Revisit large queues in a heterogeneous multi-server queue using a piecewise deterministic Markov process and a harmonic function

Masakiyo Miyazawa
Tokyo University of Science
April 9, 2015 (Adelaide)
Outline

1 Heterogeneous multi-server queue
2 Background, motivations and problems
3 Main results
   - A tractable form for the stationary equation
   - Tail asymptotics of the stationary distribution
   - Heavy traffic approximation
4 Concluding remarks
It has a single queue and $k$ servers, numbered $1, 2, \ldots, k$.

- **Exogeneous arrivals** is a renewal process, interarrival distribution $F_e$ ($\hat{F}_e$ for its MGF).
  - $\lambda_e$: arrival rate, $\sigma_e^2$: the variance of $F_e$ is finite.

- **Service** is FCFS, and their times at server $i$ are i.i.d. with distribution $F_{s,i}$ ($\hat{F}_{s,i}$ for its MGF).
  - $\lambda_{s,i}$: service rate, $\sigma_{s,i}^2$: the variance of $F_{s,i}$ is finite.

- An arriving customer who finds idles servers chooses server $i \in A$ with probability $r(A, i)$, where $A$ is the set of idle servers.

**Traffic intensity**: $\rho = \lambda_e / (\sum_i \lambda_{s,i}) < 1$. 
Heterogeneous $k$ server queue for $k = 4$

$R_e(t)$
$R_{s,1}(t) = 0$
$R_{s,2}(t) > 0$
$R_{s,3}(t) > 0$
$R_{s,4}(t) > 0$

$F_e$
$F_{s,1}$

$F_{s,2}$

$F_{s,3}$

$F_{s,4}$

$Q(t)$

FCFS based service with an arbitrarily given probabilities for selecting idle servers
The heterogeneous multi-server queue is just to the next of a single server queue such as $GI/G/1$ queue. Because of analytical difficulty, asymptotics are highlighted.

- The tail asymptotics of the stationary waiting time distribution in either light or heavy tail regime.
- A process limit of a sequence of the queue length processes in heavy traffic under two different regimes.

Both are less studied for the stationary distribution of the queue length. We attack its asymptotics.
Literature on asymptotic analysis

Tail asymptotics of the stationary waiting time distribution of a multi-server queue

(Light) Neuts & Takahashi [7], Sadowsky & Szpankowski [9]

(Heavy) Foss & Korshunov [4] for identical servers

(Mixed) Sadowsky [8] only for logarithmic rate

Heavy traffic approximation

(Diffusion) Chen & Ye [2] for proces limit

Single server case of GJN: Budhiraja and Lee [1]

GJN = Generalized Jackson Network
Motivations

The original motivation was to see the speed of convergence and quality of diffusion approximation in heavy traffic (particularly, for queueing network).

- Which characteristics are **continuous** in the diffusion approximation?
- How **fast** the stationary distribution weakly converges to that of a diffusion limit in heavy traffic?

In considering them, I became more ambitious. Both the tail asymptotics and heavy traffic approximation concern **large queues**, so they can be studied in a **unified way**?
Problems

We work on the stationary equation of a Markov process describing the heterogeneous multi-server queue.

(a) Can we derive a **tractable stationary equation** for both the asymptotic problems?

(b) Can we include **large deviations information** (on light and heavy tail) in the stationary equation?

We positively answer these questions choosing appropriate **test functions** for the stationary equation, which wash out complicated terms due to arrivals and departures.
Let $K = \{1, 2, \ldots, k\}$. For $A \in 2^K$, $A \neq \emptyset$,
\[
S_A = \{(0, \mathbf{x}) \in \mathbb{R}^{k+1}_+; x_i = 0, i \in A, x_j > 0, j \in K \setminus A\},
\]
For $A = \emptyset$, $S_{\emptyset} = \{(x_0, \mathbf{x}) \in \mathbb{Z}^{k+1}_+; x_0 \geq 0, \mathbf{x} > \mathbf{0}\}$, and
\[
S = \bigcup_{A \in 2^K} S_A.
\]
Let $X(t)$ be a Markov process with state space $S$ for describing the heterogeneous multi-server queue. The stationary distribution $\pi$ is characterized by $\pi T(t) f = \pi f$
for a sufficiently large class of test function $f$’s, where
\[
T(t) f(\mathbf{x}) \equiv \mathbb{E}(f(X(t))|X(0) = \mathbf{x}).
\]
Faces for a heterogeneous 2-server queue
Piecewise deterministic Markov process (PDMP)

- $Q(t)$: the number of customers waiting for service.
- $R_{e}(t)$: the residual exogenous arrival time.
- $R_{s,i}(t)$: the residual service time at server $i$.

Define counting processes:
- $N_{e}(t)$: arriving customers by time $t$,
- $N_{d,i}(t)$: departing customers from server $i$ by time $t$.
- $N_{s,i}(t)$: renewal process with interval distribution $F_{s,i}$.

Markov process $X(t) \equiv (Q(t), R_{e}(t), R_{s}(t))$ is deterministic between jumps of $N_{e} + \sum_{i \in K} N_{d,i}$, which is a PDMP termed by Davis [3].
Extended generator $\mathcal{A}$ for the PDMP

For $f \in C^1(S)$, let $Y(t) = f(X(t))$, $\mathcal{F}_t$ is a filtration generated by $\{X(t)\}$ and $\mathcal{A}f(X(t)) = \partial_t f(X(t))$, then, using Dvais [3]'s formulation,

$$Y(t) - Y(0) - \int_0^t \mathcal{A}f(X(u))du$$

$$- \int_0^t \mathbb{E}(\Delta Y(u)|\mathcal{F}_{u^-})\left(N_e(du) + \sum_{i \in K} N_{d,i}(du)\right)$$

is a $\mathcal{F}_t$-martingale, where $\Delta Y(u) = Y(u^+) - Y(u^-)$.

If the **integrations** by the point processes are dropped, then $T(t)f = f$ is equivalent to $\mathcal{A}f = 0$. 
Test functions $f_A$ and $f$

Let $g_i(\zeta, x)$ be a function from $\mathbb{R} \times \mathbb{R}_+$ to $\mathbb{R}$ for $i = 0, 1, 2, \ldots, k$. We assume

(a) $g_i(\zeta, x)$ is positive valued, continuous and strictly increasing in $\zeta$ and $x$ separately.

(b) $g_i(0, x) = g_i(\zeta, 0) = 1$.

For $A \in 2^K$ and $t \geq 0$, let, for $\theta, \eta \in \mathbb{R}, \xi \in \mathbb{R}^k$,

$$f_A(u, x_0, \mathbf{x}) = e^{\theta \zeta} g_0(\eta, x_0) \prod_{i \in A} g_i(\xi_i, x_i) 1(\mathbf{x} \in S_A),$$

$$f(u, x_0, \mathbf{x}) = \sum_{A \in 2^K} w_A(\theta) f_A(u, x_0, \mathbf{x}),$$

where $w_A(\theta)$ is determined later.
Choice of $\eta$, $\xi_i$ and $w_A$ as functions of $\theta \in \mathbb{R}$

\[
\hat{G}_e(\eta) = \mathbb{E}(g(\eta, T_e)), \quad \hat{G}_{s,i}(\xi_i) = \mathbb{E}(g(\xi_i, T_{s,i})),
\]

which exist but may be infinite by (a). We define $\eta$, $\xi_i$ as the solutions of the following equations, and set $w_A(\theta)$ as

\[
\hat{G}_e(\eta) = e^{-\theta}, \quad \hat{G}_{s,i}(\xi_i) = e^\theta, \quad w_A(\theta) = e^{\theta|A|}, \quad (1)
\]

where $|A|$ is the cardinality of set $A$. These $\eta$ and $\xi_i$ are uniquely determined as functions of $\theta$ as long as $\hat{G}_e^{-1}(e^{-\theta})$ and $\hat{G}_{s,i}^{-1}(e^\theta)$ are finite. Note that $\eta(\theta)$ (or $\xi_i(\theta)$) vanishes if and only if $\theta = 0$. We denote the domain on which all $\eta(\theta)$ and $\xi_i(\theta)$ are finite by $\mathcal{D}_g$. 

Ideas behind those choices of $\eta, \xi_i, w$

- $\eta, \xi_i, w$ are considered as functions of $\theta$ to be controlled for a tractable stationary equation.

- For this, they are chosen so that the state changes due to jumps of the remaining times (by arrivals and departures) to be cancelled.

⇒ As a result, not only the complicated terms are dropped, but also the remaining times are well controlled for service times.
Logarithmic rate functions for counting processes

For the light tail case, let \( g_i(\zeta, x) = e^{\zeta x} \), \( \gamma_e(\theta) = -\eta(\theta) \), then it follows from Glynn and Whitt [5] that

\[
\gamma_e(\theta) = \lim_{t \to \infty} \frac{1}{t} \log \mathbb{E}(e^{\theta N_e(t)}), \quad \theta \in \mathcal{D}_g,
\]

which is called a logarithmic rate function. Similarly,

\[
\gamma_{s,i}(\theta) \equiv -\xi_i(\theta) = \lim_{t \to \infty} \frac{1}{t} \log \mathbb{E}(e^{\theta N_{s,i}(t)}).
\]

These rate functions are related to the tail asymptotics. For the heavy tail case, we can not use \( e^{\zeta x} \) for \( g_i \). So, we put \( g_i(\zeta, x) = (1 + x)^\zeta \).
Key lemmas: $Y_A(t) = f_A(X(t))$ and $Y_i(t) = \sum_{A \ni i} Y_A(t)$

**Lemma 1 (Simple form for the generator)**

For $Y(t) = \sum_{A \in 2^K} Y_A(t)$, we have, for $\theta \in \mathcal{D}_g$,

$$Y(t) - Y(0) - \int_0^t Y'(u)du \quad \text{is a } \mathcal{F}_t\text{-martingale}.$$  

Let $\Phi(\theta) = \mathbb{E}_\pi(Y(0))$, $\Phi_{i|0}(\theta) = \mathbb{E}_\pi(Y_i(0))$ for any time invariant distribution $\pi$.

**Lemma 2 (Tractable stationary equation)**

If $F_e, F_{s,i}$ have light tails, then, for $\theta \in \mathcal{D}_g$, \[
\left(\gamma_e(\theta) + \sum_{i \in K} \gamma_{s,i}(-\theta)\right)\Phi(\theta) - \sum_{i \in K} \gamma_{s,i}(-\theta)\Phi_{i|0}(\theta) = 0,
\]

as long as $\Phi(\theta)$ is finite.
Test functions for the mixed tail case

Let $\mathcal{L}$ and $\mathcal{H}$ be the sets of distributions $F_{s,i}$ ($F_{s,0} \equiv F_e$) which are light and heavy tailed, respectively. Let

$$g^{(\mathcal{L})}(\zeta, x) = e^{\zeta x}, \quad g^{(\mathcal{H})}_i(\zeta, x) = (1 + x)^\zeta.$$ 

We redefine $f_A$ with $\xi_0 = \eta$ as

$$f_A(u, x_0, \mathbf{x}) = e^{\theta u} \prod_{j \in (A \cup \{0\}) \cap \mathcal{L}} g^{(\mathcal{L})}(\xi_i, x_i) \times \prod_{j \in (A \cup \{0\}) \cap \mathcal{H}} g^{(\mathcal{H})}(\xi_i, x_i) 1(\mathbf{x} \in S_A),$$

Using this $f_A$, we define $Y_A(t), Y(t)$. 
Stationary equations for the mixed tail case

\[ \Phi_{i}^{(M)}(\theta) = \begin{cases} 
\mathbb{E}(Y(0)), & i \in \mathcal{L}, \\
\mathbb{E}\left(\frac{1}{1+R_{s,i}(0)}Y(0)\right), & i \in \mathcal{H}, 
\end{cases} \]

\[ \Phi_{i|0}^{(M)}(\theta) = \begin{cases} 
\mathbb{E}(Y_{i}(0)), & i \in \mathcal{L}, \\
\mathbb{E}\left(\frac{1}{1+R_{s,i}(0)}Y_{i}(0)\right), & i \in \mathcal{H}. 
\end{cases} \]

Lemma 3

If the GJN has a stationary distribution, then, for \( \theta \in \mathcal{D}_{g} \),

\[ \gamma_{e}^{(M)}(\theta)\Phi_{0}^{(M)}(\theta) + \sum_{i \in K} \gamma_{s,i}^{(M)}(-\theta)\Phi_{i}^{(M)}(\theta) - \sum_{i \in K} \gamma_{s,i}^{(M)}(-\theta)\Phi_{i|0}^{(M)}(\theta) = 0, \]

as long as \( \Phi_{i}^{(M)}(\theta) \) and \( \Phi_{i|0}^{(M)}(\theta) \) are finite.
Tail asymptotics of the stationary distribution for $0 \in \mathcal{L}$

\[
\bar{\alpha} = \sup \left\{ \theta \geq 0; \gamma_e(\theta) + \sum_{i \in \mathcal{L} \setminus \{0\}} \gamma_{s,i}(-\theta) \leq 0 \right\},
\]

\[
\alpha = \sup \left\{ \theta \geq 0; \gamma_e(\theta) + \sum_{i \in \mathcal{L} \setminus \{0\}} \gamma_{s,i}(-\theta) + \sum_{i \in \mathcal{H}} \gamma^{(\mathcal{H})}_{s,i}(-\theta) \leq 0 \right\}.
\]

**Theorem 4**

\[
-\alpha \leq \liminf_{x \to \infty} \frac{1}{x} \log P(L > x) \leq \limsup_{x \to \infty} \frac{1}{x} \log P(L > x) \leq -\bar{\alpha}.
\]
Kingman’s heavy traffic approximation

Kingman [6] derived a heavy traffic approximation for the stationary waiting time distribution of $GI/G/1$ queue, in which time is not scaled. We here consider the queue length process $Q(t)$. Index the systems by numbers $1, 2, \ldots$, and denote characteristics of the $n$-th system by superscript $(n)$ like $X^{(n)}(t), Y^{(n)}(t)$. We assume:

\[
\sum_{i \in K} \lambda_{s,i}^{(n)} - \lambda_e^{(n)} = \frac{1}{\sqrt{n}} c > 0, \quad (2)
\]

\[
\lim_{n \to \infty} \sigma_{e}^{(n)} = \sigma_e, \quad \lim_{n \to \infty} \sigma_{s,i}^{(n)} = \sigma_{s,i}, \quad (3)
\]

where $c$ is a positive constant.
Lemma 5

If all $F_e, F_{s,i}$ are light tailed, then we have

\[
\gamma_e^{(n)} \left( \frac{\theta}{\sqrt{n}} \right) = \lambda_e^{(n)} \frac{\theta}{\sqrt{n}} + \frac{1}{2} (\lambda_e^{(n)})^3 (\sigma_e^{(n)})^2 \frac{\theta^2}{n} + o\left( \frac{\theta^2}{n} \right),
\]

\[
\gamma_{s,i}^{(n)} \left( - \frac{\theta}{\sqrt{n}} \right) = -\lambda_{s,i}^{(n)} \frac{\theta}{\sqrt{n}} + \frac{1}{2} (\lambda_{s,i}^{(n)})^3 (\sigma_{s,i}^{(n)})^2 \frac{\theta^2}{n} + o\left( \frac{\theta^2}{n} \right).
\]

Remark: If some of $F_e, F_{s,i}$ are heavy tailed, then we use $g_i^{(H)}$ instead of $g_i^{(L)}$. In this case, we still have the same Taylor expansions, with is recently obtained by Braverman, Dai and Miyazawa (2015) (working paper).
**Theorem 6**

For a sequence of the stable heterogeneous multi-server queues satisfying the heavy traffic assumptions (2) and (3), we have

\[
\lim_{n \to \infty} \mathbb{P}\left( \frac{1}{\sqrt{n}} L^{(n)} \leq x \right) = 1 - \exp \left( - \frac{2}{\lambda_e^2 (\sigma_e^2 + \sum_{i \in K} \sigma_{s,i}^2)} x \right), \quad x \geq 0,
\]

where \( L^{(n)} \) is a random variable subject to the stationary distribution of \( L^{(n)}(t) \).
Concluding remarks

What can be argued by this approach?

- The speed of convergence for heavy traffic approximation.
- For heavy traffic approximation:
  - Generalized Jackson network (jointly with Braverman and Dai) Tightness becomes a big issue.
  - Multi-server queue with reneging (jointly with Kobayashi and Sakuma)
- We hope this approach is also useful for queues with state space collapse.

M. H. A. Davis.
Piecewise deterministic Markov processes: a general class of non-diffusion stochastic models.

Sergey Foss and Dimitry Korshunov.
On large delays in multi-server queues with heavy tails.

P. Glynn and W. Whitt.
Logarithmic asymptotics for steady-state tail probabilities in a single-server queue.
J. F. C. Kingman.
On queues in heavy traffic.

M.F. Neuts and Y. Takahashi.
Asymptotic behavior of the stationary distributions in the GI/PH/c queue with heterogeneous servers.
References IV

J.S. Sadowsky.
The probability of large queue lengths and waiting times in a heterogeneous multiserver queue II: Positive recurrence and logarithmic limits.

J.S. Sadowsky and W. Szpankowski.
The probability of large queue lengths and waiting times in a heterogeneous multiserver queue I: Tight limits.