Markov-modulated fluid priority queues

Waiting time and queue length analysis

Gábor Horváth

June 27, 2016

Budapest University of Technology and Economics
Dept. of Networked Systems and Services

Table of contents

1. Introduction
2. The Busy Period of Markovian Fluid Models
3. Analysis of the Priority Queue
4. Numerical Example
5. Conclusion

Introduction

Model Description

The model studied:

- Continuous (fluid) queue
- Multiple fluid types, priority service
- Fluid arrival process:
  - Common CTMC generator: $\mathbf{Q}$
  - Fluid rates of class k: $\mathbf{R}^{(k)}$
- Fluid service process
  - Constant service rate d

Results:

- Waiting time
  - LST, moments, distribution function (approx.)
- Queue length
  - LST, moments, distribution function (approx.)
- We can't obtain (by this approach):
  - Properties of the joint distribution of the queues

Concept of the Solution

- Based on the tagged customer approach
- For a class k fluid drop:
    - $< k$ priority classes can be omitted
    - Before it can leave:
        - The server has to accomplish the class $k+$ workload present at arrival
        - ...plus the higher pr. workload arrived while waiting
    - Performance measures:
        - Waiting time: this duration
        - Queue length at fluid drop departures: The amount of class k fluid arriving over this duration
- Ingredients of the solution:
    - Construction of special fluid flows for the perf. measures
    - Waiting time & queue length at dep. $\rightarrow$ related to a reward accumulation problem during a busy period of this fluid model
    - Queue length at arbitrary time $\rightarrow$ from the one at departures

The Busy Period of Markovian Fluid Models

Stationary Solution of Markovian Fluid Models

- Parameters:

$$\mathbf{F} = \begin{bmatrix} \mathbf{F}_{++} & \mathbf{F}_{+-} & \mathbf{F}_{+0} \\ \mathbf{F}_{-+} & \mathbf{F}_{--} & \mathbf{F}_{-0} \\ \mathbf{F}_{0+} & \mathbf{F}_{0-} & \mathbf{F}_{00} \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} \mathbf{C}_{+} & & \\ & \mathbf{C}_{-} & \\ & & \mathbf{0} \end{bmatrix},$$

- Differential equation:

$$\frac{d}{dx}\pi(x)\mathbf{C} = \pi(x)\mathbf{F}, \qquad \pi(0)\mathbf{C} = p\,\mathbf{F}, \qquad p_i = 0, \quad \forall i : c_i > 0.$$

- Matrix-analytic solution:

$$\pi(x) = \kappa\, e^{\mathbf{K}x} \begin{bmatrix} \mathbf{I} & \mathbf{\Psi} \end{bmatrix} \begin{bmatrix} \mathbf{C}_{+} & \\ & |\mathbf{C}_{-}| \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{F}_{+0}(\mathbf{F}_{00})^{-1} \\ \mathbf{0} & \mathbf{I} & \mathbf{F}_{-0}(\mathbf{F}_{00})^{-1} \end{bmatrix},$$

- Two important matrices:
  - $\mathbf{\Psi}$: phase transition probs. over the busy period
  - $\mathbf{K}$: $e^{\mathbf{K}x}$ is the expected number of crossings of level x

The properties of the busy period

- Matrix $\boldsymbol{\Psi}$: the solution of a Riccati equation

$$\boldsymbol{\Psi}|\mathbf{C}_{-}|^{-1}\mathbf{F}^{\bullet}_{-+}\boldsymbol{\Psi} + \boldsymbol{\Psi}|\mathbf{C}_{-}|^{-1}\mathbf{F}^{\bullet}_{--} + \mathbf{C}_{+}{}^{-1}\mathbf{F}^{\bullet}_{++}\boldsymbol{\Psi} + \mathbf{C}_{+}{}^{-1}\mathbf{F}^{\bullet}_{+-} = \mathbf{0},$$

- Matrix $\boldsymbol{\Psi}^{*}(s)$: LST of the busy period distribution + phase transition probs. $\rightarrow$ solution of a Riccati equation
- Moments of the busy period duration: from the derivatives of $\boldsymbol{\Psi}^{*}(s)$ (solution of 1 Riccati and n Sylvester type equations)
- Approximation for $\boldsymbol{\Psi}(t)$: based on Erlangization:

$$\boldsymbol{\Psi}_{n}(t) = \int_{0}^{\infty} f_{\mathcal{E}(n,n/t)}(u) \cdot \boldsymbol{\Psi}(u)\, du$$

  - Interpretation: $\boldsymbol{\Psi}_{n}(t) = P$(the busy period is shorter than an Erlang)
  - Solution: count the number of Erlang phases during the busy period $\rightarrow$ recursions for $\boldsymbol{\Psi}_{n}(t)$ (Riccati+Sylvester equations)

The busy period with non-zero initial fluid

- If the initial fluid level is x, the LST of the busy period duration is:

$$
\begin{aligned}
\mathbf{G}^*_{+-}(s,x) &= \mathbf{\Psi}^*(s)\mathbf{G}^*_{--}(s,x), \\
\mathbf{G}^*_{--}(s,x) &= e^{\mathbf{H}^*_G(s)x}, \\
\mathbf{G}^*_{0-}(s,x) &= (s\mathbf{I}-\mathbf{F}_{00})^{-1}\mathbf{F}_{0+}\mathbf{G}^*_{+-}(s,x) + (s\mathbf{I}-\mathbf{F}_{00})^{-1}\mathbf{F}_{0-}\mathbf{G}^*_{--}(s,x)
\end{aligned}
$$

- Important point: matrix-exponential in x!

Accumulated reward during the busy period

- Besides $\mathbf{F}, \mathbf{C}$ we now have (diagonal) $\mathbf{D}$ as well
- $\mathbf{D}_{ii}$: reward accumulation rate in i
- We are interested in the distribution of the reward accumulated during the busy period $\mathbf{\Phi}(y)$
- LST: $\mathbf{\Phi}^*(v)$, from Riccati equation, very similar to $\mathbf{\Psi}^*(s)$
- Moments: from the derivatives of $\mathbf{\Phi}^*(v)$ (1 Riccati, n Sylvester)
- Approximation for $\mathbf{\Phi}(t)$: Erlangization can be adapted
- Non-zero initial fluid $x \to \mathbf{B}(y, x)$
- The LST $\mathbf{B}^*(v, x)$ is

$$\mathbf{B}^*_{+-}(v,x) = \mathbf{\Phi}^*(v)\mathbf{B}^*_{--}(v,x),$$
$$\mathbf{B}^*_{--}(v,x) = e^{\mathbf{H}^*_B(v)x},$$
$$\mathbf{B}^*_{0-}(v,x) = (v\mathbf{D}_0 - \mathbf{F}_{00})^{-1}\mathbf{F}_{0+}\mathbf{B}^*_{+-}(v,x) + (v\mathbf{D}_0 - \mathbf{F}_{00})^{-1}\mathbf{F}_{0-}\mathbf{B}^*_{--}(v,x)$$

- Again, matrix-exponential in x

Analysis of the Priority Queue

Workload at fluid drop arrival

Question: how much class $k+$ workload is present in the system when a fluid drop arrives?

Workload process analysis:

- Workload process $\neq$ queue length process
- In state i, over Δ amount of time the service requirement brought into the system is: $\Delta r_i^{k+}/d$
- The workload decreases with slope 1
- Resulting diff. equations:

$$\frac{\partial}{\partial t} v(t,x) + \frac{\partial}{\partial x} v(t,x)(\mathbf{R}^{(k+)}/d - \mathbf{I}) = v(t,x)\mathbf{Q}$$

- $\rightarrow$ Markovian fluid flow model!
- Stationary solution: $v(x) = \kappa e^{\mathbf{K}x}\mathbf{A}$
- Solution at drop arrival instants: $v_A(x) = \frac{1}{\lambda^{(k)}}\kappa e^{\mathbf{K}x}\mathbf{A}\mathbf{R}^{(\mathbf{k})}$

Waiting Time of Fluid Drops

- We create a special fluid flow model
- Goal: the accumulated reward over the busy period (initiated at level 0) = waiting time of fluid drops
- Parameters of the special system:

$$\mathbf{F} = \begin{bmatrix} \mathbf{K} & \mathbf{A}\mathbf{R}^{(k)}/\lambda^{(k)} \\ & \mathbf{Q} \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} \mathbf{I} & \\ & \mathbf{R}^{((k+1)+)}/d - \mathbf{I} \end{bmatrix}, \quad \mathbf{D} = \begin{bmatrix} \mathbf{0} & \\ & \mathbf{I} \end{bmatrix}$$

- Purpose of the first state group:
  - To set the workload seen by an arriving drop (recall: $v_A(x) = \frac{1}{\lambda^{(k)}} \kappa e^{\mathbf{K}x} \mathbf{A}\mathbf{R}^{(\mathbf{k})}$)
- Purpose of the second state group:
  - Fluid drop joined the queue, server is processing the workload
  - Time spent here is measured by the reward rate
- Properties of the waiting time: from $\mathbf{\Phi}^*(v)$, its derivatives, and its Erlangization based approximation

Queue Length at Departure Instants

- An other similar special fluid model with reward is created:

$$\mathbf{F} = \begin{bmatrix} \mathbf{K} & \mathbf{A}\mathbf{R}^{(k)}/\lambda^{(k)} \\ & \mathbf{Q} \end{bmatrix}, \ \mathbf{C} = \begin{bmatrix} \mathbf{I} & \\ & \mathbf{R}^{((k+1)+)}/d - \mathbf{I} \end{bmatrix}, \ \mathbf{D} = \begin{bmatrix} \mathbf{0} & \\ & \mathbf{R}^{(k)} \end{bmatrix}$$

- Purpose of the first state group:
  - To set the workload seen by an arriving drop
- Purpose of the second state group:
  - Fluid drop joined the queue, server is processing the workload
  - The amount of arriving class k fluid is measured by the reward rate
- Properties of the queue length at departures: from $\mathbf{\Phi}^*(v)$, its derivatives, and its Erlangization based approximation

Queue Length at Random Point in Time

- Known: queue length at departures $\mathcal{X}$,
  Question: at random point in time $\mathcal{Y}$
- Relation is well known for discrete queues, but not for fluid
- Relation (f: pdf, F: cdf, p mass at 0):

$$\underline{f_{\mathcal{Y}}}(u)\mathbf{R}^{(k)} - \underline{F_{\mathcal{Y}}}(u)\mathbf{Q} = \lambda^{(k)}\underline{f_{\mathcal{X}}}(u), \qquad \lambda^{(k)}\underline{p_{\mathcal{X}}} = \underline{p_{\mathcal{Y}}}\mathbf{R}^{(k)}$$

- Proof: by simple balance equations
- the rate at which state $(\mathcal{Y}(t) = x, \mathcal{J}(t) = j)$ is left (for $x > 0$):

$$\lambda^{(k)}/\delta \cdot \underline{f_{\mathcal{X}}}_j(x-\delta) + \underline{f_{\mathcal{Y}}}_j(x)r_j^{(k)}/\delta + \underline{f_{\mathcal{Y}}}_j(x)(-q_{jj})$$

- the rate at which the system enters state $(\mathcal{Y}(t) = x, \mathcal{J}(t) = j)$ is

$$\lambda^{(k)}/\delta \cdot \underline{f_{\mathcal{X}}}_j(x) + \underline{f_{\mathcal{Y}}}_j(x-\delta)r_j^{(k)}/\delta + \sum_{i \neq j} \underline{f_{\mathcal{Y}}}_i(x)q_{ij}$$

- Equating the two and $\delta \to 0$ provides the theorem
- The relation in LST domain is: $\underline{f_{\mathcal{Y}}^*}(s)(s\,\mathbf{R}^{(k)} - \mathbf{Q}) = \lambda^{(k)}\,s\,\underline{f_{\mathcal{X}}^*}(s)$
- From the moments of $\mathcal{X}$ $\to$ the moments of $\mathcal{Y}$: easy
- From the Erlangization of $\mathcal{X}$ $\to$ the one of $\mathcal{Y}$: less easy

Numerical Example

Example 1

- MATLAB implementation
- Solving Riccati: ADDA, solving Sylvester: lyap (Hessenberg-Schur)

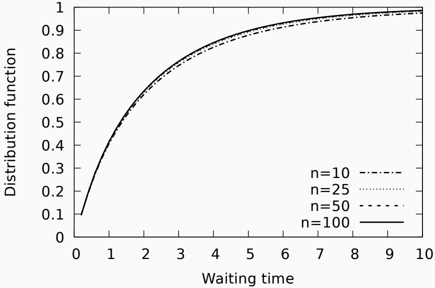
$$\mathbf{Q} = \begin{bmatrix} -8 & 5 & 0 & 3 \\ 3 & -4 & 0 & 1 \\ 4 & 6 & -10 & 0 \\ 2 & 3 & 10 & -15 \end{bmatrix}, \qquad \mathbf{R}^{(1)} = \begin{bmatrix} 1 & & & \\ & 0 & & \\ & & 2 & \\ & & & 1 \end{bmatrix},$$

$$\mathbf{R}^{(2)} = \begin{bmatrix} 2 & & & \\ & 0 & & \\ & & 4 & \\ & & & 1 \end{bmatrix}, \qquad \mathbf{R}^{(3)} = \begin{bmatrix} 4.5 & & & \\ & 1 & & \\ & & 0 & \\ & & & 2 \end{bmatrix},$$

- $d = 4 \rightarrow$ utilization $= 0.875$
- Computing 3 moments of the waiting time and queue length
    - Prompt response

Example 1

- Accuracy of the Erlangization based approximation:



- Computation times of class 1 waiting time distribution in 50 points:

	$r_3^{(2)} = 4$	$r_3^{(2)} = 4.1$
$n = 10$	$0.523\,s$	$0.189\,s$
$n = 25$	$4.771\,s$	$0.543\,s$
$n = 50$	$33.255\,s$	$1.427\,s$
$n = 100$	$249.72\,s$	$4.616\,s$

Example 2

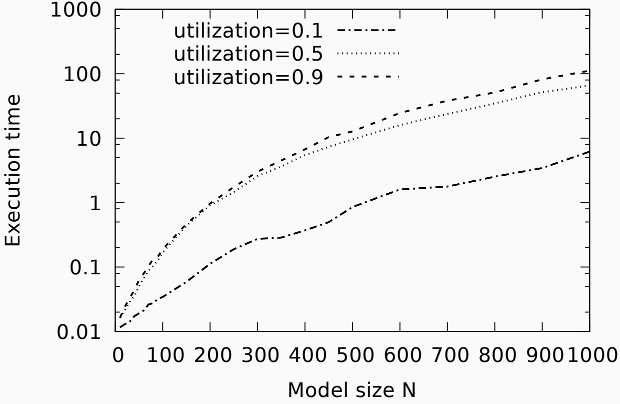
- N determines the size of the model

$$\mathbf{Q}_N = \begin{bmatrix} \bullet & N\nu & & & & \\ \gamma & \bullet & (N-1)\nu & & & \\ & 2\gamma & \bullet & (N-2)\nu & & \\ & & \ddots & \ddots & \ddots & \\ & & & (N-1)\gamma & \bullet & \nu \\ & & & & N\gamma & \bullet \end{bmatrix}, \mathbf{R}_N = \begin{bmatrix} 0 & & & & \\ & \rho/N & & & \\ & & 2\rho/N & & \\ & & & \ddots & \\ & & & & \rho \end{bmatrix},$$

- Per-class fluid rates:
  $\mathbf{R}_N^{(1)} = 0.1 \cdot \mathbf{R}_N, \mathbf{R}_N^{(2)} = 0.3 \cdot \mathbf{R}_N$ and $\mathbf{R}_N^{(3)} = 0.6 \cdot \mathbf{R}_N$
- 5 moments are computed, N varies, utilization varies

Example 1

- Execution time vs. model size:



- Size 1000 models in 10 − −100 seconds!

Conclusion

Conclusion

- Some new results are presented on fluid queues
- They are glued together with existing ones to enable the efficient analysis of fluid priority queues
- Priority queues with huge background process can be analyzed
  - Using standard math frameworks (MATLAB)
  - Without any numerical difficulties
  - With reasonable computation times