Proactive Care with Degrading Class Types

Yue Hu (DRO, Columbia Business School)

Joint work with
Prof. Carri Chan (DRO, Columbia Business School)
and
Prof. Jing Dong (DRO, Columbia Business School)

Motivation

Customer slowdown describes the phenomenon that a customer's service requirement increases with experienced delay

Motivation

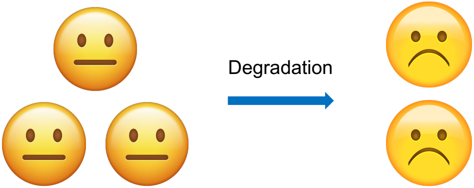
In healthcare settings, delays in receiving appropriate care can result in adverse effects, e.g., increased LOS in ICU

Motivation

The snowball effect: a delayed patient that requires a longer service time increases the overall workload of the system, therefore causing longer delays for other patients, who in turn might require longer service

Motivation

Consider a typical health-care setting with moderate and urgent patients:

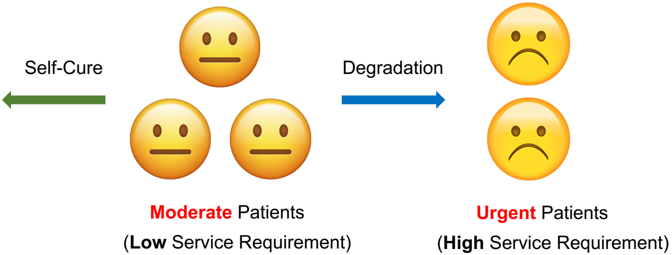


Moderate Patients
(**Low** Service Requirement)

Urgent Patients
(**High** Service Requirement)

Motivation

Consider a typical health-care setting with moderate and urgent patients:

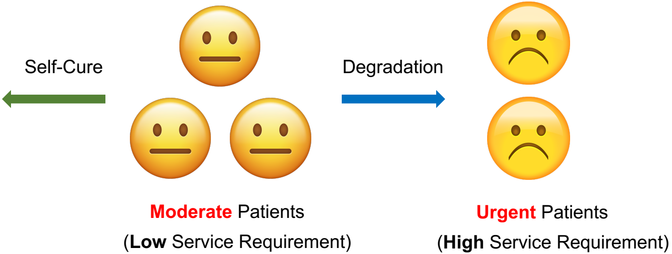


When proactive service is an option, we face the tension!

Motivation

Research questions:

- Is it worth initiating proactive care for moderate patients?
- How should we allocate resources to achieve good system performance?



Related Literature

The Snowball Effect of Customer Slowdown:

Sheridan et al. (1999), Richardson (2002), Liew et al. (2003), Siegmeth et al. (2005), Chalfin et al. (2007), Chan et al. (2008), Renaud et al. (2009), Selen et al. (2015), Dong et al. (2015), Chan et al. (2017)

Proactive Service and Queueing with Future Information:

Gans and Zhou (2002, 2007), Gans et al. (2003), Whitt (2006), Askin et al. (2007), Gurvich et al. (2010), Spencer et al. (2014), Xu (2015), Xu and Chan (2016), Hu et al. (2017), Delana et al. (2018)

Queueing Models with Dynamic Priority:

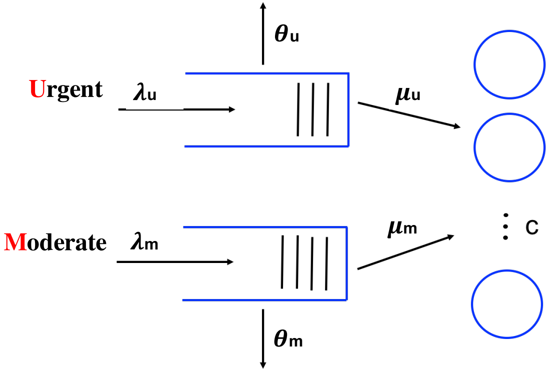
He and Neuts (2002), Wang (2004), Gómez- Corral et al. (2005), Maertens et al. (2006), Down and Lewis (2010), He et al. (2012), Girard et al. (2017), Xie et al. (2017)

Optimization and Optimal Control:

Altman et al. (2001), Harrison and Zeevi (2003), Larrañaga et al. (2013), Atar et al. (2011), Cao and Xie (2016)

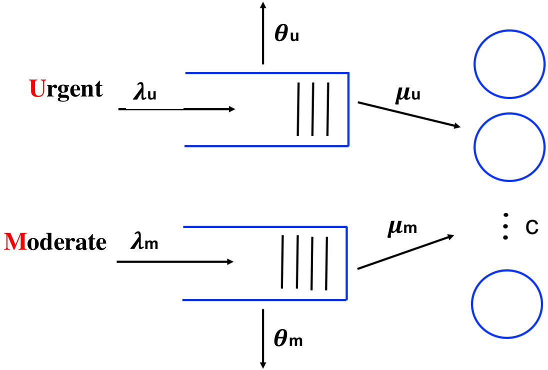
The Model

A stochastic queueing network where two queues are served by c servers



The Model

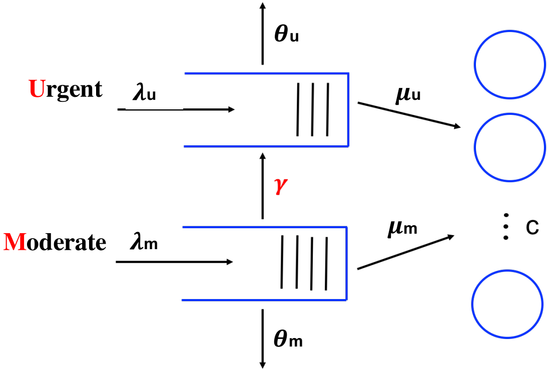
A stochastic queueing network where two queues are served by c servers



- Stationary arrival process of jobs with rate λ_i to queue i, $i \in \{u, m\}$
- IID service times with rate μ_i at queue i, $\mu_u < \mu_m$
- Type-i patients abandon the queue according to a stationary point process at rate θ_i

The Model

A stochastic queueing network where two queues are served by c servers



- Delayed moderate patients become urgent at rate γ according to a stationary point process

Goal

Our goal is to find a service control (in staffing and scheduling) that minimizes long run average costs (formal definition follows), assuming a linear unit-time holding cost h_i is incurred in queue i, and unit staffing cost s

Goal

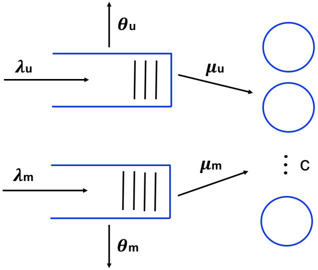
Our goal is to find a service control (in staffing and scheduling) that minimizes long run average costs (formal definition follows), assuming a linear unit-time holding cost h_i is incurred in queue i, and unit staffing cost s

We consider admissible controls that are

- non-anticipatory
- preemptive
- non-idling

Challenges

- Overloaded regime: the $c\mu/\theta$ rule is optimal (Atar et al. (2011))
- Limiting heavy-traffic regime:
  - the optimal control is the solution to the associated Hamilton-Jacobi-Bellman equation (Harrison and Zeevi (2003), Atar et al. (2004))
- Special Case: two-queue fluid system: (Larrañaga et al. (2013))



A Fluid Approximation to Simplify the Problem

Consider a piecewise affine dynamical system characterized by

$$dx_u(t) = \lambda_u + \gamma(x_m(t) - \alpha_m(t))^+ - \theta_u(x_u(t) - \alpha_u(t)) - \mu_u \alpha_u(t)$$
$$dx_m(t) = \lambda_m - (\gamma + \theta_m)(x_m(t) - \alpha_m(t))^+ - \mu_m \alpha_m(t)$$

where $\alpha_i(t)$ is the amount of capacity devoted to serving type-i customers, $0 \leq \alpha_i(t) \leq x_i(t)$, $\alpha_u(t) + \alpha_m(t) \leq c$

A Fluid Approximation to Simplify the Problem

Consider a piecewise affine dynamical system characterized by

$$dx_u(t) = \lambda_u + \gamma(x_m(t) - \alpha_m(t))^+ - \theta_u(x_u(t) - \alpha_u(t)) - \mu_u\alpha_u(t)$$
$$dx_m(t) = \lambda_m - (\gamma + \theta_m)(x_m(t) - \alpha_m(t))^+ - \mu_m\alpha_m(t)$$

where $\alpha_i(t)$ is the amount of capacity devoted to serving type-i customers, $0 \leq \alpha_i(t) \leq x_i(t)$, $\alpha_u(t) + \alpha_m(t) \leq c$

Recast the problem to the fluid model:

$$\min_{\{c,\alpha_u(t),\alpha_m(t)\}} \lim_{T\to\infty} \frac{1}{T}\int_0^T (h_u q_u(t) + h_m q_m(t) + sc)dt$$

Strict Priority Rules

We limit to a subset of admissible controls and consider the strict priority rules P_u and P_m, where P_i assigns strict priority to type-i customers.

Strict Priority Rules

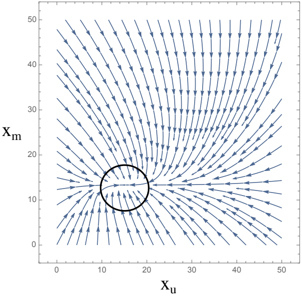
We limit to a subset of admissible controls and consider the strict priority rules P_u and P_m, where P_i assigns strict priority to type-i customers.

What is the long-run behavior of the system under P_u and P_m? Can we characterize the equilibria, if any?

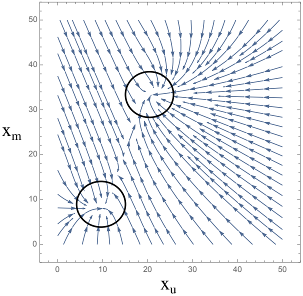
Strict Priority Rules

We limit to a subset of admissible controls and consider the strict priority rules P_u and P_m, where P_i assigns strict priority to type-i customers.

What is the long-run behavior of the system under P_u and P_m? Can we characterize the equilibria, if any?



(a) Globally asymptotically stable equilibrium

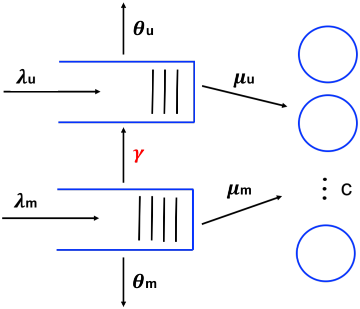


(b) Locally asymptotically stable equilibrium

Strict Priority Rules

The equilibrium behavior of the system depends on two parameter cases

- Case 1: $\mu_u > \frac{\gamma}{\gamma+\theta_m}\mu_m$
- Case 2: $\mu_u < \frac{\gamma}{\gamma+\theta_m}\mu_m$



Strict Priority Rules

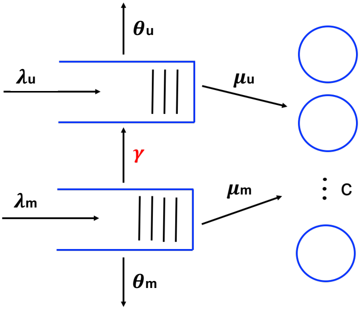
The equilibrium behavior of the system depends on two parameter cases

- Case 1: $\mu_u > \frac{\gamma}{\gamma+\theta_m}\mu_m$

  $\frac{1}{\mu_m} > \frac{\gamma}{\gamma+\theta_m}\frac{1}{\mu_u}$

  There is less work if a moderate patient degrades

- Case 2: $\mu_u < \frac{\gamma}{\gamma+\theta_m}\mu_m$



Strict Priority Rules

The equilibrium behavior of the system depends on two parameter cases

- Case 1: $\mu_u > \frac{\gamma}{\gamma+\theta_m}\mu_m$

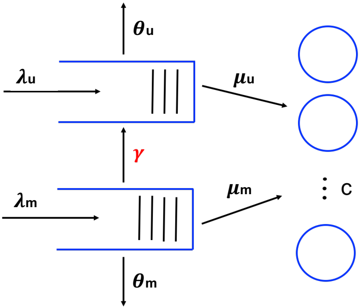
  $\frac{1}{\mu_m} > \frac{\gamma}{\gamma+\theta_m}\frac{1}{\mu_u}$

  There is less work if a moderate patient degrades

- Case 2: $\mu_u < \frac{\gamma}{\gamma+\theta_m}\mu_m$

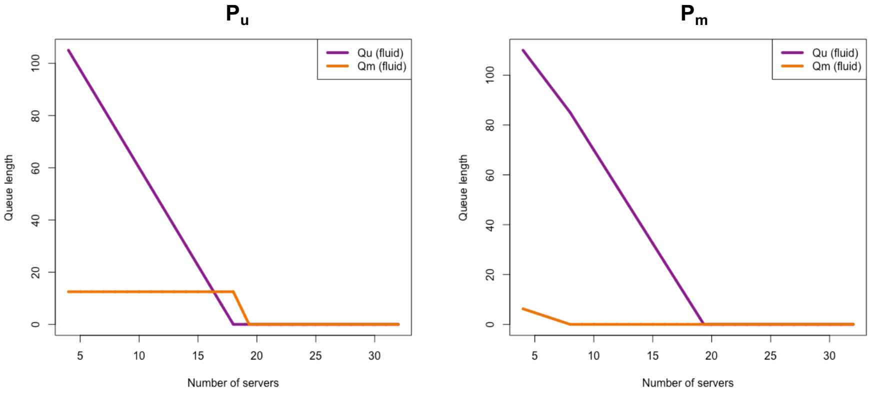
  $\frac{1}{\mu_m} < \frac{\gamma}{\gamma+\theta_m}\frac{1}{\mu_u}$

  There is less work if a moderate patient does **not** degrade



Strict Priority Rules

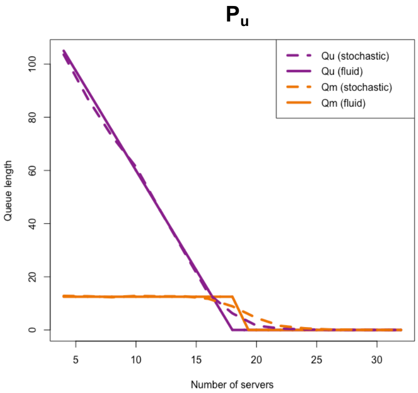
Fluid equilibrium in Case 1: $\mu_u > \frac{\gamma}{\gamma+\theta_m}\mu_m$

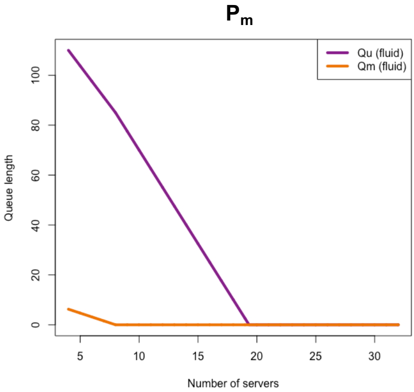


$$\lambda_u = 17, \lambda_m = 20, \mu_u = 1.5, \mu_m = 2.5, \theta_u = 0.2, \theta_m = 0.8, \gamma = 0.8$$

Strict Priority Rules

Fluid equilibrium in Case 1: $\mu_u > \frac{\gamma}{\gamma+\theta_m}\mu_m$

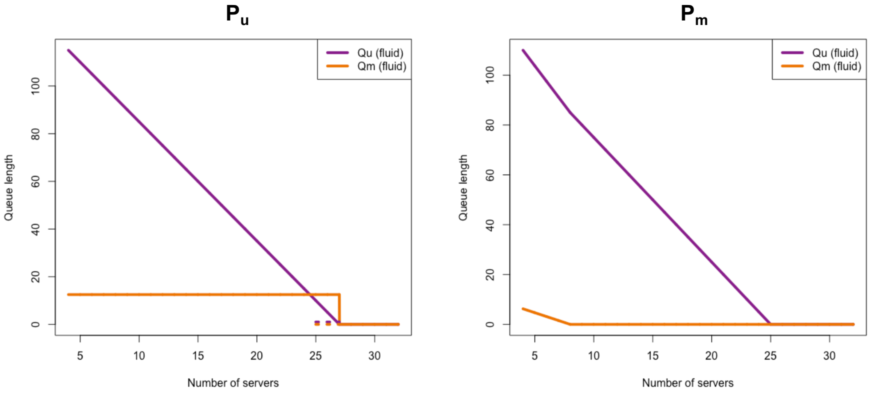




$$\lambda_u = 17, \lambda_m = 20, \mu_u = 1.5, \mu_m = 2.5, \theta_u = 0.2, \theta_m = 0.8, \gamma = 0.8$$

Strict Priority Rules

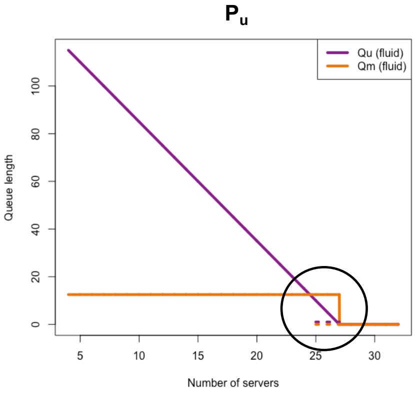
Fluid equilibrium in Case 2: $\mu_u < \frac{\gamma}{\gamma+\theta_m}\mu_m$

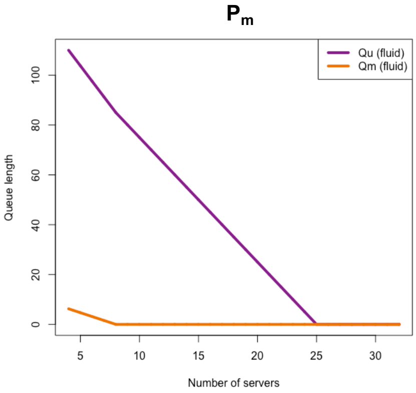


$\lambda_u = 17, \lambda_m = 20, \mu_u = 1, \mu_m = 2.5, \theta_u = 0.2, \theta_m = 0.8, \gamma = 0.8$

Strict Priority Rules

Fluid equilibrium in Case 2: $\mu_u < \frac{\gamma}{\gamma+\theta_m}\mu_m$

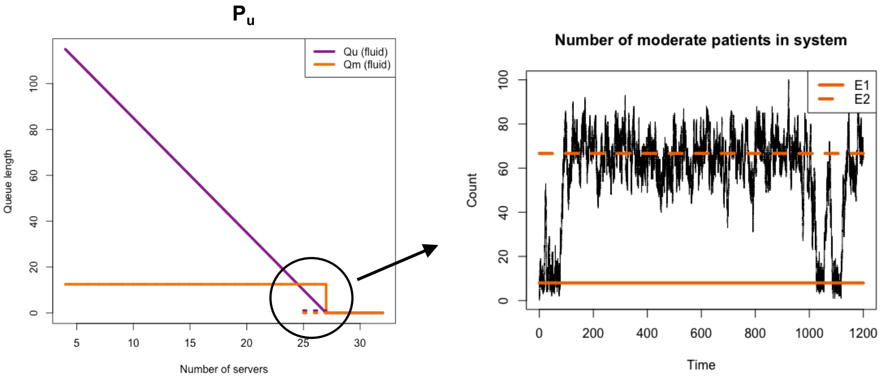




$\lambda_u = 17, \lambda_m = 20, \mu_u = 1, \mu_m = 2.5, \theta_u = 0.2, \theta_m = 0.8, \gamma = 0.8$

Strict Priority Rules

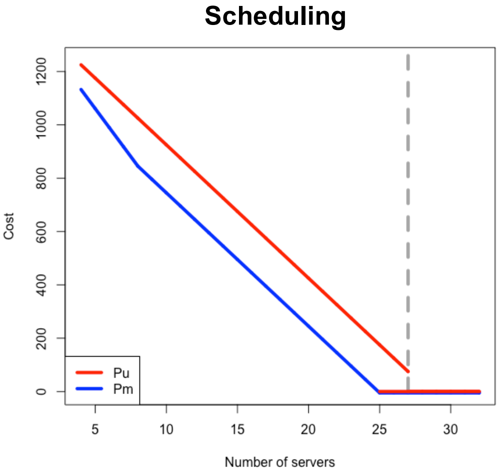
Fluid equilibrium in Case 2: $\mu_u < \frac{\gamma}{\gamma+\theta_m}\mu_m$



$\lambda_u = 17, \lambda_m = 20, \mu_u = 1, \mu_m = 2.5, \theta_u = 0.2, \theta_m = 0.8, \gamma = 0.8$

Strict Priority Rules

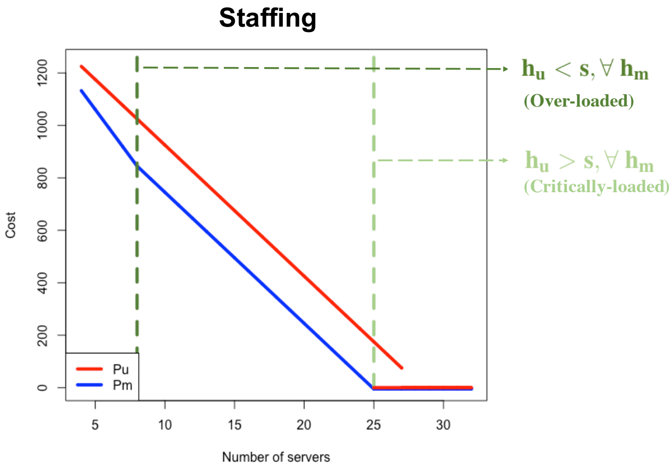
Minimizing the long-run average holding cost in Case 2: $\mu_u < \frac{\gamma}{\gamma+\theta_m}\mu_m$



$h_u = 10, h_m = 6$

Strict Priority Rules

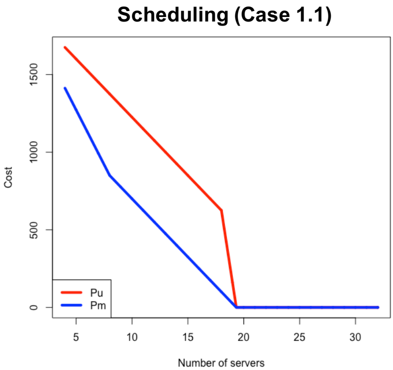
Minimizing the long-run average holding cost in Case 2: $\mu_u < \frac{\gamma}{\gamma+\theta_m}\mu_m$



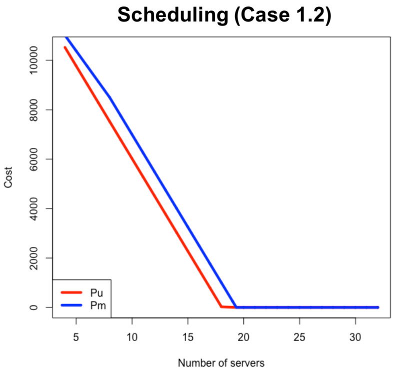
$h_u = 10, h_m = 6$

Strict Priority Rules

Minimizing the equilibrium holding cost in Case 1: $\mu_u > \frac{\gamma}{\gamma+\theta_m}\mu_m$



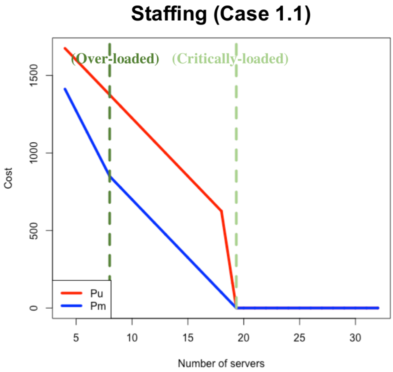
(a) $\frac{h_u}{h_m} < \frac{\mu_m}{\gamma+\theta_m}\frac{\theta_u}{\mu_u-\frac{\gamma}{\gamma+\theta_m}\mu_m}$



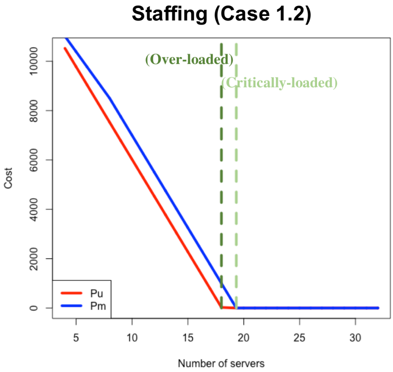
(b) $\frac{h_u}{h_m} > \frac{\mu_m}{\gamma+\theta_m}\frac{\theta_u}{\mu_u-\frac{\gamma}{\gamma+\theta_m}\mu_m}$

Strict Priority Rules

Minimizing the equilibrium holding cost in Case 1: $\mu_u > \frac{\gamma}{\gamma+\theta_m}\mu_m$



(a) $\frac{h_u}{h_m} < \frac{\mu_m}{\gamma+\theta_m}\frac{\theta_u}{\mu_u - \frac{\gamma}{\gamma+\theta_m}\mu_m}$



(b) $\frac{h_u}{h_m} > \frac{\mu_m}{\gamma+\theta_m}\frac{\theta_u}{\mu_u - \frac{\gamma}{\gamma+\theta_m}\mu_m}$

Summary

- We propose a two-class multi-server queueing model to study the potential of proactive care with degrading class types
- We consider a fluid approximation and obtain optimality results in staffing and scheduling w.r.t. the long-run average cost
- Ongoing work:
  - Relating the fluid optimality results to the stochastic system
  - Studying transient fluid dynamics
- Future direction:
  - Diffusion control

Thank You