Delay-based Scheduling to Create Service Differentiation in Multiclass Queues

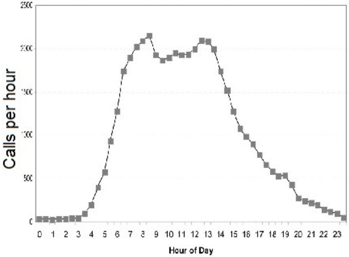
Yunan Liu

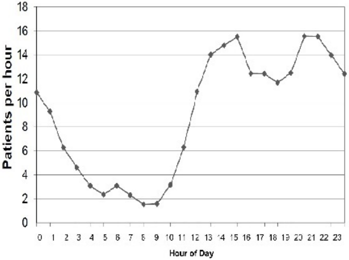
joint work with Kyle Hovey (NCSU) and Xu Sun (Columbia)

Department of Industrial and Systems Engineering, NC State University

May 2019

Time-Varying Arrival Rates





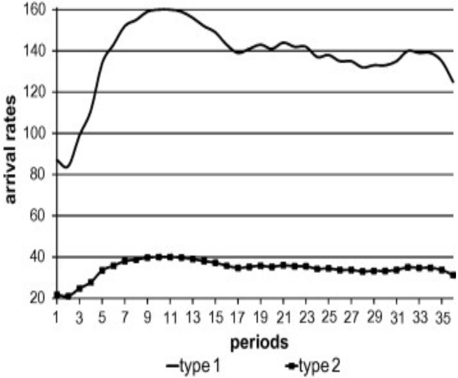
call center

Green et al. (2007)

emergency room

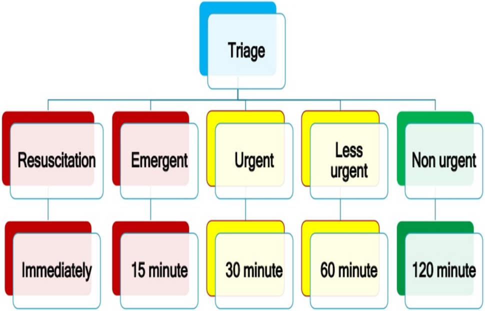
Yom-Tov and Mandelbaum (2011)

Differentiated Service Levels



- 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?

Differentiated Service Levels



Canadian triage and acuity scale (CTAS, 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

$$(w_1, w_2, w_3, w_4, w_5) = (0, 15, 30, 60, 120),$$
$$(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5) = (0.98, 0.95, 0.9, 0.85, 0.8).$$

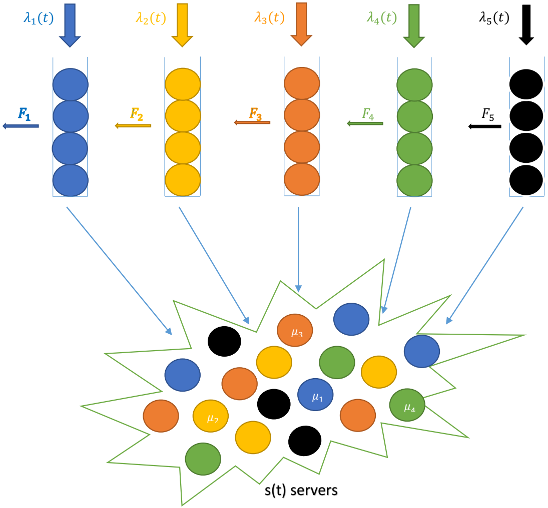
Other Examples of Service Differentiation

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.

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 Scheduling Structure

- Use real-time class-i head-of-line waiting time (HWT) $H_i(t)$ to devise a dynamic control policy;

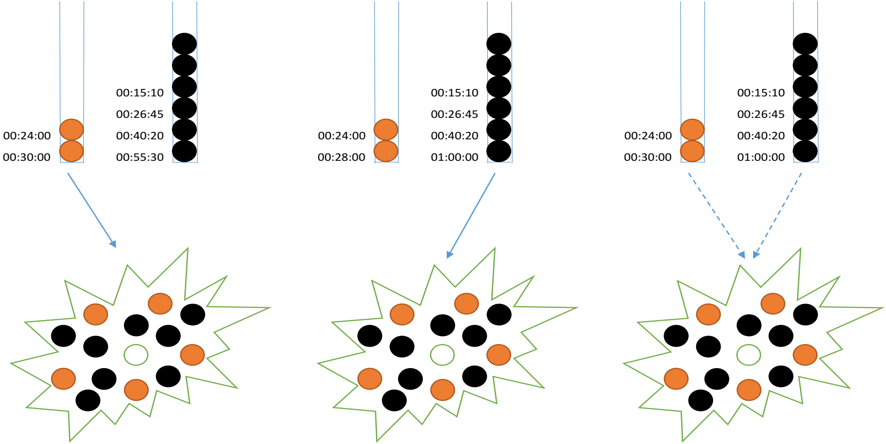
- A delay-based **time-varying dynamic prioritization scheduling** (TV-DPS) rule: Assigns the next available server to the HoL customer from class i^* satisfying

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

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-DPS Rule Works



Goal: HoL delay ratio 1/2

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:
  $$i^* \in \arg\max_{1 \le i \le K} \left\{ \frac{H_i^n(t)}{w_i} + \frac{1}{\sqrt{n}} \; \kappa_i(t) \right\}.$$
  - Service rates and abandonment distributions are fixed.
- HWT and PWT processes:
$$\widehat{H}_i^n(t) \equiv n^{1/2}\left(H_i^n(t) - w_i\right) \quad \text{and} \quad \widehat{W}_i^n(t) \equiv n^{1/2}\left(W_i^n(t) - w_i\right).$$
(Scaled "errors" of HWT and PWT)

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 (SSC): All $2K$ HWT and PWT processes degenerates to a one-dimensional process $\widehat{H}$.

2. The SVE admits a unique (strong) 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 $C_{\widehat{H}}(t,s) \equiv \text{Cov}(\widehat{H}(t), \widehat{H}(s))$;
     - ⋆ Variance $\sigma^2_{\widehat{H}}(t) = C_{\widehat{H}}(t,t)$ relies only on model parameters (independent with control functions);
     - ⋆ Control functions $c(\cdot)$ and $\kappa_i(\cdot)$ appear in the term $K(\cdot)$ only.
   - If $\mu_i = \mu$
     - ⋆ SVE degenerates to a time-varying Ornstein–Uhlenbeck (OU) process;
     - ⋆ SVE has an closed-form solution (so do $m_{\widehat{H}}(t)$ and $C_{\widehat{H}}(t,s)$).

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

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

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

$$\mathbb{P}(H_i^n(t) > w_i) = \mathbb{P}(\widehat{H}_i^n(t) > 0) \overbrace{\approx}^{n \text{ large}} \mathbb{P}(\widehat{H}_i(t) > 0)$$
$$\overbrace{=}^{\text{SSC}} \mathbb{P}(w_i(\widehat{H}(t) - \kappa_i(t)) > 0) \overbrace{=}^{\text{Gaussian dist.}} \mathbb{P}\left(\mathcal{N}\left(m_{\widehat{H}}(t), \sigma^2_{\widehat{H}}(t)\right) > \kappa_i(t)\right)$$
$$= \mathbb{P}\left(\mathcal{N}(0,1) > \frac{\kappa_i(t) - m_{\widehat{H}}(t)}{\sigma_{\widehat{H}}(t)}\right) \overbrace{\equiv}^{\text{want}} \alpha_i.$$

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

- Uniquely obtain the asymptotically "feasible" control functions:

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

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

- Recall: delay-based **time-varying dynamic prioritization scheduling** rule:

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

  where

$$\kappa_i(t) = \underbrace{z_{\alpha_i}}_{\text{quantile}} \cdot \underbrace{\sigma_{\widehat{H}}(t)}_{\text{variability}}, \qquad 1 \le i \le K, \quad 0 \le t \le T.$$

- Some insights:
  - Sign of κ_i: when $\alpha_i \le 0.5$ (> 0.5), then $\kappa_i \ge 0$ (< 0)
    Adding (removing) some weight to make class i more (less) important;
  - Difference of κ_i: if $\alpha_1 < 0.5 < \alpha_2$, then $\kappa_1 > 0 > \kappa_2$, and $\kappa_1(t) - \kappa_2(t) = (z_{\alpha_1} - z_{\alpha_2})\sigma_{\hat{H}}(t)$;
    In a more (less) random system, need more (less) contrasting κ_i's;
  - TV-DPS is both time-dependent (accounting for time variability) and state-dependent (capturing stochasticity).

Step VI: Prove Asymptotic Optimality

The asymptotically "feasible" 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 $c^*(\cdot)$ 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 $0 < t \le T, \ 1 \le i \le K.$

Review of the Approach

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

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

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

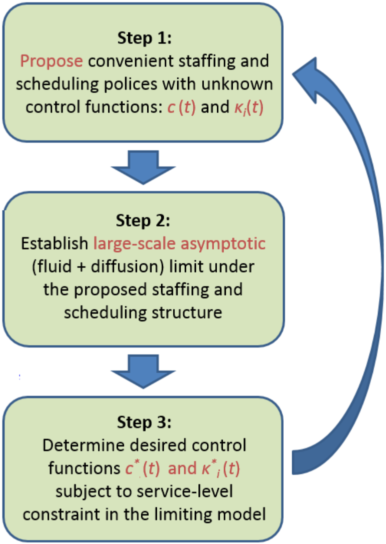
$$\widehat{H}_i(t) \equiv w_i(\widehat{H}(t) - \kappa_i(t)) \qquad \text{SSC}$$

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

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

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

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



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: $n = 50$

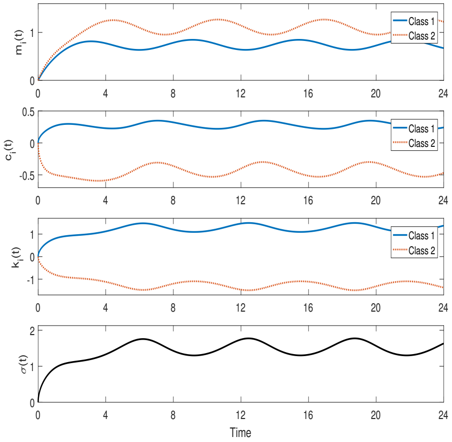
- QoS parameters
  - Delay targets $w_1 = 0.5, w_2 = 1$;
  - Probability targets $\alpha_1 = 0.2, \alpha_2 = 0.8$.

  Hope to achieve:

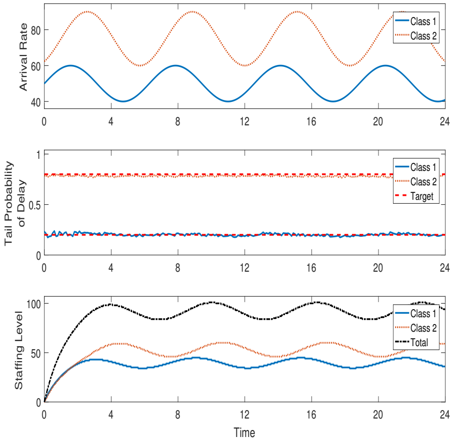
  $\mathbb{P}(W_1(t) > 0.5) \approx 20\%, \quad \mathbb{P}(W_2(t) > 1) \approx 80\%.$ Class-1 more important!

- Quality of service (QoS) increases as α_i and w_i decreases.
- Monte Carlo simulations of $\mathbb{P}(W_i(t) > w_i)$ with 5000 independent runs.

Base Case

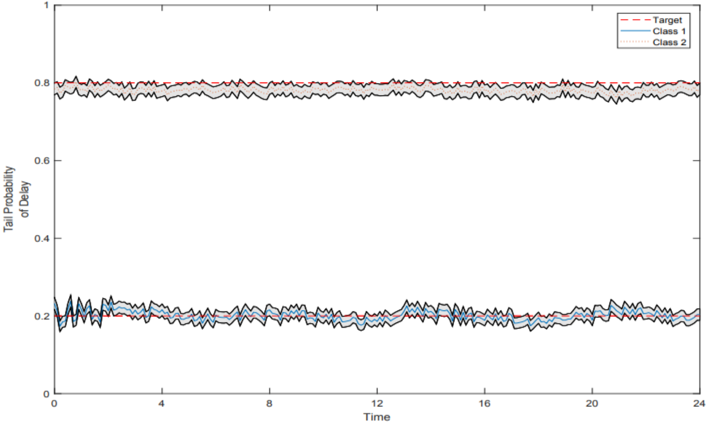


(a) Control functions



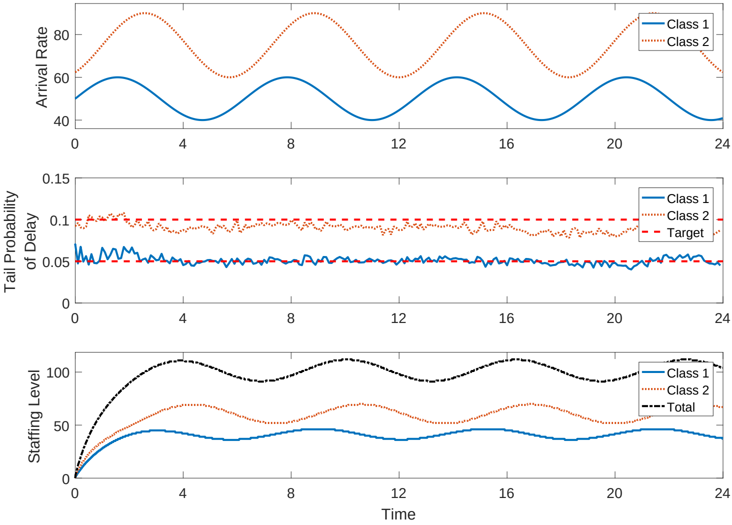
(b) Simulations

Base Case



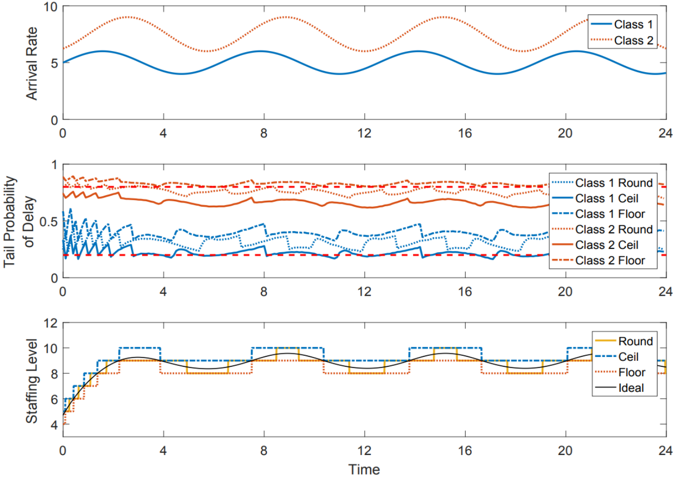
- Monte Carlo estimator: $\widehat{\mathbb{P}(W_i(t_j) > w_i)} \equiv \frac{1}{N}\sum_{k=1}^{N} \mathbf{1}(W_i^k(t_j) > w_i)$, $i = 1, 2$.
- $N = 5000$ samples, 99% confidence intervals, step size $\Delta t \equiv t_{j+1} - t_j = 0.01$.

High Quality of Service (Smaller α_i)



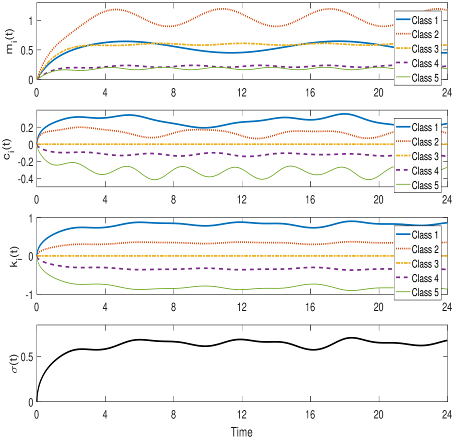
Good performance when $\alpha_i \leq 10\%$.

A Small System ($n = 5$)

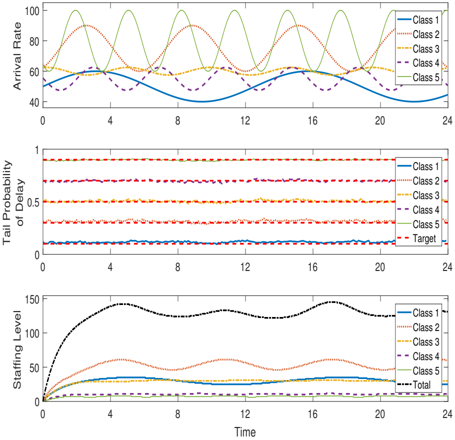


- When n is small, adding/removing a server causes bigger bumps;
- Error is attributed to discretization of staffing levels;
- Ok to apply large-scale results to a smaller scale system.

Five-Class Example



(c) Control solutions



(d) Simulations

Future Directions

- 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 (information retrieval cost, implementation convenience, and server saving).
- Dependent arrival streams (modeling, simulation, and staffing).
- Nonexponential service distributions (non-Brownian driven limits).
- Multiple service pools (skill-based routing).
- Inflexible staffing functions (incorporating costs to change staffing levels).
- Customers switch classes (e.g., patient's health deteriorates/improves).

Queue-Length based Scheduling

- Implementation convenience (discrete counters)
- Information retrieval cost (dimensionality reduction)
- May require fewer servers??
  - QL provide more holistic understanding of current system state and is more forward looking (HoL delays only capture partial information and short-term dynamics); [1]
  - Under QL-based policy, system may be less variable, thus requiring less servers?

[1] Ibrahim, R. and W. Whitt. Real-Time Delay Estimation in Overloaded Multiserver Queues with Abandonment. *Management Science*, 55(10), 2009.

Queue-Length based Scheduling

- Same goal: $\mathbb{P}\left(W_i(t) > w_i\right) \approx \alpha_i$.
- A time-varying square-root staffing (TV-SRS) rule

$$s(t) = \underbrace{m(t)}_{\text{first order}} + \underbrace{\sqrt{\lambda^\star}c(t)}_{\text{second order}}, \qquad c(t) \text{ is TBD}$$

- A queue-length-based scheduling rule:
  Assigns the next available server to the HoL customer from class i^* satisfying

$$i^* \in \arg\max_{1\le i\le K} \left\{ \frac{Q_i^n(t)}{Q_\Sigma^n(t)} - \frac{q_i(t) + n^{-1/2}\hat{q}_i(t)}{q_\Sigma(t) + n^{-1/2}\hat{q}_\Sigma(t)} \right\}.$$

  - $Q_\Sigma(t) \equiv \sum_{i=1}^K Q_i(t)$, $Q_i(t)$ is the class-i waiting queue length.
  - $q_\Sigma(t) = \sum_{i=1}^K q_i(t)$, $q_i(t)$ is the class-i fluid queue length

$$q_i(t) = \int_{t-w_i}^{t} F_i^c(t-u)\lambda_i(u)\mathrm{d}u.$$

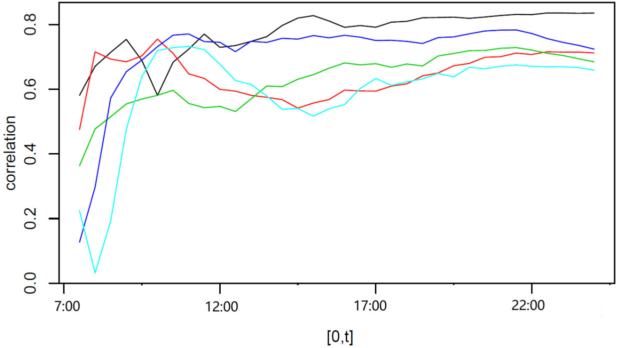
  - $\hat{q}_\Sigma(t) = \sum_{i=1}^K \hat{q}_i(t)$, $\hat{q}_i(t)$ is the second-order regulator (TBD).

Dependent Arrival Streams

Dependent Arrival Streams

- Data source: An American call center;
- Two customer types: VIP and regular;
- Time-dependent (positive) correlation for five weekdays

$$\rho_{1,2}(t) \equiv \frac{Cov(N_1(t), N_2(t))}{\sqrt{Var(N_1(t))Var(N_2(t))}} > 0.$$

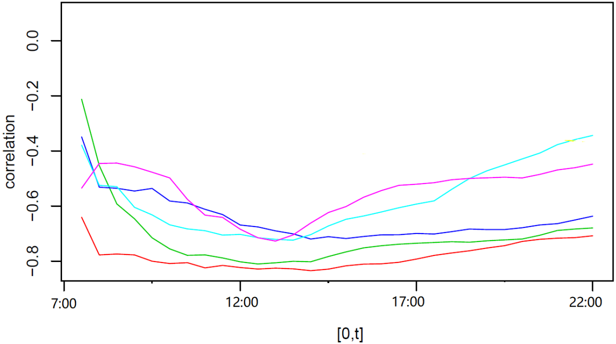


- Explanation: hidden (random) environmental process drives both classes.

Dependent Arrival Streams

- Data source: An Israeli call center;
- Two customer types: different service options;
- Time-dependent (negative) correlation on weekdays and weekends

$$\rho_{1,2}(t) \equiv \frac{Cov(N_1(t), N_2(t))}{\sqrt{Var(N_1(t))Var(N_2(t))}} < 0.$$



- Explanation: two classes share the same (finite) customer pool.

Dependent Arrival Streams

Interesting research questions:

- How to model dependent arrival process (each with time-varying rate)?
  - For example: given parameters $\lambda_i(t)$, $\rho_{i,j}(t)$, $1 \leq i, j \leq m$, how to properly define m-dim process $\{(N_1(t), \ldots, N_m(t)), 0 \leq t \leq T\}$?
- What are relevant statistical procedures to fit the model parameters from real data?
- How to effectively simulate such a model?
- What are the impacts of the dependent arrivals on
  - system performance (e.g., queue length, delay, etc.)
  - operational decisions (e.g., staffing and scheduling policies subject to service-level constraints).
    - ★ For positive/negative correlated arrival streams, should we add/remove some servers to meet a service target?
    - ★ If so, how many should we add and how does this dependent on $\rho_{i,j}$?

THANK YOU!

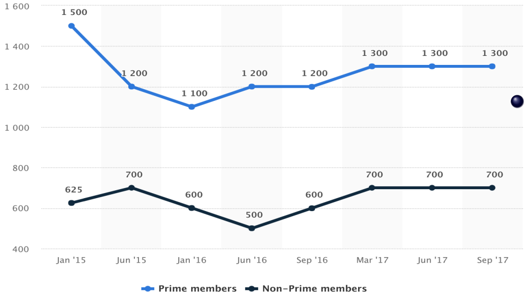
References

- 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. https://yunanliu.wordpress.ncsu.edu/files/2019/01/SLED_OR-20180103.pdf
- 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.

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:

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

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:

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

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

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

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 $\text{Var}(\hat{H}(t)) = C(t,t)$, for $0 \leq t \leq T$.

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

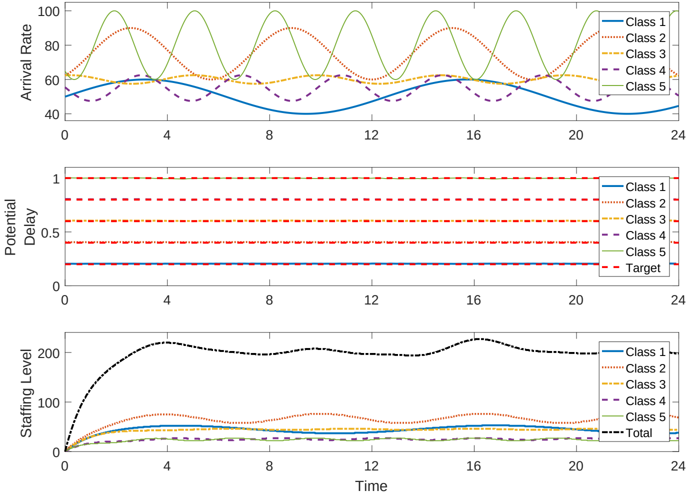
- Sinusoidal arrival rates $\lambda_i(t) = n\bar{\lambda}_i \left(1 + r_i \sin(\gamma_i t + \phi_i)\right)$;
- Exponential service times;
- Exponential abandonment times;
- Scale $n = 50$.

	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

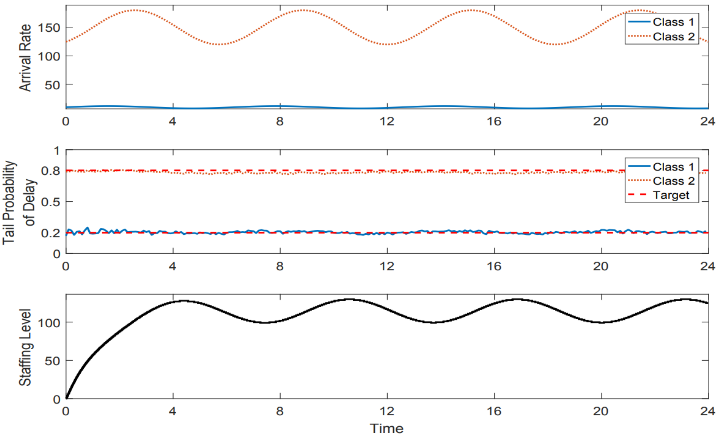
The priority decreases in i, $1 \leq i \leq 5$.

Five-Class Example

- 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 $\alpha_i = 1/2$, $1 \leq i \leq 5$.

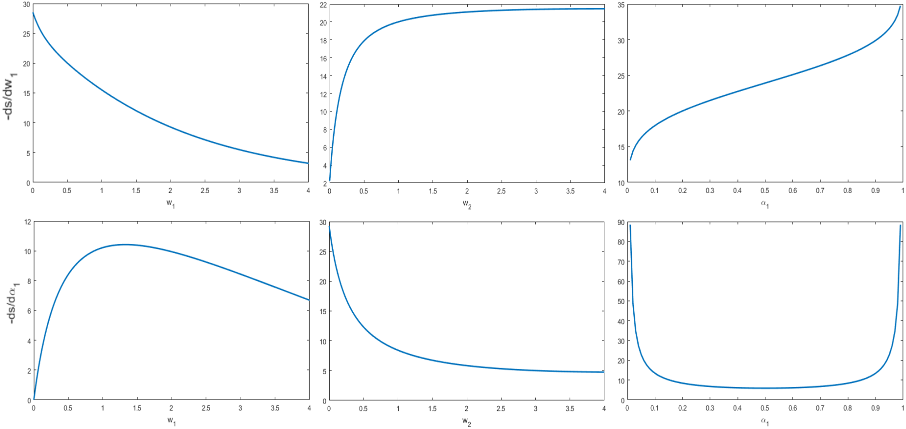


Mixed Arrival



Base Case: Constant Arrival Rate

$\lambda_1 = n = 50$, $\lambda_2 = 1.5n = 75$, $(w_1, w_2, \alpha_1, \alpha_2) = (0.5, 1, 0.2, 0.8)$.

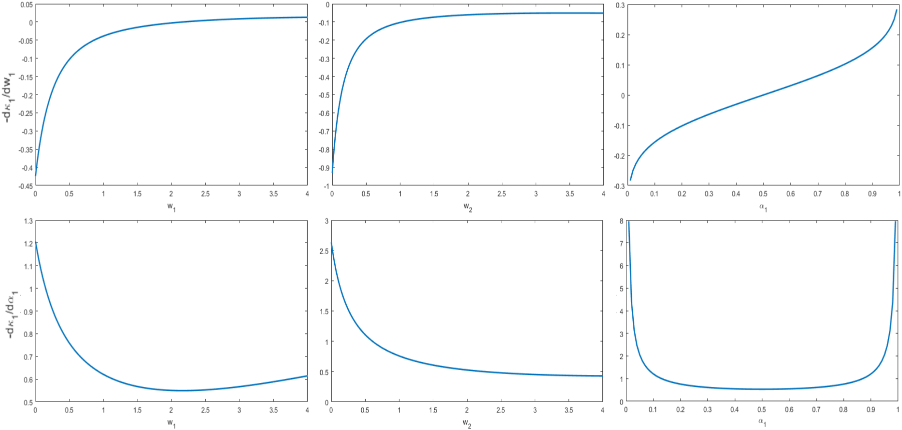


For example, reducing w_1 from 0.5 to 0.4, need to add

$$\Delta s \approx \left(-\frac{\partial s}{\partial w_1}\right) \times 0.1 \approx 20 \times 0.1 = 2.$$

Base Case: Constant Arrival Rate

$\lambda_1 = n = 50$, $\lambda_2 = 1.5n = 75$, $(w_1, w_2, \alpha_1, \alpha_2) = (0.5, 1, 0.2, 0.8)$.

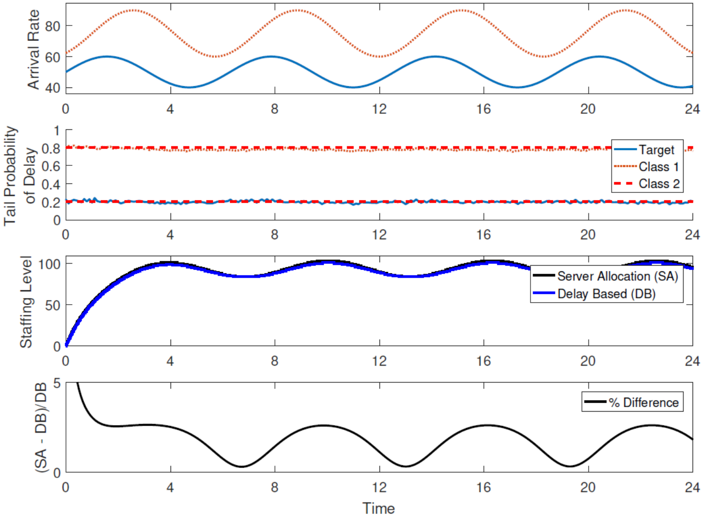


For example, reducing α_1 from 0.2 to 0.15, need to increase κ_1 by

$$\Delta\kappa_1 \approx \left(-\frac{\partial\kappa_1}{\partial\alpha_1}\right) \times 0.05 \approx 0.75 \times 0.05 = 0.0375.$$

Server-Based Scheduling

- Compute $s_1(t)$ and $s_2(t)$ using single-class formulas;
- Schedule customers based on server ratio $s_1(t)/s_2(t)$.



Special Case: Class-Independent Service $\mu_i = \mu$ and Constant Arrival $\lambda_i(t) = \lambda_i$

- 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

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

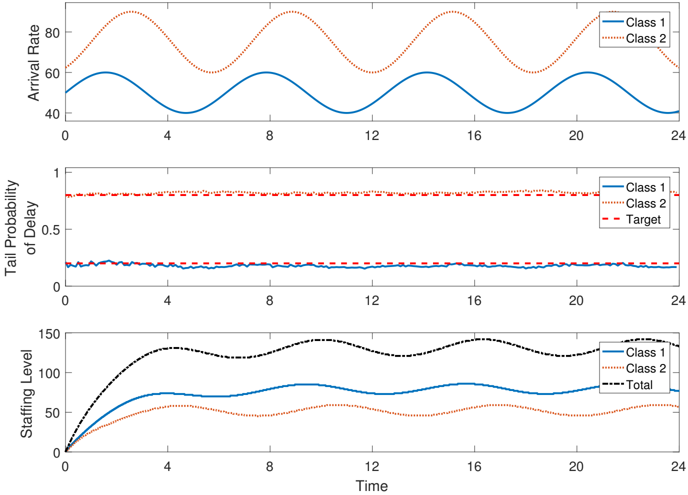
These formulas can be used to estimate

- the required server capacity 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 Δs are needed and how to adjust the scheduling threshold κ_i?

$$\Delta s = \left(-\frac{\partial s(\mathbf{w}, \boldsymbol{\alpha})}{\partial w_i}\right) \Delta w_i, \qquad \Delta \kappa_i = \left(-\frac{\partial \kappa_i(\mathbf{w}, \boldsymbol{\alpha})}{\partial \alpha_i}\right) \Delta \alpha_i,$$

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



- $\sigma^2_{\widehat{H}}(t)$ is numerically computed using our algorithm.
- It takes 42 iterations to converge with $\epsilon = 10^{-6}$.

Many-Server Limit of Waiting Times and SSC

Assumption: Exponential abandonment-time distributions with rate θ_i.

Under SRS staffing and QL-based scheduling:

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

with total queue length $\hat{\mathbf{X}} = \left(\hat{X}_1, \ldots, \hat{X}_K\right)$ solving a K-dim time-varying OU SDE

$$\mathrm{d}\hat{\mathbf{X}}(t) = \mathbf{A}(t)\hat{\mathbf{X}}(t)\mathrm{d}t + \mathbf{B}(t)\mathrm{d}\boldsymbol{\mathcal{W}}(t) + \mathbf{C}(t)dt, \qquad \hat{X}_\Sigma(t) = \sum_{i=1}^{K} \hat{X}_i(t),$$

$$\widehat{H}_i(t) = \frac{e^{\theta_i w_i}}{\lambda_i(t)}\left[\left(p_i(t)(\hat{X}_\Sigma(t) - c(t) - \hat{q}_\Sigma(t)) + \hat{q}_i(t)\right) - \int_{t-w_i}^{t} e^{-\theta_i(t-u)}\sqrt{\lambda_i(t-u)}\mathrm{d}W_{\lambda_i}(u) - \int_{t-w_i}^{t} \sqrt{e^{-\theta_i(t-u)}(1-e^{-\theta_i(t-u)})}\sqrt{\lambda_i(t-u)}\mathrm{d}W_{\theta_i}(u)\right].$$

Many-Server Limit of Waiting Times and SSC

$$\mathrm{d}\hat{\mathbf{X}}(t) = \mathbf{A}(t)\hat{\mathbf{X}}(t)\mathrm{d}t + \mathbf{B}(t)\mathrm{d}\boldsymbol{\mathcal{W}}(t) + \mathbf{C}(t)dt,$$

where $\hat{\mathbf{X}}(t) \equiv [\hat{X}_1(t), \ldots, \hat{X}_K(t)]^T$,

$$\mathbf{A}(t) \equiv -\begin{bmatrix} \mu_1 & & \\ & \ddots & \\ & & \mu_K \end{bmatrix} + \begin{bmatrix} (\mu_1-\theta_1)p_1(t) & \cdots & (\mu_1-\theta_1)p_1(t) \\ \vdots & \vdots & \vdots \\ (\mu_K-\theta_K)p_K(t) & \cdots & (\mu_K-\theta_K)p_K(t) \end{bmatrix},$$

$$\mathbf{B}(t) \equiv \begin{bmatrix} \sqrt{\lambda_1(t)+\mu_1 m_1(u)+\theta_1 q_1(u)} & & \\ & \ddots & \\ & & \sqrt{\lambda_m(t)+\mu_K m_K(u)+\theta_K q_K(u)} \end{bmatrix},$$

$$\mathbf{C}(t) \equiv -(c(t)+\hat{q}_\Sigma(t)) \begin{bmatrix} (\mu_1-\theta_1)p_1(t) \\ \vdots \\ (\mu_K-\theta_K)p_K(t) \end{bmatrix} + \begin{bmatrix} (\mu_1-\theta_1)\hat{q}_1(t) \\ \vdots \\ (\mu_K-\theta_K)\hat{q}_K(t) \end{bmatrix},$$

$$p_i(t) \equiv \frac{q_i(t)}{q_\Sigma(t)},$$

and $\boldsymbol{\mathcal{W}}(t) \equiv [\mathcal{W}_1(t), \ldots, \mathcal{W}_K(t)]^T$ is a K-dim independent BM.

Many-Server Limit of Waiting Times and SSC

Next, matching the TPoD to desired probability target

$$\mathbb{P}(H_i^n(t) > w_i) = \mathbb{P}(\widehat{H}_i^n(t) > 0) \approx \mathbb{P}(\widehat{H}_i(t) > 0) = \cdots = \alpha_i,$$

yields "unique" control functions

- scheduling regulators

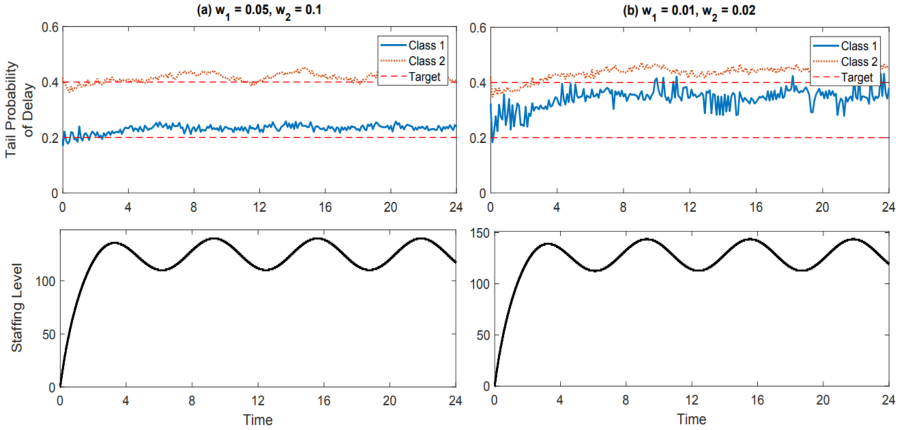
$$\hat{q}_i(t) = z_{\alpha_i}\ \tilde{\sigma}(t), \qquad 1 \le i \le K.$$

- our desired second-order staffing function c^* solves a *fixed-point equation*

$$c = \Psi(c), \quad \text{where} \quad \Psi(c)(t) \equiv \beta(t) + \mathbf{e} \int_0^t e^{\int_u^t \mathbf{A}(v)\mathrm{d}v} \mathbf{C}(u)\mathrm{d}u.$$

Remark: When $\mu_i = \mu$, $c(t)$ has a closed-form solution.

Very High Quality of Service (Smaller w_i)



- If $w_i \approx 0$, TPoD degenerates to probability of delay (PoD) $\mathbb{P}(W_i(t) > 0)$.
- Our method will eventually break down...

Fixed Staffing Intervals

Average staffing level (ASL) and *max staffing level* (MSL)

$$s^{\mathrm{ASL}}(t) \equiv \sum_{i=1}^{\lceil T/\Delta_s \rceil} \bar{s}_i \mathbf{1}_{\{t \in [(i-1)\Delta_s, i\Delta_s\}}, \qquad \bar{s}_i \equiv \frac{1}{\Delta_s} \int_{(i-1)\Delta_s}^{i\Delta_s \wedge T} s(u)\mathrm{d}u,$$

$$s^{\mathrm{MSL}}(t) \equiv \sum_{i=1}^{\lceil T/\Delta_s \rceil} s_i^{\uparrow} \mathbf{1}_{\{t \in [(i-1)\Delta_s, i\Delta_s\}}, \qquad s_i^{\uparrow} \equiv \sup_{(i-1)\Delta_s \le u \le i\Delta_s \wedge T} s(u),$$