# A Mean Field Model for a Join-the-Shortest-Queue Network

Yiqiang Q. Zhao

School of Mathematics and Statistics
Carleton University
Ottawa, Ontario, Canada

at the 13th Workshop on Markov Processes and Related Topics
(organized by WHU and BNU, July 17–21, 2017)

Carleton UNIVERSITY

(Based on joint work with Don Dawson and Jiashan Tang)

# Outline

1. Motivations
2. Focus
3. Main Result
4. Convergence Theorem ( LLN )
5. Stationary Distribution
6. Justification
7. Conclutions

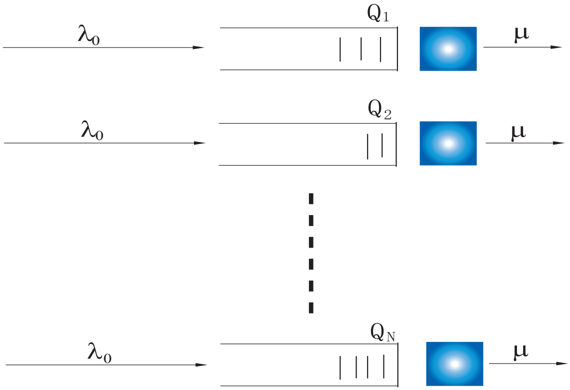
# Applications

JSQ with a large number of queues (nodes):

1. ATM switches where per flow queueing is supported (number of queues can be easily in hundreds)
2. Internet server clusters (having a large number of processors)
3. Local computer networks (connected by several 10's or even more machines)
4. Distributed/parallel networks (each of the nodes can have hundreds of links)

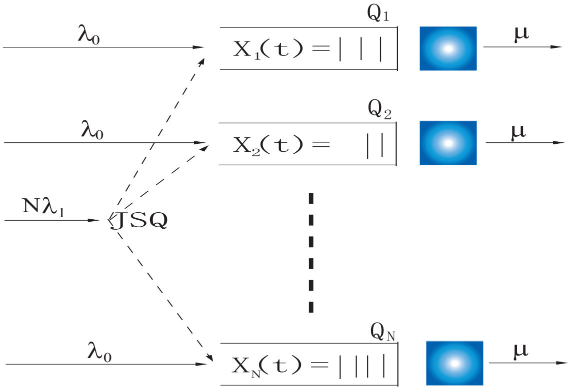
See reference list for more information.

# Parallel queues without interaction



# Parallel queues with interaction

A network with mean field interaction: JSQ with large $N$.



# Balancing techniques

1. Changing the arrival and/or service rates
2. Joining different queues
   - JSQ models
   - Extra arrival sources may choose different queues to join
   - Dobrushin's mean-field model
3. Jockeying
   - Periodically redistributing customers in all queues
   - $r$ difference jockeying

# Our Focus

- Focus: Stationary behaviour of a "typical queue" in JSQ with large $N$;
- Tools: Using the stationary behavious of a "typical queue" in JSQ with $N = \infty$;
- Wanted:

**(1)** Limiting process (mean field limit) as $N \to \infty$
**(2)** Stationary behaviour of the mean field limit
**(3)** Justification of:

$$\lim_{t} \lim_{N} \text{ (JSQ with } N) = \lim_{N} \lim_{t} \text{ (JSQ with } N)$$

# Stationary behaviour of the limiting "typical queue" I

## Theorem

*(1) For JSQ mean field interaction network (=JSQ($\infty$)), if $\lambda_0 + \lambda_1 < \mu$, then the unique stationary distribution of the "typical queue" of the interaction network is*

$$\pi_0^{JSQ} = 1 - \frac{\lambda_0+\lambda_1}{\mu},$$

$$\pi_k^{JSQ} = \frac{\lambda_0+\lambda_1}{\mu}\left(1 - \frac{\lambda_0}{\mu}\right)\left(\frac{\lambda_0}{\mu}\right)^{k-1}, k \geq 1.$$

# Stationary behaviour of the limiting “typical queue” II

## Theorem

*(2) If* $\lambda_1 = 0$ *then*

$$\pi_k^{JSQ} = \left(1 - \frac{\lambda_0}{\mu}\right)\left(\frac{\lambda_0}{\mu}\right)^k, \qquad k \geq 0.$$

*(3) If* $\lambda_0 = 0$, *then*

$$\begin{aligned} \pi_0^{JSQ} &= 1 - \tfrac{\lambda_1}{\mu}, \\ \pi_1^{JSQ} &= \tfrac{\lambda_1}{\mu}, \\ \pi_k^{JSQ} &= 0 \qquad \textit{for all} \qquad k \geq 2. \end{aligned}$$

# Notation I

$\mathbb{R}$: the set of all real numbers;
$\mathbb{R}_+$: the set of all nonnegative numbers;
$E = \{0, 1, 2, \cdots\}$: equipped with discrete topology;
$C_b(E)$: the set of bounded continuous functions in $E$;
$D_T(E)$ ($D_\infty(E)$): the set of functions from $[0, T]$ ($[0, \infty)$) to $E$, which are right-continuous with a left limit;
$X(t, \omega) = X_t(\omega) = \omega(t) = X(t)$: process with $w \in D_\infty(E)$;
$\mathcal{F}_t$: $\sigma\{X(s), 0 \leq s \leq t\}$; $\mathcal{F}$: $\sigma\{X(s), s \geq 0\}$;
$\mathcal{P}(D_\infty(E), \mathcal{F})$: the set of the probability measures on $(D_\infty(E), \mathcal{F})$ with the usual weak topology;
$\langle \nu, f \rangle = \int f(x)\nu(dx)$;
$\mathcal{P}(E)$ ($\mathcal{P}_p(E)$): the set of probability measures on $E$ (with a finite $p$th moment), equipped with the Vasershtein metric ($L^p$-analogue).

# Interaction function I

Interaction is caused by JSQ. For the original model, the arrival rate to a shortest queue is described by

$$q^{(N)}_{x,y} = \lambda_0 + \frac{N\lambda_1 \mathbf{1}_{\min\{x_1,\cdots,x_N\}}(x_k)}{\#\{j : x_j = \min\{x_1,\cdots,x_N\}, j = 1,\cdots,N\}}$$
$$= \lambda_0 + \frac{\lambda_1 \mathbf{1}_{\min\{x_1,\cdots,x_N\}}(x_k)}{\text{proportion of SQs in } N},$$

where $y = (x_1, \cdots, x_{k-1}, x_k + 1, x_{k+1}, \cdots, x_N)$, $\#A$ denotes the cardinality of the set $A$, and $\mathbf{1}_x(\cdot)$ is the indicator function on a single point $x$.

# Interaction function II

Define an *interaction function* $h : E \times \mathcal{P}(E) \to \mathbb{R}_+$ as the following (extra arrival rate to a SQ):

$$h(x, \nu) = \frac{\lambda_1}{\nu(\{ms(\nu)\})} \delta_{ms(\nu)}(x).$$

where $ms(\nu) = \inf\{x \geq 0, \nu(\{x\}) > 0\}$ is the minimum point of the support of the probability measure $\nu$.

# Master equation

For the above interaction function, define operator:

$$\Omega_{h,u(t)} f(i) = [\lambda_0 + h(i, u(t))(f(i+1) - f(i)] + \mu[f(i-1) - f(i)]$$

The nonlinear *master equation* has the following form

$$\frac{\mathrm{d}\langle u(t), f\rangle}{\mathrm{d}t} = \langle u(t), \Omega_{h,u(t)} f\rangle, \qquad f \in C_b(E),$$

where $u(\cdot)$ is a measure-valued function from $[0, +\infty)$ to $\mathcal{P}(E)$.

# Definition of $q$-solution

## Definition

Let $u \in \mathcal{P}(E)$, $P \in \mathcal{P}(D_\infty(E), \mathcal{F})$ is called a solution of the master equation with initial value $u$ if its marginal distribution $u_t(\cdot) = P \circ X_t^{-1}(\cdot)$ satisfies the master equation and $u_0 = u$. Moreover, $P$ is called a $q-$solution if, in addition, it is Markovian in the sense of McKean(Funaki(1984)), i.e. for any $j \in E$,

$$P(X_{t+s} = j|\mathcal{F}_t) = p(t, X_t, t+s, j), \qquad P-a.s.$$

where transition function $p(s, i, t, j)$ satisfies that

$$\frac{\mathrm{d}}{\mathrm{d}s} p(t, i, t+s, j) = \sum_{k \in E} p(t, i, t+s, k) \Omega_{h, u_{t+s}} I_{\{j\}}(k), t \geq 0.$$

# Empirical probability measure

Let $X_j(t)$ (notation abused here) be the queue length of queue $j$ at time $t$, define

$$U_N(t) := \frac{1}{N} \sum_{j=1}^{N} \delta_{X_j(t)} \tag{1}$$

which is the empirical distribution of queue length of the $N$ queues at time $t$.

# Convergence(LLN)

**Theorem**

*Let $U_N(t)$ satisfies*

$$\sup_N E^{(N)}\langle U_N(0)(\mathrm{d}x), x\rangle < \infty,$$

$$U_N(0) \overset{weakly}{\longrightarrow} U(0), \quad \langle U(0)(\mathrm{d}x), x^2\rangle < \infty.$$

*Then the sequence $\{U_N\}_{N=1}^{\infty}$ converges in the sense of weakly convergence of measure-valued stochastic processes to a q-solution of the nonlinear master equation. Moreover, if $\lambda_0 + \lambda_1 < \mu$ and $U(0)(\{0\}) > 0$, then the solution of the master equation is unique.*

# Stationary distribution

**Definition**

$\pi \in \mathcal{P}_p(E)$ is called a stationary distribution of the $q$-solution of the master equation if $P \circ X_0^{-1} = \pi$ implies that for all $t \geq 0$, $P \circ X_t^{-1} = \pi$.

# Q-matrix of limiting ‘typical queue”

## Theorem

*(1) Under the conditions of the convergence theorem, let $t \to \infty$, then the Q-matrix of a “typical queue” of the interaction queue is*

$$Q^{JSQ} = \begin{pmatrix} -(\lambda_0 + \frac{\lambda_1}{\pi_0}) & \lambda_0 + \frac{\lambda_1}{\pi_0} & 0 & \cdots \\ \mu & -(\lambda_0 + \mu) & \lambda_0 & \cdots \\ 0 & \mu & -(\lambda_0 + \mu) & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

*where $\pi = (\pi_0, \pi_1, \cdots)$ is the unique stationary distribution.*
*(2) The unique stationary distribution is $\pi_0^{JSQ} = 1 - \frac{\lambda_0+\lambda_1}{\mu}$,*

$$\pi_k^{JSQ} = \frac{\lambda_0 + \lambda_1}{\mu}\left(1 - \frac{\lambda_0}{\mu}\right)\left(\frac{\lambda_0}{\mu}\right)^{k-1}, k \geq 1$$

# Join infinity queues randomly ($J_\infty Q$) I

## Theorem

*(1) If the extra customer can join all queues randomly, then the corresponding Q-matrix will be that*

$$Q^{J_\infty Q} = \begin{pmatrix} -(\lambda_0+\lambda_1) & \lambda_0+\lambda_1 & 0 & \cdots \\ \mu & -(\lambda_0+\lambda_1+\mu) & \lambda_0+\lambda_1 & \cdots \\ 0 & \mu & -(\lambda_0+\lambda_1+\mu) & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

*which is equivalent to that of an $M/M/1$ queue with arrival and service rate are $\lambda_0+\lambda_1$ and $\mu$ respectively.*

# Join infinity queues randomly ($J_\infty Q$) II

## Theorem

*(2) If we let $\lambda_0 + \lambda_1 < \mu$, then this queue will be stable and the stationary distribution satisfies that*

$$\pi_k^{J_\infty Q} = \left(1 - \frac{\lambda_0 + \lambda_1}{\mu}\right)\left(\frac{\lambda_0 + \lambda_1}{\mu}\right)^k, \qquad k \geq 0 \tag{2}$$

# Comparison of stationary distributions between JSQ and $J_\infty$Q I

**(1)** $\pi_0^{JSQ} = \pi_0^{J_\infty Q}$, which means that since the average arrival rate and service rate are the same, the idle probability of the servers are the same;

**(2)** The tail of $\pi_\cdot^{J_\infty Q}$ is something like $const \cdot (\frac{\lambda_0+\lambda_1}{\mu})^k$, while that of $\pi_\cdot^{JSQ}$ is something like $const \cdot (\frac{\lambda_0}{\mu})^k$;

**(3)** The average queue length of JSQ is shorter than that of $J_\infty$Q: $\sum_{k=0}^{\infty} k\pi_k^{JSQ} = \frac{\lambda_0+\lambda_1}{\mu-\lambda_0} < \frac{\lambda_0+\lambda_1}{\mu-(\lambda_0+\lambda_1)} = \sum_{k=0}^{\infty} k\pi_k^{J_\infty Q}$;

**(4)** If $\lambda_1 = 0$, then $\pi_k^{JSQ} = \pi_k^{J_\infty Q}$, $k \geq 0$. Because in this case, they all are equivalent to $M(\lambda_0)/M(\mu)/1$ queue.

# Comparison of stationary distributions between JSQ and $J_\infty$Q II

**(5)** As we know that the tail of $\pi_{\cdot}^{JSQ}$ is depending on $\lambda_0$, if we let $\lambda_0 = 0$, then we have: $\pi_0^{JSQ} = 1 - \frac{\lambda_1}{\mu}$, $\pi_1^{JSQ} = \frac{\lambda_1}{\mu}$ and $\pi_k^{JSQ} = 0$ for all $k \geq 2$.

**(6)** $\lambda_1 \uparrow (\mu - \lambda_0)$ such that $\lambda_0 + \lambda_1 \uparrow \mu$, then the limit of the stationary distribution of the JSQ is that $\pi_0^{JSQ} \downarrow 0$, $\pi_k \uparrow (1 - \lambda_0)(\frac{\lambda_0}{\mu})^{k-1}$, $k \geq 1$, while the stationary distribution of the $J_\infty$Q does not have a limit distribuyion.

# Join the $m$-th shortest queue: $1 \leq m \leq s$ I

If the extra customer can randomly join the queue whose length is between the shortest and $s$-shortest, then convergence result similar to Theorem 2 can also be established, in this case, as the time $t$ tends to infinity, then the $Q$-matrix will be

$$q_{ij}^{J1\sim sQ} = \begin{cases} \lambda_0 + \frac{\lambda_1}{\pi_0^1 + \cdots + \pi_{s-1}^1}, & j = i+1, i = 0, \cdots, s-1 \\ \lambda_0, & j = i+1, i > s-1 \\ -\lambda_0 - \mu - \frac{\lambda_1}{\pi_0^1 + \cdots + \pi_{s-1}^1}, & j = i, i = 0, \cdots, s-1 \\ -\lambda_0 - \mu, & j = i, i > s-1 \\ \mu, & j = i-1, i \geq 1 \\ 0, & \text{others} \end{cases}$$

# Join the $m$-th shortest queue: $1 \leq m \leq s$ II

(1) For the case of $s = 2$, then the stationary distribution of the limiting typical queue is that

$$\pi_0 = 1 - \frac{\lambda_0 + \lambda_1}{\mu}$$

$$\pi_1 = \frac{1}{2}\left(\sqrt{\left(1-\frac{\lambda_0}{\mu}\right)^2\left(1-\frac{\lambda_0+\lambda_1}{\mu}\right)^2 + 4\frac{\lambda_0+\lambda_1}{\mu}\left(1-\frac{\lambda_0}{\mu}\right)\left(1-\frac{\lambda_0+\lambda_1}{\mu}\right)} - \left(1-\frac{\lambda_0}{\mu}\right)\left(1-\frac{\lambda_0+\lambda_1}{\mu}\right)\right)$$

$$\pi_k = \frac{1}{2}\left(1-\frac{\lambda_0}{\mu}\right)\left(\left(1-\frac{\lambda_0}{\mu}\right)\left(1-\frac{\lambda_0+\lambda_1}{\mu}\right) + 2\frac{\lambda_0+\lambda_1}{\mu} - \sqrt{\left(1-\frac{\lambda_0}{\mu}\right)^2\left(1-\frac{\lambda_0+\lambda_1}{\mu}\right)^2 + 4\frac{\lambda_0+\lambda_1}{\mu}\left(1-\frac{\lambda_0}{\mu}\right)\left(1-\frac{\lambda_0+\lambda_1}{\mu}\right)}\right)\left(\frac{\lambda_0}{\mu}\right)^{k-2},$$

Moreover, the average arrival rate is $\lambda_0 + \lambda_1$.

# Join the $m$-th shortest queue: $1 \le m \le s$ III

(2) If $\lambda_1 = 0$, then

$$\pi_k = \left(1 - \frac{\lambda_0}{\mu}\right)\left(\frac{\lambda_0}{\mu}\right)^k, \qquad k \ge 0.$$

(3) If $\lambda_0 = 0$, then

$$\begin{aligned}
\pi_0 &= 1 - \frac{\lambda_1}{\mu} \\
\pi_1 &= \frac{1}{2}\left(\sqrt{\left(1 - \frac{\lambda_1}{\mu}\right)^2 + 4\frac{\lambda_1}{\mu}\left(1 - \frac{\lambda_1}{\mu}\right)} - \left(1 - \frac{\lambda_1}{\mu}\right)\right) \\
\pi_2 &= \frac{1}{2}\left(1 + \frac{\lambda_1}{\mu} - \sqrt{\left(1 - \frac{\lambda_1}{\mu}\right)^2 + 4\frac{\lambda_1}{\mu}\left(1 - \frac{\lambda_1}{\mu}\right)}\right) \\
\pi_k &= 0, \qquad k \ge 3.
\end{aligned}$$

# Justification I

The stationary distribution of the limiting typical queue is the behaviour of $\lim_t \lim_N \text{JSQ}^{(N)}$. However, our original interest is $\lim_t \text{JSQ}^{(N)}$ when $N$ is large, or we would like to approximate $\pi^{(N)}$ by the behaviour of $\lim_N \lim_t \text{JSQ}^{(N)}$. In order to do so, we need to justify:

$$\lim_N \lim_t \text{JSQ}^{(N)} = \lim_t \lim_N \text{JSQ}^{(N)}$$

in some sense.

Let $\lambda + \lambda_s < \mu$. For $N \geq 1$, denote by $\mathbf{E}_N(\cdot)$ the mathematical expectation with respect to the stationary distribution $\pi^{(N)}$.

# Justification II

## Theorem

*For any integer* $k \geq 0$*,*

$$\lim_{N\to\infty} \mathbf{E}_N \langle U_N(\cdot), \{k\} \rangle = \pi_k,$$

*where the measure* $U_N(\cdot)$ *is defined in formula (1) and* $\{\pi_k, k \geq 0\}$ *is the stationary distribution of the limiting typical queue given by formula (1).*

# Conclutions

- When $N$ is large, the interaction queueing network can be studied in terms of the limiting "typical" queue
- Load-balancing described as the mean-field interaction in this talk does improve the system performance
- We expect that this method can be used to study other balancing mechanisms

# References I

Adan, I.J.B.F.; Wessels, J.; Zijm, W.H.M. (1990). Analysis of the symmetric shortest queue problem. Stochastic Models, 6(4): 691-713.

Dawson, D. A. (1983). Critical dynamics and fluctuations for a mean-field model of cooperative behavior. Journal of Statistical Physics, 31(1): 29-85.

Dawson, D. A.; Tang, J.S.; Zhao, Y. Q. (2005). Balancing queues by mean field interaction. Queueing Systems, 49(3-4): 335-361.

Dawson, D. A.; Zheng, X. G. (1991). Laws of large numbers and a central limit theorem for unbounded jump mean field models. Adv. Appl. Math., 12(3): 293-326.

Ethier, S.N.; Kurtz, T.G. (1986). Markov processes: characterization and convergence, Wiley & Sons, New York.

Carleton UNIVERSITY

# References II


- Feng, S. (1994). Large deviations for Markov processes with mean field interaction and unbounded jumps. Probab. Theory Relat. Fields, 100(2): 227-252.
- Flatto, L.; McKean, H. P. (1977). Two queues in parallel. Comm. Pure Appl. Math, 30(2): 255-263.
- Foley, R. D.; McDonald, D. R. (2001) Join the shortest queue: stability and exact asymptotics. Ann. Appl. Probab., 11 (3): 569-607.
- Funaki, T. (1984). A certain class of diffusion processes associated with nonlinear parabolic equations. Zei fur Wahr, 67(3): 331-348.
- Graham, C. (2000). Chaoticity on path space for a queueing network with selection of the shortest queue amone several. J. Appl. Prob., 37(1): 198-211.

# References III


- Graham C. (2005). Functional central limit theorems for a large network in which customers join the shortest of several queues. Prob. Theory Relat. Fields., 131(1): 97-120.
- Haight, F. A. (1958). Two queues in parallel. Biometrika, 45: 401-410.
- Halfin, S.(1985) The shortest queue problem. J. Appl. Probab., 22(4): 865–878.
- Hepp, K.; Lieb, E. H. (1973). On the superradiant phase transition for molecules in a quantized radiation field: The Dicke Maser model. Ann. Physics, 76: 360-404.
- Johri, P. K. (1989). Optimality of the shortest line discipline with state-dependent service rates. European J. Oper. Res., 41(2): 157–161.

# References IV


- Kingman, J. F. C. (1961) Two similar queues in parallel, Ann. Math. Statist., 32: 1314-1323.
- Luczak, M.J.; McDiarmid, C. (2006). On the maximum queue length in the supermarket model, Ann. Probab., 34(2): 493-527.
- Martin J. B.; Suhov Yu. M. (1999). Fast Jackson networks. Annals of Applied Probability, 9(3): 854-870.
- McDonald, D. (1995). Overloading parallel servers when arrivals join the shortest queue. Stochastic networks (New York, 1995), 169-196, Lecture Notes in Statist., 117, Springer, New York, 1996.
- McDonald, D. (1999). Asymptotics of first passage times for random walk in a quadrant. Annals of Applied Probability, 9(1): 110-145.
- Mitzenmather, M. (1996). The power of two choices in randomized load balancing. Ph.D. thesis, University of Carlifornia, Berkeley.

# References V


- Mitzenmacher, M.; Voecking, B. (2002). Selecting the shortest of two, improved. Analytic Methods in applied probability, 165-176, Amer. Math. Soc. Transl. Ser. 2, 207, Amer. Math. Soc., Providence, RI.
- Oseledets, V. I.; Khmelëv, D. V. (2000). Global stability of infinite systems of nonlinear differential equations, and nonhomogeneous countable Markov chains. Problemy Peredachi Informatsii 36(1): 60-76; translation in Probl. Inf. Transm., 36(1): 54-70.
- Oseledets, V. I.; Khmelev, D. V. (2002). Stochastic transportion networks and stability of dynamical systems. Theory of Probability and Its Applications, 46(1): 154-161.
- Roberts, G. E.; Kaufman, H. (1966). Table of Laplace transforms. Saunders W. B. Company, Philadelphia and London

# References VI


- Vvedenskaya, N. D.; Dobурshin R. L.; Karpelevich, F. I. (1996). Queueing system with selection of the shortest of two queues: an asymptotic approach. Problems of Information Transmission, 32(1): 15-27.
- Vvedenskaya, N. D.; Suhov, Yu. M. (1997). Dobrushin's mean-field approximation for a queue with dynamic routing. Markov Processes and Related Fields, 3(4): 493-526.
- Vvedenskaya, N. D.; Suhov, Yu. M. (2002). Fast Jackson networks with dynamic routing. Problems of Information Transmission, 38(2): 136-153.

# Thanks You!

Carleton UNIVERSITY