Overflow trigger times

- Wards usually accept patients from primary specialties
Entire hospital runs in the QED regime

- Quality- and Efficiency-Driven (QED) regime
  - Waiting time is a small fraction of service time
    - Average waiting time = 2.8 hours = 1/43 average LOS
  - Typical bed occupancy rate is 86% ~ 93%

- Multi-server pools with certain flexibility
  - 30 ~ 60 servers in each pool
  - 15 server pools (500-600 servers)

- Trade-off between waiting time and overflow fraction
Predict average queue length curve

- Tractability and relevance (Period 1)
Part 3: Two-time-scale framework

- Discrete-time queues
  - The LOS and daily arrival rate determine $\{X_k\}$, the midnight customer count, and thus determine the daily performance

- Time-varying performance
  - The arrival rate pattern and discharge timing determine the time-of-day behavior
A simplified single-pool model

- A single-pool model with $N$ servers
  - Arrival is periodic Poisson with rate function $\lambda(t)$ and period of 1 day
  - LOS follow a geometric distribution with mean $m$
  - Discharge times follow a discrete distribution
  - Allocation delay

- Service times follow the non-iid model

- Performance measure: steady-state, mean queue length curve
  $\mathbb{E}[Q(t)]$ for $0 \leq t < 1$
Step 1: daily customer count

- $X_k$ denotes the number of customers at midnight of day $k$
  \[ X_{k+1} = X_k - D_k + A_k \]
  - Discrete time queue

- Number of discharges $D_k$ only depends on $X_k$ and independent coin tosses since
  - LOS is geometric
  - LOS starts from 1 (no same-day discharge)

- Number of arrivals $A_k$ is a Poisson random variable
  - Independent of number of discharges

- $\{X_k\}$ is a discrete time Markov chain (DTMC)
  - Stationary distribution $\pi$ can be solved numerically
Step 2: hourly customer count

\[ X(t) = X(0) - D_{(0,t]} + A_{(0,t]} \]

- Conditioning on \( X(0) \), \( X(t) \) is a convolution between a Poisson r.v. (arrival) and a binomial r.v (discharge)
- The mean queue length \( \mathbb{E}[Q(t)] = \mathbb{E}[X(t) - N]^+ \)

Mean customer count can be solved via fluid equation

- \( \mathbb{E}[X(t)] = \mathbb{E}[X(0)] + \int_0^t \lambda(s) ds - \mathbb{E}[D_{(0,t]}] \)
- \( \mathbb{E}[Q(t)] \equiv \left( \mathbb{E}[Q(0)] + \int_0^t \lambda(s) ds - \mathbb{E}[D_{(0,t]}] \right)^+ \)
Related work

  - ED evolves in a much faster time scale than wards.

  - Two time scales: service times are in days; waiting times are in hours.

  - Affiliations: Department of Emergency Medicine, Northwestern University; Harvard Affiliated Emergency Medicine Residency, Brigham and Women’s Hospital–Massachusetts General Hospital, …
Allocations delays

• Each patient experiences a random delay $T$ after a bed is allocated to her

• Model the allocation queue as a $\ast/GI/\infty$ system
  • Allocation queue length at time $t$ is equivalent to mean number of customers in the infinite-server system at time $t$
  • Use the infinite-server queue theory
Numerical results

• Alloc delays follow a log-normal distribution
  • Mean alloc delay is 2.5 hours, CV=1
• Discrete discharge distribution from NUH period 1 data
• N=525; m=5.3;
  \[ \Lambda = 90.95 \]
Recall the queue length curve from the hospital model (Period 1)
Aggressive early discharge policy

Discharge distribution

NUH per 1
NUH per 2
aggressive early dis
Insights from the simplified model

- Impact of discharge policy
- Steady-state, time-of-day mean waiting time

![Graph showing waiting time over time with different discharge policies](image-url)
Simulation results

- Simulation shows NUH early discharge policy has little improvement
  - (a) hourly avg. waiting time
  - (b) 6-hour service level
Aggressive early discharge + smooth allocation delay

- Waiting time performance can be stabilized
  (a) hourly avg. waiting time
  (b) 6-hour service level
Challenges

- For a multi-pool model with “state”-dependent overflow trigger time, develop an analytical theory for
  - Performance analysis
  - Near optimal overflow policy (real time); impossible for simulation
  - Optimal capacity allocation among different wards (once every 6 months?); time consuming for simulation
  - Perry & Whitt (X-model); Pang & Yao (switch-over)

- For a single-pool model, analyze the discrete time queue under
  - General LOS distribution
  - Day-of-week model
  - Matrix analytic method, diffusion approximations
Operational Challenges

- Push early discharge
- Reduce LOS
  - AM- and PM-admissions
  - Using step-down care facilities
Questions?