Heavy-traffic analysis of the discriminatory random-order-of-service discipline

Ane Izagirre

BCAM–Basque Center for Applied Mathematics

19th October, 2011

Joint work with U. Ayesta and I.M. Verloop
Model description

$M/G/1$ queue with $K$ different classes of customers. Discriminatory random order of service (DROS) discipline.

- Service is non-preemptive.
- Each class-$k$ customer is selected to be served with probability

$$\frac{p_k}{\sum_j n_j p_j},$$

where $p_k$, $k = 1, \ldots, K$, are class-dependent weights, and $n_k$ is the number of class-$k$ customers at decision epoch.

Application in telecommunication networks.
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-\(k\) customer.

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 \(k\).
Objectives

We are interested in:

- Queue length distribution.
- Waiting time distribution.

We consider the heavy-traffic regime:

\[ \lambda \uparrow \hat{\lambda} := \frac{1}{\sum_{k=1}^{K} \alpha_k \mathbb{E}(B_k)}, \]

since then \( \rho = \lambda \sum_{k=1}^{K} \alpha_k \mathbb{E}(B_k) \uparrow 1. \)
State-space collapse

**Proposition**

The scaled queue length vector at arbitrary time epochs satisfies, as $\rho \uparrow 1$, 

$$(1 - \rho)(N_1, \ldots, N_K) \Longrightarrow (\hat{N}_1, \ldots, \hat{N}_K) \overset{d}{=} X\left(\frac{\hat{\lambda}_1}{p_1}, \frac{\hat{\lambda}_2}{p_2}, \ldots, \frac{\hat{\lambda}_K}{p_K}\right),$$

where $X$ is exponentially distributed with mean

$$\frac{1}{\nu(p)} := \frac{\sum_{k=1}^{K} \hat{\lambda}_k \mathbb{E}(B_k^2)}{2 \sum_{k=1}^{K} \frac{\hat{\lambda}_k}{p_k} \mathbb{E}(B_k)}.$$
Proof for arbitrary epochs

Define $\psi(z_1, \ldots, z_K) := \mathbb{E}(z_1^{N_1} \cdots z_K^{N_K})$.

**Theorem (Kim et al, 2011)**

The joint probability generating function $\psi(z_1, \ldots, z_K)$ of the stationary queue lengths at arbitrary time is given by

$$
\psi(z_1, \ldots, z_K) = 1 - \rho + \sum_{j=1}^{K} \lambda_j z_j \phi_j(z_1, \ldots, z_K) \frac{1 - B_j^*(\lambda - \sum_{k=1}^{K} \lambda_k z_k)}{\lambda - \sum_{k=1}^{K} \lambda_k z_k},
$$

where

$$
\phi_j(z_1, \ldots, z_K) = 1 - \rho + \frac{\lambda p_j}{\lambda_j} \frac{\partial}{\partial z_j} r(z_1, \ldots, z_K).
$$
Proof for arbitrary epochs

Change of variable: \( z_i = e^{-(1-\rho)s_i} \) with \( s_i \geq 0, i = 1, \ldots, K \).

We are interested in \( \lim_{\rho \uparrow 1} (1 - \rho)(N_1, \ldots, N_K) \), therefore we study

\[
\lim_{\rho \uparrow 1} \psi(e^{-(1-\rho)\vec{s}}) = \lim_{\rho \uparrow 1} \mathbb{E}(e^{-(1-\rho)s_1 N_1 \ldots e^{-(1-\rho)s_K N_K}}).
\]

Lemma

\[ \lim_{\rho \uparrow 1} \psi(e^{-(1-\rho)\vec{s}}) \text{ exists and it satisfies} \]

\[
\mathbb{E}(e^{-\sum_{j=1}^{K} s_j \hat{N}_j}) = \lim_{\rho \uparrow 1} \psi(e^{-(1-\rho)\vec{s}}) = \hat{\lambda} \frac{\partial \hat{r}^*(v)}{\partial v} \bigg|_{\vec{v} = \sum_{j=1}^{K} \frac{\hat{\lambda}_j}{p_j} s_j}.
\]
Proof for arbitrary epochs

We have

$$\mathbb{E}(e^{-\sum_{i=1}^{K} s_i \hat{N}_i}) = \mathbb{E}(e^{-\frac{p_1}{\hat{\lambda}_1} \hat{N}_1 \sum_{i=1}^{K} \frac{\hat{\lambda}_i}{p_i} s_i - s_2 \frac{p_2}{\hat{\lambda}_2} (\hat{N}_2 - \frac{p_1}{\hat{\lambda}_1} \hat{N}_1) - ... - s_K \frac{\hat{\lambda}_K}{p_K} (\frac{p_K}{\hat{\lambda}_K} \hat{N}_K - \frac{p_1}{\hat{\lambda}_1} \hat{N}_1)},$$

which depends on $s$ only through $\sum_{j=1}^{K} \frac{\hat{\lambda}_j}{\hat{\lambda}_j} s_j$, and this implies that

$$\frac{p_i}{\hat{\lambda}_i} \hat{N}_i = \frac{p_j}{\hat{\lambda}_j} \hat{N}_j, \text{ almost surely for all } i, j$$

and we obtain

$$(\hat{N}_1, ..., \hat{N}_K) = (\frac{\hat{\lambda}_1}{p_1}, \frac{\hat{\lambda}_2}{p_2}, ..., \frac{\hat{\lambda}_K}{p_K}) \frac{p_1}{\hat{\lambda}_1} \hat{N}_1,$$

almost surely, or equivalently

$$(\hat{N}_1, ..., \hat{N}_K) \overset{d}{=} (\frac{\hat{\lambda}_1}{p_1}, \frac{\hat{\lambda}_2}{p_2}, ..., \frac{\hat{\lambda}_K}{p_K}) X,$$

with $X$ distributed as $\frac{p_1}{\hat{\lambda}_1} \hat{N}_1.$
Proof for arbitrary epochs: Determining the common factor

- Let $V$ be the total workload in the network.

Kingman (1961):

$$(1 - \rho)V \overset{d}{\to} \hat{V},$$

where $\hat{V}$ is exponentially distributed with mean

$$\mathbb{E}(\hat{V}) = \frac{\sum_{j=1}^{K} \hat{\lambda}_j \mathbb{E}(B_j^2)}{2}.$$
The total workload can be represented as

\[ V \doteq \sum_{j=1}^{K} \sum_{h=1}^{N_j-1} B_{j,h} + \sum_{j=1}^{K} \tilde{B}_j. \]

For the scaled workload we can write

\[
\mathbb{E}(e^{-s\hat{V}}) = \lim_{\rho \uparrow 1} \mathbb{E}(e^{-(1-\rho)sV}) = \lim_{\rho \uparrow 1} \mathbb{E}(e^{-(1-\rho)s(\sum_{j=1}^{K} \sum_{h=1}^{N_j-1} B_{j,h} + \sum_{j=1}^{K} \tilde{B}_j)})
\]

\[
= \lim_{\rho \uparrow 1} \mathbb{E}(e^{-s\sum_{j=1}^{K}(1-\rho)(N_j-1)\sum_{h=1}^{N_j-1} \frac{B_{j,h}}{(N_j-1)} e^{-(1-\rho)s\sum_{j=1}^{K} \tilde{B}_j})}
\]

\[
= \mathbb{E}(e^{-s\sum_{j=1}^{K} \mathbb{E}(B_j)\hat{N}_j})
\]
Proof for arbitrary epochs: Determining the common factor

Then, we obtain

\[ \hat{V} \overset{d}{=} \sum_{j=1}^{K} \mathbb{E}(B_j) \hat{N}_j, \]

which is

\[ \hat{V} \overset{d}{=} X \sum_{j=1}^{K} \frac{\hat{\lambda}_j}{p_j} \mathbb{E}(B_j). \]

Since \( \hat{V} \) is exponentially distributed, the same is true for \( X \). Taking expectations we obtain

\[ \mathbb{E}(X) = \frac{\sum_{j=1}^{K} \hat{\lambda}_j \mathbb{E}(B_j^2)}{2 \sum_{j=1}^{K} \frac{\hat{\lambda}_j}{p_j} \mathbb{E}(B_j)}. \]
Waiting time

We denote by $W_k$ the waiting time of a class-$k$ customer.

**Proposition**

As $\rho \uparrow 1$,

$$(1 - \rho) (W_k, N_1, \ldots, N_K) \to X(Z_k, \frac{\hat{\lambda}_1}{p_1}, \ldots, \frac{\hat{\lambda}_K}{p_K}),$$

where $\to$ denotes convergence in distribution and $X$ and $Z_k$ are exponentially distributed independent random variables with $\mathbb{E}(X) = 1/\nu(p)$ and $\mathbb{E}(Z_k) = 1/p_k$. 
Theorem (Kim et al, 2011)

The joint transform of the waiting time and number of customers is denoted by

\[ T_k(u, z_1, \ldots, z_K) := \mathbb{E}[e^{-uW_k} z_1^{N_1} \cdots z_K^{N_K} 1_{\{W_k>0\}}], \]

and it satisfies

\[
\sum_{j=1}^K \frac{p_j}{p_k} \left( \frac{\partial}{\partial z_j} T_k(u, z_1, \ldots, z_K) \right) (z_j - B_j^*(u + \lambda - \sum_{r=1}^K \lambda_r z_r)) + T_k(u, z_1, \ldots, z_K) = W_k^1(u, z_1, \ldots, z_K). \]
Waiting time: Sketch of the proof

We are interested in

\[ \lim_{\rho \uparrow 1} \mathbb{E}(e^{-(1-\rho)W_k \sum_{j=1}^{K}(1-\rho)s_jQ_j}) =: \hat{T}_k(u, y) \bigg|_{y = \sum_{j=1}^{K} s_j \lambda_j}. \]

As \( \rho \uparrow 1 \) we get the following ODE:

\[ \frac{\partial \hat{T}_k(u, y)}{\partial y} - \frac{p_k}{u} \hat{T}_k(u, y) = -\frac{p_k}{u} \frac{\nu(p)}{\nu(p) + y}, \]

which solution is

\[ \hat{T}_k(u, y) = p_k \frac{\nu(p)}{u} e^{p_k \frac{\nu+y}{u}} \int_{p_k \frac{\nu(p)+y}{u}}^{\infty} \frac{e^{-l}}{l} \, dl. \]
Waiting time: Sketch of the proof

Being $Z_k$ and $X$ two exponentially distributed random variables with $\mathbb{E}(Z_k) = 1/p_k$ and $\mathbb{E}(X) = 1/\nu(p)$, the Laplace transform of $(Z_k \cdot X, \frac{\lambda_1}{p_1} X, \ldots, \frac{\lambda_K}{p_K} X)$ is given by:

$$\mathbb{E}[e^{-uZ_kX} - s_1\frac{\lambda_1}{p_1} X - \ldots - s_K\frac{\lambda_K}{p_K} X] = \mathbb{E}[e^{-uZ_kX - yX}]$$

$$= \mathbb{E}[\mathbb{E}[e^{-uZ_kX - yX} | Z_k = z_k]] = \mathbb{E}\left[\frac{\nu(p)}{\nu(p) + uZ_k + y}\right]$$

$$= \int_{0}^{\infty} p_k e^{-p_k z} \frac{\nu(p)}{\nu(p) + uZ_k + y} dZ_k = \frac{\nu(p)p_k e^{p_k \frac{\nu(p) + y}{u}}}{u} \int_{0}^{\infty} \frac{e^{-l}}{l} dl.$$  

We end up that it is the same Laplace transform as $(\hat{W}_k, \hat{Q}_1, \ldots, \hat{Q}_K)$.
Optimal selection of weights

We now investigate how to set the weights in order to minimize:

- Scaled holding cost in heavy-traffic:

\[
\lim_{\rho \uparrow 1}(1 - \rho) \sum_{j=1}^{K} c_j N_j \overset{d}{=} \sum_{j=1}^{K} c_j \frac{\hat{\lambda}_j}{p_j} X,
\]

where \( c_j \) is the cost associated to a class-\( j \) customer.

- \( m \)-th moment of the scaled waiting time of an arbitrary customer.
Let $c_j \geq 0$, $j = 1, \ldots, K$. Without loss of generality we assume that
\[
\frac{c_1}{\mathbb{E}(B_1)} \geq \frac{c_2}{\mathbb{E}(B_2)} \geq \cdots \geq \frac{c_K}{\mathbb{E}(B_K)}. \quad \text{If} \quad \frac{p_j}{\tilde{p}_{j+1}} \leq \frac{\tilde{p}_j}{\tilde{p}_{j+1}}, \quad \text{for all} \quad j = 1, \ldots, K, \quad \text{then}
\]
\[
\lim_{\rho \uparrow 1} (1 - \rho) \sum_{j=1}^{K} c_j N_j(p) \geq_{st} \lim_{\rho \uparrow 1} (1 - \rho) \sum_{j=1}^{K} c_j N_j(\tilde{p}).
\]
Moments of the scaled waiting time

Proposition

- If \( \frac{p_j}{p_{j+1}} \leq \frac{\tilde{p}_j}{\tilde{p}_{j+1}} \), for all \( j = 1, \ldots, K \) and \( E(B_1) \leq \ldots \leq E(B_K) \), then
  \[ E(\hat{W}(p)) \geq E(\hat{W}(\tilde{p})). \]

- The unique minimizer of \( E(\hat{W}^m(p)) \) if \( m > 1 \) is
  \[
  p_i^* = \left( \frac{1/E(B_i)}{\sum_{k=1}^K \frac{1}{E(B_k)}} \right)^{1/m-1}.
  \]
Conclusions and Future Work

Conclusions:
- State space collapse for the queue length and waiting time.
- Selection of optimal weights for the performance.

Future work:
- Light traffic ($\rho \to 0$).
- Extension of methodology to network of queues.
Thank you for your attention!