Sojourn time approximations in a multi-class time-sharing server

Ane Izagirre

IEEE INFOCOM 2014

Joint work with U. Ayesta and I.M. Verloop
Policy: Discriminatory-Processor-Sharing

The $K$ customer classes share a common resource of capacity one. Each class-$k$ customer, $k = 1, \ldots, K$, is served at rate

$$\frac{g_k}{\sum_{j=1}^{K} n_j g_j}$$

where $g_k > 0$ are class-dependent weights, and $n_k$ is the number of class-$k$ customers in the system.

Motivation: Useful model to analyze systems with relative priority like the Internet and TCP.
Interested in approximating

- $S_k(\lambda, b)$: Mean conditional (on the service requirement) sojourn time

- $S_k(\lambda)$: Mean unconditional sojourn time
Outline

- **Extreme regimes** are easier to analyze:
  - Heavy-traffic result \( (\rho \uparrow 1 \Leftrightarrow \lambda \uparrow \frac{1}{\mathbb{E}[B]} ) \)
  - Light-traffic result \( (\rho \downarrow 0 \Leftrightarrow \lambda \downarrow 0) \) **Now**

- Light-traffic interpolation **Now**

- Numerical results **Now**
Model description

- \( M/G/1 \) queue with \( K \) different classes of customers.
- Class-\( k \) customers arrive according to independent Poisson processes with rate \( \lambda_k \), \( k = 1, \ldots, K \).
- The overall arrival rate: \( \lambda = \sum_{k=1}^{K} \lambda_k \).
- \( B_k \): Service requirement of class-\( k \) customers (i.i.d.)
- \( \rho_k = \lambda_k \mathbb{E}[B_k] \): traffic intensity for class-\( k \) customers
- Total traffic intensity:

\[
\rho = \sum_{k=1}^{K} \rho_k = \sum_{k=1}^{K} \lambda_k \mathbb{E}[B_k] = \lambda \sum_{k=1}^{K} \alpha_k \mathbb{E}[B_k],
\]

where \( \alpha_k = \lambda_k / \lambda \) is the probability an arrival is of class \( k \).
- Stability condition: \( \rho < 1 \)
Known results


\[
\mathbb{E}\left[\lim_{\lambda \uparrow 1/\mathbb{E}[B]} (1 - \lambda \mathbb{E}[B]) S_k(\lambda, b)\right] = \frac{b}{g_k} \frac{\mathbb{E}[B^2]}{\sum_{j=1}^{K} \alpha_j \mathbb{E}[B_j^2]/g_j}.
\]


The derivatives of the mean conditional sojourn times of the various classes satisfy the following system of integro-differential equations:

\[
\overline{S}_k^{(1)} (\lambda, b) = 1 + \sum_{j=1}^{K} \int_0^{\infty} \lambda_j \frac{g_j}{g_k} \overline{S}_j^{(1)} (\lambda, y) [1 - F_j(y + \frac{g_j}{g_k} b)] dy
\]

\[
+ \int_0^{b} \overline{S}_k^{(1)} (\lambda, y) \sum_{j=1}^{K} \lambda_j \frac{g_j}{g_k} [1 - F_j(\frac{g_j}{g_k} (b - y))] dy,
\]

for \( k = 1, \ldots, K \), where \( \overline{S}_j^{(1)} (\lambda, b) := \frac{\partial \overline{S}_j(\lambda, b)}{\partial b} \). The natural boundary conditions are \( \overline{S}_k(\lambda, 0) = 0, k = 1, \ldots, K \).
Light-traffic result \((\rho \downarrow 0 \Leftrightarrow \lambda \downarrow 0)\)

From


\[
\overline{S}^L_T(k, \lambda, b) \approx \overline{S}_k(0, b) + \lambda \overline{S}'_k(0, b),
\]

where

\[
\overline{S}_k(0, b) \equiv \mathbb{E}[S_k | \text{no arrivals on } \mathbb{R}] = b
\]

\[
\overline{S}'_k(0, b) \equiv \int_{\mathbb{R}} (\mathbb{E}[S_k | \text{exactly one arrival on } \mathbb{R} \text{ at } t] - \mathbb{E}[S_k | \text{no arrivals on } \mathbb{R}])dt
\]

\[
= \mathbb{E} \left[ \frac{1}{2} \left(1 + \frac{g_k}{gU_t}\right) \min\{B_{U_t}, b\frac{gU_t}{g_k}\}^2 \right.
\]

\[
- \left( b\frac{gU_t}{g_k} + \frac{g_k}{gU_t} B_{U_t} \right) \min\{B_{U_t}, b\frac{gU_t}{g_k}\} + \frac{g_k + gU_t}{g_k} bB_{U_t} \right].
\]
Light-traffic interpolation

Interpolate

\[ t_k(\lambda) := (1 - \rho)\overline{S}_k(\lambda, b) = (1 - \lambda \mathbb{E}[B])\overline{S}_k(\lambda, b) \]

by the polynomial

\[ \hat{t}_k(\lambda) = h_0 + h_1 \lambda + h_2 \lambda^2. \]

The light-traffic conditions are

\[ \hat{t}_k^{(m)}(0) = t_k^{(m)}(0), \text{ for } m = 0, 1, \]

and when \( \lambda \) is small \( \overline{S}_k(\lambda, b) \approx \overline{S}_k^{LT}(\lambda, b) \), where

\[ \overline{S}_k^{LT}(\lambda, b) \approx \overline{S}_k(0, b) + \lambda \overline{S}_k'(0, b). \]
Light-traffic interpolation: Determining $h_0$, $h_1$

\[
\begin{align*}
\hat{t}_k(0) &= h_0 \\
t_k(0) &= S_k(0, b) = b \\
\end{align*}
\Rightarrow h_0 = b
\]

\[
\begin{align*}
\hat{t}_k^{(1)}(0) &= h_1 \\
t_k^{(1)}(0) &= -E[B]S_k(0, b) + S_k^{(1)}(0, b) \\
\end{align*}
\Rightarrow
\]

\[
h_1 = -bE[B] + E\left[\frac{1}{2} \left( 1 + \frac{g_k}{g_{U_t}} \right) \min\{B_{U_t}, b\frac{g_{U_t}}{g_k}\}^2 - \left( b\frac{g_{U_t}}{g_k} + \frac{g_k}{g_{U_t}} B_{U_t} \right) \min\{B_{U_t}, b\frac{g_{U_t}}{g_k}\} + b\frac{g_k + g_{U_t}}{g_k} B_{U_t} \right]
\]
Light-traffic interpolation: Determining $h_2$

The heavy-traffic condition is

$$\hat{t}_k((1/\mathbb{E}[B])^-) = t_k((1/\mathbb{E}[B])^-) = H_k,$$

where

$$H_k = \mathbb{E}\left[ \lim_{\rho \uparrow 1} (1 - \rho) S_k(\lambda, b) \right] = \frac{b}{g_k} \frac{\mathbb{E}[B^2]}{\sum_{j=1}^{K} \alpha_j \mathbb{E}[B_j^2] / g_j}.$$  

$$\hat{t}_k(1/\mathbb{E}[B]) = h_0 + h_1/\mathbb{E}[B] + h_2/\mathbb{E}[B]^2$$

$$t_k(1/\mathbb{E}[B]) = H_k$$

$$h_2 = \frac{b}{g_k} \frac{\mathbb{E}[B]^2 \mathbb{E}[B^2]}{\sum_{j=1}^{K} \alpha_j \mathbb{E}[B_j^2] / g_j} - \mathbb{E}[B] \left( \frac{1}{2} \mathbb{E}[(1 + \frac{g_k}{g_{Ut}}) \min\{B_{Ut}, b \frac{g_{Ut}}{g_k}\}^2] - \mathbb{E}[\left( b \frac{g_{Ut}}{g_k} + \frac{g_k}{g_{Ut}} B_{Ut}\right) \min\{B_{Ut}, b \frac{g_{Ut}}{g_k}\} + b \mathbb{E}\left[\frac{g_k + g_{Ut}}{g_k} B_{Ut}\right].\right)$$
We undo the normalization being
\[
\overline{S}_k^{\text{INT}}(\lambda, b) := \frac{\hat{t}_k(\lambda)}{(1 - \lambda \mathbb{E}[B])}, \quad 0 \leq \lambda < 1/\mathbb{E}[B]
\]
the approximation of \(\overline{S}_k(\lambda, b)\).

**Proposition:** The light-traffic interpolation (of order 2) of the mean conditional sojourn time for a tagged class-\(k\) customer is
\[
\overline{S}_k^{\text{INT}}(\lambda, b) = b(1 + \rho) + \lambda \mathbb{E}\left[\frac{1}{2} \left(1 + \frac{g_k}{g_{U_t}}\right) \min\{B_{U_t}, b \frac{g_{U_t}}{g_k}\}\right]^2
-
\left(b \frac{g_{U_t}}{g_k} + \frac{g_k}{g_{U_t}} B_{U_t}\right) \min\{B_{U_t}, b \frac{g_{U_t}}{g_k}\} + b \frac{g_{U_t}}{g_k} B_{U_t}\]
\[
+ \frac{\left(\lambda \mathbb{E}[B]\right)^2}{(1 - \lambda \mathbb{E}[B])} \frac{b}{g_k} \frac{\mathbb{E}[B^2]}{\sum_{j=1}^{K} \alpha_j \mathbb{E}[B_j^2]/g_j}.
\]
Properties:

- The case of **one class**, that is, $\alpha_i = 0$, $\forall i \neq k$ and $\alpha_k = 1$.

  $$ \bar{S}_k^{INT} (\lambda, b) = \frac{b}{1 - \rho}. $$

- We now assume all weights are the same, i.e., $g_i = g_k$, $\forall i, k = 1, \ldots, K$.

  $$ \bar{S}_k^{INT} (\lambda, b) = \frac{b}{1 - \rho}. $$

- $g_k \to \infty$. Hence, class $k$ is prioritized in the limit.

  $$ \bar{S}_k^{INT} (\lambda, b) = b(1 + \rho_k). $$
**Corollary:** The light-traffic interpolation (of order 2) of the mean unconditional sojourn time for a tagged class-$k$ customer is given by

\[
S_k^{INT}(\lambda) := \int_0^\infty S_k^{INT}(\lambda, b)\,dF_k(b)
\]

\[
= \mathbb{E}[B_k](1 + \rho) + \lambda\mathbb{E}\left[\frac{1}{2}(1 + \frac{g_k}{g_{U_t}}) \min\{B_{U_t}, B_k \frac{g_{U_t}}{g_k}\}\right]^2
\]

\[
- \left(\frac{B_k \frac{g_{U_t}}{g_k} + \frac{g_k}{g_{U_t}} B_{U_t}}{g_k} \right) \min\{B_{U_t}, B_k \frac{g_{U_t}}{g_k}\} + B_k \frac{g_{U_t}}{g_k} B_{U_t}
\]

\[
+ \frac{(\lambda\mathbb{E}[B])^2}{(1 - \lambda\mathbb{E}[B])} \frac{\mathbb{E}[B_k]}{g_k} \frac{\mathbb{E}[B^2]}{\sum_{j=1}^K \alpha_j \mathbb{E}[B_j^2]/g_j}.
\]
Numerical results: Monotonicity in the weights

Two classes \( K = 2 \) with exponentially distributed service requirements.
Parameters: \( \mathbb{E}[B_1] = 2, \mathbb{E}[B_2] = 1, \lambda_1 = 0.2, \lambda_2 = 0.3 \) and \( b = 1 \)

Figure: Mean conditional sojourn time as a function of \( g_1 \)
Numerical results: Relative error of the mean conditional sojourn time

Four classes $K = 4$ with exponentially distributed service requirements.

Figure: Service requirement $b_i$ such that $\mathbb{P}(B_i \leq b_i) = 0.01$ (left) and $\mathbb{P}(B_i \leq b_i) = 0.99$ (right).
Numerical results: Relative error of the mean unconditional sojourn time

Four classes $K = 4$ with exponentially distributed service requirements.

Figure: Relative error for the mean unconditional sojourn time.
Conclusions and Future Work

Conclusions:

- Making an interpolation and combining the heavy-traffic and light-traffic results we obtain an approximation of the system for any $\rho$.
- Accuracy of the interpolation result.

Future work:
- To evaluate the accuracy of the light-traffic approximations with Pareto distributions
- The sojourn time distribution
- Light-traffic interpolation technique for other complex systems: Power-of-two (work in progress)
Conclusions and Future Work

Conclusions:

- Making an interpolation and combining the heavy-traffic and light-traffic results we obtain an approximation of the system for any $\rho$.
- Accuracy of the interpolation result.

Future work:

- To evaluate the accuracy of the light-traffic approximations with Pareto distributions
- The sojourn time distribution
- Light-traffic interpolation technique for other complex systems: Power-of-two (*work in process*)
Thank you for your attention!