Skills-based Routing under Demand Surges

Jinsheng Chen
(Joint work with Jing Dong and Pengyi Shi)

Monday 22nd February, 2021

Motivation

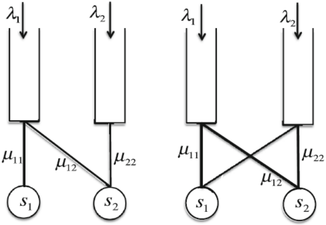
- Many service and manufacturing systems involve multiple customer classes and server types
- Different servers may serve different customer classes at different rates, and may have a preferred or ‘primary’ customer class
- Hospital setting: inpatient wards grouped according to specialty
- Patients have a ‘primary’ ward that they are best served in
- Patients may be placed in non-primary wards if necessary (‘overflowed’), which may lead to service slowdown and other costs
- Recent survey paper: J, J. Dong, and P. Shi. “A survey on skills-based routing with applications to service operations management,” *Queueing Systems*, Oct. 2020.

Motivation

- Arrival rates in general not fixed, but are prone to sudden surges, e.g. pandemic or seasonality
- Can often anticipate future arrival rates
- Want to make use of future arrival rate information, e.g. do we prioritize a customer class now if its arrival rate is going to increase but hasn't yet?

Stochastic Model

- Markovian N and X models
- Preferred pool i for each class i (of size N_i)
- Holding costs h_i, overflow costs $\phi_{ij} \geq 0$
- Arrival rates $\lambda_i(t)$, service rates μ_{ij}



Stochastic Model

- Want to minimize

$$\mathbb{E}_\pi \left[\int_0^T \left(\sum_i h_i X_i(t) + \sum_{i \neq j} \phi_{ij} Z_{ij}(t) \right) dt \right]$$

  for some 'large' T
- Headcounts $X_i(t)$, $Z_{ij}(t)$ class i customers in pool j service
- Expectation depends on the scheduling policy π

Fluid Model

Associated fluid control problem:

$$\min_z \int_0^T \left(\sum_i h_i x_i(t) + \sum_{i,j} \phi_{ij} z_{ij}(t) \right) dt$$

$$\text{s.t. } x_i'(t) = \lambda_i(t) - \sum_j \mu_{ij} z_{ij}(t)$$

$$x_i(t) \geq 0$$

$$z_{ij}(t) \geq 0$$

$$\sum_i z_{ij}(t) \leq N_j,$$

The Arrival Rate $\lambda(t)$

Assumption (Initial high arrival rate)

For $i = 1, 2$*, there exists* $K_i \in [0, \infty)$ *such that* $\lambda_i(t) \geq N_i\mu_{ii}$ *for* $t < K_i$ *and* $\lambda_i(t) < N_i\mu_{ii}$ *for* $t > K_i$*.*

Assumption (Regularity)

$\lambda_i(t)$ *is piecewise monotone and* $\int_0^\infty (N_i\mu_{ii} - \lambda_i(s))\, ds = \infty$*.*

T is large enough that $\int_0^T (N_i\mu_{ii} - \lambda_i(s))\, ds > X_i(0)$

Definition

For each $t \geq 0$, the function $G_i^t : \mathbb{R}_+ \to \mathbb{R}_+$ is defined by $\int_t^{t+G_i^t(x)} (N_i\mu_{ii} - \lambda_i(s))\, ds = x$. It is a continuous strictly increasing bijection with $G_i^t(0) = 0$. Note that if $\lambda_i(t) \equiv \lambda_i$ is constant, then $G_i^t(x) = \frac{x}{N_i\mu_{ii} - \lambda_i}$ for all $t \geq 0$.

The 'look-ahead' function G_i^t

Definition

For each $t \geq 0$, the function $G_i^t : \mathbb{R}_+ \to \mathbb{R}_+$ is defined by

$$G_i^t(x) = \inf \left\{ s \geq (K_i - t)^+ : \int_t^{t+s} (N_i \mu_{ii} - \lambda_i(s))\, ds \geq x \right\}.$$

- $G_i^t(x)$ is how long it takes for the class i queue of length x to be emptied using only the primary pool i, starting at time t.
- If $\lambda_i(t) \equiv \lambda_i$ is constant, then $G_i^t(x) = \frac{x}{N_i \mu_{ii} - \lambda_i}$ for all $t \geq 0$.
- $G_i^t(x)$ can be large even if x is small

N Model Optimal Control

Theorem

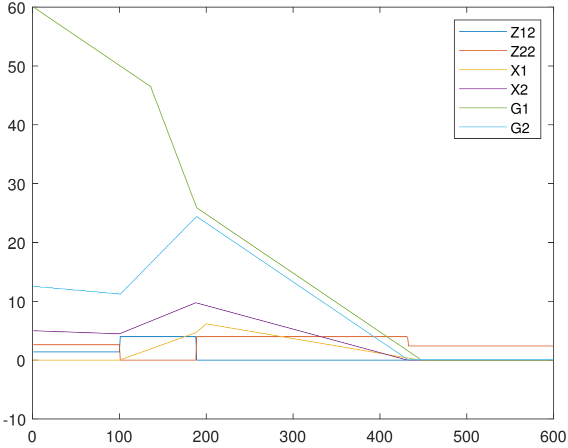
The following control is optimal for the fluid model:

I. *When $h_1\mu_{12} > h_2\mu_{22}$, pool 2 gives priority to class 1 when queue 1 is large enough relative to queue 2. In particular,*
   a. *If $h_1\mu_{12}G_1^t(x_1(t)) - \phi_{12} > h_2\mu_{22}G_2^t(x_2(t))$, $z_{12}^*(t) = N_2$ and $z_{22}^*(t) = 0$.*
   b. *Otherwise, $z_{12}^*(t) = 0$ and*
      $z_{22}^*(t) = N_2 1\{x_2(t) > 0\} + \frac{\lambda_2(t)}{\mu_{22}} 1\{x_2(t) = 0\}$.

II. *When $h_1\mu_{12} < h_2\mu_{22}$, pool 2 gives priority to class 2 and will help queue 1 when $x_2(t) = 0$ and $x_1(t)$ is large enough. In particular,*
   a. *If $x_2(t) = 0$ and $h_1\mu_{12}G_1^t(x_1(t)) - \phi_{12} > 0$,*
      $z_{12}^*(t) = N_2 - \frac{\lambda_2(t)}{\mu_{22}}$ *and* $z_{22}^*(t) = \frac{\lambda_2(t)}{\mu_{22}}$.
   b. *Otherwise, $z_{12}^*(t) = 0$ and*
      $z_{22}^*(t) = N_2 1\{x_2(t) > 0\} + \frac{\lambda_2(t)}{\mu_{22}} 1\{x_2(t) = 0\}$.

N Model Optimal Control

Example: $h_1 = 1.5, h_2 = 1, \phi_{12} = 1, \mu_{11} = \mu_{22} = .25, \mu_{12} = .18, x(0) = (0, 5), N = (3, 4)$, and $\lambda_2(t) \equiv 0.6$ and $\lambda_1(t) = 1$ for $0 \leq t < 10$, 2 for $10 \leq t < 20$, and 0.5 for $t \geq 20$.



X Model Optimal Control

Assumption ('Basic Inefficient Sharing Condition' (Perry and Whitt 2009))

$\mu_{11}\mu_{22} > \mu_{12}\mu_{21}$

Theorem (The case $h_1\mu_{12} > h_2\mu_{22}$ and $h_1\mu_{11} > h_2\mu_{21}$)

The following policy is optimal for the fluid model. Pool 1 prioritizes class 1, and partially helps class 2 if $G_1^t(q_1(t)) = 0$ *and* $G_2^t(q_2(t)) > \frac{\phi_{21}}{h_2\mu_{21}}$*. Pool 2 prioritizes class 1 if the following holds, and serves only its own class otherwise. Let* $\tau_i = G_i^t(q_i(t))$ *be the remaining time to empty queue* $i = 1, 2$*, and let* $\tau = \inf\{s \geq 0 : G_2^{\tau_1+t+s}(q_2(\tau_1 + t + s)) \leq \frac{\phi_{21}}{h_2\mu_{21}}\}$ *be the partial help duration after* τ_1*. Then,*

$$\begin{aligned}\phi_{12} <& h_1\mu_{12}\tau_1 + h_2\frac{\mu_{12}\mu_{21}}{\mu_{11}}\tau \\ & - h_2\mu_{22}\left(\tau_2 1\{\tau = 0\} + \left(\tau_1 + \tau + \frac{\phi_{21}}{h_2\mu_{21}}\right)1\{\tau > 0\}\right).\end{aligned}$$

X Model Optimal Control

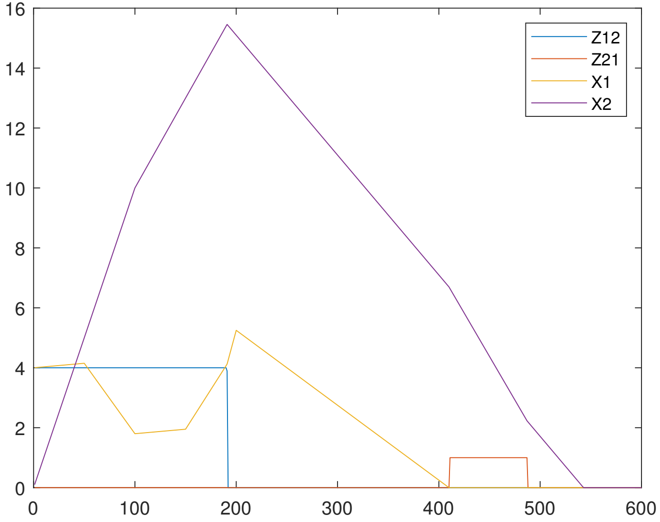
Consider the setting $N = (3,4), x = (4,0)$, service rates $\mu_{ii} = 0.25, \mu_{ij} = 0.18$ and arrival rates

$$\lambda_1(t) = \begin{cases} 1.5 & t \in [0,5) \\ 1 & t \in [5,10) \\ 1.5 & t \in [10,15) \\ 2 & t \in [15,20) \\ 0.5 & t \in [20,\infty) \end{cases}$$

$$\lambda_2(t) = \begin{cases} 1 & t \in [0,10) \\ 0.6 & t \in [10,\infty) \end{cases}$$

Costs are $h = (2,1)$ and $\phi_{21} = 10\phi_{12} = 1$.

X Model Optimal Control



Tracking Policies

- The optimal fluid control can be defined by four constants T_1, T_2, T_3, T_4
- In $[0, T_1)$, both pools fully help class 1, and any idle pool 2 servers (due to insufficient class 1 customers) serve class 2.
- In $[T_1, T_2)$, both pools each serve their own class only, until queue 1 is emptied at time T_2.
- In $[T_2, T_3)$, both pools each fully serve their own class, and any idle pool 1 servers serve class 1.
- In $[T_3, \infty)$, both pools each serve their own class only. Queue 2 is emptied at time T_4.

X Model Optimal Control

Theorem (The case $h_1\mu_{12} < h_2\mu_{22}$ and $h_2\mu_{21} < h_1\mu_{11}$)

The following policy is optimal for the fluid model. Each pool i prioritizes its own class i, and partially helps the other class $j \neq i$ if $G_i^t(q_i(t)) = 0$ and $G_j^t(q_j(t)) > \frac{\phi_{ji}}{h_j\mu_{ji}}$.

Asymptotic Regime

- We consider sequences indexed by $n \to \infty$
- Let $X^n(t)$ be the headcount process defined on $t \in [0, nT]$ with arrival rate $\lambda^n(t) := \lambda(t/n), t \in [0, nT]$ and initial state
- A scheduling policy π^n for the nth system is defined on $[0, nT]$ so that $Z^n(t) = \pi^n_t(X^n(t))$.
- We say that a sequence (X^n, Z^n) is asymptotically optimal if

$$\lim_{n\to\infty} \frac{1}{n} \int_0^{nT} \left(\sum_i \frac{h_i}{n} X_i^n(t) + \sum_{i \neq j} \phi_{ij} Z_{ij}^n(t) \right) dt = V^F(x),$$

where $V^F(x)$ is the optimal cost for the fluid problem with initial state x.

Tracking Policies

We define the policy σ^n by

$$Z_{11}^n(X;t) = N_1 \wedge X_1$$

$$Z_{21}^n(X;t) = \begin{cases} 0 & t \in [0, nT_2) \\ (N_1 - X_1)^+ \wedge (X_2 - (N_2 \wedge X_2)) & t \in [nT_2, nT_3) \\ 0 & t \in [nT_3, \infty) \end{cases}$$

$$Z_{22}^n(X;t) = \begin{cases} N_2 - (N_2 \wedge (X_1 - N_1)^+) & t \in [0, nT_1) \\ N_2 \wedge X_2 & t \in [nT_1, \infty) \end{cases}$$

$$Z_{12}^n(X;t) = \begin{cases} N_2 \wedge (X_1 - N_1)^+ & t \in [0, nT_1) \\ 0 & t \in [nT_1, \infty) \end{cases}$$

Tracking Policies

Theorem

The tracking policy is asymptotically optimal.

Tracking Policies

Previous example

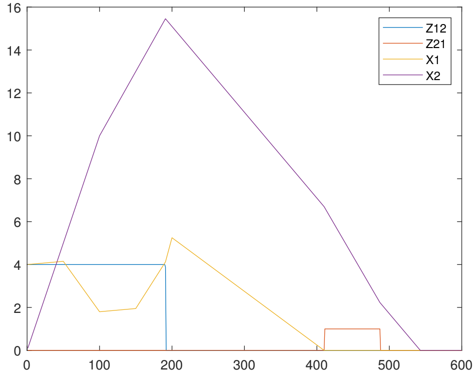


Figure: Fluid trajectory

Tracking Policies

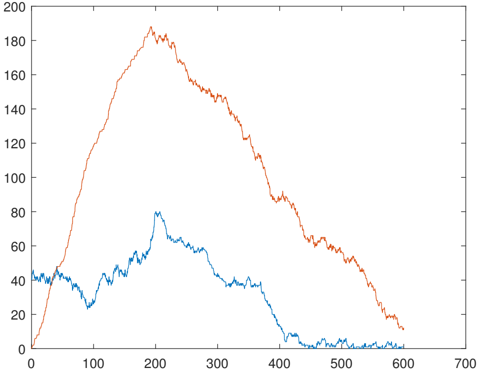


Figure: Stochastic sample path ($n = 10$)

Tracking Policies

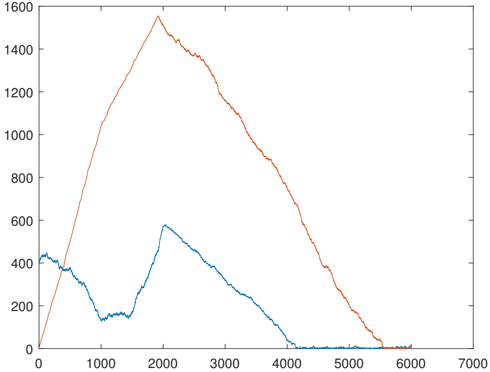


Figure: Stochastic sample path ($n = 100$)

Tracking Policies

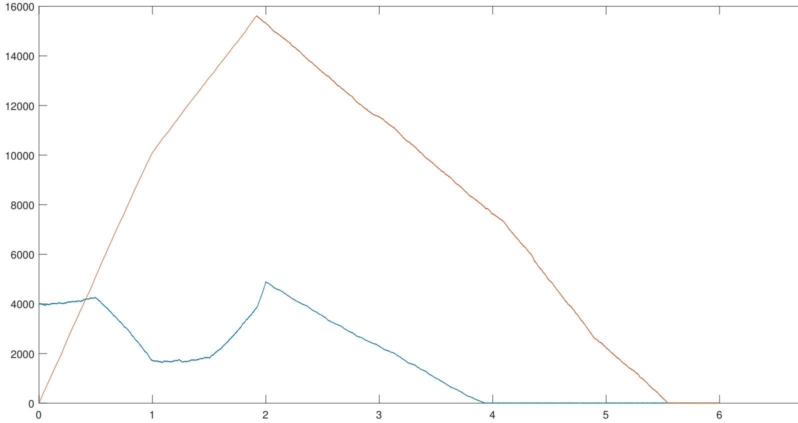


Figure: Stochastic sample path ($n = 1000$)

Ongoing Work

- Other adaptations of the fluid control policy to the pre-limit setting
- More general parallel server systems