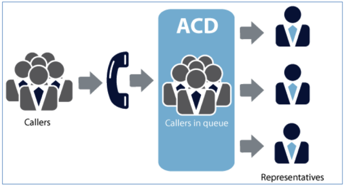
Staffing and Scheduling Multi-class Service Systems with Flexible Servers

Jinsheng Chen
Columbia University
(Joint work with Jing Dong)

Monday 30th September, 2019

Motivation

- Many service systems involve multiple customer classes:
  - Hospitals with multiple types of wards, e.g. cardiology and neurology
  - Call centers with customers that require service in different languages
- Servers can be dedicated (monolingual) or flexible (multilingual)

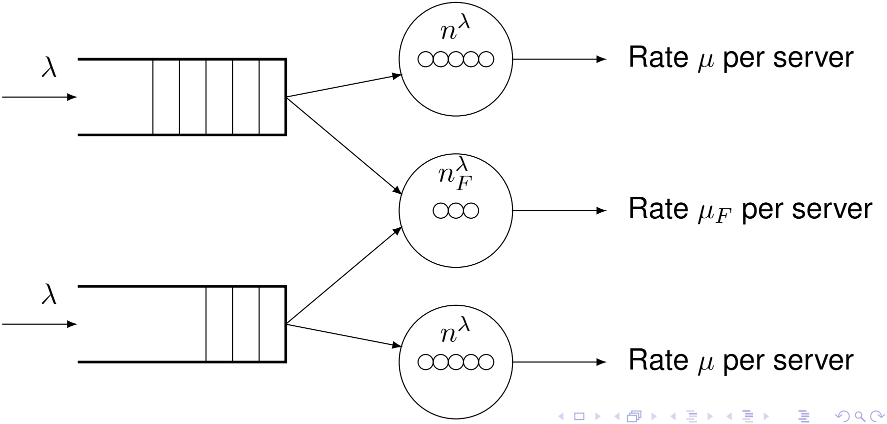


Motivation

- How to match customers with available servers?
- How much server flexibility is optimal?
- Advantages
  - Flexible servers can continue to work if one customer class has no queue
  - Helps buffer the system against random imbalances
- Disadvantages
  - Multi-skilled servers may be costly or even infeasible (e.g. many languages) to train
  - Flexible servers may also be less efficient than dedicated servers

Model

- Symmetric queueing model with two customer classes
- Poisson arrivals of rate λ
- Exponential service with rates $\mu \geq \mu_F$
- n^λ dedicated servers per class and n_F^λ flexible servers



Scheduling Policy

- Markovian scheduling policy ν^λ
- Policy choice may depend on staffing levels n^λ and n_F^λ
- The policy ν^λ specifies how to assign servers to customers as a function of the total-in-system state $(X_1^\lambda(t), X_2^\lambda(t))$

Objective

- Choose staffing levels n^λ and n_F^λ and scheduling policy ν^λ to minimize the total staffing and holding cost

$$\Pi_\lambda^{\nu^\lambda}(n^\lambda, n_F^\lambda; \nu^\lambda) := 2c^\lambda n^\lambda + c_F^\lambda n_F^\lambda + \gamma E[Q_\Sigma^\lambda(\infty; n^\lambda, n_F^\lambda; \nu^\lambda)]$$

- Let Π_λ^* be the optimal cost
- Costs c^λ, c_F^λ can vary with λ to model e.g. economies of scale
- Assume $\Delta^\lambda := c_F^\lambda - c^\lambda \geq 0$

A heavy-traffic asymptotic approach

- Exact analysis very difficult
- Numerical solution not very insightful, e.g. does not reveal how optimal n^λ and n_F^λ vary with the parameters
- Thus we use a heavy-traffic asymptotic approach and let $\lambda \to \infty$
- Assume limits $c = \lim_{\lambda\to\infty} c^\lambda > 0$ and $c_F = \lim_{\lambda\to\infty} c_F^\lambda$ exist
- Hence $\Delta = \lim_{\lambda\to\infty} \Delta^\lambda \geq 0$

Some related literature

Halfin-Whitt Regime

- Halfin and Whitt (1984)
- Puhalskii and Reiman (2000)
- Garnett et al. (2002)

Scheduling

- Harrison and Zeevi (2004), Atar et al. (2004)
- Armony (2005), Gurvich and Whitt (2008, 2009), Dai and Tezcan (2008, 2010)

Staffing

- Borst et al. (2004)
- Wallace and Whitt (2005)
- Bassamboo et al. (2012)

Staffing and scheduling

- Armony and Mandelbaum (2011)

Optimal Scheduling Policy

Consider a ‘maximum pressure’ scheduling policy:

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

and for the flexible pool of servers, if $X_1^\lambda(t) \geq X_2^\lambda(t)$,

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

otherwise,

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

Z_i^λ (Z_{Fi}^λ) is the number of dedicated (flexible) servers serving class i customers

Optimal Scheduling Policy

Key points of the policy:

- Dedicated servers have priority over flexible servers
- Flexible servers prioritize more congested customer class

Theorem

The maximum pressure scheduling policy MP is optimal. That is, for any preemptive deterministic Markovian scheduling policy M, we have

$$\Pi_\lambda^{MP}(n^\lambda, n_F^\lambda) \leq \Pi_\lambda^{M}(n^\lambda, n_F^\lambda).$$

Policy is optimal in the pre-limit (i.e. not just as asymptotically). So, we can fix this policy for the rest of the talk.

Asymptotic Optimality

Let $R^\lambda = \lambda/\mu$ be the 'minimum staffing level'.

Lemma

There exist constants $0 < K_l < K_u$ *such that*

$$2c^\lambda R^\lambda + K_l\sqrt{\lambda} + o(\sqrt{\lambda}) < \Pi_\lambda^* < 2c^\lambda R^\lambda + K_u\sqrt{\lambda} + o(\sqrt{\lambda})$$

Definition

A sequence of staffing policies (n^λ, n_F^λ) is asymptotically optimal if

$$\limsup_{\lambda\to\infty} \frac{\Pi_\lambda(n^\lambda, n_F^\lambda) - 2c^\lambda R^\lambda}{\Pi_\lambda^* - 2c^\lambda R^\lambda} = 1$$

Two Regimes

1. 'Complete Resource Pooling' – $\Delta = 0, \mu_F = \mu$
   - Optimal number of flexible servers is of order strictly larger than $\sqrt{\lambda}$
   - System exhibits 'state-space collapse' where dedicated servers are essentially always busy
   - System behaves as if all servers are flexible
2. 'Partial Resource Pooling' – $\Delta > 0$ or $\mu_F < \mu$
   - Optimal number of flexible servers is of order exactly $\sqrt{\lambda}$
   - No state-space collapse and system behaves approximately as a two-dimensional diffusion process

Complete Resource Pooling Regime

Let $\alpha^* = \arg\min_{\alpha>0}\{c\alpha + \gamma h(-\alpha)/(\alpha h(-\alpha) + \alpha^2)\}$.

Theorem

Suppose $\Delta = 0$ *and* $\mu_F = \mu$. *A sequence of staffing policies* (n^λ, n_F^λ) *is asymptotically optimal if and only if*

1. $2n^\lambda + n_F^\lambda = 2R^\lambda + \alpha^*\sqrt{2R^\lambda} + o(\sqrt{R^\lambda})$
2. $\liminf_{\lambda\to\infty} \frac{n_F^\lambda}{\sqrt{\lambda}} = \infty$
3. $\limsup_{\lambda\to\infty} \frac{n_F^\lambda \Delta^\lambda}{\sqrt{\lambda}} = 0$

Complete Resource Pooling Regime (Explanation)

- Dedicated servers are essentially always busy and only flexible servers can become idle
- The scaled number-of-customers-in-system process behaves asymptotically as the one-dimensional diffusion process

$$d\hat{X}_c(t) = (-\alpha\mu + \mu\hat{X}_c(t)^-)\,dt + \sqrt{2\mu}\,dB(t).$$

- α^* is the solution of $\min_{\alpha>0} c\alpha + \gamma E[\hat{X}_c(\infty;\alpha)^+]$

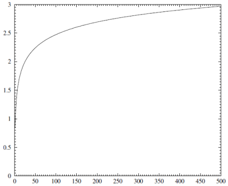
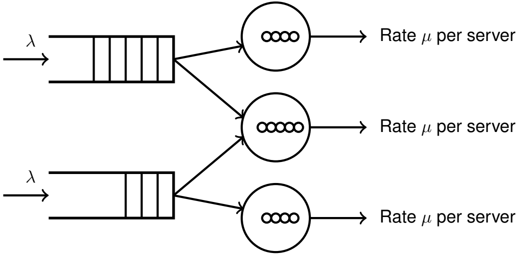
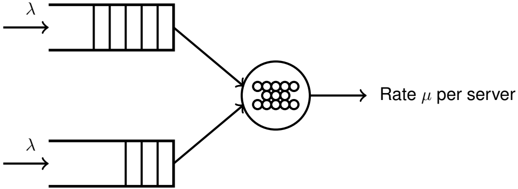


Figure: α^* as a function of γ/c (Borst et al. 2004)

Complete Resource Pooling Regime (Explanation)



Above system performs (almost) as well as the below system even though only a fraction of servers are flexible



Partial Resource Pooling Regime

Let β^* and β_F^* denote the solution of

$$\min_{\beta,\beta_F} 2c\beta + c_F\beta_F + \gamma E[k_{\beta_F}(\hat{X}_p(\infty;\beta,\beta_F))]$$

Theorem

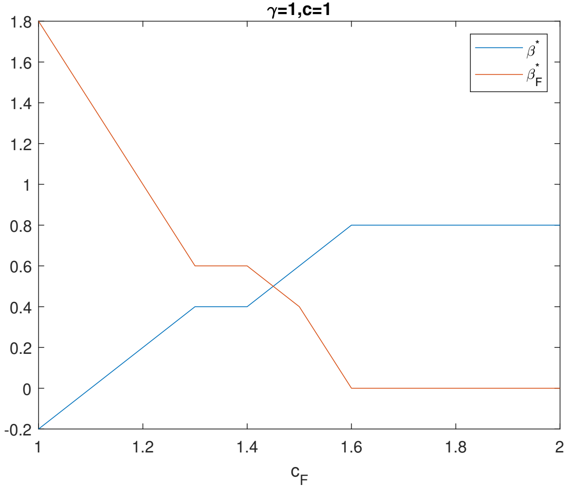
Suppose $\Delta > 0$ or $\mu_F < \mu$. A sequence of staffing policies (n^λ, n_F^λ) is asymptotically optimal if and only if

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

$\sqrt{R^\lambda}$ order of flexible servers is not sufficient to achieve complete resource pooling

Partial Resource Pooling Regime

$\mu = 1, \mu_F = 0.9$



Partial Resource Pooling Regime (Explanation)

- The scaled number-in-system process is asymptotically a two-dimensional diffusion $\hat{X}_p(t)$:

$$d\hat{X}_{pi}(t) = \Big(-\beta\mu + \mu\hat{X}_{pi}(t)^- - \mu_F g_i(\hat{X}_{p1}(t), \hat{X}_{p2}(t))\Big)\, dt + \sqrt{2\mu}\, dB_i(t),$$

for $i = 1, 2$, where

$$g_1(x_1, x_2) = \begin{cases} x_1^+ \wedge \beta_F & \text{if } x_1 \geq x_2 \\ x_1^+ \wedge \left(\beta_F - x_2^+\right)^+ & \text{if } x_1 < x_2 \end{cases}$$

and

$$g_2(x_1, x_2) = \begin{cases} x_2^+ \wedge \left(\beta_F - x_1^+\right)^+ & \text{if } x_1 \geq x_2 \\ x_2^+ \wedge \beta_F & \text{if } x_1 < x_2 \end{cases}$$

- $k_{\beta_F}(x, y) = (x^+ + y^+ - \beta_F)^+$ gives the queue length

Summary

- When conditions favor increased flexibility ($\Delta = 0, \mu_F = \mu$), optimal flexible pool size is of order greater than $\sqrt{\lambda}$
    - 'State-space collapse' with dedicated servers always busy
    - System achieves 'complete resource pooling' and performs as if all servers were flexible
- Otherwise, optimal flexible pool size is of order exactly $\sqrt{\lambda}$
    - Only 'partial resource pooling' is attained
- In each regime, staffing levels together with MP policy give asymptotically optimal joint staffing and scheduling policies

Other Scheduling Policies (CRP)

- CRP is attained under any non-idling policy that prioritizes dedicated servers
- So, can consider other scheduling policies
- Consider queue-ratio scheduling: for $r_1 \in [0,1]$ and $r_2 = 1 - r_1$, servers serve i such that $Q_i^\lambda(t) - r_i Q_\Sigma^\lambda(t)$ is maximum

Theorem

Suppose $\hat{Q}_1^\lambda(0) - r_1 \hat{Q}_\Sigma^\lambda(0) \to 0$ as $\lambda \to \infty$. Then, under queue ratio scheduling, we have for $i = 1, 2$ that $\hat{Q}_i^\lambda - r_i \hat{Q}_\Sigma^\lambda \Rightarrow 0$ in D as $\lambda \to \infty$.

So, have $(\hat{Q}_1^\lambda, \hat{Q}_2^\lambda) \Rightarrow (r_1 \hat{Q}_\Sigma, r_2 \hat{Q}_\Sigma)$ as $\lambda \to \infty$, where $\hat{Q}_\Sigma = \hat{X}_c^+$.

Other Scheduling Policies (PRP)

- Same techniques can be used to obtain diffusion limits under alternative scheduling policies
- However, diffusion limits depend on choice of policy
- Other policies likely to be sub-optimal

Asymmetric Systems

Can consider general case with arrival rates $a_i\lambda$ where $a_i > 0$ and $a_1 + a_2 = 2$. Here, choose $n_1^\lambda, n_2^\lambda, n_F^\lambda$.

Complete Resource Pooling

- Still optimal to have $> O(\sqrt{\lambda})$ flexible servers, which still achieves CRP
- Similar conditions for asymptotic optimality, e.g. same choice of α^*

Partial Resource Pooling

- Diffusion limits are highly sensitive to choice of scheduling policy
- Optimal scheduling policy is an open problem