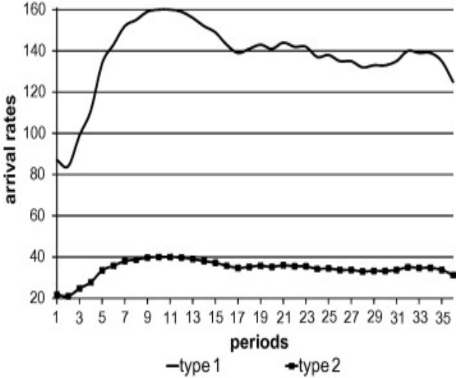
Staffing and Scheduling to Differentiate Service in Multiclass Time-Varying Service Systems

Xu Sun
joint work with advisor Ward Whitt, Yunan Liu (NCSU) and Kyle Hovey (NCSU)

Department of Industrial Engineering & Operations Research

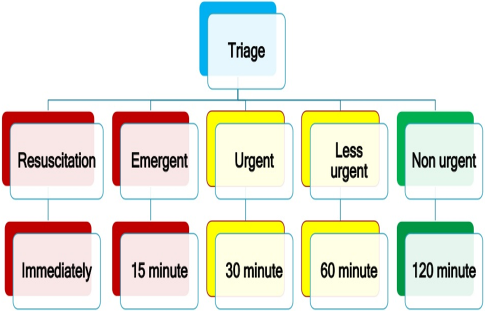
September 22, 2018

Motivating Example 1 - Call Center



- 80% of type 1 calls need to be answered within 20 seconds ("80-20 rule")
- 50% of type 2 calls need to be answered within 60 seconds
- How many servers are needed over the course of day?
- How to assign a newly idle agents to one of these queues?

Example 2 - Canadian Triage and Acuity Scale (CTAS)



According to CTAS guideline Ding et al. (2018), "CTAS level i patients need to be seen by a physician within w_i minutes $100\alpha_i$% of the time", with

$$(w_1, w_2, w_3, w_4, w_5) = (0, 15, 30, 60, 120),$$

$$(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5) = (0.98, 0.95, 0.9, 0.85, 0.8).$$

Other Examples of Multi-class Settings

Modeling Real Service Systems

Model with these features are difficult to analyze:

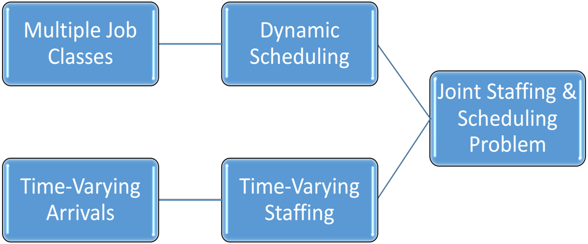
- Time-varying arrivals
- Customer abandonment
- Non-exponential service and abandonment distributions
- Multiple customer classes

Goals:

- Break through fundamental barriers holding back the community;
- Bring more practical models within range of tractability;
- Provide performance analysis and decision support tools.

Objective of Study and Approach

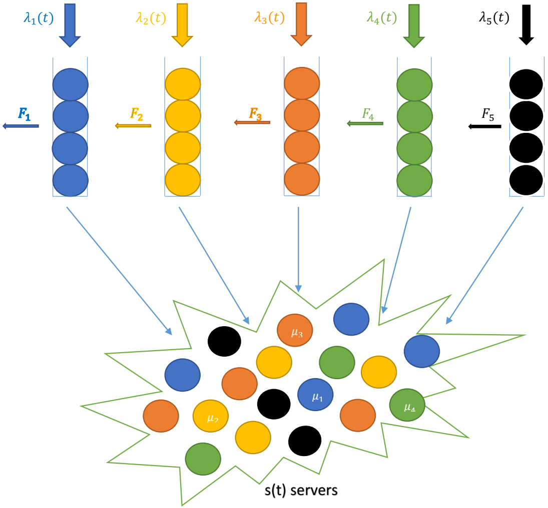
- Goal: achieve acceptable service level for different classes
- Method:
  - Effective planning of service capacity to cope with time-varying demand (staffing);
  - Timely allocation of service resources to every customer class (scheduling).



Existing Works and Contributions

- Literature review
  - Service differentiation
    Gurvich, Armony & Mandelbaum (08); Gurvich & Whitt (10); Soh & Gurvich (16); Kim, Randhawa & Ward (2017)
    All assume a critical-loading system and the demand to be stationary
  - Performance stabilization of time-varying queues
    Jennings et al. (96); Feldman et al. (08); Pender & Massey (17); Liu & Whitt (12,14); Liu (18)
    All consider single-class models
- Our contribution:
  studying service differentiation with time-varying demand and class-dependent services, focusing on overloaded systems.

A Multi-Class V Model



- Class-dependent arrival rate $\lambda_i(t)$ (non-homogeneous Poisson)
- Class-dependent abandonment-time distribution F_i
- A large time-varying number of servers $s(t)$
- Exponential service times with class-dependent service rate μ_i
- First-Come First-Served within each class

Problem Statement

- Model parameters

$$\mathcal{P} \equiv (\underbrace{\lambda_i(t), F_i, \mu_i,}_{\text{customer behavior}} \quad \underbrace{w_i, \alpha_i,}_{\text{service level}} \; 1 \le i \le K, \; 0 \le t \le T)$$

- Obtain convenient staffing and scheduling rules (in terms of $\mathcal{P}$), such that the *tail probability of delay* (TPoD)

$$\mathbb{P}\left(W_i(t) > w_i\right) \le \alpha_i, \qquad 1 \le i \le K, \quad 0 \le t \le T,$$

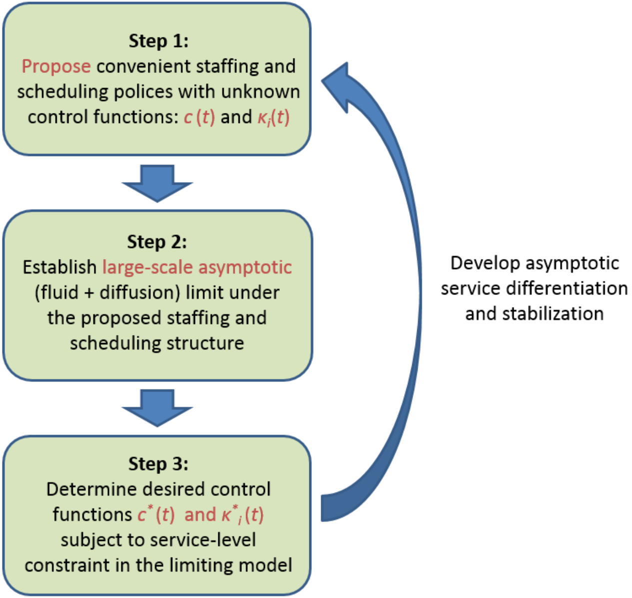
$$\text{or} \quad \mathbb{P}\left(W_i(t) > w_i\right) \approx \alpha_i$$

for any

  - $w_i > 0$ (delay target).
  - $\alpha_i \in (0, 1)$ (probability target: fraction of excessive delay).

$W_i(t)$: potential waiting time of class i at time t, i.e., offered delay to a class-i arrival at t assuming infinitely patient.

Main Steps of Our Approach



Step I: Proposed Staffing Formula

1. Use offered-load (OL) to determine the nominal service capacity
   - No. of busy servers $B(t)$ in $M_t/GI/\infty \sim$ Poisson r.v. with mean
   $$m(t) \equiv \mathbb{E}[B(t)] = \int_0^t \lambda(t-s)G^c(s)\mathrm{d}s.$$
   - Here, for the i^{th} class, the OL is
   $$m_i(t) = \int_0^t \underbrace{F_i^c(w_i)\lambda_i(u-w_i)}_{\text{effective arr. rate}} \underbrace{e^{-\mu_i(t-u)}}_{\text{exp. service dist.}} \mathrm{d}u.$$

   OL: mean No. of busy servers needed to serve all customers who are willing to wait (excluding an acceptable faction of abandonment).

2. A **time-varying square-root staffing** (TV-SRS) rule
$$\boxed{s(t) = \underbrace{m(t)}_{\text{first order}} + \underbrace{\sqrt{\lambda^\star}c(t)}_{\text{second order}}} \quad \text{for} \quad m(t) \equiv m_1(t) + \cdots + m_K(t)$$
where $c(t)$ is a control function (TBD), and $\lambda^\star$ is the system's scale, i.e.,
$$\lambda^\star \equiv \frac{1}{T}\int_0^T \lambda(t)\mathrm{d}t, \qquad \text{with} \qquad \lambda(t) \equiv \lambda_1(t) + \cdots + \lambda_K(t).$$

Step I: Proposed Control Structure

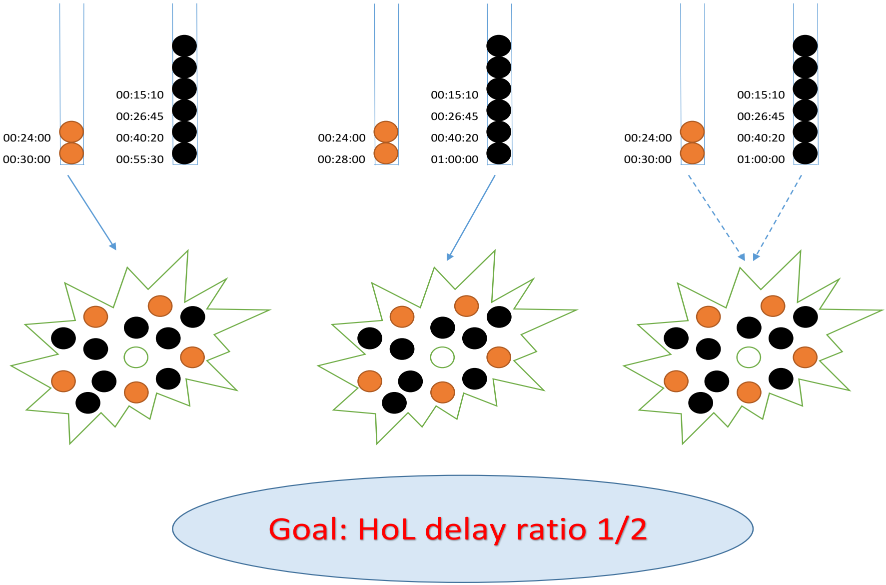
- Use real-time class-i head-of-line waiting time (HWT) $H_i(t)$ to devise a dynamic control policy;
- A **time-varying dynamic prioritization scheduling** (TV-DPS) rule: Assigns the next available server to the HoL customer from class i^* satisfying

$$\boxed{i^* \in \arg\max_{1 \le i \le K} \left\{ \underbrace{\frac{H_i(t)}{w_i}}_{\textbf{normalized HWT}} + \frac{1}{\sqrt{\lambda^\star}} \kappa_i(t) \right\}.}$$

where $\kappa_i(t)$ is a control function (TBD).

- Main ideas of TV-DPS:
  - $\tilde{H}_i(t) \equiv H_i(t)/w_i$ focuses on the delay target w_i;
  - $\kappa_i(t)$ helps accomplish the class-dependent probability target α_i;
  - TV-DPS is both time-dependent (accounting for time variability) and state-dependent (capturing stochasticity).

An Illustration of How TV-DP Rule Works



Step II: Large-Scale Asymptotic Analysis

- Exact analysis is difficult; hence do asymptotic analysis as scale grows (realistic for large-scale systems).
- Use n in place of $\lambda^\star$ and consider a sequence of models indexed by n.
- In the n^{th} model:
  - Arrival rate $\lambda_i^n(t) \equiv n\lambda_i(t)$;
  - Staffing level:
  $$s^n(t) = nm(t) + \sqrt{n}c(t).$$
  - Scheduling rule:
  $$i^* \in \underset{1\le i\le K}{\arg\max}\left\{\frac{H_i^n(t)}{w_i} + \frac{1}{\sqrt{n}}\,\kappa_i(t)\right\}.$$
  - Service rates and abandonment distributions are fixed.
- Scaled HWT and PWT processes:
$$\widehat{H}_i^n(t) \equiv n^{1/2}\left(H_i^n(t) - w_i\right) \quad \text{and} \quad \widehat{W}_i^n(t) \equiv n^{1/2}\left(W_i^n(t) - w_i\right).$$

Limit of Waiting Times and State-Space Collapse

Under the TV-SRS and TV-DPS policy, the CLT-scaled waiting time processes

$$\left(\widehat{H}_1^n, \ldots, \widehat{H}_K^n, \widehat{W}_1^n \ldots, \widehat{W}_K^n\right) \Rightarrow \left(\widehat{H}_1, \ldots, \widehat{H}_K, \widehat{W}_1 \ldots, \widehat{W}_K\right) \quad \text{in} \quad \mathcal{D}^{2K} \quad \text{as} \quad n \to \infty,$$

with all HWT and PWT limits in terms of a one-dimensional process $\widehat{H}(\cdot)$, where

$$\widehat{H}_i(t) \equiv w_i(\widehat{H}(t) - \kappa_i(t)), \qquad \widehat{W}_i(t) = w_i(\widehat{H}(t + w_i) - \kappa_i(t + w_i)).$$

The process $\widehat{H}$ uniquely solves the following *stochastic Volterra equation* (SVE)

$$\widehat{H}(t) = \int_0^t L(t,s)\widehat{H}(s)ds + \int_0^t J(t,s)d\mathcal{W}(s) + K(t),$$

where $\mathcal{W}$ is a standard Brownian motion,

$$L(t,s) \equiv \frac{\sum_{i=1}^K \eta_i(s)e^{\mu_i(s-t)}\left(\mu_i - h_{F_i}(w_i)\right)}{\eta(t)}, \quad J(t,s) \equiv \frac{\sqrt{\sum_{i=1}^K e^{2\mu_i(s-t)}\left(F_i^c(w_i)\lambda_i(s-w_i) + \mu_i m_i(s)\right)}}{\eta(t)},$$

$$K(t) \equiv \frac{\sum_{i=1}^K \left(\eta_i(t)\kappa_i(t) - \int_0^t \eta_i(s)e^{\mu_i(s-t)}\left(\mu_i - h_{F_i}(w_i)\right)\kappa_i(s)ds\right) - c(t)}{\sum_{i=1}^K \eta_i(t)}.$$

for $\eta_i(t) \equiv w_i\lambda_i(t-w_i)F_i^c(w_i)$.

Limit of Waiting Times and State-Space Collapse

1. State-space collapse:
   All limiting HWT and PWT processes degenerates to a one-dimensional process $\widehat{H}$.

2. The SVE admits a unique solution which is a Gaussian process
   - If $\mu_i \neq \mu$
     - ★ SVE has NO analytic solution;
     - ★ We gave effective algorithms (geometrically fast) to compute $m_{\widehat{H}}(t) \equiv \mathbb{E}[\widehat{H}(t)]$ and variance $C_{\widehat{H}}(t,s) \equiv \text{Cov}(\widehat{H}(t), \widehat{H}(s))$.
   - If $\mu_i = \mu$, SVE has an closed-form solution (so do $m_{\widehat{H}}(t)$ and $C_{\widehat{H}}(t,s)$).

$$\widehat{H}(t) = \frac{1}{R(t)} \left(\int_0^t \tilde{J}(u) d\mathcal{W}(u) + \int_0^t \tilde{R}(u) dK(u) + \int_0^t \tilde{K}(u) dR(u) \right).$$

3. Variance $\sigma^2_{\widehat{H}}(t) = \mathrm{Var}(\widehat{H}(t))$ relies only on model parameters (independent with control functions).

4. Control functions $c(\cdot)$ and $\kappa_i(\cdot)$ appear in the term $K(\cdot)$ only.

Step III: Solve $c^*(t)$ and $\kappa_i^*(t)$ subject to Service-Level Constraints

- Main ideas: when n is large, at each time $t \in [0, T]$, we hope

$$
\begin{aligned}
\alpha_i &\equiv \mathbb{P}(H_i^n(t) > w_i) = \mathbb{P}(\widehat{H}_i^n(t) > 0) \\
&\approx \mathbb{P}(\widehat{H}_i(t) > 0) = \mathbb{P}(w_i(\widehat{H}(t) - \kappa_i(t)) > 0) \\
&= \mathbb{P}\left(\mathcal{N}\left(m_{\widehat{H}}(t), \sigma^2_{\widehat{H}}(t)\right) > \kappa_i(t)\right) = \mathbb{P}\left(\mathcal{N}(0,1) > \frac{\kappa_i(t) - m_{\widehat{H}}(t)}{\sigma_{\widehat{H}}(t)}\right)
\end{aligned}
$$

Recall that $m_{\widehat{H}}(t)$ is a function of $\kappa_i(t)$ and $c_i(t)$.

- Obtain the asymptotically "optimal" control functions:

$$
\begin{aligned}
c(t) &= \sum_{i=1}^{K} \left(\eta_i(t)\kappa_i(t) - \int_0^t \eta_i(s) e^{\mu_i(s-t)} \left(\mu_i - h_{F_i}(w_i)\right) \kappa_i(s) ds \right), \\
\kappa_i(t) &= z_{\alpha_i} \sigma_{\widehat{H}}(t), \qquad 1 \le i \le K, \quad 0 \le t \le T.
\end{aligned}
$$

where $z_\alpha = \Phi^{-1}(1-\alpha)$, $\eta_i(t) \equiv w_i \lambda_i(t - w_i) F_i^c(w_i)$.

Step III: Solve $c^*(t)$ and $\kappa_i^*(t)$ subject to Service-Level Constraints

The asymptotically "optimal" control functions:

$$c(t) = \sum_{i=1}^{K} \left(\eta_i(t)\kappa_i(t) - \int_0^t \eta_i(s) e^{\mu_i(s-t)} \left(\mu_i - h_{F_i}(w_i)\right) \kappa_i(s) ds \right), \qquad (1)$$

$$\kappa_i(t) = z_{\alpha_i} \sigma_{\widehat{H}}(t), \qquad 1 \le i \le K, \quad 0 \le t \le T. \qquad (2)$$

Theorem (Asymptotic service differentiation)

Under our staffing and scheduling rule with $c_i(\cdot)$ and $\kappa_i(\cdot)$ in (1) and (2),

(i) *Mean PWT and HWT are both asymptotically differentiated and stabilized:*

$$\mathbb{E}[W_i^n(t)] \to w_i \quad \text{and} \quad \mathbb{E}[H_i^n(t)] \to w_i \qquad \text{as} \quad n \to \infty, \qquad \text{for} \quad 0 < t \le T, \ 1 \le i \le K.$$

(ii) *TPoDs for PWT and HWT are both asymptotically differentiated and stabilized:*

$$\mathbb{P}(W_i^n(t) > w_i) \to \alpha_i \quad \text{and} \quad \mathbb{P}(H_i^n(t) > w_i) \to \alpha_i \qquad \text{as} \quad n \to \infty$$

for $0 < t \le T, \ 1 \le i \le K.$

Constant arrival rates

When $\lambda_i(t) = \lambda_i$

- Staffing

$$m_i(t) \sim m_i \equiv \frac{\lambda_i F_i^c(w_i)}{\mu}, \qquad c(t) \sim c \equiv \sum_{i=1}^{K} \frac{w_i \lambda_i f_i(w_i)}{\mu} \kappa_i,$$

- Scheduling

$$\kappa_i(t) \sim \kappa_i \equiv z_{\alpha_i} \cdot \underbrace{\sqrt{\frac{\sum_{j=1}^{K} \lambda_j F_j^c(w_i)}{\left(\sum_{j=1}^{K} \lambda_j f_j(w_j) w_j\right)\left(\sum_{j=1}^{K} \lambda_j F_j^c(w_j) w_j\right)}}}_{\text{independent with } \alpha_i}.$$

These formulas can be used to estimate

- the required average number of servers and scheduling threshold;
- the *marginal price of staffing and scheduling* (MPSS):

  To improve the service to the next level ($w_i \to w_i - \Delta w_i$ or $\alpha_i \to \alpha_i - \Delta \alpha_i$), how many extra servers are need and how much should the scheduling thresholds be adjusted?

Review of the Approach

$$s(t) = m(t) + (\lambda^\star)^{1/2} c(t)$$

$$i^* \in \arg\max_{1 \le i \le K} \left\{ H_i(t)/w_i + \frac{1}{\sqrt{\lambda^\star}} \kappa_i(t) \right\}$$

$$\left(\widehat{H}_1^n, \ldots, \widehat{H}_K^n\right) \Rightarrow \left(\widehat{H}_1, \ldots, \widehat{H}_K\right)$$

$$\widehat{H}_i(t) \equiv w_i(\widehat{H}(t) - \kappa_i(t)) \qquad \text{SSC}$$

$$\widehat{H}(t) = \int_0^t L(t,s)\widehat{H}(s)ds + \int_0^t J(t,s)dW(s) + K(t)$$

Analyzing $\widehat{H}(t)$ to obtain

$$\kappa_i^*(t) = z_{\alpha_i}\sigma_{\widehat{H}}(t)$$

$$c^*(t) = \cdots\cdot$$

Step 1:
Propose convenient staffing and scheduling polices with unknown control functions: $c(t)$ and $\kappa_i(t)$

Step 2:
Establish large-scale asymptotic (fluid + diffusion) limit under the proposed staffing and scheduling structure

Step 3:
Determine desired control functions $c^*(t)$ and $\kappa^*_i(t)$ subject to service-level constraint in the limiting model

Numerical Examples

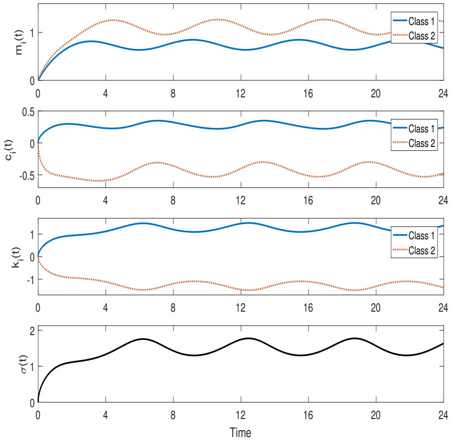
Base Case - A Two-Class V Model

- Model parameters
  - Sinusoidal arrival rates $\lambda_i(t) = n\bar{\lambda}_i\,(1 + r_i \sin(\gamma_i t + \phi_i))$
    $\bar{\lambda}_1 = 1, \bar{\lambda}_2 = 1.5, r_1 = 0.2, r_2 = 0.3, \gamma_1 = \gamma_2 = 1, \phi_1 = 0, \phi_2 = -1$
  - Service rates $\mu_1 = \mu_2 = 1$ (later extend to class-dependent case)
  - Exponential abandonment times with rates $\theta_1 = 0.6, \theta_2 = 0.3$.
  - System scale: $n = 50$
- QoS parameters
  - Delay targets $w_1 = 0.5, w_2 = 1$;
  - Probability targets $\alpha_1 = 0.2, \alpha_2 = 0.8$.

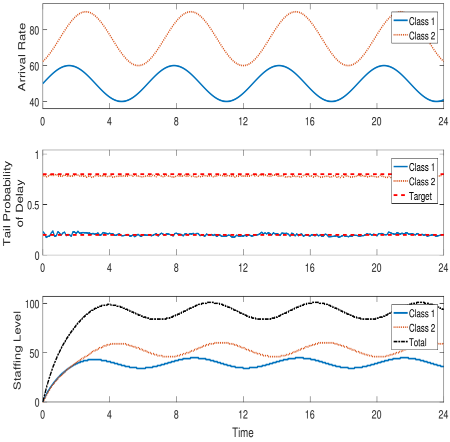
  Hope to achieve:

  $\mathbb{P}(W_1(t) > 0.5) \approx 20\%, \quad \mathbb{P}(W_2(t) > 1) \approx 80\%.$ Class-1 more important!
- Monte Carlo simulation with 5000 independent runs.

Base Case

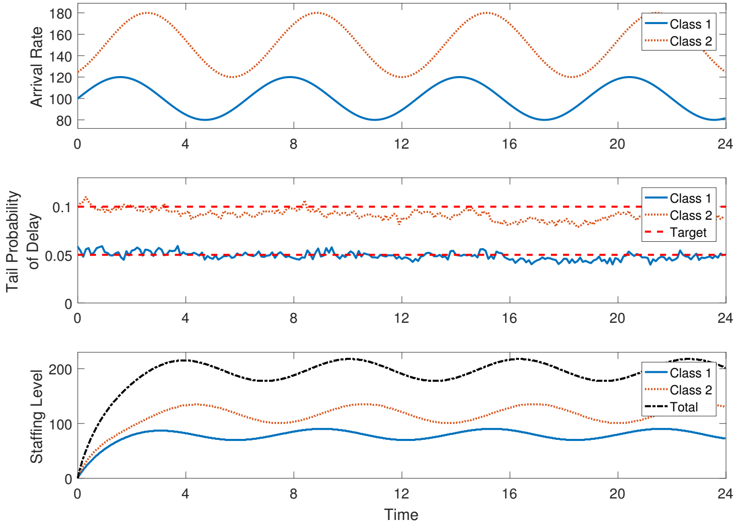


(a) Control functions



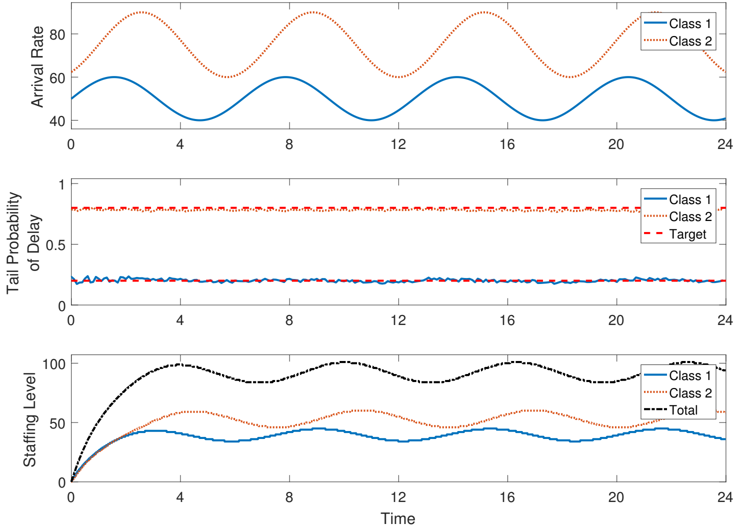
(b) Simulations

High Quality of Service ($\alpha_1 = 0.05, \alpha_2 = 0.1$)



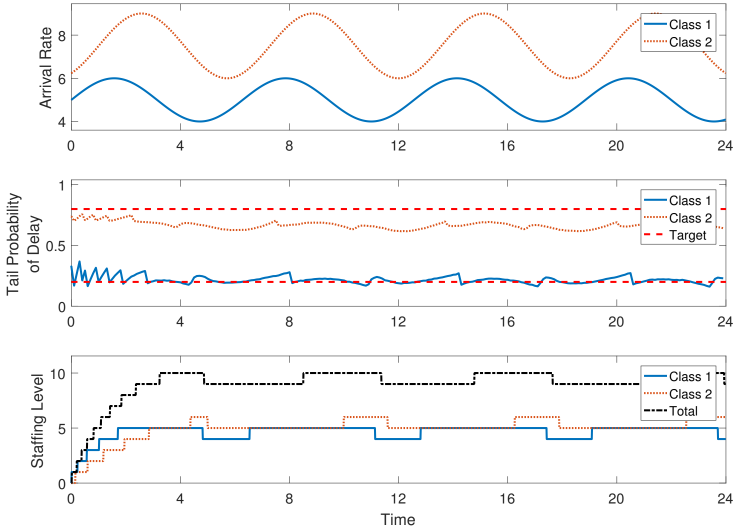
Good performance when $\alpha_i \approx 0$.

Very High Quality of Service ($w_1 = 0.05, w_2 = 0.1$)



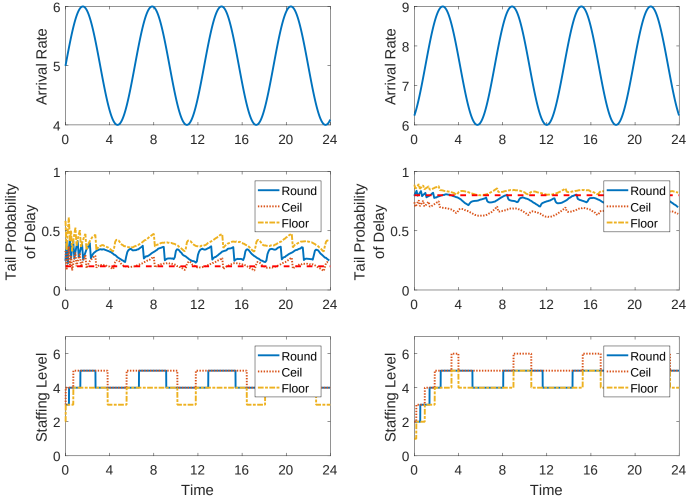
If $w_i \approx 0$, TPoD degenerates to probability of delay (PoD) $\mathbb{P}(W_i(t) > 0)$.

A Small System ($n = 5$)



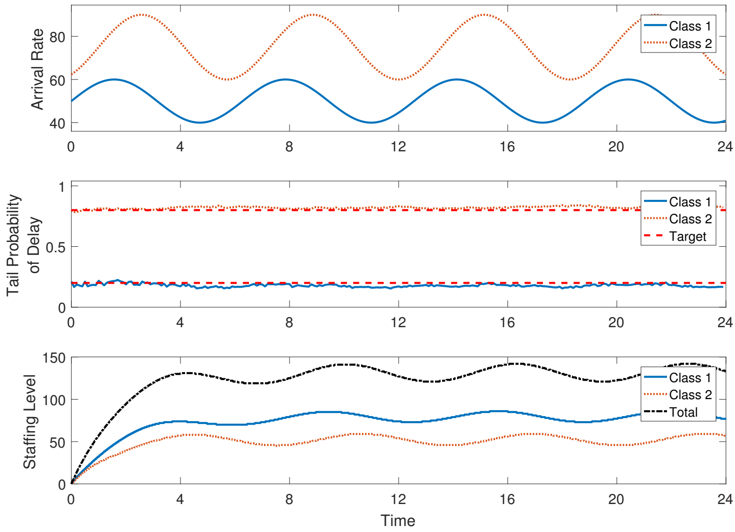
It is ok to apply $n \to \infty$ results to a system with a very small n.

A Small System ($n = 5$)



- When n is small, adding/removing a server causes bigger bumps;
- Error is attributed to discretization of staffing levels.

Class-Dependent Service Rates ($\mu_1 = 0.5, \mu_2 = 1$)



When $\mu_1 \neq \mu_2$, $\sigma^2_{\widehat{H}}(t)$ is numerically computed using our algorithm.

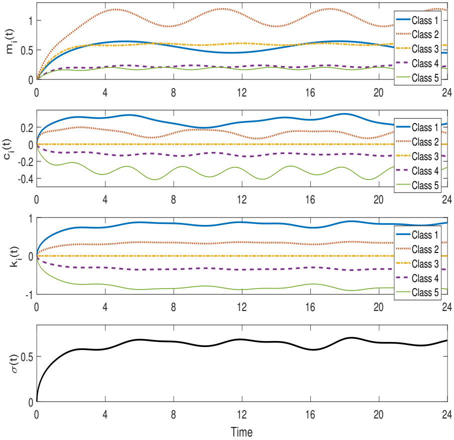
A Five-Class V Model: Parameters and QoS Targets

- Sinusoidal arrival rates $\lambda_i(t) = n\bar{\lambda}_i\,(1 + r_i \sin(\gamma_i t + \phi_i))$;
- Exponential service times;
- Exponential abandonment times;
- Scale $n = 50$.

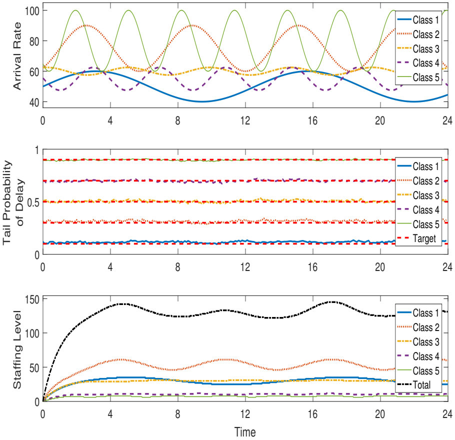
	Arrival Parameters				Abandonment rates	Service rates	Service levels	
Class	$\bar{\lambda}_i$	r_i	γ_i	ϕ_i	θ_i	μ_i	w_i	α_i
1	1.0	0.20	1	0	0.6	1	0.2	0.1
2	1.5	0.30	1	-1	0.3	1	0.4	0.3
3	1.2	0.05	1	1	0.5	1	0.6	0.5
4	1.1	0.15	1	-2	1.0	1	0.8	0.7
5	1.6	0.40	1	2	1.2	1	1.0	0.9

The priority decreases in i, $1 \leq i \leq 5$.

Five-Class Example



(c) Control solutions



(d) Simulations

Conclusions

Summary

- Propose a time-varying staffing and dynamic scheduling policy for a multi-class V-model;
- Prove asymptotic stability for

$$\mathbb{E}[W_i(t)] \approx w_i, \qquad \mathbb{P}(W_i(t) > w_i) \approx \alpha_i, \qquad 0 < t \le T,\ 1 \le i \le K.$$

- Engineering confirmation via simulations.

Future works

- Differentiate PoD $\mathbb{P}(W_i(t) > 0) \approx \alpha_i$ (QED). Our scheduling rule

$$i^* \in \underset{1 \le i \le K}{\arg\max} \left\{ \frac{H_i^n(t)}{w_i} + \frac{1}{\sqrt{n}}\, \kappa_i(t) \right\} \quad \text{breaks down when } w_i = 0!$$

- Scheduling policies based on other states (e.g., queue length).
- Nonexponential service distributions.
- Multiple service pools.

References

- Liu, Y. (2018). Staffing to stabilize the tail probability of delay in service systems with time-varying demand. Operations Research, 66(2), 514–534.
- Hovey, K., Liu, Y., & Sun, X. (2018). Staffing and Scheduling to Differentiate Service in Multiclass Time-Varying Service Systems. Submitted to Operations Research.
- Ding, Y., Park, E., Nagarajan, M., & Grafstein, E. (2018) Patient Prioritization in Emergency Department Triage Systems: An Empirical Study of Canadian Triage and Acuity Scale (CTAS). Manufacturing & Service Operations Management.
- Eick, S. G., Massey, W. A., & Whitt, W. (1993). The physics of the $M_t/G/\infty$ queue. Operations Research, 41(4), 731–742.
- Feldman, Z., Mandelbaum, A., Massey, W. A., & Whitt, W. (2008). Staffing of time-varying queues to achieve time-stable performance. Management Science, 54(2), 324–338.
- Gurvich, I., Armony, M., & Mandelbaum, A. (2008). Service-level differentiation in call centers with fully flexible servers. Management Science, 54(2), 279–294.

References

Gurvich, I., & Whitt, W. (2010). Service-level differentiation in many-server service systems via queue-ratio routing. Operations research, 58(2), 316–328.

Kim, J., Randhawa, R., Ward, A. R. (2018). Dynamic Scheduling in a Many-Server Multi-Class System: the Role of Customer Impatience in Large Systems. Manufacturing & Service Operations Management, 20(2), 285–301.

Liu, Y., & Whitt, W. (2012). Stabilizing customer abandonment in many-server queues with time-varying arrivals. Operations research, 60(6), 1551–1564.

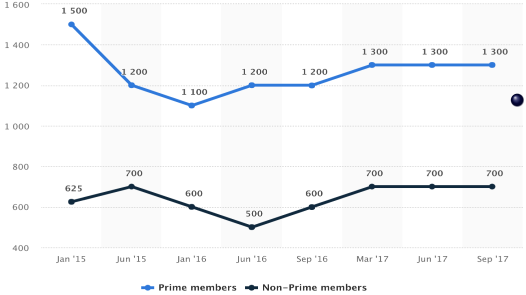
Liu, Y., & Whitt, W. (2014). Many-server heavy-traffic limit for queues with time-varying parameters. The Annals of Applied Probability, 24(1), 378–421.

Pender, J., & Massey, W. A. (2017). Approximating and stabilizing dynamic rate Jackson networks with abandonment. Probability in the Engineering and Informational Sciences, 31(1), 1–42.

Soh, S. B., & Gurvich, I. (2016). Call center staffing: Service-level constraints and index priorities. Operations Research, 65(2), 537–555.

THANK YOU!

Another Motivating Example - Electronic Commerce



- Delivery guarantee
  - Prime member: within 24 hours
  - Regular member: within 4 days

- Non-stationary demand
  - How to determine the fleet size?
  - How to schedule shipment date?

Full Description of the Main Theorem

Suppose the system operates under the TV-SRS staffing and TV-DP scheduling rule. Then there is a joint convergence for the CLT-scaled processes:

$$\left(\widehat{H}_1^n, \dots, \widehat{H}_K^n, \widehat{V}_1^n \dots, \widehat{V}_K^n, \hat{X}_1^n, \dots, \hat{X}_K^n, \hat{Q}_1^n, \dots, \hat{Q}_K^n\right)$$
$$\Rightarrow \left(\widehat{H}_1, \dots, \widehat{H}_K, \widehat{V}_1 \dots, \widehat{V}_K, \hat{X}_1, \dots, \hat{X}_K, \hat{Q}_1, \dots, \hat{Q}_K\right) \quad \text{in} \quad \mathcal{D}^{4K} \quad \text{as} \quad n \to \infty,$$

where all limiting waiting-time processes can be expressed in terms of a one-dimensional process $\widehat{H}(\cdot)$:

$$\widehat{H}_i(t) \equiv w_i(\widehat{H}(t) - \kappa_i(t)), \qquad \widehat{V}_i(t) = w_i(\widehat{H}(t + w_i) - \kappa_i(t + w_i));$$

the process $\hat{H}$ uniquely solves the following *stochastic Volterra equation*

$$\widehat{H}(t) = \int_0^t L(t,s)\widehat{H}(s)ds + \int_0^t J(t,s)dW(s) + K(t),$$

where W is a standard Brownian motion,

$$L(t,s) \equiv \frac{\sum_{i=1}^K \eta_i(s) e^{\mu_i(s-t)} \left(\mu_i - h_{F_i}(w_i)\right)}{\eta(t)}, \quad J(t,s) \equiv \frac{\sqrt{\sum_{i=1}^K e^{2\mu_i(s-t)} (F_i^c(w_i)\lambda_i(s - w_i) + \mu_i m_i(s))}}{\eta(t)},$$

$$K(t) \equiv \frac{\sum_{i=1}^K \left(\eta_i(t)\kappa_i(t) - \int_0^t \eta_i(s) e^{\mu_i(s-t)} \left(\mu_i - h_{F_i}(w_i)\right) \kappa_i(s) ds\right) - c(t)}{\eta(t)}$$

for $\eta_i(t) \equiv w_i \lambda_i(t - w_i) F_i^c(w_i)$ and $\eta(t) \equiv \sum_{i \in \mathcal{I}} \eta_i(t)$.

Functional Weak Law of Large Numbers

The limit for each queue-length process can be decomposed into three terms:

$$\begin{aligned}
\hat{Q}_i(t) &\equiv \hat{Q}_{i,1}(t) + \hat{Q}_{i,2}(t) + \hat{Q}_{i,3}(t) \\
\hat{Q}_{i,1}(t) &\equiv \int_{t-w_i}^{t} F_i^c(t-u)\sqrt{\lambda_i(u)}\mathrm{d}W_{\lambda_i}(u), \\
\hat{Q}_{i,2}(t) &\equiv \int_{t-w_i}^{t} \sqrt{F_i^c(t-u)F_i(t-u)\lambda_i(u)}\mathrm{d}W_{\theta_i}(s), \\
\hat{Q}_{i,3}(t) &\equiv \lambda_i(t-w_i)F_i^c(w_i)\widehat{H}_i(t),
\end{aligned}$$

for $W_{\lambda_i}, W_{\theta_i}, W_{\mu_i}$ being independent standard Brownian motions. Finally, the limits for number in system is given by $\hat{X}_i(t) = \hat{B}_i(t) + \hat{Q}_i(t)$.

As an immediate consequence of the FCLT result, we have

$$\begin{aligned}
&(\bar{B}_1^n, \ldots, \bar{B}_K^n, \bar{Q}_1^n, \ldots, \bar{Q}_K^n, \bar{X}_1^n, \ldots, \bar{X}_K^n, H_1^n, \ldots, H_K^n, V_1^n, \ldots, V_K^n) \\
\Rightarrow &(m_1, \ldots, m_K, q_1, \ldots, q_K, x_1, \ldots, x_K, w_1\mathfrak{e}, \ldots, w_K\mathfrak{e}, w_1\mathfrak{e}, \ldots, w_K\mathfrak{e}) \quad \text{in} \quad \mathcal{D}^{5K}
\end{aligned}$$

as $n \to \infty$ where $\mathfrak{e}$ denotes constant function of one.

Computing $C(t, s)$

Algorithm:

(i) Pick an initial candidate $C^{(0)}(\cdot, \cdot)$;

(ii) In the k^{th} iteration, let $C^{(k+1)} = \Theta\left(C^{(i)}\right)$ with Θ given by

$$\Theta(C_{\widehat{H}})(t, s) = -\int_0^t \int_0^s L(t, u)L(s, v)C_{\widehat{H}}(u, v)dvdu + \int_0^t L(t, u)C_{\widehat{H}}(u, s)du$$
$$+ \int_0^s L(s, v)C_{\widehat{H}}(t, v)dv + \int_0^{s\wedge t} J(t, u)J(s, u)du.$$

Here $\Theta(\cdot)$ is a contraction operator.

(iii) If $\|C^{(k+1)} - C^{(k)}\|_T < \epsilon$, stop; otherwise, $k = k + 1$ and go back to step (ii).

According to the Banach contraction theorem, this algorithm should converge exponentially fast. Finally, we take $\mathrm{Var}(\hat{H}(t)) = C(t, t)$, for $0 \le t \le T$.

Part I - Single Class Model

$M_t/M/s_t + GI$

- Nonhomogenous Process arrivals (easily extendable)
- I.I.D. exponential service times with rate μ (great difficulty arises when extended to general services)
- Time-varying staffing level (TBD)
- I.I.D. abandonment times $\sim F(x) \equiv \mathbb{P}(A \leq x)$ (the $+GI$)
- First-Come First-Served
- Unlimited waiting capacity

Performance functions

- $Q(t)$ and $B(t)$: number in queue and in service at time t
- $X(t) \equiv Q(t) + B(t)$: total number in system at time t
- $V(t)$: potential waiting time at time t

Staffing to Reduce Excessive Delay

- Objective: $\mathbb{P}(V(t) > w) \approx \alpha \in (0,1)$
- Key idea: $V(t)$ is approximately normal for λ and s large (L&W14)
- Propose staffing:
$$s(t) = \lceil m(t) + \tilde{c}(t) \rceil \tag{3}$$
- Detailed Formula:

$$m(t) = F^c(w) \int_0^t e^{-\mu(t-u)} \lambda(u-w) \mathrm{d}u \quad \text{(offered-load process)} \tag{4}$$

$$c(t) = z_\alpha e^{-\mu t} \left(Z(t) - (\mu - h_F(w)) \int_0^t Z(s) ds \right) \tag{5}$$

$$\text{for} \quad Z(t) \equiv e^{(\mu - h_F(w))t} \sqrt{\int_0^t e^{2h_F(w)} \left(F^c(w)\lambda(u-w) + \mu m(u) \right)}, \tag{6}$$

$z_\alpha = \Phi^{-1}(1-\alpha)$ and $h_F(x) \equiv f(x)/F^c(x)$.

- Formula (4) was derived by L&W12 and (5) - (6) came from L18.

Heuristic Derivation of $m(\cdot)$

1. Consider an $M_t/GI/\infty$ model with arrival from time zero. Then number in system $Q(t) \sim$ Poisson r.v. with mean (EMW93)

$$m(t) \equiv \mathbb{E}[Q(t)] = \int_0^t \lambda(u) G^c(t-u)\mathrm{d}u = \int_0^t \lambda(t-s) G^c(s)\mathrm{d}s.$$

   With exponential services we have $G^c(x) = e^{-\mu x}$, and so

$$m(t) = \int_0^t \lambda(u) e^{-\mu(t-u)}\mathrm{d}u. \tag{7}$$

2. If the mean waiting time is stabilized at the target w, then on average a customer (if not abandon) will wait w time units before entering service.
3. Hence the "effective arrival" rate $\tilde{\lambda}(t) \equiv \lambda(t-w)F^c(w)$. Replacing $\lambda(t)$ in (7) with $\tilde{\lambda}(t)$ yields (4), as desired!
4. In summary, the offered load $m(t)$ is the mean number of busy servers needed to serve all customers who are willing to wait (hence excluding an acceptable faction of customer abandoned).

A Five-Class V Model: Parameters and QoS Targets

- Sinusoidal arrival rates $\lambda_i(t) = n\bar{\lambda}_i\,(1 + r_i \sin(\gamma_i t + \phi_i))$;
- Exponential service times;
- Exponential abandonment times;
- Scale $n = 50$.

Class	Arrival Parameters				Abandonment rates	Service rates	Service levels	
	$\bar{\lambda}_i$	r_i	γ_i	ϕ_i	θ_i	μ_i	w_i	α_i
1	1.0	0.20	1	0	0.6	1	0.2	0.1
2	1.5	0.30	1	-1	0.3	1	0.4	0.3
3	1.2	0.05	1	1	0.5	1	0.6	0.5
4	1.1	0.15	1	-2	1.0	1	0.8	0.7
5	1.6	0.40	1	2	1.2	1	1.0	0.9

The priority decreases in i, $1 \leq i \leq 5$.

Five-Class Example

- Goal: Stabilizing mean waiting time $\mathbb{E}[W_i(t)] = w_i$, $(w_1, w_2, w_3, w_4, w_5) = (0.2, 0.4, 0.6, 0.8, 1)$.
- Apply our staffing and scheduling rule with $\alpha_i = 1/2$, $1 \le i \le 5$.