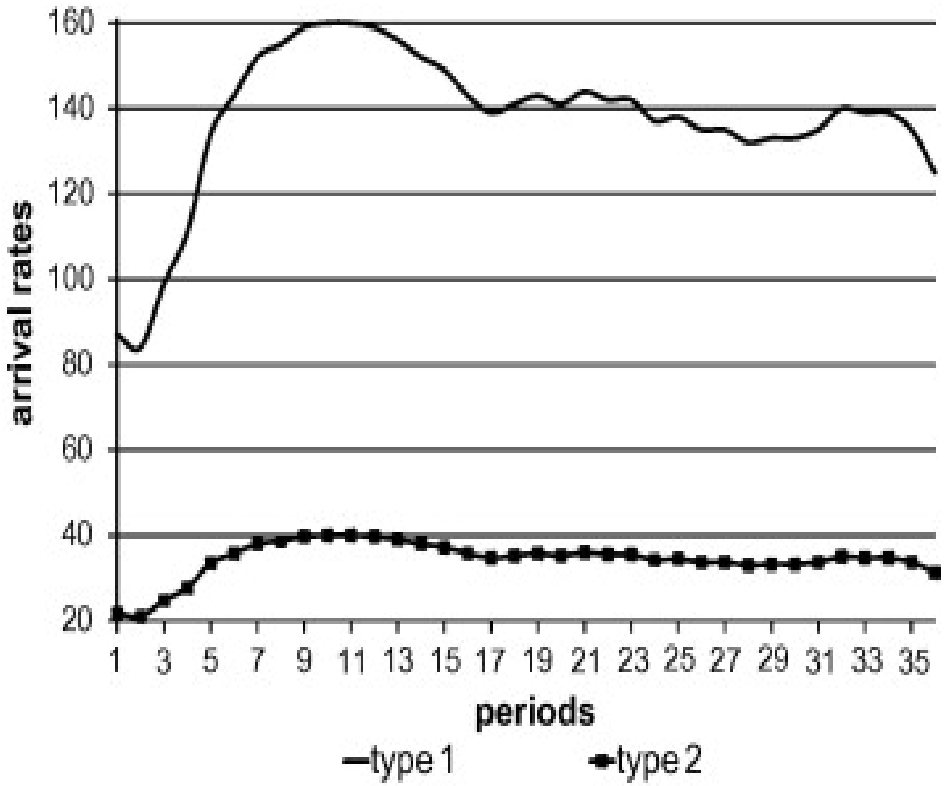
Staffing and Scheduling to Achieve Service Differentiation in Multi-Class Service Systems with Time-Varying Demand

Xu Sun
(with advisor Ward Whitt)

Dept. of Industrial Engineering & Operations Research
Columbia University

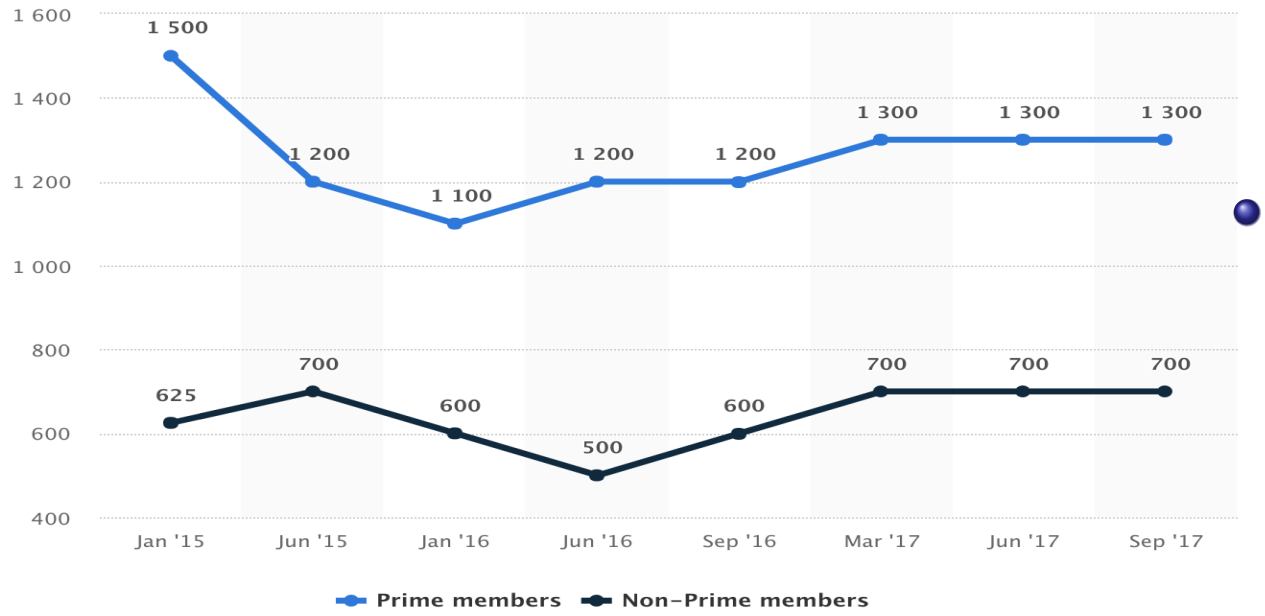
September 22, 2018

Motivating Example 1 - Call Center



- 80% of type 1 calls need to be answered within 20 seconds
- 80% 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?

Motivating Example 2 - Electronic Commerce

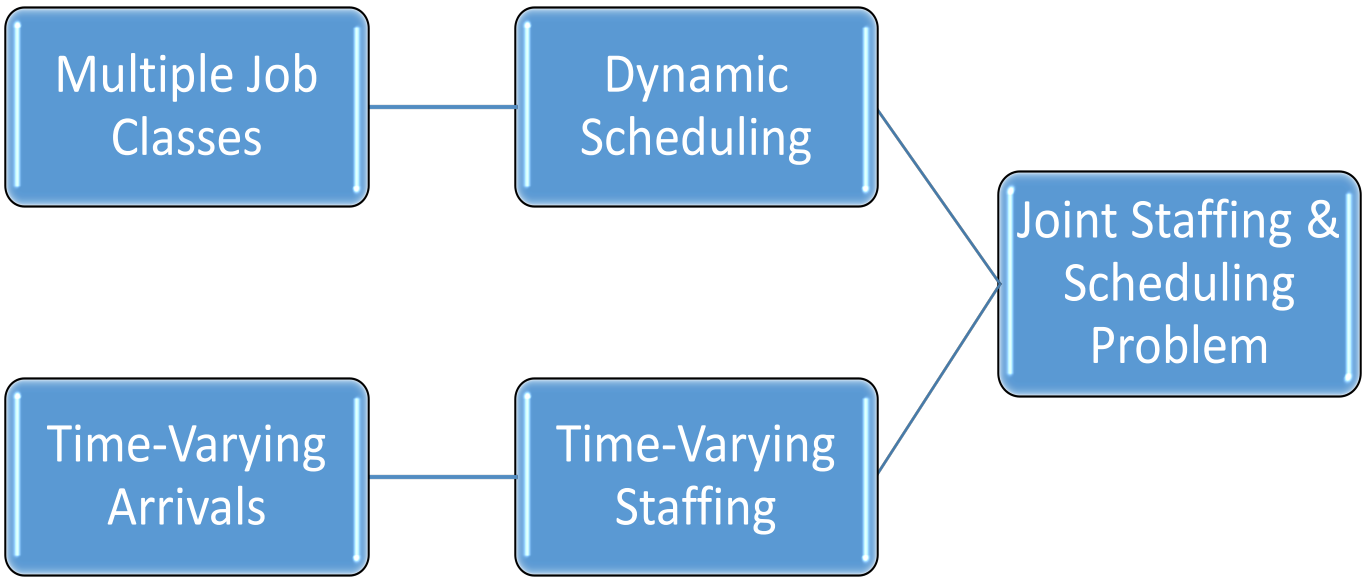


- Delivery guarantee
  - Prime member: delivered within 24 hours
  - regular member: delivered within 5 days

- Non-stationary demand
  - How many fleet vehicles are needed?
  - How to schedule shipment date?

Common Features

- Multiple customer/job classes - need to make scheduling decisions (assigning a newly idle server to one of the classes)
- Time-varying (TV) demand - need to determine the service capacity over time so as to match it with service demand



Our Contribution

- Dynamic scheduling with many servers

  cost minimization - Harrison & Zeevi (2004); Atar, Mandelbaum & Reiman (2004); Atar (2005)
  service differentiation - Gurvich, Armony & Mandelbaum (2008); Gurvich & Whitt (2010)
  All assume the demand to be stationary.

- Performance analysis of TV queues

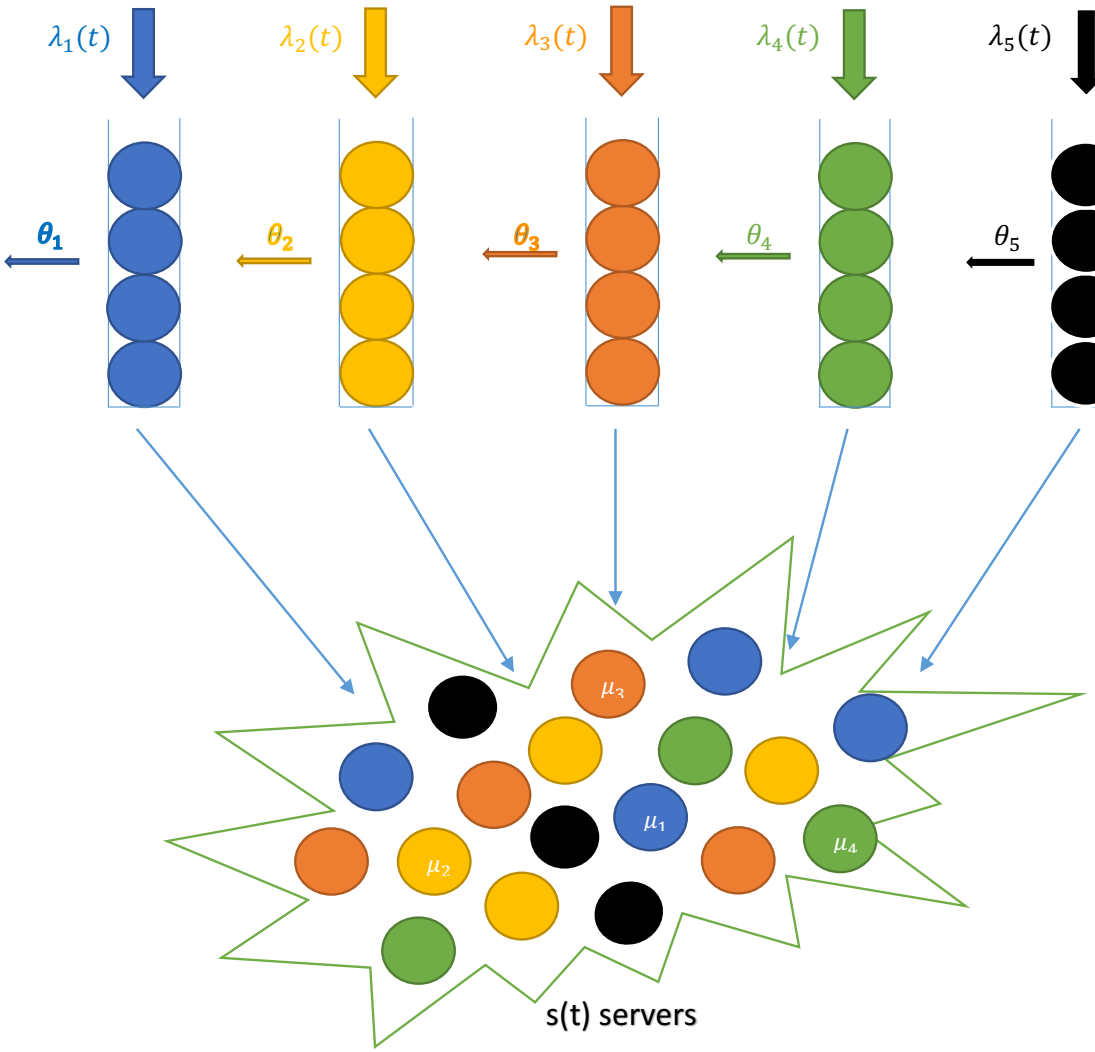
  Koopman (1972); Rothkopf and Oren (1979); Taaffe and Ong (1987); Nelson and Taaffe (2004a, b); Kelly (1991); Massey and Pender (2013).

- Staffing TV queues

  Jennings, Mandelbaum, Massey & Whitt (1996); Green, Kolesar & Whitt; Feldman, Mandelbaum, Massey & Whitt (2008); Liu & Whitt (2012); Liu (2018)
  All consider single-class staffing problems

- First paper studying service differentiation with TV demand

The Model



- Class-dependent arrival rate $\lambda_i(t)$ (non-homogeneous Poisson)
- Exponential patience times with class-dependent abandonment rate θ_i
- TV staffing level $s(t)$
- Exponential service times with class-dependent service rate μ_i
- First-Come First-Serve within each class

Problem Statement

- $V_i(t)$ - waiting time of *a virtual customer* of class i arriving at time t
- $w_i(t)$ - delay target for class i customers at time t
- s(t) - number of servers or staffing level at time t
- π scheduling rule

mean-waiting-time formulation

$$\min_{\pi,s} \int_0^T s(t)\mathrm{d}t$$
$$\text{s.t. } \mathbb{E}[V_i(t)] \leq w_i(t)$$
$$\text{for} \quad t \leq T, \quad i \in \mathcal{I}$$

tail-probability formulation

$$\min_{\pi,s} \int_0^T s(t)\mathrm{d}t$$
$$\text{s.t. } \mathbb{P}\left(V_i(t) > w_i(t)\right) \leq \alpha$$
$$\text{for} \quad t \leq T, \quad i \in \mathcal{I}$$

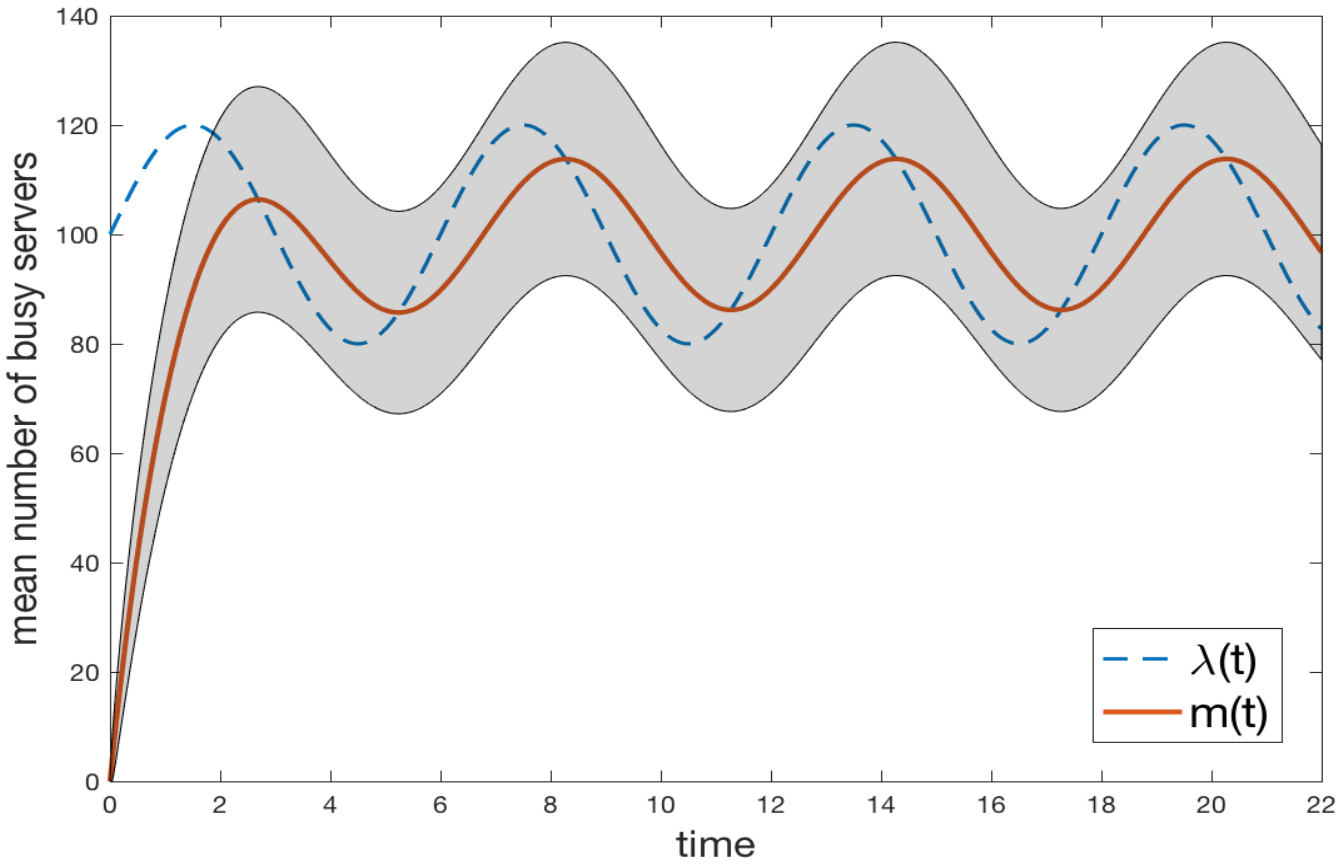
Reviewing Single-Class Staffing Problem

- Only one class of customers: single-class staffing problem
- How to set staffing requirements?
  - pretend that there is no capacity constraint and look at the number of busy servers in system; i.e., consider an $M_t/GI/\infty$ model
- The number of busy servers at time t, denoted by $X(t)$, follows a Poisson distribution with mean

$$m(t) = \int_0^t \lambda(s) G^c(t-s) \mathrm{d}s,$$

where $G^c(x) \equiv 1 - G(x)$ is the complementary cumulative distribution function of the service time. The function $m(t)$ is also called the offered-load process.

Insights From the $M_t/GI/\infty$ Model



- Time Lag in Congestion: customer delays peak after the arrival-rate peaks
- Square-Root-Staffing (SRS) rule: $s(t) = \lceil m(t) + \tilde{c}(t)\sqrt{m(t)} \rceil$

Extending SRS to Multiple Classes

1. Calculate the offer-load processes

$$\dot{m}_i(t) = \lambda_i(t) - \mu_i m_i(t) \quad \text{for} \quad i = 1, \ldots, K.$$

2. Compute the aggregate offered-load process $m(t) \equiv \sum_{i=1}^{K} m_i(t)$.

3. Set the staffing function

$$s(t) = \lceil m(t) + \tilde{c}(t)\sqrt{m(t)} \rceil, \quad t \geq 0$$

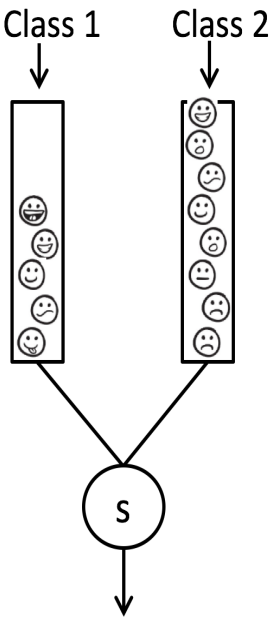
where $\tilde{c}$ is a "design function" and will be determined by the prescribed performance targets.

Scheduling

minimize staffing

s.t. $\mathbb{P}\{V_1(t) > w_1\} \leq \alpha$

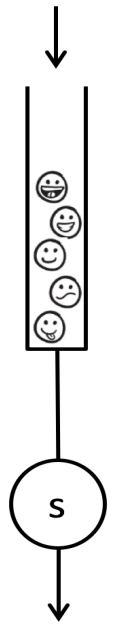
$\mathbb{P}\{V_2(t) > w_2\} \leq \alpha$



$\Rightarrow$

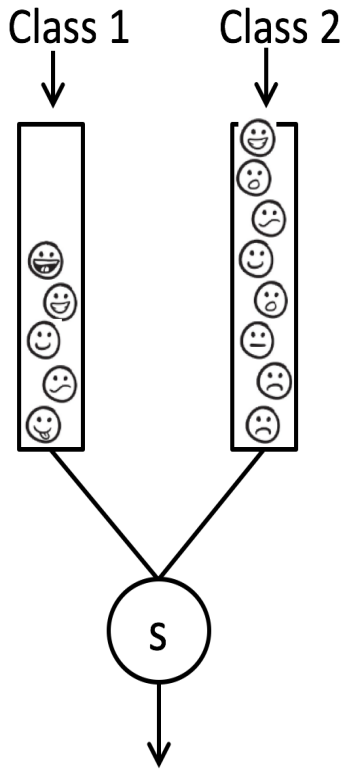
minimize staffing

s.t. $\mathbb{P}\{Q(t) > q(t)\} \leq \alpha$



Example

- Two-class $M_t/M/s_t + M$ queue
- Arrival-rate functions: $\lambda_1(t) = 60 - 20\sin(2t/5)$ and $\lambda_2 = 90 + 30\sin(2t/5)$
- Common service rate $\mu = 1$ and abandonment rate $\theta = 1$
- Target delays $w_1 = 1/6$ and $w_2 = 1/3$
- Service-level constraints: $\mathbb{P}(V_i(t) > w_i) \leq \alpha$, $i = 1, 2$



Motivation for Queue-Ratio-Control Rule

- Events $\{V_1(t) > w_1\}$ and $\{V_2(t) > w_2\}$ are equivalent if

$$V_1(t)/V_2(t) \approx w_1/w_2. \tag{1}$$

- Suppose that a TV Little's law $Q_i(t) \approx \lambda_i(t)V_i(t)$ holds. Then in order to achieve the delay ratio given in (1), one would want

$$\frac{Q_1(t)}{Q_2(t)} \approx \frac{\lambda_1(t)V_1(t)}{\lambda_2(t)V_2(t)} \approx \frac{\lambda_1(t)w_1}{\lambda_2(t)w_2},$$

  or equivalently

$$\frac{Q_i(t)}{Q_1(t)+Q_2(t)} \approx \frac{\lambda_i(t)w_i}{\lambda_1(t)w_1+\lambda_2(t)w_2} \equiv r_i(t).$$

- Make assignments in such a way that a desired queue ratio is maintained.

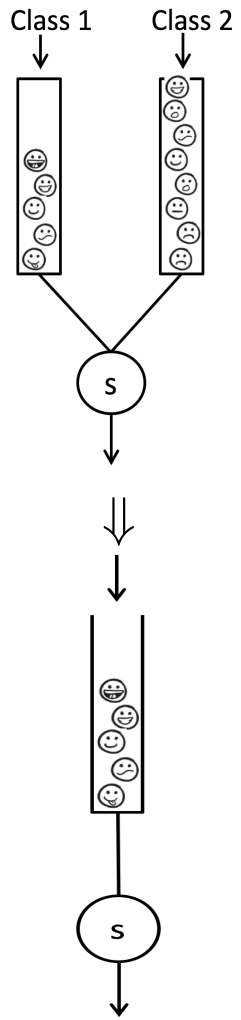
Time-Varying Queue-Ratio (TVQR)

- Given ratio functions $r_1(t), r_2(t)$ $(r_1(t) + r_2(t) = 1)$: aim for

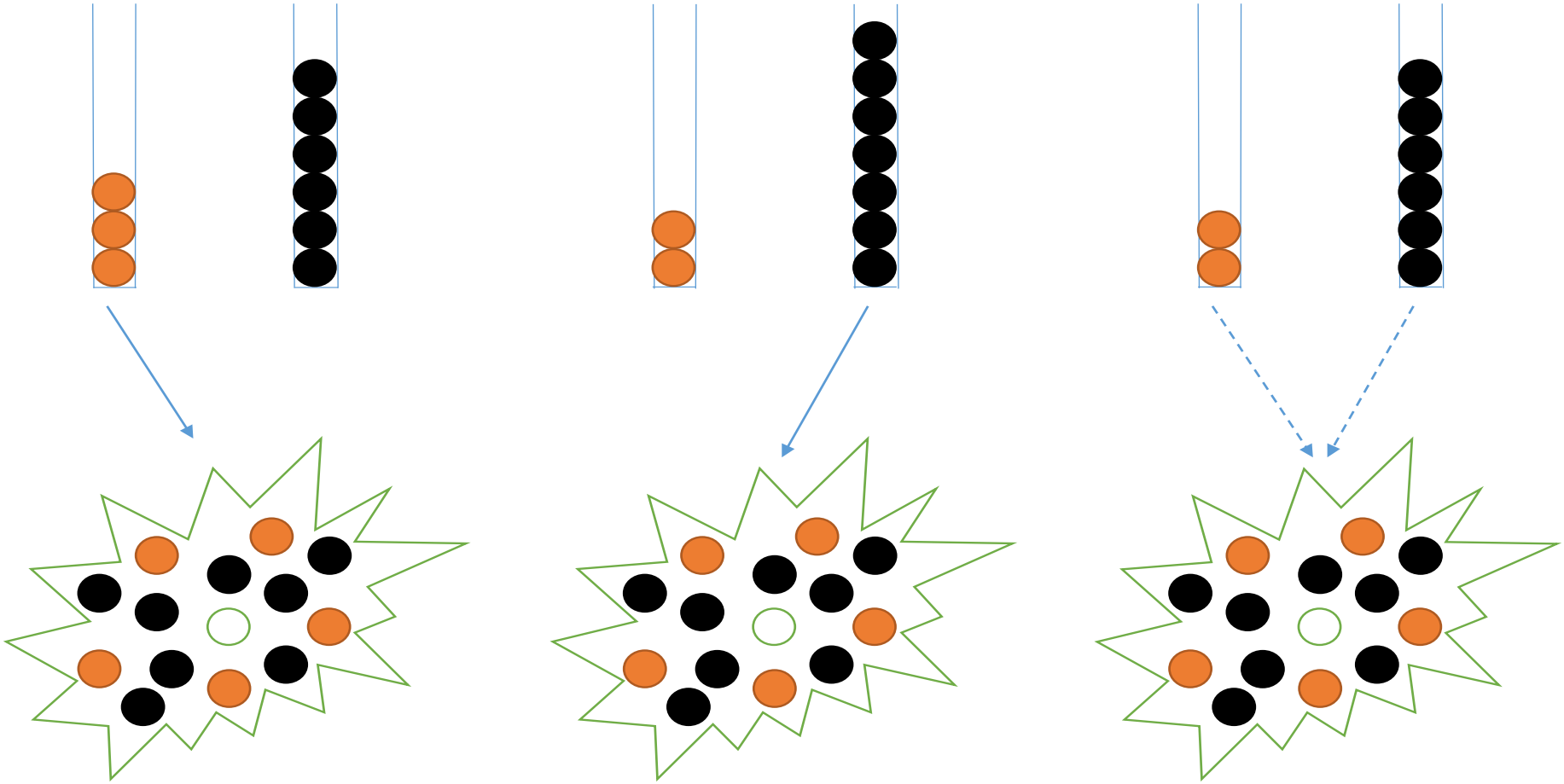
$$\frac{Q_i(t)}{Q_1(t) + Q_2(t)} \approx r_i(t)$$

- Serve class i with greatest queue imbalance

$$\frac{Q_i(t)}{Q_1(t) + Q_2(t)} - r_i(t)$$



TVQR - An Illustration



Goal: queue ratio 1/3

Simplification Through Ratio-Control Rules

$$\mathbb{P}\{V_i(t) > w_i\} \approx \mathbb{P}\{Q_i(t) > \lambda_i(t) w_i\} \quad [\text{TV Little's Law } Q_t(t) \approx \lambda_i(t) V_i]$$

$$\approx \mathbb{P}\left\{ r_i(t) \left[\sum_{k=1}^{2} Q_k(t) \right] > \lambda_i(t) w_i \right\} \quad \left[\frac{Q_i(t)}{Q_1(t) + Q_2(t)} \approx r_i(t) \right]$$

$$= \mathbb{P}\left\{ \left[\sum_{k=1}^{2} Q_k(t) \right] > \sum_{k=1}^{2} \lambda_i(t) w_i \right\} \quad \left[r_i(t) = \frac{\lambda_i(t) w_i}{\lambda_1(t) w_1 + \lambda_2(t) w_2} \right]$$

$$\approx \mathbb{P}\{Q(t) > q(t)\} \qquad [q(t) = \lambda_1(t) w_1 + \lambda_2(t) w_2]$$

Proposed Solution via TVQR

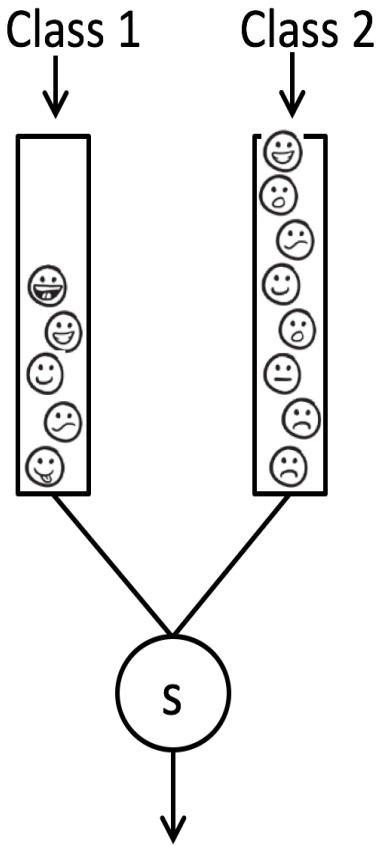
Fundamental idea: decouple staffing and scheduling

- Staffing: for $q(t) \equiv \lambda_1(t)w_1 + \lambda_2(t)w_2$, *choose a staffing function* $s(t)$ that makes

$$\mathbb{P}\{Q(t) > q(t)\} = \alpha.$$

- Scheduling: use the TVQR rule with ratio functions

$$r_i(t) = \frac{\lambda_i(t)w_i}{\lambda_1(t)w_1 + \lambda_2(t)w_2}, \quad i = 1, 2.$$



Many-Server Heavy-Traffic (MSHT) Analysis

- Since exact analysis is difficult, we do asymptotic analysis as scale grows (realistic for large-scale systems).
- Consider a sequence of systems by n.
- Service and abandonment rates are fixed; service demand and capacity grow: $\lambda_i^n(t) \equiv n\lambda_i(t)$, so that the offered load $m_i^n(t) = nm_i(t)$ in model n.
- Staffing function satisfies the SRS formula:

$$s^n(t) = m^n(t) + \tilde{c}(t)\sqrt{m^n(t)} = nm(t) + \sqrt{n}c(t) \quad \text{for} \quad c(t) \equiv \tilde{c}(t)\sqrt{m(t)}.$$

- HT scaling for the number-in-system processes:

$$\hat{X}_i^n(\cdot) \equiv n^{-1/2}\left(X_i^n(\cdot) - nm_i(\cdot)\right) \quad \text{and} \quad \hat{X}^n(\cdot) \equiv n^{-1/2}\left(X^n(\cdot) - nm(\cdot)\right).$$

- HT scaling for the queue-length and delay processes as well as the target delays

$$\hat{Q}_i^n(\cdot) \equiv n^{-1/2}Q_i^n(\cdot), \quad \hat{V}_i^n(t) \equiv n^{1/2}V_i^n(t) \quad \text{and} \quad w_i^n \equiv n^{-1/2}w_i.$$

Theorem (MSHT Limits for TVQR)

Suppose that the system uses SRS and operates under the TVQR scheduling rule. Then we have the joint convergence

$$\begin{aligned}&\left(\hat{X}_1^n, \ldots, \hat{X}_K^n, \hat{Q}_1^n, \ldots, \hat{Q}_K^n, \hat{V}_1^n, \ldots, \hat{V}_K^n\right)\\ &\Rightarrow \left(\hat{X}_1, \ldots, \hat{X}_K, \hat{Q}_1, \ldots, \hat{Q}_K, \hat{V}_1, \ldots, \hat{V}_K\right)\end{aligned} \tag{2}$$

in $\mathcal{D}^{3K}$ *where the diffusion limits* $\hat{X}_i(\cdot)$ *satisfy*

$$\begin{aligned}\hat{X}_i(t) = \hat{X}_i(0) &- \mu_i \int_0^t \hat{X}_i(u)\mathrm{d}u\\ &- (\theta_i - \mu_i)\int_0^t r_i(u)\left[\hat{X}(u) - c(u)\right]^+ \mathrm{d}u + \int_0^t \sqrt{\lambda_i(u) + \mu_i m_i(u)}\mathrm{d}W_i(u)\end{aligned} \tag{3}$$

where the one-dimensional process $\hat{X}(\cdot) \equiv \sum_{i=1}^K \hat{X}_i(\cdot)$ *and* $W_i(\cdot)$ *are standard Brownian motions. For each* $i \in \mathcal{I}$

$$\hat{Q}_i(\cdot) \equiv r_i(\cdot)\left[\hat{X}(\cdot) - c(\cdot)\right]^+, \quad \textit{and} \quad \hat{V}_i(\cdot) \equiv \frac{r_i(\cdot)}{\lambda_i(\cdot)} \cdot \left[\hat{X}(\cdot) - c(\cdot)\right]^+. \tag{4}$$

Important Insights

- TV Little's law

$$Q_i^n(t) \approx \lambda_i^n(t) V_i^n(t)$$

- State-space collapse

$$\frac{Q_i^n(t)}{Q_1^n(t) + Q_2^n(t)} \approx r_i(t)$$

- Let $q(t) \equiv \lambda_1(t) w_1 + \lambda_2(t) w_2$. By the theorem

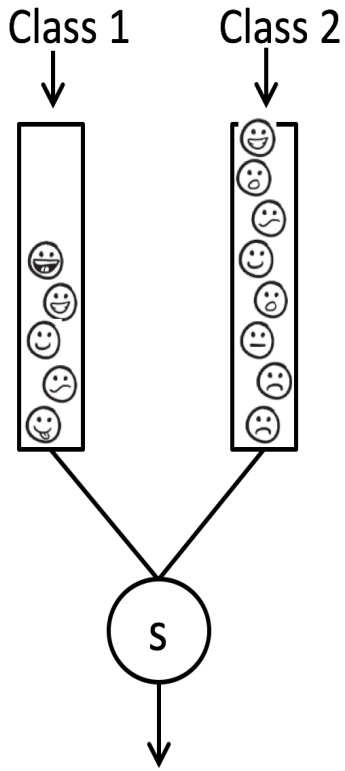
$$\mathbb{P}\{V_i^n(t) > w_i^n\} \to \mathbb{P}\left\{\hat{Q}(t) > q(t)\right\} = \mathbb{P}\left\{\hat{X}(t) > c(t) + q(t)\right\}.$$

- Thus, to make $\mathbb{P}\{V_i^n(t) > w_i^n\} \leq \alpha$, if suffices to choose $c(t)$ that satisfies

$$\mathbb{P}\left\{\hat{X}(t) > c(t) + q(t)\right\} = \alpha.$$

Example

- Two-class $M_t/M/s_t + M$ queue
- Arrival-rate functions: $\lambda_1(t) = 60 - 20\sin(2t/5)$ and $\lambda_2 = 90 + 30\sin(2t/5)$
- Common service rate $\mu = 1$ and abandonment rate $\theta = 1$
- Target delays $w_1 = 1/6$ and $w_2 = 1/3$
- Service-level constraints: $\mathbb{P}(V_i(t) > w_i) \leq \alpha$, $i = 1, 2$



Numerical Test I (TVQR)

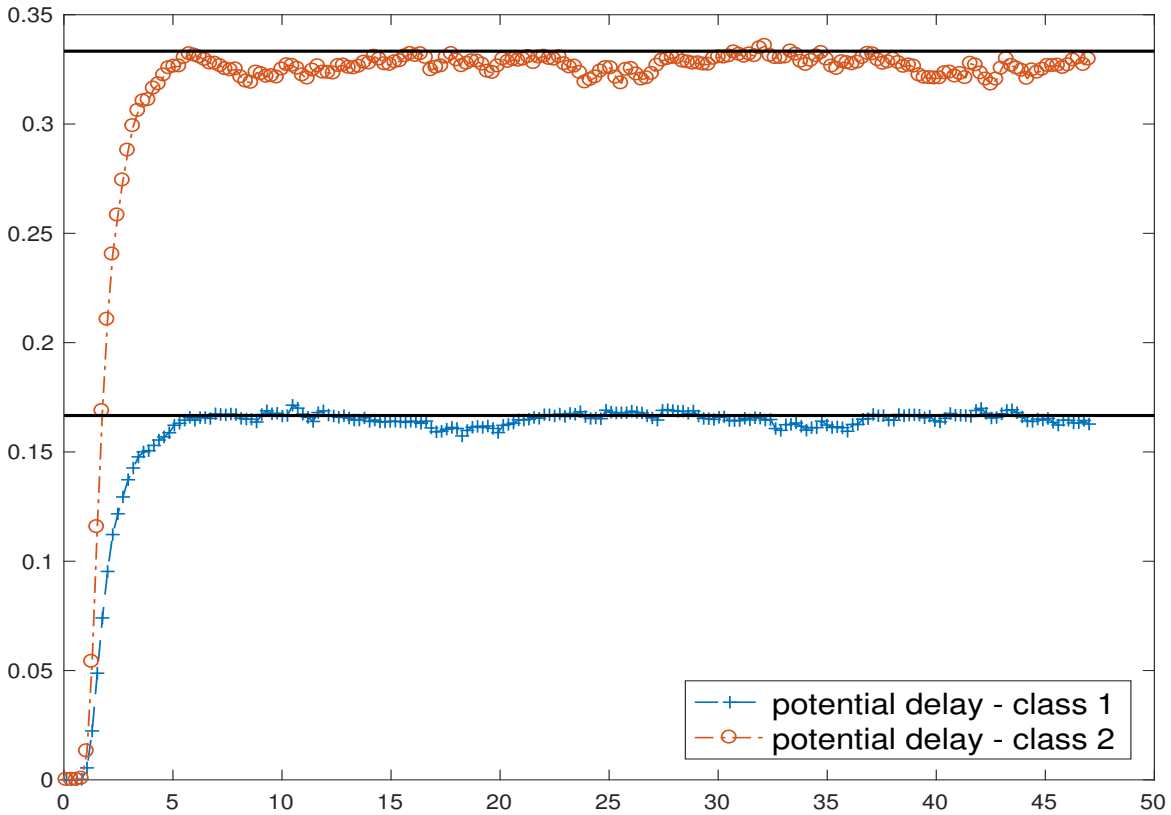


Figure: Base case with the mean-waiting-time formulation and TVQR rule.

Numerical Test II (TVQR)

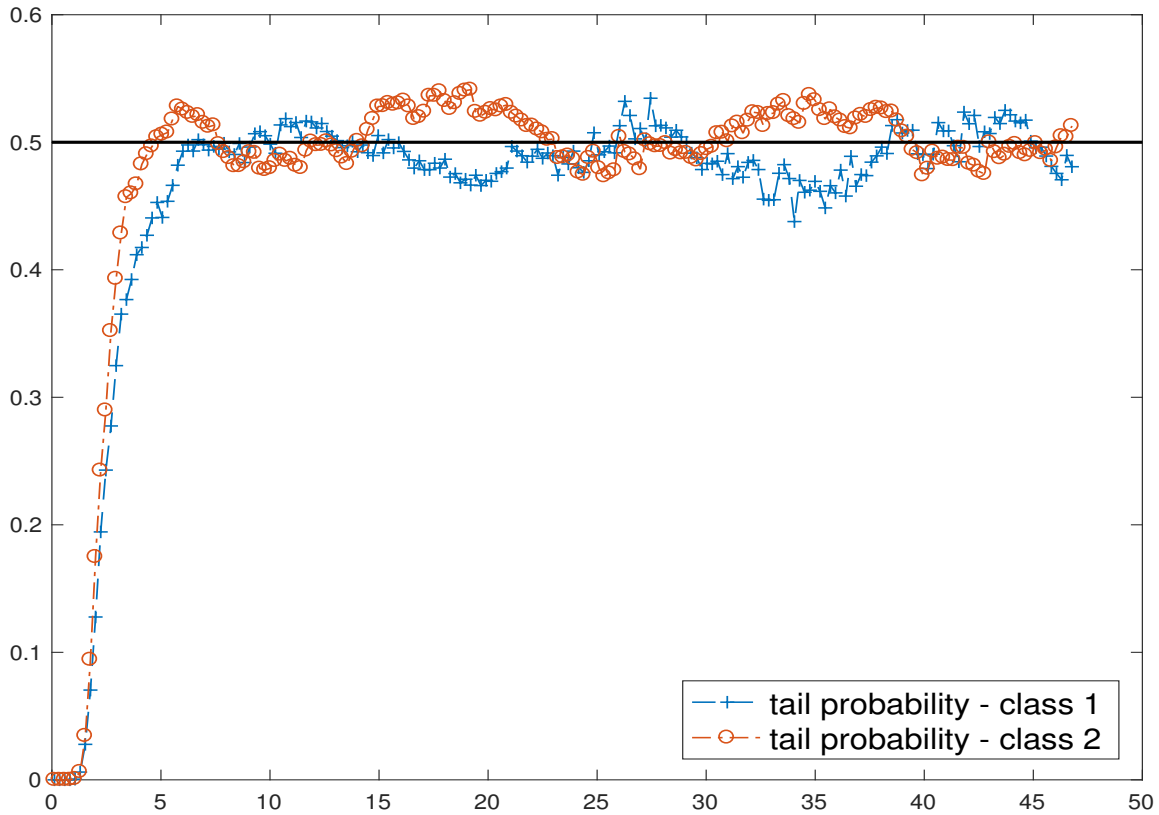


Figure: Base case with the tail-probability formulation ($\alpha = 0.5$) and TVQR rule.

Head-of-Line (HoL) Delays

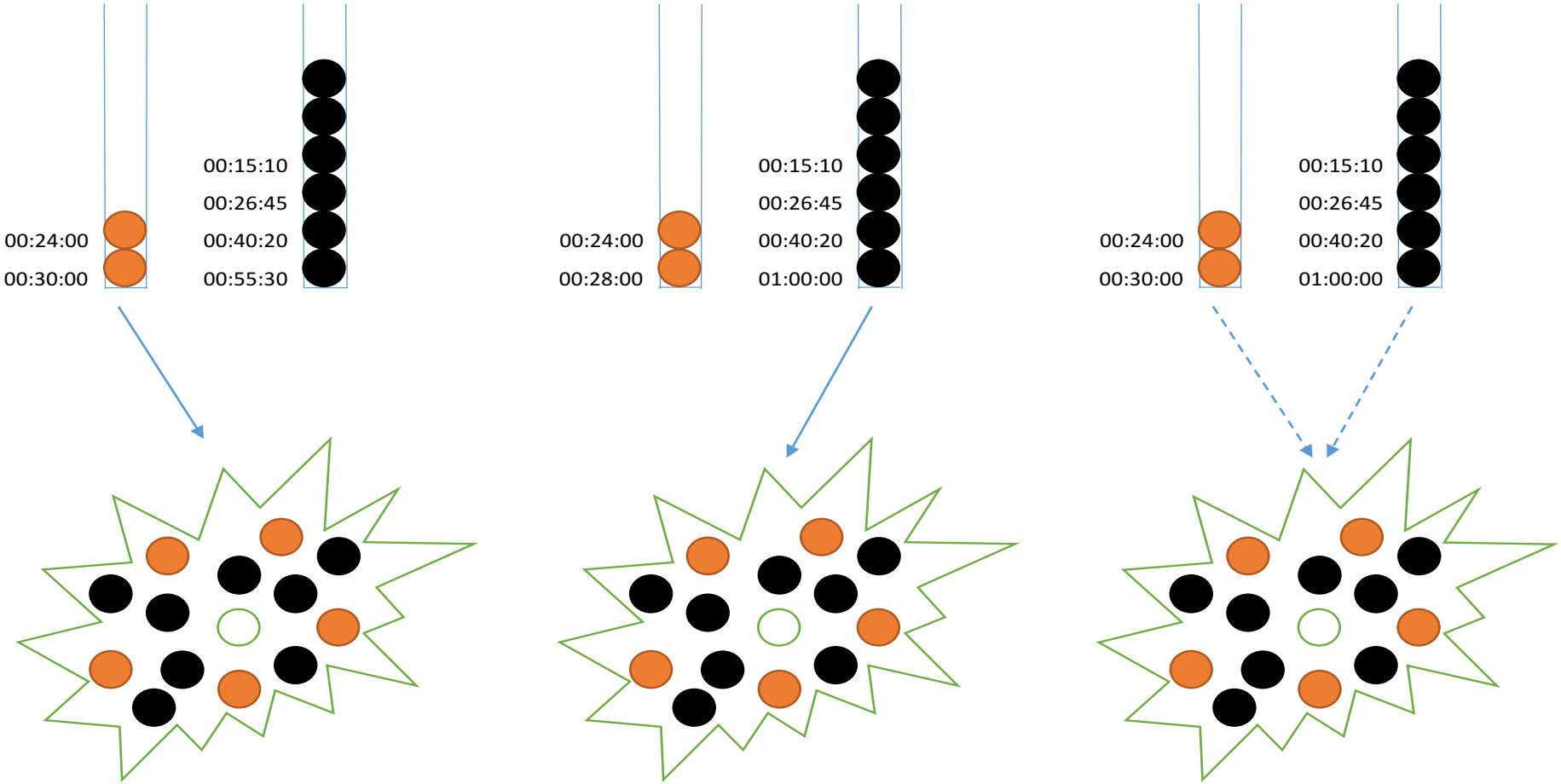
- The HoL delay, denoted by $U_i(t)$, is the current delay experienced by the customer at the head of queue i.
- HoL delays $U_i(\cdot)$ are observable at any given point in time.
- The HoL delay and virtual waiting time satisfy a simple relation

$$U_i(t) = V_i(t - U_i(t)) \quad \text{or} \quad V_i(t) = U_i(t + V_i(t)).$$

If $U_i(t)$ are small, then

$$V_i(t) \approx U_i(t).$$

Head-of-Line-Delay-Ratio (HLDR) Rule



Goal: HoL delay ratio 1/2

Theorem (MSHT Limits for HLDR)

Suppose that the system uses SRS and operates under the HLDR scheduling rule. Then we have the joint convergence

$$\begin{aligned}&\left(\hat{X}_1^n, \ldots, \hat{X}_K^n, \hat{Q}_1^n, \ldots, \hat{Q}_K^n, \hat{V}_1^n, \ldots, \hat{V}_K^n\right)\\ &\Rightarrow \left(\hat{X}_1, \ldots, \hat{X}_K, \hat{Q}_1, \ldots, \hat{Q}_K, \hat{V}_1, \ldots, \hat{V}_K\right) \quad \textit{in} \quad \mathcal{D}^{3K}\end{aligned} \tag{5}$$

as $n \to \infty$, where the diffusion limits $\hat{X}_i(\cdot)$ satisfy

$$\begin{aligned}\hat{X}_i(t) = \hat{X}_i(0) - \mu_i \int_0^t \hat{X}_i(u)\mathrm{d}u - (\theta_i - \mu_i) \int_0^t \gamma(u)^{-1} v_i(u)\lambda_i(u)\\ \times \left[\hat{X}(u) - c(u)\right]^+ \mathrm{d}u + \int_0^t \sqrt{\lambda_i(u) + \mu_i m_i(u)}\mathrm{d}W_i(u)\end{aligned} \tag{6}$$

with $\gamma(\cdot) \equiv \sum_{i\in\mathcal{I}} v_i(\cdot)\lambda_i(\cdot)$, $\hat{X} \equiv \sum_{i\in\mathcal{I}} \hat{X}_i$ and $W_i(\cdot)$ i.i.d. standard Brownian motions. For each $i \in \mathcal{I}$,

$$\begin{aligned}\hat{Q}_i(\cdot) &\equiv \gamma(\cdot)^{-1} v_i(\cdot)\lambda_i(\cdot)\left[\hat{X}(\cdot) - c(\cdot)\right]^+,\\ \hat{V}_i(\cdot) &\equiv v_i(\cdot)\cdot\gamma(\cdot)^{-1}\left[\hat{X}(\cdot) - c(\cdot)\right]^+.\end{aligned} \tag{7}$$

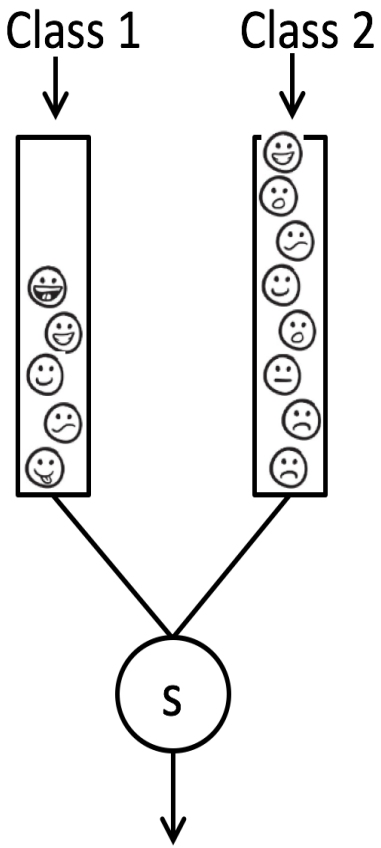
Proposed Solution via HLDR

- Proposed solution
  - Staffing: based on single class with TV target queue length

$$q(t) = \lambda_1(t)w_1 + \lambda_2(t)w_2$$

  - Scheduling: Use the HLDR rule with constant ratio functions

$$v_i(t) = w_i$$



Numerical Test I (HLDR)

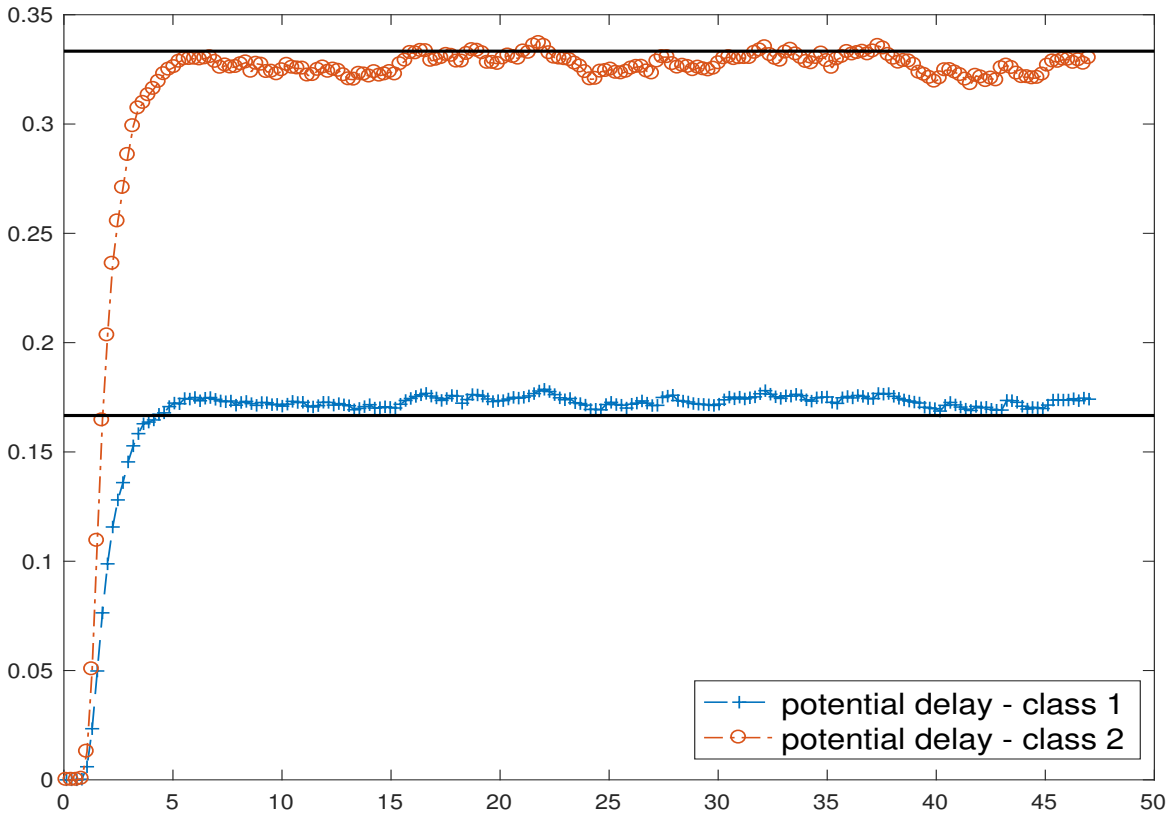


Figure: Base case with the mean-waiting-time formulation and HLDR rule.

Numerical Test II (HLDR)

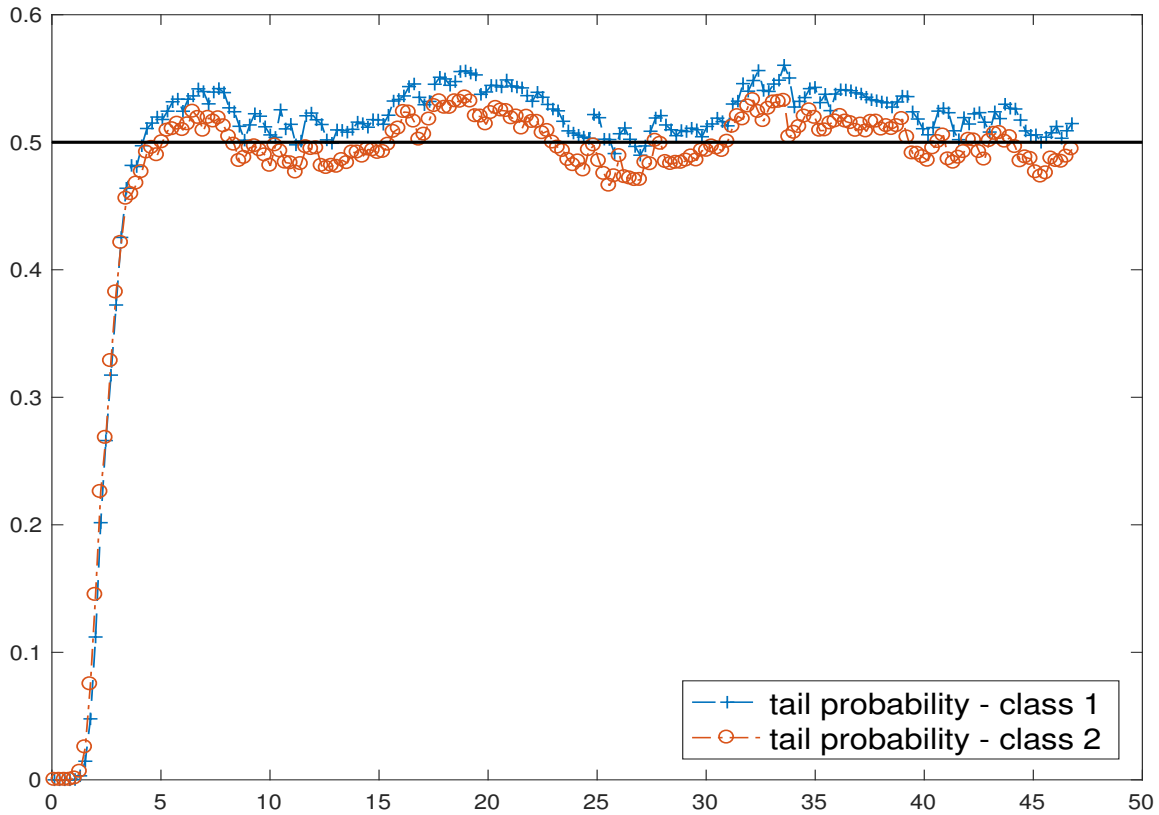


Figure: Base case with the tail-probability formulation ($\alpha = 0.5$) and HLDR rule.

Summary

- Defined the stochastic model and introduced the staffing minimization problems in a TV setting
- Proposed two ratio-control rules: TVQR and HLDR
- Established and characterized the heavy-traffic limit for two rules and extracted important insights, such as TV Little's law
- Used the proposed ratio-control rules to construct solution to the joint staffing and scheduling problem
- Showed via simulation studies that the algorithm performs well