Managing Resource Flexibility: Staffing and Scheduling

Jinsheng Chen
Columbia University

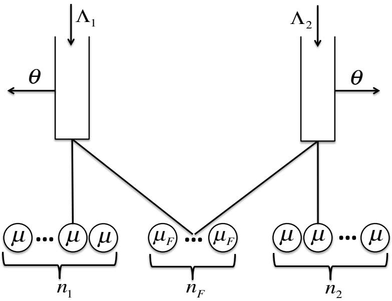
Monday 29th March, 2021

Motivation

- Many service systems involve multiple customer classes and server types
- Different servers have different skill sets
- Tradeoff between benefit (load-balancing) and cost (more expensive, inefficient) of resource flexibility
- Want to know how to staff and schedule
- Also consider random arrival rates

Model

- Queueing model with two customer classes
- Poisson arrivals of random intensity $\Lambda = (\Lambda_1, \Lambda_2)$
- Assume $\Lambda_i = p_i\lambda + \lambda^{\alpha_i} Y_i$, where $p_i > 0$, $\alpha_i \leq 1$ and $Y = (Y_1, Y_2)$ has zero mean and finite variance
- Exponential service with rates $\mu \geq \mu_F$
- Exponential abandonment with rate $\theta > 0$
- n_i dedicated servers for class i and n_F flexible servers



Objective

- Choose staffing levels n_1, n_2 and n_F and scheduling policy ν to minimize the total staffing, holding and abandonment cost

$$\begin{aligned}\Pi(n_1, n_2, n_F; \nu) :=& c(n_1 + n_2) + c_F n_F \\ &+ (h + a\theta)\mathbb{E}[Q_\Sigma(\infty; n_1, n_2, n_F; \nu)]\end{aligned}$$

- $\mathbb{E}[Q_\Sigma(\infty; n_1, n_2, n_F; \nu)] = \mathbb{E}[\mathbb{E}[Q_\Sigma(\infty; n_1, n_2, n_F; \nu)|\Lambda]]$
- Let Π^* be the optimal cost
- Assume $c/\mu < c_F/\mu_F < h/\theta + a$
- Scheduling policy ν maps headcounts X_i to assignments Z_{ij}:

$$\nu : (X_1, X_2) \mapsto Z = (Z_1, Z_2, Z_{F1}, Z_{F2})$$

The low uncertainty setting ($\alpha_i < 1/2$)

- Assume $\alpha_i < 1/2$
- Assume symmetry for tractability: $\Lambda_1 = \Lambda_2, p_i = 1$
- Have $n_1 = n_2 = n$
- Approach: derive optimal scheduling policy ν^* for any (n, n_F), then optimize over (n, n_F) using diffusion approximation for this fixed policy
- Use superscript λ for the λth system, e.g. $\Pi^{\lambda,*}, n_i^\lambda, n_F^\lambda$
- Let $R^\lambda = \lambda/\mu$

Optimal Scheduling Policy $\nu^{\lambda,*}$

Dedicated servers have priority:

$$Z_i^\lambda(t) = \min\{n^\lambda, X_i^\lambda(t)\} \text{ for } i = 1, 2;$$

Flexible servers prioritize the more congested class: if $X_1^\lambda(t) \geq X_2^\lambda(t)$,

$$Z_{F1}^\lambda(t) = \min\{n_F^\lambda, (X_1^\lambda(t) - n^\lambda)^+\}$$
$$Z_{F2}^\lambda(t) = \min\{n_F^\lambda - Z_{F1}^\lambda(t), (X_2^\lambda(t) - n^\lambda)^+\}$$

Similar if $X_1^\lambda(t) < X_2^\lambda(t)$.

Theorem

Suppose $\theta \leq \mu_F$. For any Markovian scheduling policy ν^λ,

$$\mathbb{E}[Q_\Sigma^\lambda(\infty; n^\lambda, n_F^\lambda; \nu^\lambda)] \geq \mathbb{E}[Q_\Sigma^\lambda(\infty; n^\lambda, n_F^\lambda; \nu^{\lambda,*})],$$

which implies that $\Pi^\lambda(n^\lambda, n_F^\lambda; \nu^\lambda) \geq \Pi^\lambda(n^\lambda, n_F^\lambda; \nu^{\lambda,*})$.

Optimal Scheduling Policy $\nu^{\lambda,*}$

- Proof is by coupling
- $\theta \leq \mu_F$ means that flexible servers prioritizing the more congested class is load-balancing
- Can define scheduling policy $\phi^{\lambda,*}$ with reverse priority: if $X_1^\lambda(t) \leq X_2^\lambda(t)$,

$$Z_{F1}^\lambda(t) = \min\{n_F^\lambda, (X_1^\lambda(t) - n^\lambda)^+\}$$
$$Z_{F2}^\lambda(t) = \min\{n_F^\lambda - Z_{F1}^\lambda(t), (X_2^\lambda(t) - n^\lambda)^+\}$$

Theorem

Suppose $\theta \geq \mu = \mu_F$. For any deterministic Markovian scheduling policy ν^λ,

$$\mathbb{E}[Q_\Sigma^\lambda(\infty; n^\lambda, n_F^\lambda; \nu^\lambda)] \geq \mathbb{E}[Q_\Sigma^\lambda(\infty; n^\lambda, n_F^\lambda; \phi^{\lambda,*})],$$

which implies that $\Pi^\lambda(n^\lambda, n_F^\lambda; \nu^\lambda) \geq \Pi^\lambda(n^\lambda, n_F^\lambda; \phi^{\lambda,*})$.

Optimal Staffing

Can focus on case where $n^\lambda = R^\lambda + O(\sqrt{\lambda})$ and $n_F^\lambda = O(\sqrt{\lambda})$:

Lemma

We have $\Pi^{\lambda,*} = 2cR^\lambda + O(\sqrt{\lambda})$*. Moreover, for* $(n^{\lambda,*}, n_F^{\lambda,*})$*,*

$$-\infty < \liminf_{\lambda\to\infty} \frac{n^{\lambda,*} - R^\lambda}{\sqrt{\lambda}} \le \limsup_{\lambda\to\infty} \frac{n^{\lambda,*} - R^\lambda}{\sqrt{\lambda}} < \infty$$

and

$$\limsup_{\lambda\to\infty} \frac{n_F^{\lambda,*}}{\sqrt{\lambda}} < \infty.$$

Optimal Staffing

Let $\hat{X}_i^\lambda(\cdot) = \frac{X_i^\lambda(\cdot) - n^\lambda}{\sqrt{\lambda}}$, $\hat{X}^\lambda = (\hat{X}_1^\lambda, \hat{X}_2^\lambda)$ and $\hat{Q}_\Sigma^\lambda = \frac{Q_\Sigma^\lambda}{\sqrt{\lambda}}$.

Theorem

Suppose $n^\lambda = R^\lambda + \beta\sqrt{R^\lambda} + o(\sqrt{R^\lambda})$ *and* $n_F^\lambda = \beta_F\sqrt{R^\lambda} + o(\sqrt{R^\lambda})$*, where* $\beta \in \mathbb{R}, \beta_F \geq 0$*, and if* $\theta = 0$*,* $2\beta\mu + \beta_F\mu_F > 0$*. Then, if* $\hat{X}^\lambda(0) \Rightarrow \hat{X}(0)$ *as* $\lambda \to \infty$*,*

$$\hat{X}^\lambda \Rightarrow \hat{X} \text{ in } D^2 \text{ as } \lambda \to \infty,$$

where $\hat{X}$ *is a two-dimensional diffusion process. Moreover,*

$$\mathbb{E}[\hat{Q}_\Sigma^\lambda(\infty)] \to \mathbb{E}[(\hat{X}_1(\infty)^+ + \hat{X}_2(\infty)^+ - \beta_F/\sqrt{\mu})^+] \text{ as } \lambda \to \infty.$$

Optimal Staffing

Get the approximate diffusion problem

$$\min_{(\beta,\beta_F)} \hat{V}_p(\beta,\beta_F) := 2c\beta/\sqrt{\mu} + c_F\beta_F/\sqrt{\mu}$$
$$+ (h+a\theta)\mathbb{E}\left[\left(\hat{X}_1(\infty;\beta,\beta_F)^+ + \hat{X}_2(\infty;\beta,\beta_F)^+ - \beta_F/\sqrt{\mu}\right)^+\right]$$

Theorem

For $\theta \le \mu_F \le \mu$, assuming $\arg\min_{(\beta,\beta_F)} \hat{V}_p(\beta,\beta_F)$ is finite, a sequence of staffing policies (n^λ, n_F^λ) is $o(\sqrt{\lambda})$-optimal if and only if the following two conditions hold:

1. $n^\lambda = R^\lambda + \beta^\lambda\sqrt{R^\lambda} + o(\sqrt{R^\lambda})$
2. $n_F^\lambda = \beta_F^\lambda\sqrt{R^\lambda} + o(\sqrt{R^\lambda})$

where $(\beta^\lambda, \beta_F^\lambda) \in \arg\min_{(\beta,\beta_F)} \hat{V}_p(\beta,\beta_F)$.

Sensitivity

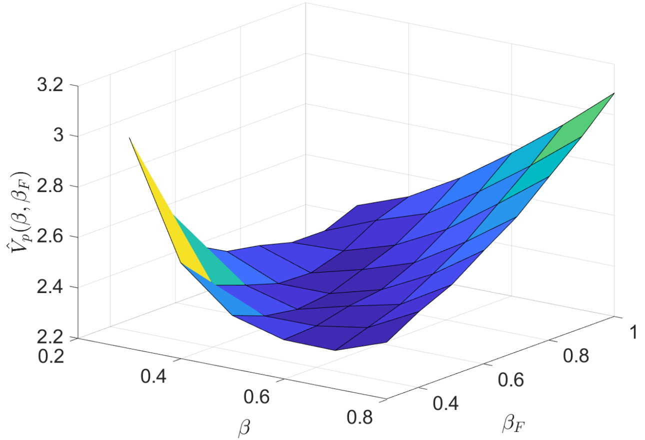


Figure: $\hat{V}_p(\beta, \beta_F)$ as a function of β and β_F.
$(\mu = 1, \mu_F = 0.85, \theta = 0, c = 1, c_F = 1.4, h = 1)$

Sensitivity

Set $h = c = 1, \mu = 1, \theta = 0$

c_F	1	1.2	1.4	1.6	1.8
β^*	-0.2	0.2	0.5	0.9	0.9
β_F^*	1.9	1.1	0.5	0	0

Table: Sensitivity of (β^*, β_F^*) with respect to c_F when $\mu_F = 0.85$

μ_F	0.55	0.65	0.75	0.85	0.95
β^*	0.8	0.8	0.7	0.5	0.4
β_F^*	0	0.1	0.2	0.5	0.6

Table: Sensitivity of (β^*, β_F^*) with respect to μ_F when $c_F = 1.4$

Numerical Illustration

c_F	$(\hat{n}^\lambda, \hat{n}_F^\lambda)$	$(n^{\lambda,*}, n_F^{\lambda,*})$	$\Pi^{\lambda,*}$	Gap
	$\lambda = 25$			
1	(27,10)	(26,11)	65.91	0.17
1.2	(28,7)	(28,7)	67.76	0
1.4	(29,5)	(30,4)	69.12	0.05
	$\lambda = 100$			
1	(103,20)	(102,22)	230.94	0.08
1.2	(106,15)	(106,15)	234.79	0
1.4	(108,11)	(108,10)	237.27	0.19
	$\lambda = 400$			
1	(406,40)	(405,42)	861.42	0.16
1.2	(412,30)	(413,27)	868.71	0.25
1.4	(416,22)	(416,21)	873.85	0.01

Table: Performance of $(\hat{n}^\lambda, \hat{n}_F^\lambda)$ for systems with different scales, λ's. $(\mu = 1, \mu_F = 0.85, \theta = 0, h = 8, c = 1)$

The high uncertainty setting ($\alpha_i > 1/2$)

- Approach follows Harrison and Zeevi (2005)
- The rate of customer abandonment can be expressed as $\theta \mathbb{E}[Q_\Sigma(\infty; n_1, n_2, n_F; \nu)]$.
- By rate conservation, the rate of customer abandonment can also be approximated by $\mathbb{E}\left[((\Lambda_1 - n_1\mu)^+ + (\Lambda_2 - n_2\mu)^+ - n_F\mu_F)^+\right]$
- This suggests the stochastic-fluid optimization problem:

$$\min_{\tilde{n}_1 \geq 0, \tilde{n}_2 \geq 0, \tilde{n}_F \geq 0} \tilde{\Pi}(\tilde{n}_1, \tilde{n}_2, \tilde{n}_F) := c(\tilde{n}_1 + \tilde{n}_2) + c_F \tilde{n}_F$$
$$+ (h/\theta + a)\mathbb{E}\left[((\Lambda_1 - \tilde{n}_1\mu)^+ + (\Lambda_2 - \tilde{n}_2\mu)^+ - \tilde{n}_F\mu_F)^+\right]$$

The stochastic-fluid optimization problem solution

- Let $c_P := h/\theta + a$ and let q_i solve $\mathbb{P}(Y_i > q_i) = \frac{c}{c_P \mu}$.
- If $\mathbb{P}(Y_1 > q_1$ or $Y_2 > q_2) > \frac{c_F}{c_P \mu_F}$, let $r_1, r_2 \in \mathbb{R}$, and $r_F > 0$ solve:

$$\mathbb{P}(Y_1 > r_1, Y_1 - r_1 + (Y_2 - r_2)^+ > r_F) = \frac{c}{c_P \mu},$$
$$\mathbb{P}((Y_1 - r_1)^+ + (Y_2 - r_2)^+ > r_F) = \frac{c_F}{c_P \mu_F}.$$

Lemma

Suppose $\alpha_1 = \alpha_2 = \alpha$.
If $\mathbb{P}(Y_1 > q_1$ *or* $Y_2 > q_2) \le \frac{c_F}{c_P \mu_F}$, $\tilde{n}_i^* = (p_i \lambda + q_i \lambda^\alpha)/\mu$ *for* $i = 1, 2$, *and* $\tilde{n}_F^* = 0$.
If $\mathbb{P}(Y_1 > q_1$ *or* $Y_2 > q_2) > \frac{c_F}{c_P \mu_F}$, $\tilde{n}_i^* = (p_i \lambda + r_i \lambda^\alpha)/\mu$ *for* $i = 1, 2$, *and* $\tilde{n}_F^* = r_F \lambda^\alpha / \mu_F$.

The scheduling policy $\tilde{\nu}$

- Given a realization of the arrival rate $\Lambda = \gamma := (\gamma_1, \gamma_2)$, let $\delta(\gamma) \in [0,1]$ solve

$$\begin{aligned} & ((\gamma_1 - n_1\mu)^+ + (\gamma_2 - n_2\mu)^+ - n_F\mu_F)^+ \\ = \; & (\gamma_1 - n_1\mu - \delta n_F\mu_F)^+ + (\gamma_2 - n_2\mu - (1-\delta)n_F\mu_F)^+. \end{aligned}$$

- Under $\tilde{\nu}$, we allocate $\lfloor \delta(\gamma) n_F \rfloor$ flexible servers to class 1 and the remaining $\lceil (1-\delta(\gamma)) n_F \rceil$ flexible servers to class 2
- Choice of δ minimizes total approximate abandonment rate
- Dedicated servers are prioritized over the flexible servers.
- Upon each realization of the arrival rates $\Lambda = \gamma$, the policy $\tilde{\nu}$ turns the M-model into two independent inverted-V models that follow the fastest-server-first policy.

Two other scheduling policies

- $\tilde{\nu}_R$: same as $\tilde{\nu}$, but follow slowest-server-first policy for each inverted-V model
- $\tilde{\nu}_I$: same as $\tilde{\nu}$, but flexible servers now give priority to their assigned class (instead of only serving that class)
- Have
  $\Pi(n_1, n_2, n_F; \tilde{\nu}_I) \leq \Pi(n_1, n_2, n_F; \tilde{\nu}) \leq \Pi(n_1, n_2, n_F; \tilde{\nu}_R)$

Asymptotic Optimality

Lemma

For any scheduling policy ν, $\tilde{\Pi}(n_1, n_2, n_F) \leq \Pi(n_1, n_2, n_F; \nu)$.

Theorem

Assume $\alpha_1 \geq \alpha_2 > 1/2$. *For* $\nu^\lambda \in \{\tilde{\nu}^\lambda, \tilde{\nu}_R^\lambda, \tilde{\nu}_I^\lambda\}$,

$$\Pi(\lceil \tilde{n}_1^{\lambda,*} \rceil, \lceil \tilde{n}_2^{\lambda,*} \rceil, \lfloor \tilde{n}_F^{\lambda,*} \rfloor; \nu^\lambda) = \Pi^{\lambda,*} + O(\lambda^{1-\alpha_2}).$$

Sensitivity

Y_1, Y_2 standard bivariate normal with correlation ρ, $\alpha_1 = \alpha_2$

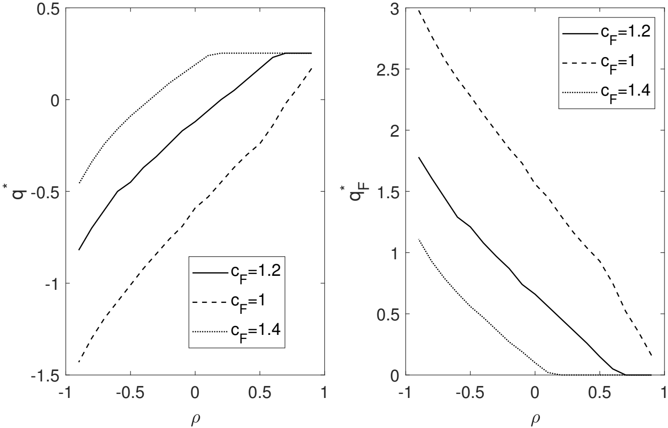


Figure: How q^* and q_F^* vary with ρ when $\mu_F = 1.2$ and $c_F \in \{1, 1.2, 1.4\}$

Sensitivity

Y_1, Y_2 standard bivariate normal with correlation ρ, $\alpha_1 = \alpha_2$

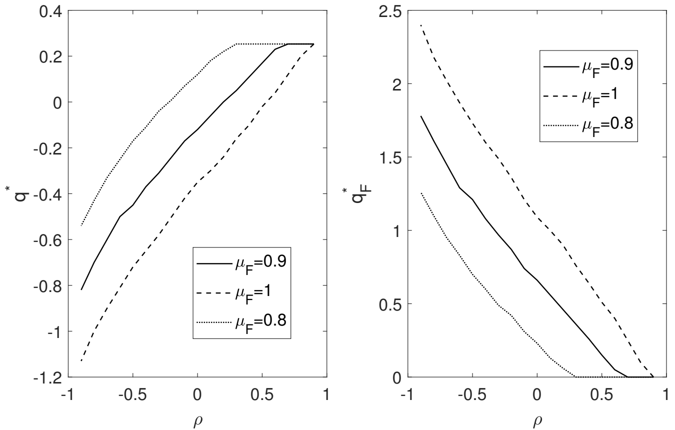


Figure: How q^* and q_F^* vary with ρ when $c_F = 1.2$ and $\mu_F \in \{0.8, 0.9, 1\}$

Numerical Illustration

Approximate gap is AG $= \Pi(\lceil \tilde{n}_1^{\lambda,*} \rceil, \lceil \tilde{n}_2^{\lambda,*} \rceil, \lfloor \tilde{n}_F^{\lambda,*} \rfloor; \tilde{\nu}^{\lambda}) - \tilde{\Pi}^{\lambda,*}$

	$\lambda = 25$		$\lambda = 50$		$\lambda = 100$		$\lambda = 200$	
α	$\tilde{\Pi}^{\lambda,*}$	AG	$\tilde{\Pi}^{\lambda,*}$	AG	$\tilde{\Pi}^{\lambda,*}$	AG	$\tilde{\Pi}^{\lambda,*}$	AG
0.6	78.4	9.6	143.0	12.2	265.2	14.7	498.9	18.6
0.8	104.1	5.7	194.1	6.7	363.9	7.2	685.3	7.9
1	152.9	4.4	305.8	4.6	611.7	4.5	1223.3	4.2

Table: Performance of $(\lceil \tilde{n}_1^{\lambda,*} \rceil, \lceil \tilde{n}_2^{\lambda,*} \rceil, \lfloor \tilde{n}_F^{\lambda,*} \rfloor; \tilde{\nu}^{\lambda})$ for systems with different values of λ and α.
$(c = 1, c_F = 1.2, h = a = 8, \mu = 1, \mu_F = 0.9, \theta = 0.5, \rho = 0.5)$

Numerical Illustration

	$\lambda = 25$			$\lambda = 50$		
α	$\tilde{\nu}_I$	$\tilde{\nu}$	$\tilde{\nu}_R$	$\tilde{\nu}_I$	$\tilde{\nu}$	$\tilde{\nu}_R$
0.6	86.3	88.0	88.3	153.0	155.2	155.5
0.8	108.3	109.8	110.0	199.5	200.8	201.0
1	156.2	157.3	157.5	309.6	310.4	310.6
	$\lambda = 100$			$\lambda = 200$		
α	$\tilde{\nu}_I$	$\tilde{\nu}$	$\tilde{\nu}_R$	$\tilde{\nu}_I$	$\tilde{\nu}$	$\tilde{\nu}_R$
0.6	276.8	279.9	280.3	513.6	517.5	518.1
0.8	369.2	371.1	371.4	691.2	693.2	693.5
1	614.3	616.2	616.4	1226.6	1227.5	1227.7

Table: The cost under scheduling policies $\nu \in \{\tilde{\nu}_I, \tilde{\nu}, \tilde{\nu}_R\}$ for different values of λ and α. ($c = 1, c_F = 1.2, h = a = 8, \mu = 1, \mu_F = 0.9$, $\theta = 0.5$, $\rho = 0.5$)

Summary

- Exactly optimal scheduling policy for symmetric M model via coupling construction
- Exactly optimal non-standard scheduling policy under high abandonment rates via coupling construction and 'dual approach'
- Diffusion limit of M model when there is only partial resource pooling
- Establish that sizing of flexible pool should match degree of uncertainty
- Establish near-optimality of stochastic-fluid approximation for M model under random demand, and sufficiency of simple scheduling policies

Future Work

- Asymmetric systems under low demand uncertainty
- The intermediate $\alpha_i = 1/2$ case
- More general systems