 + R^2 A_0$$

- Algorithms for R (all implemented in SMCSolver in Matlab)
  - Functional iterations (FI), (Neuts, Latouche)
  - Logarithmic Reduction (LR), (Latouche, Ramaswami)
  - Newton Iteration (NI), (Latouche)
  - Cyclic Reduction (CR), (Bini, Meini)
  - Invariant Subspace (IS), (Akar, Sohraby)
- Typically compute $G = A_0 + A_1 G + A_2 G^2$ and use $R = A_2(I - A_1 - A_2 G)^{-1}$

Universiteit Antwerpen

Step 1b: Reduction to QBD

Transition matrices

- Basic idea: split transition of A_i into i steps.
- Matrix geometric form of A_i, for $i > 0$ allows reduction to a *small* QBD.
- In our case: it suffices to add m_a states to each level.
- QBD is characterized by A_0^* (down), A_1^*, A_2^* (up)

$$A_2^* = \begin{bmatrix} 0 & 0 \\ 0 & A_0 \end{bmatrix}, \; A_1^* = \begin{bmatrix} 0 & L \\ 0 & A_1 \end{bmatrix}, \; A_0^* = \begin{bmatrix} D_0 & 0 \\ s_{ser} \otimes D_0 & 0 \end{bmatrix}.$$

- Steady state distribution of GI/M/1-type MC can be obtained from QBD steady state by censoring

Step 1: Concluding remarks

Possible Generalizations

- Semi-Markovian arrivals (no QBD reduction)
- General customer impatience (level-dependent QBD)
- Correlation between service and inter-arrival times
- Multiple-servers (small number): age of youngest in service
- Batch arrivals (messy business)

Some related papers (discrete-time):

- *The delay distribution of a type k customer in a FCFS MMAP[K]/PH[K]/1 queue*, B. Van Houdt and C. Blondia, Journal of Applied Probability (JAP), Vol. 39, No 1, March 2002.
- *The waiting time distribution of a type k customer in a discrete-time FCFS MMAP[K]/PH[K]/c (c=1,2) queue using QBDs*, B. Van Houdt and C. Blondia, Stochastic Models, Vol 20, no 1, pp. 55-69, 2004.
- *Response time distribution in a D-MAP/PH/1 queue with general customer impatience*, J. Van Velthoven, B. Van Houdt and C. Blondia, Stochastic Models, Vol 21, pp. 745-765, 2005.
- *Age Process, Workload Process, Sojourn Times, and Waiting Times in a Discrete Time SM[K]/PH[K]/1/FCFS Queue*, Qi-Ming He, Queueing Systems, Vol 49, pp. 363-403, 2005.
- *Queues with correlated inter-arrival and service times and its application to optical buffers*, J. Lambert , B. Van Houdt and C. Blondia, Stochastic Models, Vol 22(2), pp. 233-251, 2006.

Step 2: Solution method

Step 2a

- Construct a MC with a matrix exponential (ME) distribution that allows us to compute
  - Waiting/sojourn time distributions
  - Queue length distributions

  from its steady state distribution.

Step 2b

- Reduce the MC with ME distribution to a (Markov modulated) fluid queue compute the steady state distribution more efficiently (in terms of time and memory usage).

MCs with ME distribution

Definition

- $\{(X_t, N_t)\}_{t \geq 0}$ with $N_t \in \{1, \ldots, b\}$ and $X_t \geq 0$.
- X_t increases at rate 1 and makes occasional downward jumps:
  - while X_t increases, N_t evolves according to $b \times b$ matrix D (subgenerator)
  - at rate $(-De)_i$ downward jumps occur in state (x, i) for any x.
  - given a jump from (x, i): probability $(P(u))_{i,j}$ that we jump to state (y, j) with $y \in [x - u, x)$.
- MC is characterized by D and $dA(u) = diag(-De)dP(u)$.

Universiteit Antwerpen

MCs with ME distribution: Main results

Positive recurrence and stationary vector π

- Markov process is positive recurrent if and only if

$$\theta \int_0^\infty u dA(u) > 1,$$

  with $\theta A = 0$ and $A = D + \int_0^\infty dA(u)$.

- Stationary vector $\pi(x) \in \mathbb{R}^b$ for $x > 0$ obeys,

$$\pi(x) = \pi(0) \exp(Tx)$$

  with $\pi(0) = -\theta T$.

Universiteit Antwerpen

MCs with ME distribution: Main results

Key Equation

- Minimal solution to non-linear integral equation

$$T = D + \int_0^{\infty} \exp(Tu)dA(u).$$

- In general very few algorithms for T (linear convergence).
- In some cases numerical integration can be avoided by solving a Sylvester matrix equation.

Universiteit Antwerpen

Step 2a: the MC with ME distribution

Matrices D and $dA(u)$

- Markov process defined as in discrete-time case.
- Matrix $D = S_{ser} \otimes I_{m_a}$ with

$$S_{ser} = \begin{bmatrix} S_1 & & \\ & \ddots & \\ & & S_K \end{bmatrix}.$$

- Densities $dA(u) = (s_{ser} \otimes I_{m_a}) \exp(D_0 u) L$ with

$$s_{ser} = -S_{ser} e,$$

and

$$L = \left[(\alpha_1 \otimes D_1) \quad \dots \quad (\alpha_K \otimes D_K) \right].$$

Markov modulated fluid queue

Definition

- $\{(X_t, N_t)\}_{t\geq 0}$ with $N_t \in \{1, \ldots, b\}$ and $X_t \geq 0$.
- Define $S^+ = \{1, \ldots, a\}$ and $S^- = \{a+1, \ldots, b\}$.
- X_t increases at rate 1 if $N_t \in S^+$.
- X_t decreases at rate 1 if $N_t \in S^-$ (unless $X_t = 0$).
- N_t changes state according to CTMC

$$F = \begin{bmatrix} F_{++} & F_{+-} \\ F_{-+} & F_{--} \end{bmatrix}.$$

- Fluid queue is fully characterized by F.

Universiteit Antwerpen

Fluid queues: Main results

Positive recurrence and stationary vector π

- Markov process is positive recurrent if and only if

$$\xi_+ e < \xi_- e,$$

  with $\xi F = 0$ and $\xi = (\xi_+, \xi_-)$.
- Stationary vector $\pi(x) = (\pi_+(x), \pi_-(x)) \in \mathbb{R}^b$ for $x > 0$ obeys,

$$(\pi_+(x), \pi_-(x)) = \pi_+(0) \exp(Kx)[I, \Psi],$$

  with $K = F_{++} + F_{+-}\Psi$ and

$$\pi_+(0) = p_-(0) F_{-+},$$
$$p_-(0)(F_{--} + F_{-+}\Psi) = 0,$$
$$p_-(0)e + \int_0^\infty \pi(x)e = 1.$$

Universiteit Antwerpen

Fluid queues: Main results

Key Equation

- The smallest non-negative solution of the algebraic Riccati equation

$$F_{+-} + \Psi F_{--} + F_{++}\Psi + \Psi F_{-+}\Psi = 0.$$

- Many algorithms with quadratic convergence
  - Reduction to QBD (Ramaswami, 1999)
  - Newton Iteration, (Guo, 2001)
  - SDA: Structure-preserving Doubling Algorithm, (Guo, Iannazzo, Meini, 2007)
  - ADDA: Alternating-Directional Doubling Algorithm, (Wang, Wang, Li, 2012)

$\Rightarrow$ SDA and ADDA compute the Ψ matrix of the fluid queue and the level reversed fluid queue.

Universiteit Antwerpen

Step 2b: the reduction to the fluid queue

The matrix F

- Replace immediate downward jumps of size u by interval of length u during which the level decreases at rate 1.
- As $dA(u) = (s_{ser} \otimes I_{m_a})\exp(D_0 u)L$, we have

$$
\begin{aligned}
F_{++} &= S_{ser} \otimes I_{m_a}, \\
F_{+-} &= s_{ser} \otimes I_{m_a}, \\
F_{--} &= D_0, \\
F_{-+} &= L.
\end{aligned}
$$

- Matrix T of the MC with ME distribution is equal to matrix K of fluid queue.

Universiteit Antwerpen

Step 2: Generalizations and references

Generalizations

- Batch arrivals, semi-Markovian arrivals, multiple servers, correlation between service and inter-arrival times, etc.

Some related papers (continuous-time):

- *Markov processes whose steady state distribution is matrix exponential with an application to the GI/PH/1 queue*, B. Sengupta, Advances in Applied Probability, Vol. 21, pp. 159-180, 1989.
- *Analysis of a Continuous Time SM[K]/PH[K]/1/FCFS Queue: Age Process, Sojourn Times, Waiting Times, and Queue Lengths*, Qi-Ming HE, Journal of Systems Science and Complexity (JSSC), Vol. 25, pp. 133-155, 2012.
- *A matrix geometric representation for the queue length distribution of multitype semi-Markovian queues*, B. Van Houdt, Performance Evaluation, Vol. 69, no 7-8, pp. 299-314, 2012.

Step 3: Solution method

Finite support impatience

- Assume for now that impatience distributions have finite support $\{d_1, \ldots, d_r\}$.
- Denote $a_{i,k}$ as the probability that the patience of a type k customer is at least d_i.

Step 3a

- Construct a jump process that allows us to compute
  - Waiting/sojourn time distributions
  - Probability of abandonment

  from its steady state distribution.

Step 3b

- Reduce the jump process to a fluid queue with thresholds to compute the steady state distribution more efficiently (in terms of time and memory usage).

Step 3a: The jump process

Definition

- Workload process $\{(V_t, M_t)\}_{t\geq 0}$ with $M_t \in \{1, \ldots, m_a\}$ and $V_t \geq 0$.
  - V_t: workload in the queue at time t,
  - M_t: state of the MMAP[K] at time t.

Step 3a: The jump process

Evolution

- V_t decreases at rate 1 and makes occasional upward jumps.
- Jump rate from (x, j) with $x \in (d_{i-1}, d_i]$ to (y, j) with $y \in (x + u, x + u + du)$:

$$\sum_{k=1}^{K} (D_k)_{j,j'} a_{i,k} (\alpha_k \exp(S_k u) s_k) du + o(du).$$

- While V_t decreases, M_t evolves according to $m_a \times m_a$ matrix

$$D_0 + \sum_{k=1}^{K} D_k (1 - a_{i,k}).$$

Markov modulated fluid queue with thresholds

Definition

- $\{(X_t, N_t)\}_{t \geq 0}$ with $N_t \in \{1, \ldots, b\}$ and $X_t \geq 0$.
- Define $S^+ = \{1, \ldots, a\}$ and $S^- = \{a+1, \ldots, b\}$.
- X_t increases at rate 1 if $N_t \in S^+$.
- X_t decreases at rate 1 if $N_t \in S^-$ (unless $X_t = 0$).
- Thresholds $0 = d_0 < d_1 < d_2 < \ldots < d_r < d_{r+1} = \infty$.
- N_t changes state according to CTMC

$$F^{(i)} = \begin{bmatrix} F^{(i)}_{++} & F^{(i)}_{+-} \\ F^{(i)}_{-+} & F^{(i)}_{--} \end{bmatrix},$$

  when $X_t \in (d_{i-1}, d_i]$.
- Fluid queue is fully characterized by $F^{(i)}$ matrices.

Universiteit Antwerpen

Fluid queues with thresholds: Main results

Positive recurrence and stationary vector π

- Markov process is positive recurrent if and only if

$$\xi_+^{(r+1)} e < \xi_-^{(r+1)} e,$$

with $\xi^{(r+1)} F^{(r+1)} = 0$ and $\xi^{(r+1)} = (\xi_+^{(r+1)}, \xi_-^{(r+1)})$.

- Stationary vector $\pi(x) = (\pi_+(x), \pi_-(x)) \in \mathbb{R}^b$ expressed via
  - matrices $\Psi^{(i)}$ and $\tilde{\Psi}^{(i)}$ for $i = 1, \ldots, r+1$.
  - boundary densities $\pi(d_i) = (\pi_+(d_i), \pi_-(d_i))$.
  - probability vector $p_-(0)$.

Universiteit Antwerpen

Fluid queues with thresholds: Main results

Computing $\pi(x)$

- The smallest non-negative solution of the algebraic Riccati equation

$$F^{(i)}_{+-} + \Psi^{(i)} F^{(i)}_{--} + F^{(i)}_{++} \Psi^{(i)} + \Psi^{(i)} F^{(i)}_{-+} \Psi^{(i)} = 0,$$

while $\tilde{\Psi}^{(i)}$ solves the above equation if exchange $+$ and $-$.

- Densities $\pi(d_i) = (\pi_+(d_i), \pi_-(d_i))$ can be computed via a structured linear system in time and memory complexity that is linear in r.

Universiteit Antwerpen

Step 3b: the reduction to the fluid queue with thresholds

The matrices $F^{(i)}$

- Replace immediate upward jumps of size u by interval of length u during which the level decreases at rate 1.
- Introduce $m_a \sum_{k=1}^{K} m_{s,k}$ phases that form S^+.
- One finds

$$
\begin{aligned}
F_{++}^{(i)} &= S_{ser} \otimes I_{m_a}, \\
F_{+-}^{(i)} &= s_{ser} \otimes I_{m_a}, \\
F_{--}^{(i)} &= D_0 + \sum_{k=1}^{K} D_k(1 - a_{i,k}), \\
F_{-+}^{(i)} &= L^{(i)}.
\end{aligned}
$$

with $$L^{(i)} = \left[(\alpha_1 \otimes D_1) a_{i,1} \quad \dots \quad (\alpha_K \otimes D_K) a_{i,K} \right].$$

Step 3: General customer impatience

Approximation method

- Use step functions that lower and upper bound the impatience distributions
- Increasing the number of steps increases the accuracy
- Can solve systems with as many as $2^{16} = 65536$ thresholds $\Rightarrow$ accurate results even for heavy tailed impatience distributions.

Universiteit Antwerpen

Step 3: Generalizations and references

Generalizations

- Adaptive arrivals, impatience while in service, etc.

Some related papers:

- *Matrix-analytic methods for fluid queues with finite buffers*, A. da Silva Soares and G. Latouche, Performance Evaluation, Vol. 63, pp. 295-314, 2006.
- *Fluid queues with level dependent evolution*, A. da Silva Soares and G. Latouche, European Journal of Operational Research (EJOR), Vol. 196, pp. 1041-1048, 2009.
- *Analysis of the adaptive MMAP[K]/PH[K]/1 queue: a multi-type queue with adaptive arrivals and general impatience*, B. Van Houdt, European Journal of Operational Research (EJOR), Vol. 220, no 3, pp. 695-704, 2012.
- *A multi-layer fluid queue with boundary phase transitions and its application to the analysis of multi-type queues with general customer impatience*, G. Horvath and B. Van Houdt, Proceedings of QEST 2012, London (UK), 2012.

Questions?

Further info and MATLAB software

- Visit my webpage at `http://win.ua.ac.be/~vanhoudt`

Universiteit Antwerpen