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 $\nu$ to minimize the total staffing, holding and abandonment cost

$$\Pi(n_1, n_2, n_F; \nu) := c(n_1 + n_2) + c_F n_F$$

$$+ (h + a\theta) E[Q_\Sigma(\infty; n_1, n_2, n_F; \nu)]$$

• $E[Q_\Sigma(\infty; n_1, n_2, n_F; \nu)] = E[E[Q_\Sigma(\infty; n_1, n_2, n_F; \nu) | \Lambda]]$

• Let $\Pi^*$ be the optimal cost

• Assume $c/\mu < c_F/\mu_F < h/\theta + a$

• Scheduling policy $\nu$ 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 $\nu^*$ for any $(n, n_F)$, then optimize over $(n, n_F)$ using diffusion approximation for this fixed policy
- Use superscript $\lambda$ for the $\lambda$th 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:

$$Z^\lambda_i(t) = \min \{n^\lambda, X^\lambda_i(t)\} \text{ for } i = 1, 2;$$

Flexible servers prioritize the more congested class: if $X^\lambda_1(t) \geq X^\lambda_2(t)$,

$$Z^\lambda_{F1}(t) = \min \{n^\lambda_F, (X^\lambda_1(t) - n^\lambda)^+\}$$
$$Z^\lambda_{F2}(t) = \min \{n^\lambda_F - Z^\lambda_{F1}(t), (X^\lambda_2(t) - n^\lambda)^+\}$$

Similar if $X^\lambda_1(t) < X^\lambda_2(t)$.

Theorem
Suppose $\theta \leq \mu_F$. For any Markovian scheduling policy $\nu^\lambda$,

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

which implies that $\Pi^\lambda(n^\lambda, n^\lambda_F; \nu^\lambda) \geq \Pi^\lambda(n^\lambda, n^\lambda_F; \nu^{\lambda,*})$. 
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 $X^\lambda_1(t) \leq X^\lambda_2(t)$,

$$Z^\lambda_{F1}(t) = \min\{n^\lambda_F, (X^\lambda_1(t) - n^\lambda)^+\}$$
$$Z^\lambda_{F2}(t) = \min\{n^\lambda - Z^\lambda_{F1}(t), (X^\lambda_2(t) - n^\lambda)^+\}$$

**Theorem**

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

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

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

\[-\infty < \liminf_{\lambda \to \infty} \frac{n^\lambda,* - R^\lambda}{\sqrt{\lambda}} \leq \limsup_{\lambda \to \infty} \frac{n^\lambda,* - R^\lambda}{\sqrt{\lambda}} < \infty \]

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 \(\theta \leq \mu_F \leq \mu\), assuming \(\arg \min_{(\beta, \beta_F)} \hat{V}_p(\beta, \beta_F)\) is finite, a sequence of staffing policies \((n^\lambda, n_F^\lambda)\) 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 \((\beta^\lambda, \beta_F^\lambda) \in \arg \min_{(\beta, \beta_F)} \hat{V}_p(\beta, \beta_F)\).
Figure: $\hat{V}_p(\beta, \beta_F)$ as a function of $\beta$ and $\beta_F$.
($\mu = 1, \mu_F = 0.85, \theta = 0, c = 1, c_F = 1.4, h = 1$)
Sensitivity

Set \( h = c = 1, \mu = 1, \theta = 0 \)

<table>
<thead>
<tr>
<th>( c_F )</th>
<th>1</th>
<th>1.2</th>
<th>1.4</th>
<th>1.6</th>
<th>1.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta^* )</td>
<td>-0.2</td>
<td>0.2</td>
<td>0.5</td>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td>( \beta^*_{F} )</td>
<td>1.9</td>
<td>1.1</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table: Sensitivity of \((\beta^*, \beta^*_{F})\) with respect to \(c_F\) when \(\mu_F = 0.85\)

<table>
<thead>
<tr>
<th>( \mu_F )</th>
<th>0.55</th>
<th>0.65</th>
<th>0.75</th>
<th>0.85</th>
<th>0.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta^* )</td>
<td>0.8</td>
<td>0.8</td>
<td>0.7</td>
<td>0.5</td>
<td>0.4</td>
</tr>
<tr>
<td>( \beta^*_{F} )</td>
<td>0</td>
<td>0.1</td>
<td>0.2</td>
<td>0.5</td>
<td>0.6</td>
</tr>
</tbody>
</table>

Table: Sensitivity of \((\beta^*, \beta^*_{F})\) with respect to \(\mu_F\) when \(c_F = 1.4\)
### Numerical Illustration

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

**Table:** Performance of $(\hat{n}^\lambda, \hat{n}_F^\lambda)$ for systems with different scales, $\lambda$'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
  $\mathbb{E} [((\Lambda_1 - n_1 \mu)^+ + (\Lambda_2 - n_2 \mu)^+ - n_F \mu_F)^+]$
- This suggests the stochastic-fluid optimization problem:

$$
\min_{\tilde{n}_1 \geq 0, \tilde{n}_2 \geq 0, \tilde{n}_F \geq 0} \tilde{\Pi}(\tilde{n}_1, \tilde{n}_2, \tilde{n}_F) := c(\tilde{n}_1 + \tilde{n}_2) + c_F \tilde{n}_F + (h/\theta + a) \mathbb{E} [((\Lambda_1 - \tilde{n}_1 \mu)^+ + (\Lambda_2 - \tilde{n}_2 \mu)^+ - \tilde{n}_F \mu_F)^+]
$$
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 $\mathbb{P}(Y_1 > q_1 \text{ or } Y_2 > q_2) > \frac{c_F}{c_P\mu_F}$, 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* $\mathbb{P}(Y_1 > q_1 \text{ or } Y_2 > q_2) \leq \frac{c_F}{c_P\mu_F}$, $\tilde{n}_i^* = (p_i \lambda + q_i \lambda^\alpha)/\mu$ *for* $i = 1, 2$, *and* $\tilde{n}_F^* = 0$.

*If* $\mathbb{P}(Y_1 > q_1 \text{ or } Y_2 > q_2) > \frac{c_F}{c_P\mu_F}$, $\tilde{n}_i^* = (p_i \lambda + r_i \lambda^\alpha)/\mu$ *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

$$
\left((\gamma_1 - n_1 \mu)^+ + (\gamma_2 - n_2 \mu)^+ - n_F \mu_F\right)^+ = (\gamma_1 - n_1 \mu - \delta n_F \mu_F)^+ + (\gamma_2 - n_2 \mu - (1 - \delta) n_F \mu_F)^+.
$$

- 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 $\delta$ 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 $\nu$, $\tilde{\Pi}(n_1, n_2, n_F) \leq \Pi(n_1, n_2, n_F; \nu)$.

**Theorem**

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

$$\Pi([\tilde{n}_1^\lambda,^*], [\tilde{n}_2^\lambda,^*], [\tilde{n}_F^\lambda,^*]; \nu^\lambda) = \Pi^\lambda,^* + O(\lambda^{1-\alpha_2}).$$
Sensitivity

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

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

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

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

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

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>$\lambda = 25$</th>
<th>$\lambda = 50$</th>
<th>$\lambda = 100$</th>
<th>$\lambda = 200$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6</td>
<td>78.4</td>
<td>143.0</td>
<td>265.2</td>
<td>498.9</td>
</tr>
<tr>
<td>0.8</td>
<td>104.1</td>
<td>194.1</td>
<td>363.9</td>
<td>685.3</td>
</tr>
<tr>
<td>1</td>
<td>152.9</td>
<td>305.8</td>
<td>611.7</td>
<td>1223.3</td>
</tr>
</tbody>
</table>

Table: Performance of $([\tilde{n}_1^\lambda,*], [\tilde{n}_2^\lambda,*], [\tilde{n}_F^\lambda,*]; \tilde{\nu}^\lambda)$ for systems with different values of $\lambda$ and $\alpha$.

($c = 1$, $c_F = 1.2$, $h = a = 8$, $\mu = 1$, $\mu_F = 0.9$, $\theta = 0.5$, $\rho = 0.5$)
### Numerical Illustration

<table>
<thead>
<tr>
<th></th>
<th>$\lambda = 25$</th>
<th>$\lambda = 50$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>$\tilde{\nu}_I$</td>
<td>$\tilde{\nu}$</td>
</tr>
<tr>
<td>0.6</td>
<td>86.3</td>
<td>88.0</td>
</tr>
<tr>
<td>0.8</td>
<td>108.3</td>
<td>109.8</td>
</tr>
<tr>
<td>1</td>
<td>156.2</td>
<td>157.3</td>
</tr>
<tr>
<td></td>
<td>$\lambda = 100$</td>
<td></td>
</tr>
<tr>
<td>$\lambda = 100$</td>
<td>$\tilde{\nu}_I$</td>
<td>$\tilde{\nu}$</td>
</tr>
<tr>
<td>0.6</td>
<td>276.8</td>
<td>279.9</td>
</tr>
<tr>
<td>0.8</td>
<td>369.2</td>
<td>371.1</td>
</tr>
<tr>
<td>1</td>
<td>614.3</td>
<td>616.2</td>
</tr>
<tr>
<td></td>
<td>$\lambda = 200$</td>
<td></td>
</tr>
</tbody>
</table>

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