

Lecture Notes, Math 289, Monday November 24, 2008

Notetaker: Vladimir Pesic

Single Class Queueing Networks

A single class queueing network consists of n infinite capacity buffers and n non identical servers, exactly one server for each buffer. The service discipline is **FIFO**-first in first out. Probabilistic Markovian routing is given by an $n \times n$ matrix P where P_{ij} is the probability that a job just completed at server i goes next to buffer/server j . We assume that P is transient on $\{1, \dots, n\}$ in other words

$$\text{spectral radius of } P < 1 \text{ or } P^m \rightarrow 0 \text{ as } m \rightarrow \infty.$$

The transience of P corresponds to the fact that jobs eventually leave the system.

Stochastic Primitives

Exogenous Arrivals

For each buffer i , there is an arrival process E_i such that $E_i(t)$ is the number of exogenous arrivals up to time t . This is a counting process that goes with a sequence of positive i.i.d. inter-arrival times. We assume that inter-arrival times have finite mean $\frac{1}{\alpha_i}$ (arrival rate is α_i) and variance a_i^2 .

Service

For $i = 1, \dots, n$ $S_i(\cdot)$ is a renewal process that goes with service times for buffer i , where $S_i(t)$ is the number of service completions by time t if server i never idles. We assume that service times for buffer i are strictly positive i.i.d. with mean $\frac{1}{\mu_i}$ (service rate is μ_i) and variance b_i^2 .

Routing vectors

For each buffer j there is a sequence $\{\varphi^j(l)\}_{l=1}^{\infty}$ of n -dimensional random vectors such that

$$P(\varphi^j(l) = e_i) = P_{ji},$$

where e_i is the i -th unit basis vector. $1 - \sum_i P_{ji}$ is the probability that jobs leave the system from buffer/server j immediately after service there.

Performance Measure

$Q_i(t)$ is the number of jobs in buffer i at time t . $Q_i(t)$ can be expressed as follows

$$Q_i(t) = Q_i(0) + E_i(t) - D_i(t) + \sum_{j=1}^n \Phi_i^j(D_j(t)),$$

where $D_i(t)$ is the number of jobs served by server i up to time t . $D_i(t) = S_i(T_i(t))$ where $T_i(t)$ is the amount of time up to time t that server i is busy, in other words

$$T_i(t) = \int_0^t 1_{\{Q_i(s) > 0\}} ds.$$

$\Phi^j(m) = \sum_{l=1}^m \varphi^j(l)$, so $\Phi_i^j(D_j(t))$ is the number of departures from buffer j routed to buffer i up to time t .

We assume that $E_1, \dots, E_n, S_1, \dots, S_n, \Phi^1, \dots, \Phi^n$ are all independent of each other.

Note: If all inter-arrival times and service times are exponential the network described above is called a Jackson Network and Q is a Markov process. The current model is called a Generalized Jackson Network and Q need not be a Markov process.

It is difficult to analyze Q exactly, so we consider a more tractable approximation.

Fluid Scaling

We let r go to infinity through some sequence of positive real numbers and we

let $\bar{Q}^r(t) = \frac{Q^r(rt)}{r}$ (Q^r may have an initial condition that depends on r). Then

$$\begin{aligned}\bar{Q}_i^r(t) &= \bar{Q}_i^r(0) + \bar{E}_i^r(t) - \bar{D}_i^r(t) - \sum_{j=1}^n \frac{\Phi_i^j(D_j(rt))}{r} \\ &= \bar{Q}_i^r(0) + \bar{E}_i^r(t) - \bar{D}_i^r(t) + \sum_{j=1}^n \bar{\Phi}_i^{r,j}(\bar{D}_j(t)),\end{aligned}$$

where

$$\begin{aligned}\bar{D}_i^r(t) &= \frac{S_i(T_i^r(rt))}{r} = \frac{S_i(r\bar{T}_i^r(t))}{r} = \bar{S}_i^r(\bar{T}_i^r(t)), \\ \bar{T}_i^r(t) &= \frac{T_i^r(rt)}{r} = \frac{\int_0^{rt} 1_{\{Q_i(s)>0\}} ds}{r} = \int_0^t 1_{\{\bar{Q}_i^r(s)>0\}} ds, \\ \bar{\Phi}_i^{r,j}(t) &= \frac{\sum_{l=1}^{[rt]} \varphi^j(l)}{r} = \frac{\Phi^j([rt])}{r}.\end{aligned}$$

By the **FLLN** ($\bar{E}^r(\cdot), \bar{S}^r(\cdot) \Rightarrow (\alpha(\cdot), \mu(\cdot))$) where $\alpha(t) = \alpha t$ and $\mu(t) = \mu t$. Assume that $\bar{Q}^r(0) \Rightarrow q_0$ as $r \rightarrow 0$, where q_0 is deterministic.

What about $\bar{D}^r$?

Since $\{\bar{T}^r\}$ are uniformly Lipschitz continuous, with a bound independent of r , it is easy to see that $\{\bar{T}^r\}$ is C -tight. Also by **FLLN**,

$$(\bar{\Phi}^{r,1}(\cdot), \dots, \bar{\Phi}^{r,n}(\cdot)) \Rightarrow (P^1(\cdot), \dots, P^n(\cdot))$$

where,

$$P^j(t) = \mathbb{E}[\varphi^j(1)]t \text{ and } P_i^j(t) = \mathbb{E}[\varphi_i^j(1)]t = P_{ji}t.$$

Putting this all together with the random time change theorem we get that

$$(\bar{E}^r, \bar{D}^r, \bar{S}^r, \bar{T}^r, \bar{\Phi}^{r,1}, \dots, \bar{\Phi}^{r,n}, \bar{Q}^r(\cdot)) \text{ is } C\text{-tight,}$$

and any weak limit point

$$(\alpha(\cdot), \bar{D}, \mu(\cdot), \bar{T}, P^1(\cdot), \dots, P^n(\cdot), \bar{Q}(\cdot))$$

satisfies,

$$\bar{Q}_i(t) = q_{0,i} + \alpha_i t - \bar{D}_i(t) + \sum_{j=1}^n P_{ji} \bar{D}_j(t)$$

where $\bar{D}_i(t) = \mu_i \bar{T}_i(t)$ (remember that $\bar{D}_i^r(t) = \bar{S}_i^r(\bar{T}_i^r(t))$).

$\bar{I}_i^r(t) = t - \bar{T}_i^r(t)$ is the idle time of server i . As $r \rightarrow \infty$, $\bar{I}^r \Rightarrow \bar{I}$, where $\bar{I}_i(\cdot) = t - \bar{T}_i(\cdot)$. $\bar{I}_i(t)$ can increase only when $\bar{Q}_i(t) = 0$.

In vector form,

$\bar{Q}(t) = q_0 + \alpha t - (I - P')\bar{D}(t)$ where P' is the transpose of P and I is the $n \times n$ identity matrix,

$$\bar{D}(t) = \text{diag}(\mu)\bar{T}(t),$$

$\bar{I}(t) = \mathbf{1}t - \bar{T}(t)$ where $\mathbf{1}$ is the n -dim'l vector of ones.

The above equations are referred to as the **Fluid Model** equations. These equations can be expressed as

$$\bar{Q}(t) = q_0 + \alpha t - (I - P')\mu t + (I - P')\text{diag}(\mu)\bar{I}(t),$$

$$\int_0^\infty 1_{\{\bar{Q}_i(s) > 0\}} d\bar{I}_i(s) = 0 \text{ for every } i.$$

By the uniqueness of solutions of the Skorokhod problem $\bar{Q}, \bar{I}$ are unique given q_0, α, μ, P . Then it follows that $\bar{Q}^r \Rightarrow \bar{Q}$ as $r \rightarrow \infty$.

Theorem (DAI, 1995): Under mild assumptions on exogenous arrival processes (including “spread out assumptions”), if all fluid model solutions reach the origin in finite time and stay there, then the original queueing network has a Markovian state descriptor that is positive Harris recurrent. The head of the line service discipline is assumed.