5^{th} **Workshop on Markov Processes & Related Topics**

Beijing, China, July 14-18, 2007

Jingshi Hotel, Beijing Normal University, 16 Jul, 09:00-09:30

Performance Analysis of Joining the Shortest Queue Model among a Large Number of Queues*

JIASHAN TANG

College of Mathematics and Physics,
Nanjing University of Posts and Telecommunications

*This is a joint work with Don Dawson and Yiqiang Q Zhao

Outline

- ***Motivation***
- ***Modeling***
- ***Convergence Theorem (LLN)***
- ***Stationary Distribution***
- ***Conclutions***

1 Motivation

1.1. Balancing techniques

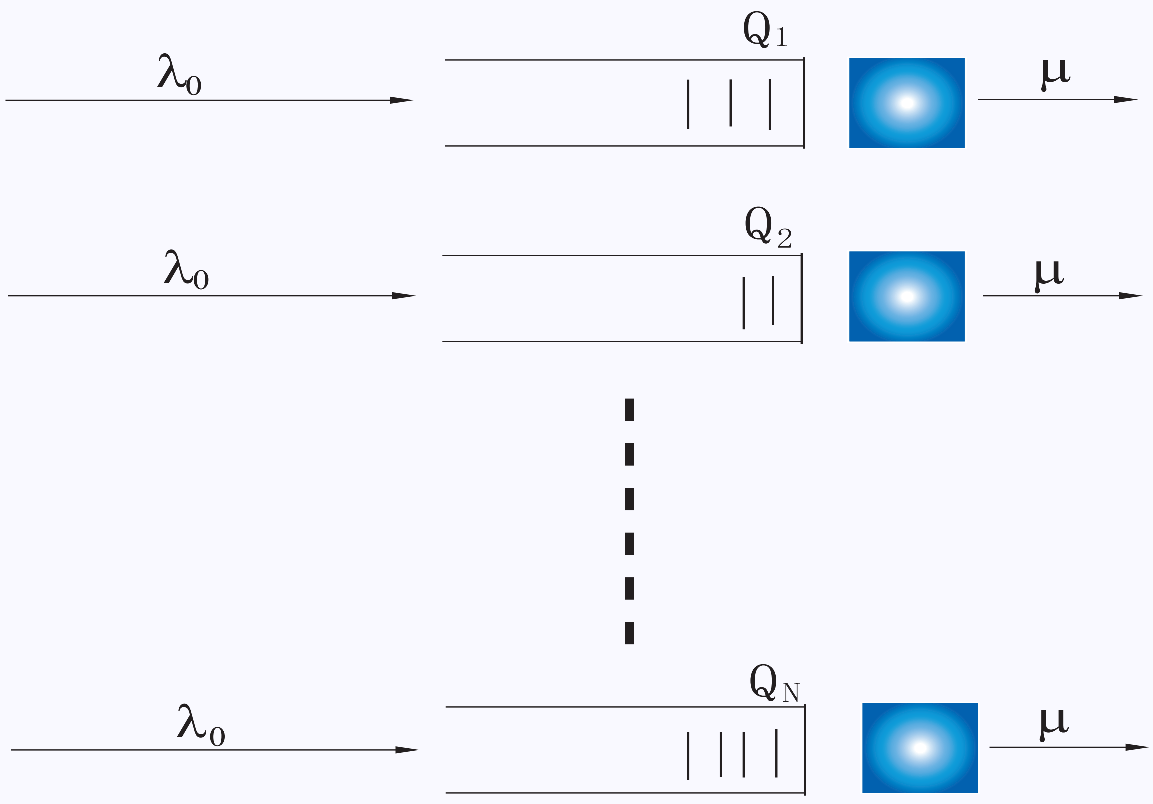
- **Changing the arrival and/or service rates**
  - ▷ Changing the arrival rate
  - ▷ Changing the service rate
- **Joining different queues**
  - ▷ JSQ model
  - ▷ Extra arrival source may choose different queues to join
  - ▷ Dobrushin's mean-field model

1.1 Balancing techniques(con't)

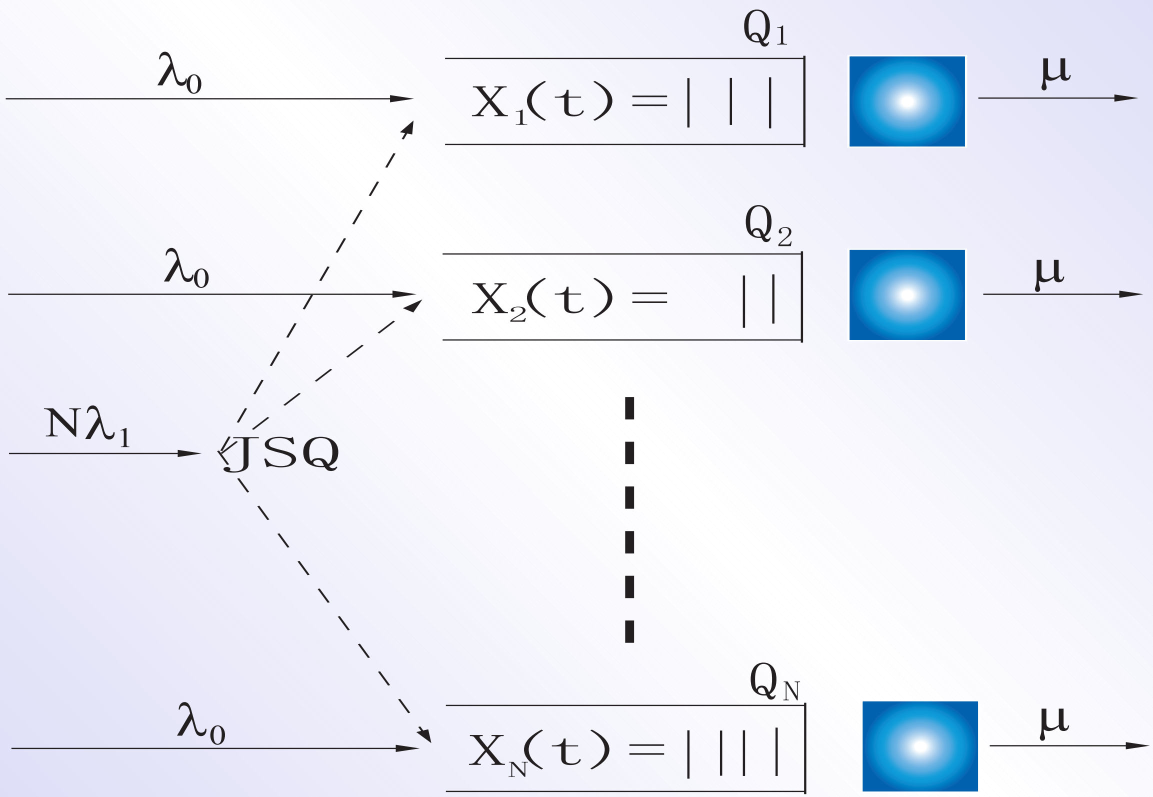
- **Jockeying**
  - ▷ Periodic redistributing customers in all queues
  - ▷ r difference jockeying

2 Modeling

2.1. Noninteraction parallel network



2.2. Mean field interaction parallel network: JSQ with large N



2.3. Description of the main result

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

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

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

2.3 Description of the main result(con't)

(2) If $\lambda_1 = 0$ then

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

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

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

$$\pi_1^{JSQ} = \frac{\lambda_1}{\mu},$$

$$\pi_k^{JSQ} = 0 \quad \text{for all} \quad k \geq 2.$$

2.4. Empirical probability measure

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

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

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

2.5. Interaction function

Define an *interaction function* $h : \mathbf{R}_+ \times \mathcal{P}(E) \to \mathbf{R}$ as the following:

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

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

2.6. 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

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

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

2.7. Definition of q-solution

Definition 1. 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 $j \in E$,

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

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

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

3 Convergence theorem (LLN)

Theorem 2. Let $U_N(t)$ satisfies

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

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

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.

4 Stationary distribution

4.1. Definition

Definition 3. $\pi \in \mathcal{P}_p(E)$ is called a stationary distribution of the q-solution of the master equation if $P \circ X_0^{-1} = \pi$ implies that for all $t \geq 0$, $P \circ X_t^{-1} = \pi$.

4.2. Join the shortest queue

Theorem 4. (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.

4.2 Join the shortest queue(con't)

(2) The unique stationary distribution is

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

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

4.3. Join infinity queues randomly

Theorem 5. (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 μ respectively.

4.3 Join infinity queues randomly(con't)

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

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

4.4. Comparison of stationary distributions between JSQ and $J_\infty Q$

(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.

4.4 Comparison of stationary distributions between JSQ and J_∞Q(con't)

(5) As we know that the tail of $\pi_{\cdot}^{JSQ}$ is depending on λ_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_∞Q does not have the limit.

From these it is very easy to see that the performance of the JSQ system has been improved.

4.5. Join the m-th shortest queue: $1 \leq m \leq s$

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

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

4.5 Join the m-th shortest queue: $1 \leq m \leq s$(con't)

One can calculate the stationary distribution corresponds to Q-matrix, as an example, we have

Theorem 6 (1) For the case of $s = 2$, then the stationary distribution of the limiting typical queue is that

$$
\begin{aligned}
\pi_0 &= 1 - \frac{\lambda_0 + \lambda_1}{\mu} \\
\pi_1 &= \frac{1}{2}\left(\sqrt{\left(1 - \frac{\lambda_0}{\mu}\right)^2 \left(1 - \frac{\lambda_0 + \lambda_1}{\mu}\right)^2 + 4\frac{\lambda_0 + \lambda_1}{\mu}\left(1 - \frac{\lambda_0}{\mu}\right)\left(1 - \frac{\lambda_0 + \lambda_1}{\mu}\right)} \right. \\
&\qquad \left. - \left(1 - \frac{\lambda_0}{\mu}\right)\left(1 - \frac{\lambda_0+\lambda_1}{\mu}\right)\right) \\
\pi_k &= \frac{1}{2}\left(1 - \frac{\lambda_0}{\mu}\right)\left(\left(1 - \frac{\lambda_0}{\mu}\right)\left(1 - \frac{\lambda_0 + \lambda_1}{\mu}\right) + 2\frac{\lambda_0 + \lambda_1}{\mu}\right. \\
&\qquad \left. - \sqrt{\left(1 - \frac{\lambda_0}{\mu}\right)^2 \left(1 - \frac{\lambda_0+\lambda_1}{\mu}\right)^2 + 4\frac{\lambda_0+\lambda_1}{\mu}\left(1 - \frac{\lambda_0}{\mu}\right)\left(1 - \frac{\lambda_0+\lambda_1}{\mu}\right)}\right)\left(\frac{\lambda_0}{\mu}\right)^{k-2}, \qquad k \geq 2
\end{aligned}
$$

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

4.5 Join the m-th shortest queue: $1 \le m \le s$(con't)

(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

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

5 Conclusions

- When N is large, the interaction queueing network can be studied in terms of "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

Thank you

E-mail: jiashant@yahoo.ca

Notation

$\mathbf{R}$ ——— the set of all real numbers,

$\mathbf{R}_+$ —— the set of all nonnegative numbers,

$(E,\mathcal{E})$ — $E=\{0,1,2,\cdots\}$ and be equipped with discrete topology,

$C_b(E)$ — the set of bounded continuous functions in E,

$D_\infty(E)$ - the set of functions from $[0,\infty)$ to E which are càdlàg,

$X(t,w)$ – coordinate process with $w\in D_\infty(E)$,

$\mathcal{F}_t$ ——— $\sigma\{X(s), 0\leq s\leq t\}$,

$\mathcal{F}$ ——— $\sigma\{X(s), s\geq 0\}$,

$\mathcal{P}(E)$ —— the set of probability measures on E,

$\mathcal{P}_p(E)$ — space of elements in $\mathcal{P}(E)$ with finite pth moment,

$\mathcal{P}(D_\infty(E),\mathcal{F})$ — probability measures on $(D_\infty(E),\mathcal{F})$.

$$\langle \nu, f\rangle = \int f(x)\nu(dx)$$