Stochastic Network Models for Hospital Inpatient Flow Management

Jim Dai

School of ORIE, Cornell University

(on leave from Georgia Institute of Technology)

Team members

- Georgia Tech
  - Pengyi Shi

- University of International Business & Economics, Beijing
  - Ding Ding

- National University of Singapore (NUS)
  - Jame Ang, Mabel Chou

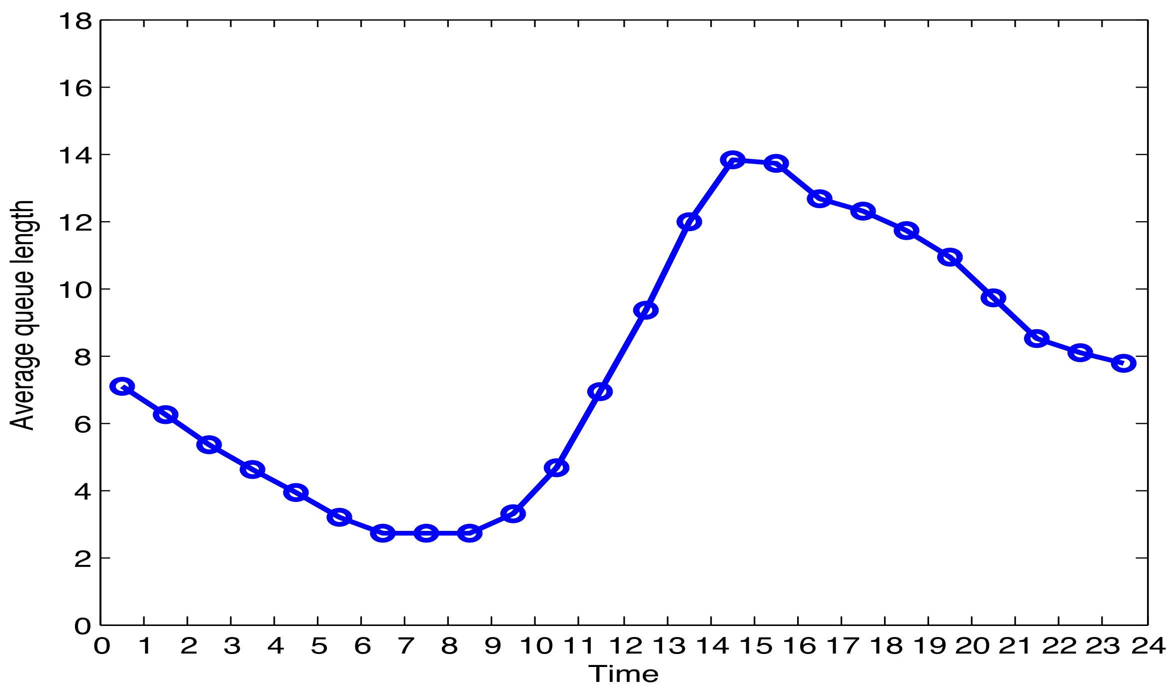
- National University Hospital (NUH)
  - Jin Xin, Joe Sim

Outline

- Part 1: Empirical observations
- Part 2: Stochastic network models
- Part 3: Two-time-scale framework
- Part 4: Managerial insights & future research

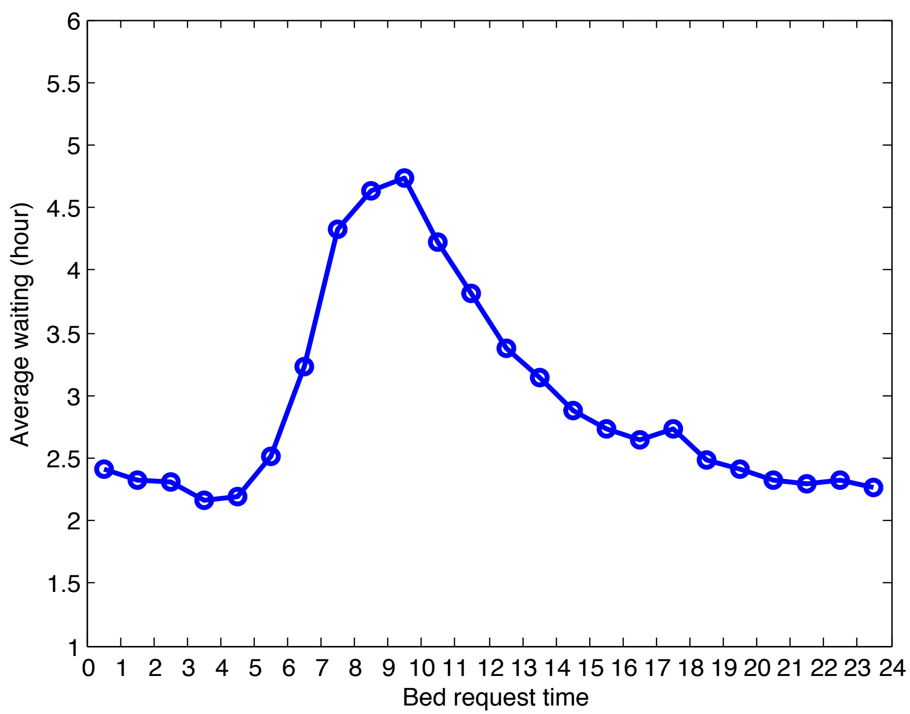
Empirical observation at NUH

- Average queue length curve over 547 days
  - # of patients who are waiting for inpatient beds from the emergency department (ED)
  - Can we build a model and find methods to predict the curve?

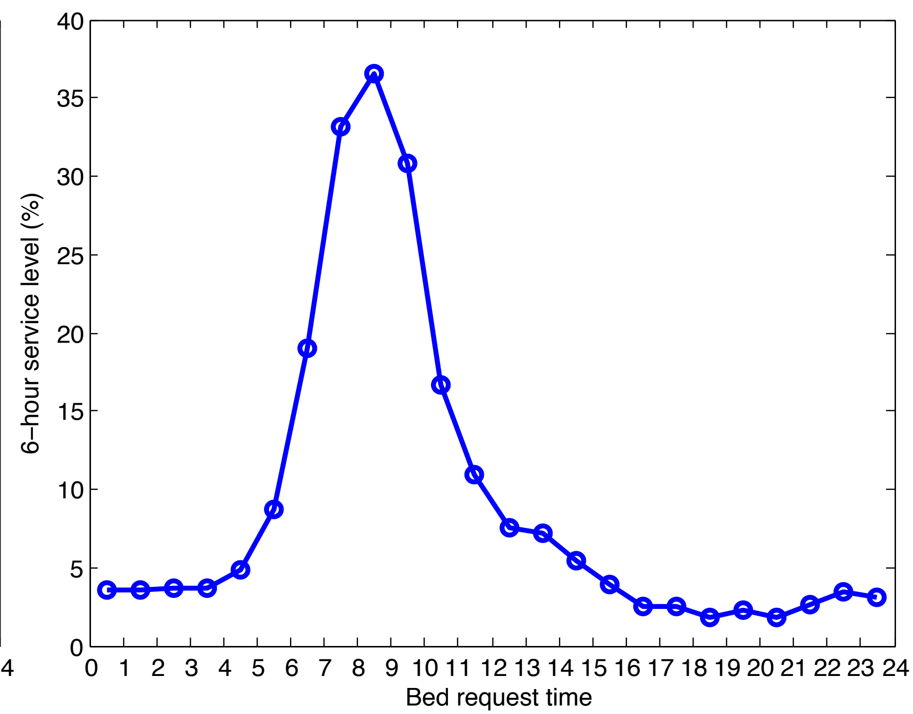


Waiting time statistics: Period 1

Average waiting time



Fraction of patients who wait at least 6 hours



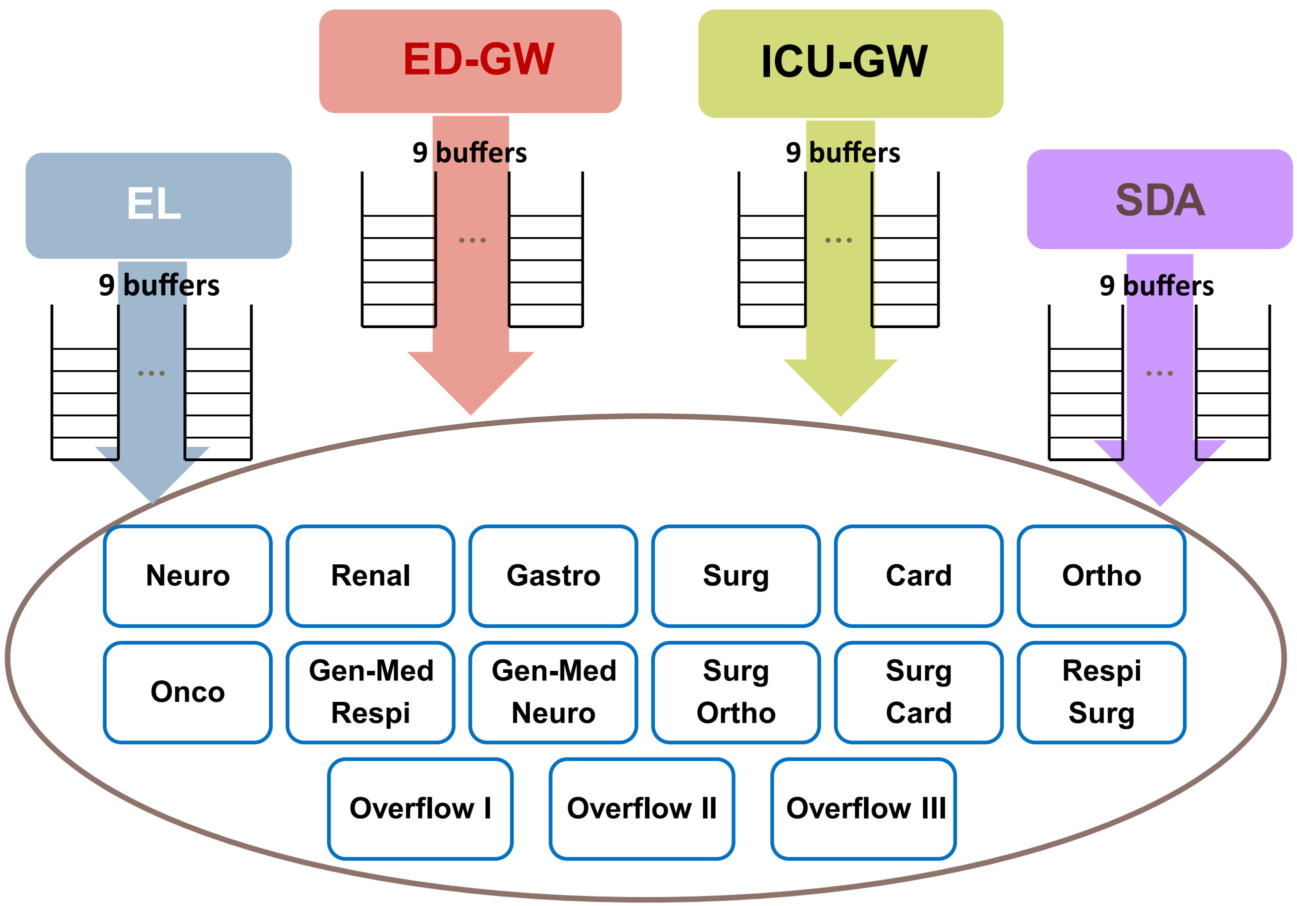
Can we flatten the curve?

Part 2: Stochastic network models

- Time-varying queues
  - Massey (1981), non-stationary queues
  - Whitt (1991)
  - Green-Kolesar (1991, 1997)
  - Massey, Mandelbaum and Reiman (1998)
  - Feldman-Mandelbaum-Massey-Whitt (2008), "Staffing of Time-Varying Queues to Achieve Time-Stable Performance."
  - Liu-Whitt (2011,2012)
  - $M_t/GI/N$ framework

A new stochastic network model

- Multi-server pools serving multi-class customers



New features

- Endogenous service times
- Allocation delays
- Overflow trigger times
- Missing any one of these features makes the model less relevant

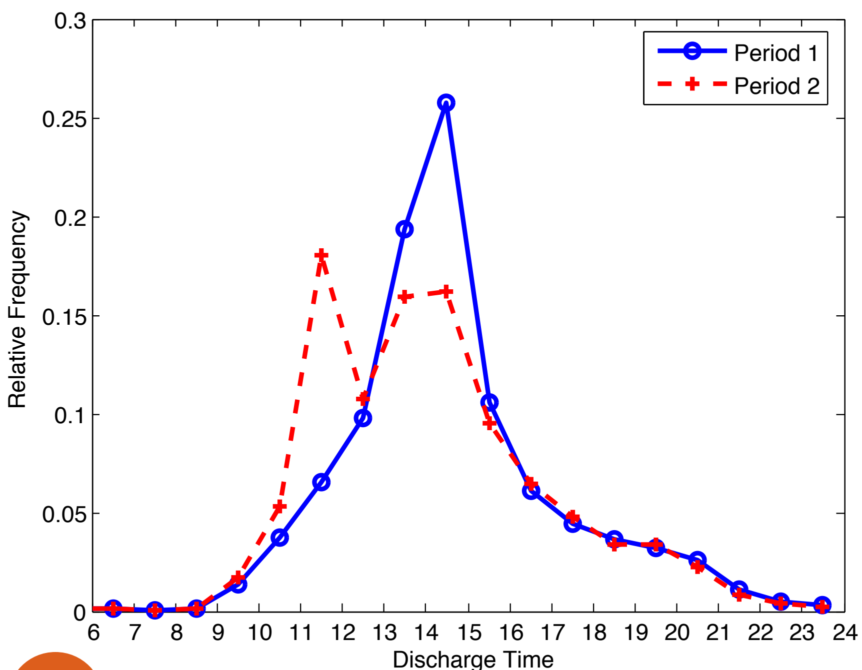
Endogenous service times

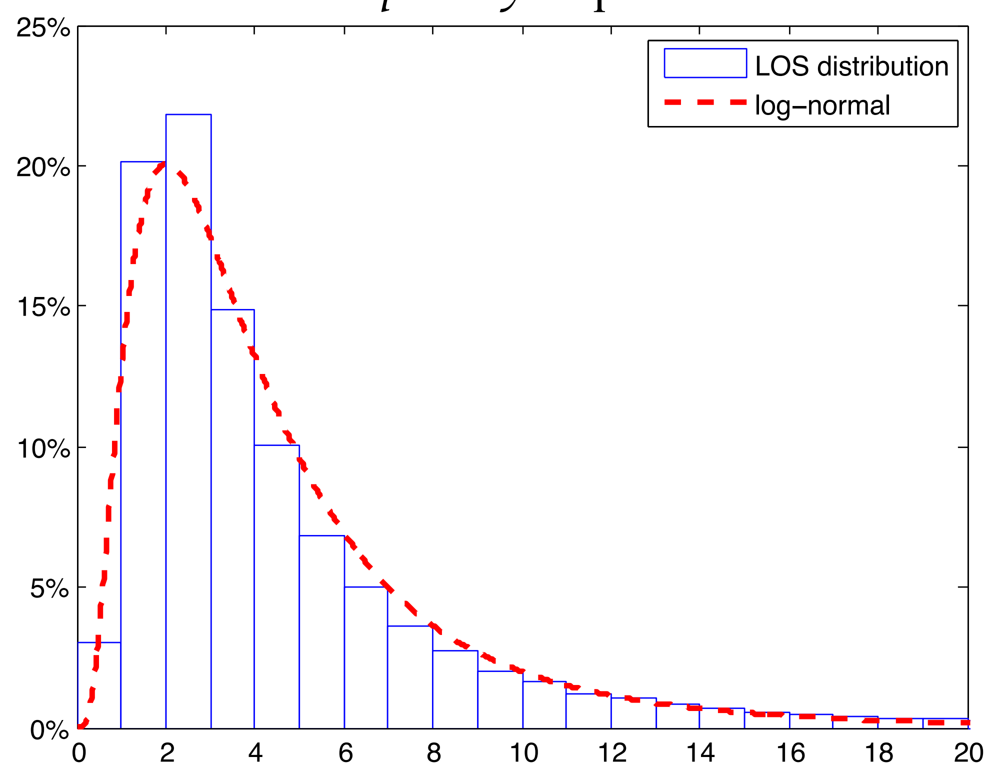
Service time = Discharge time – Admission time

= **LOS + Dis hour – Adm hour**

Length-of-stay (LOS) = number of nights in hospital

- LOS distribution
  - Average is ~ 5 days; *admission source* and *medical specialty* dependent

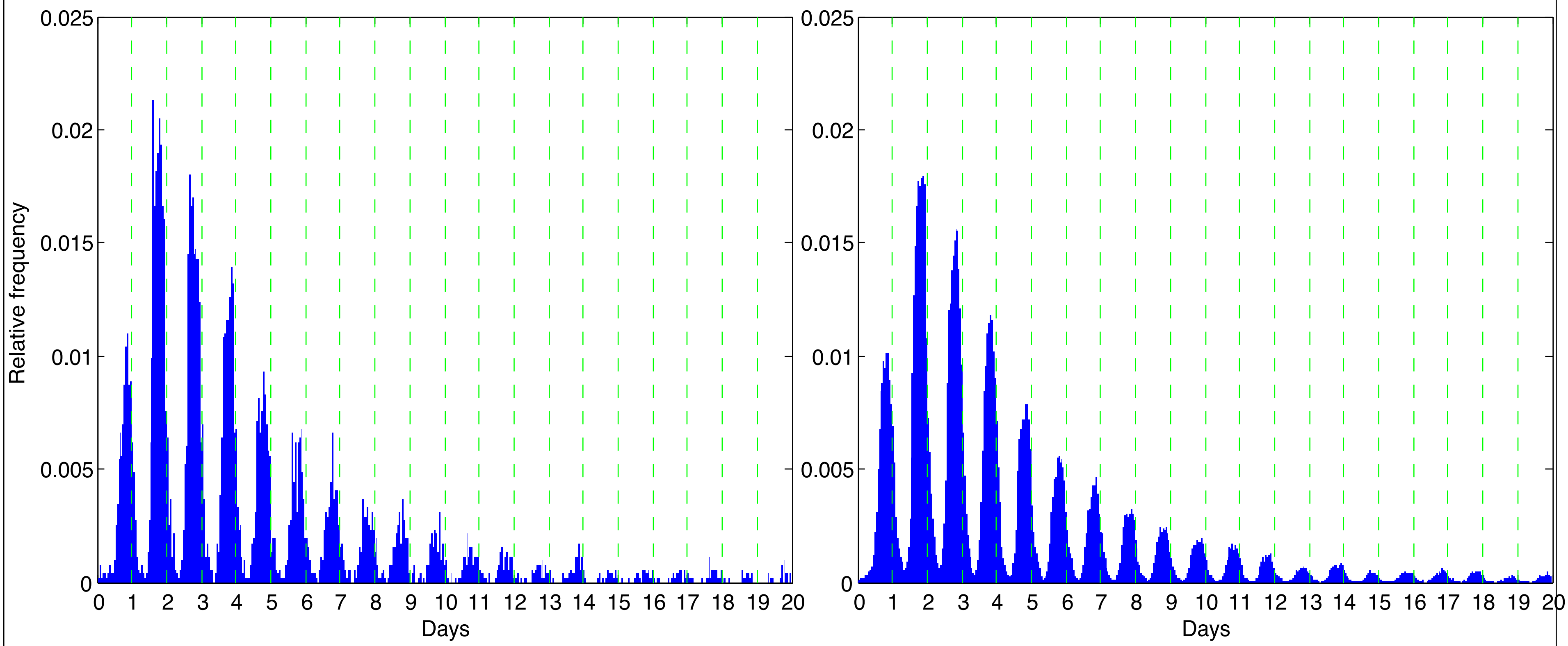




Checking the service time model

(a) Empirical

(b) Simulation output



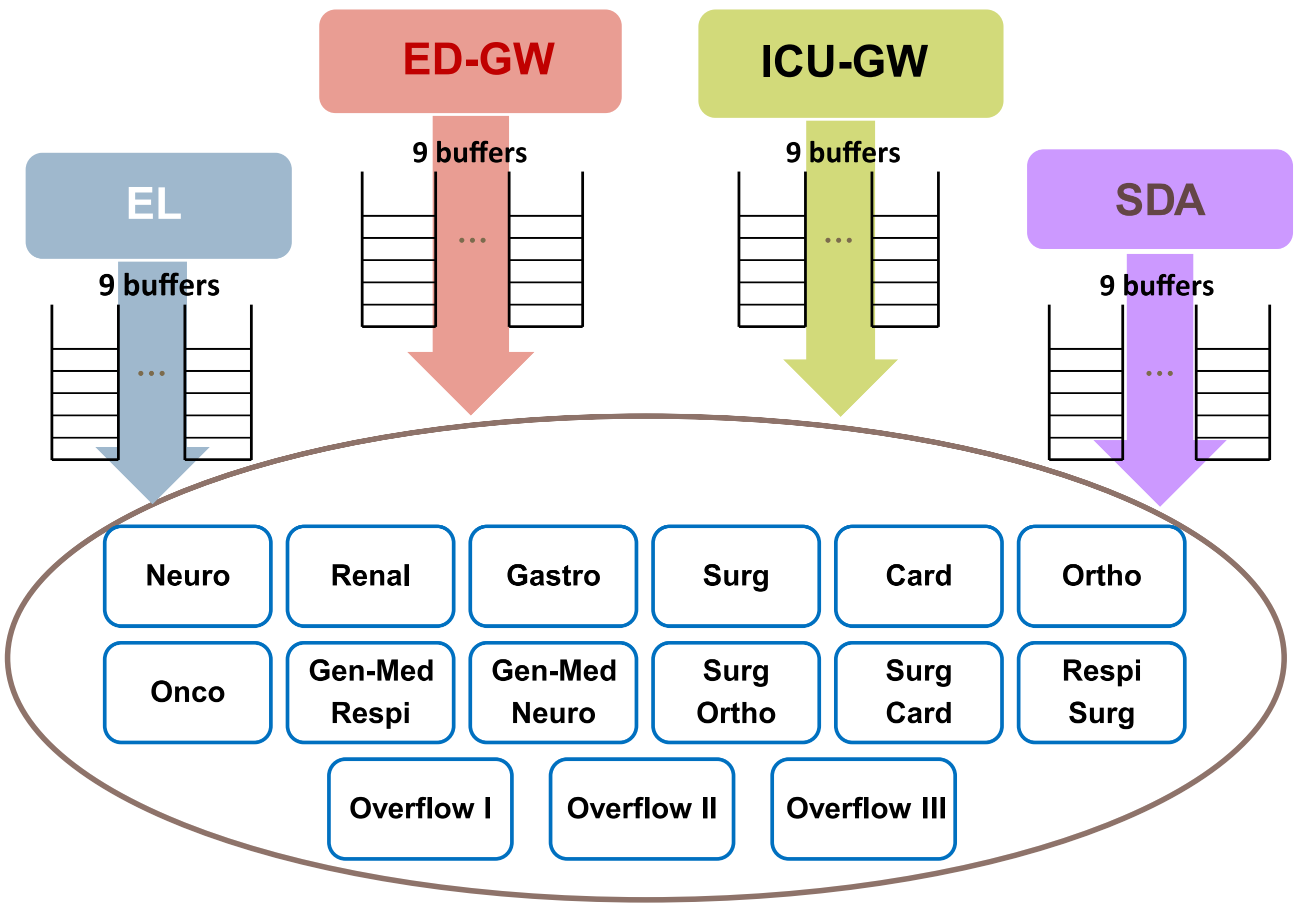
Allocation delays

- Getting a bed is a process
  - Pre-allocation delay
    - Bed management unit searches/negotiates for beds
  - Post-allocation delay
    - Delays in ED discharge
    - Delays in transportation
    - Delays in ward admission

- In our model: each patient experiences a random delay T after a bed is allocated to her

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

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 π 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

- $\mathbb{E}[X(t)]=\mathbb{E}[X(0)] + \int_0^t \lambda(s)ds - \mathbb{E}[D_{(0,t]}]$
- $\mathbb{E}[Q(t)] \stackrel{?}{=} \mathbb{E}[Q(0)] + \int_0^t \lambda(s)ds - \mathbb{E}[D_{(0,t]}]$

Related work

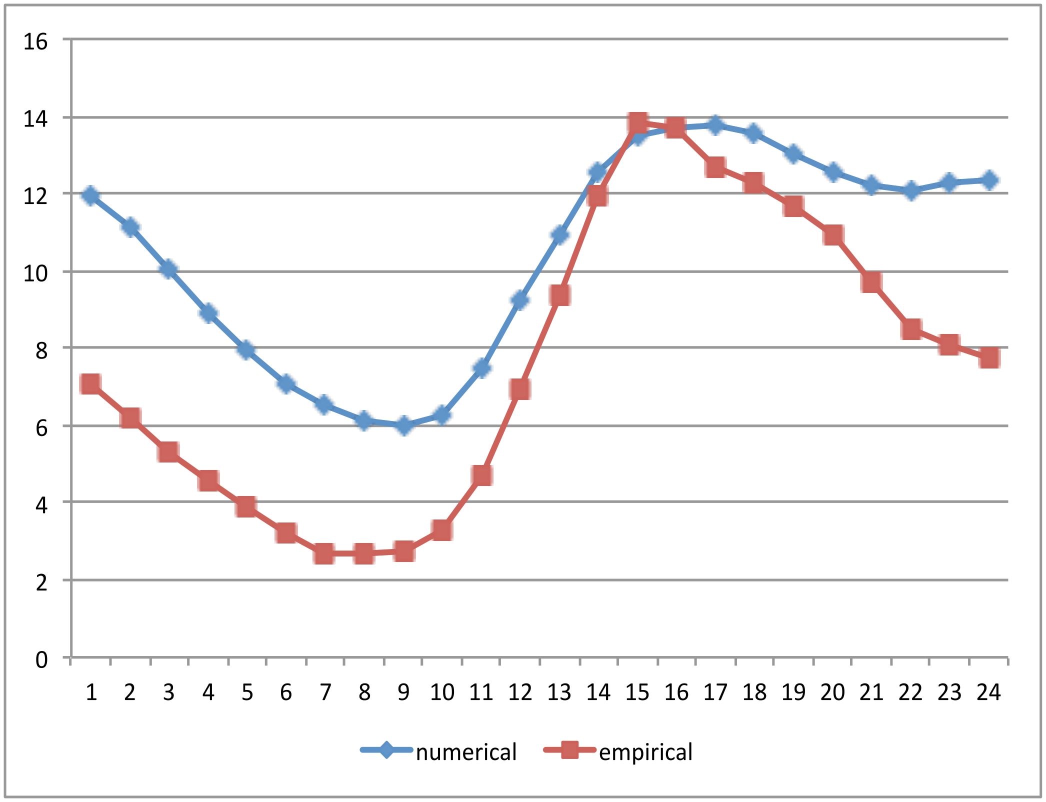
- M. Ramakrishnan, D. Sier, P. Taylor (2005), "A two-time-scale model for hospital patient flow", *IMA Journal of Management Mathematics.*
  - ED evolves in a much faster time scale than wards.

- A. Mandelbaum, P. Momcilovic, Y. Tseytlin (2012), "On Fair Routing from Emergency Departments to Hospital Wards: QED Queues with Heterogeneous Servers", *Management Science*.
  - Two time scales: service times are in days; waiting times are in hours.

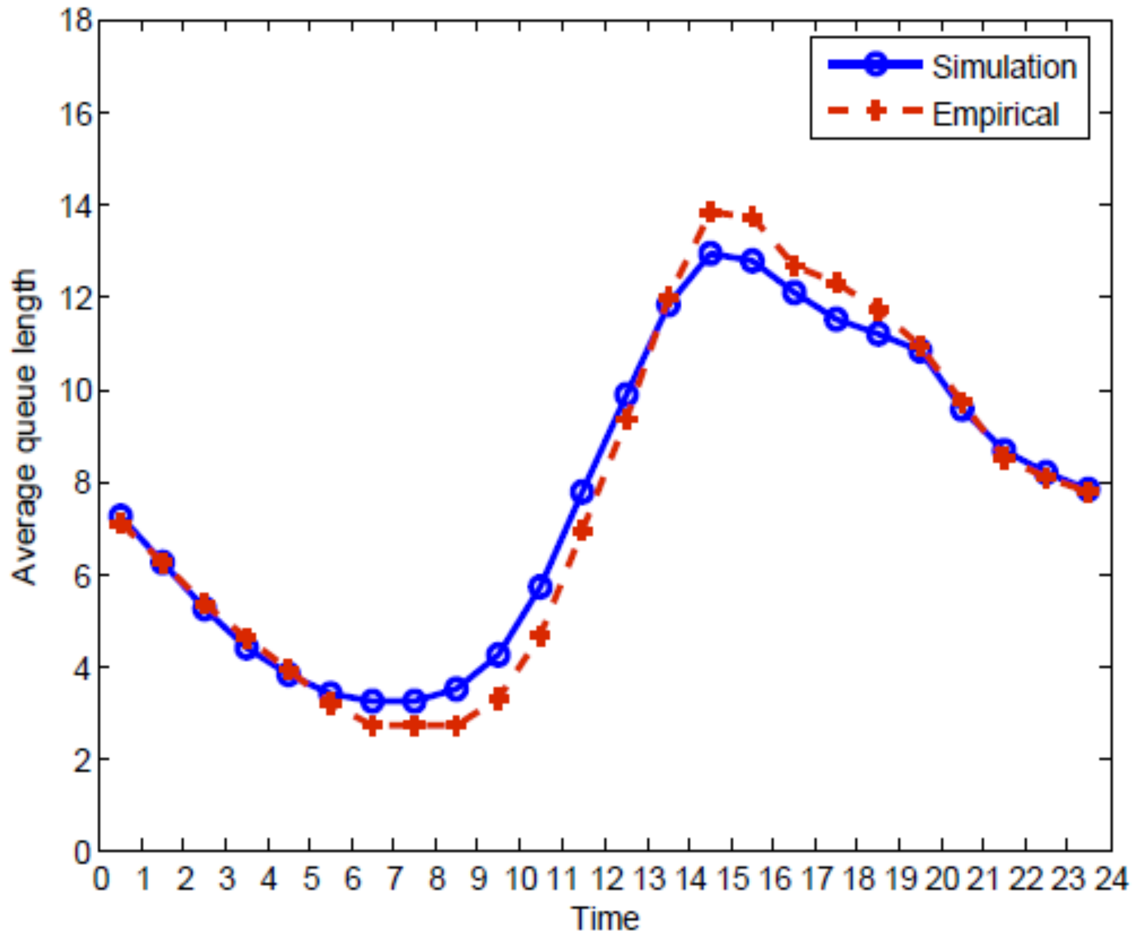
- E. S. Powell et al. (2012), "The relationship between inpatient discharge timing and emergency department boarding", *The Journal of Emergency Medicine*
  - Affiliations: Department of Emergency Medicine, Northwestern University; Harvard Affiliated Emergency Medicine Residency, Brigham and Women's Hospital–Massachusetts General Hospital, …

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;
  $\Lambda = 90.95$



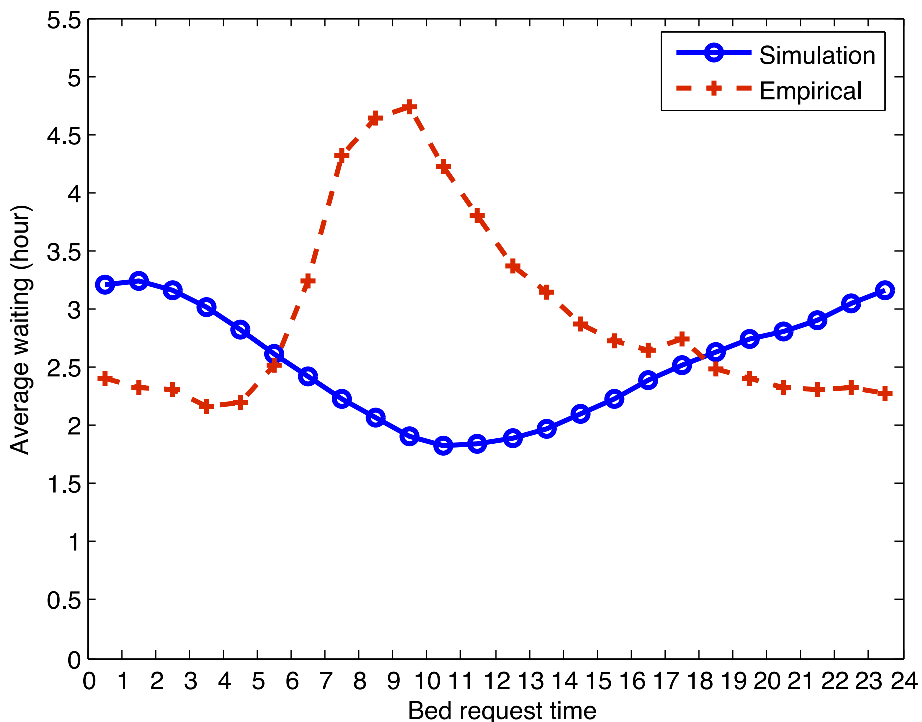
Queue length curve from the FULL hospital model (Period 1)



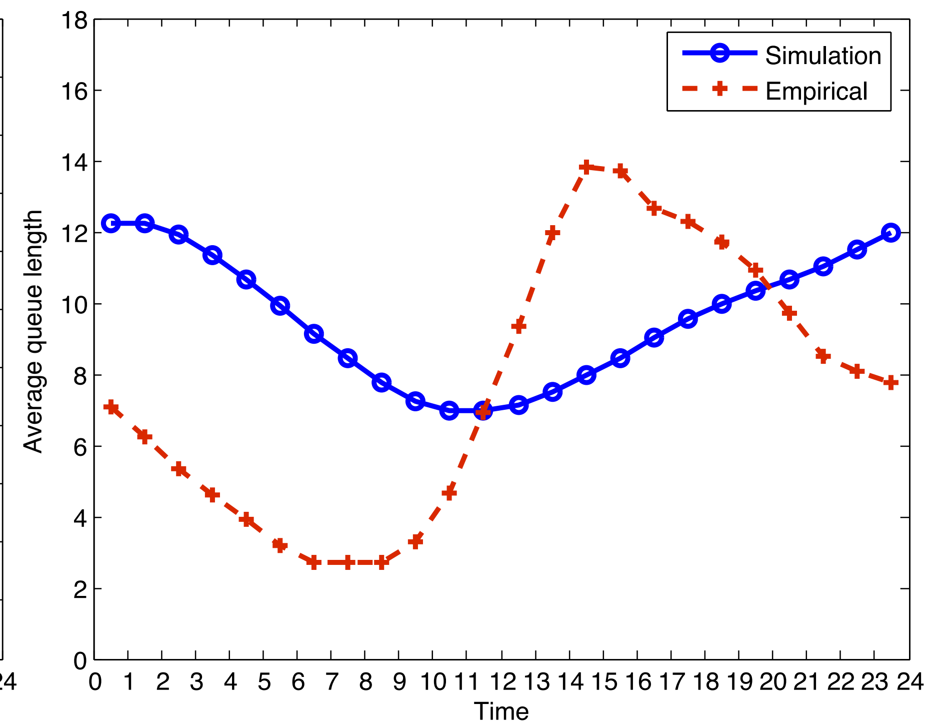
$M_t/GI/N$ queues fail to capture

- Simulation results from an $M_{\text{peri}}/\text{lognormal}/N$ system

avg waiting time

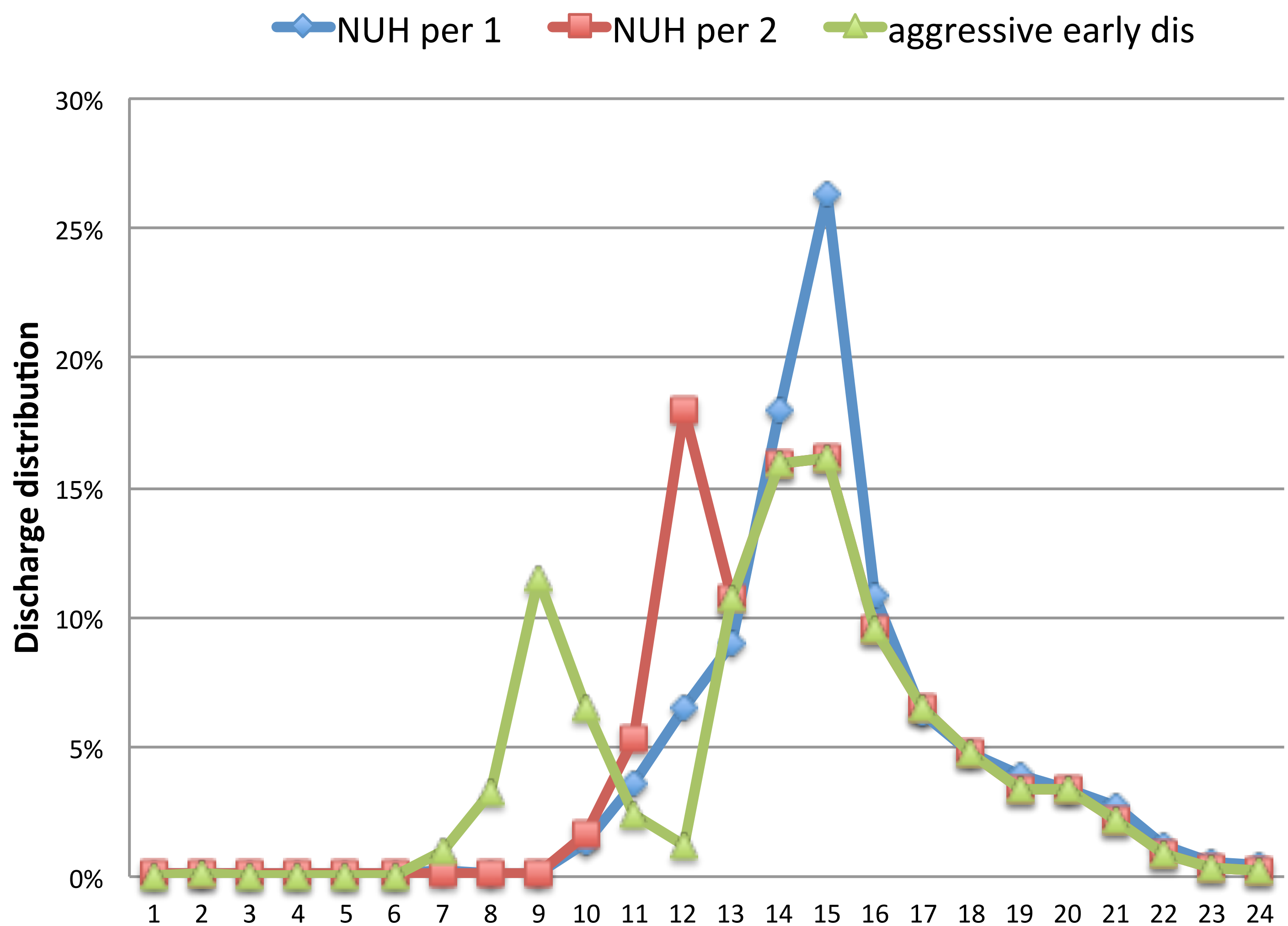


avg queue length



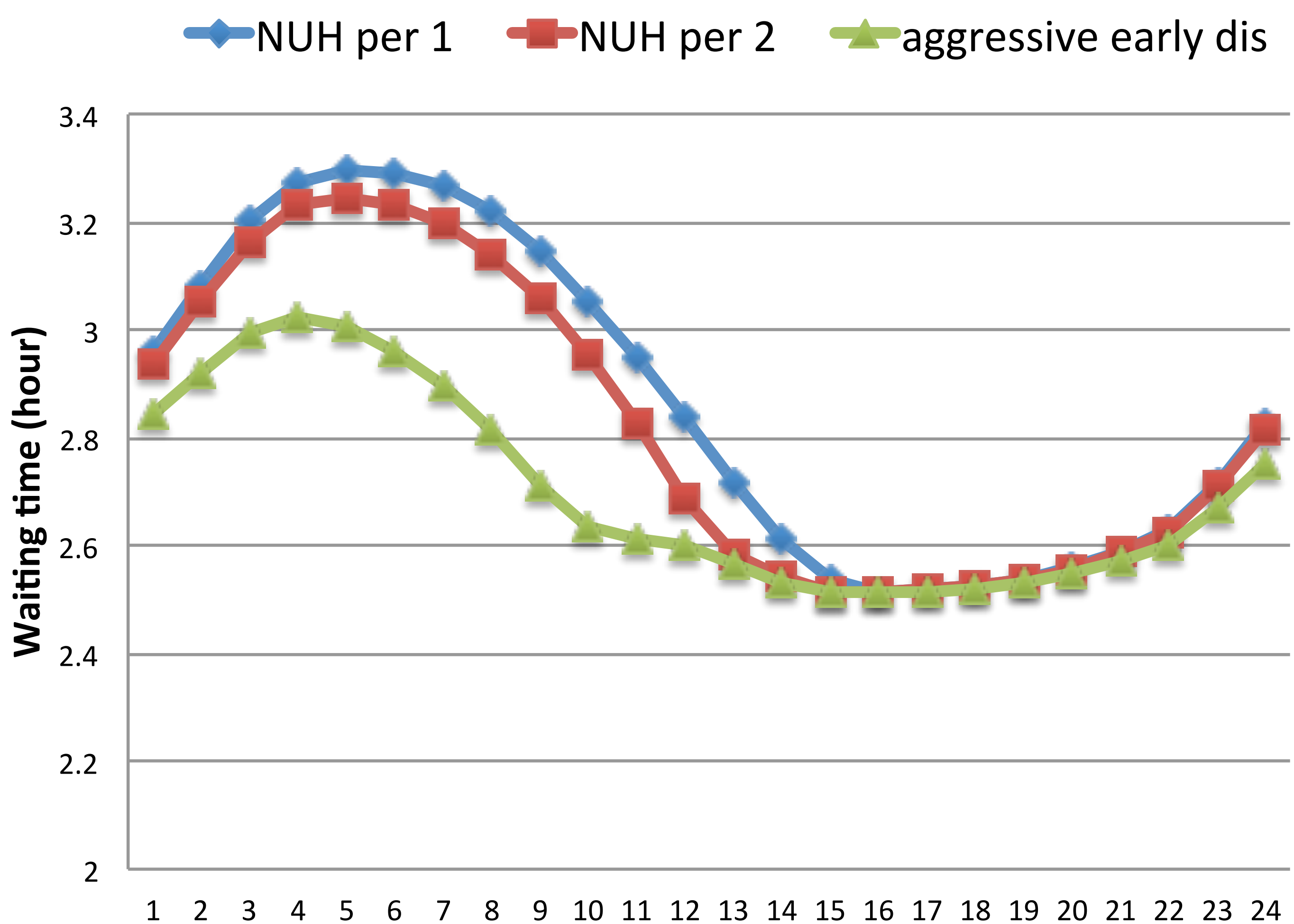
Part 4: Insights & challenges

Aggressive early discharge policy



Insights from the simplified model

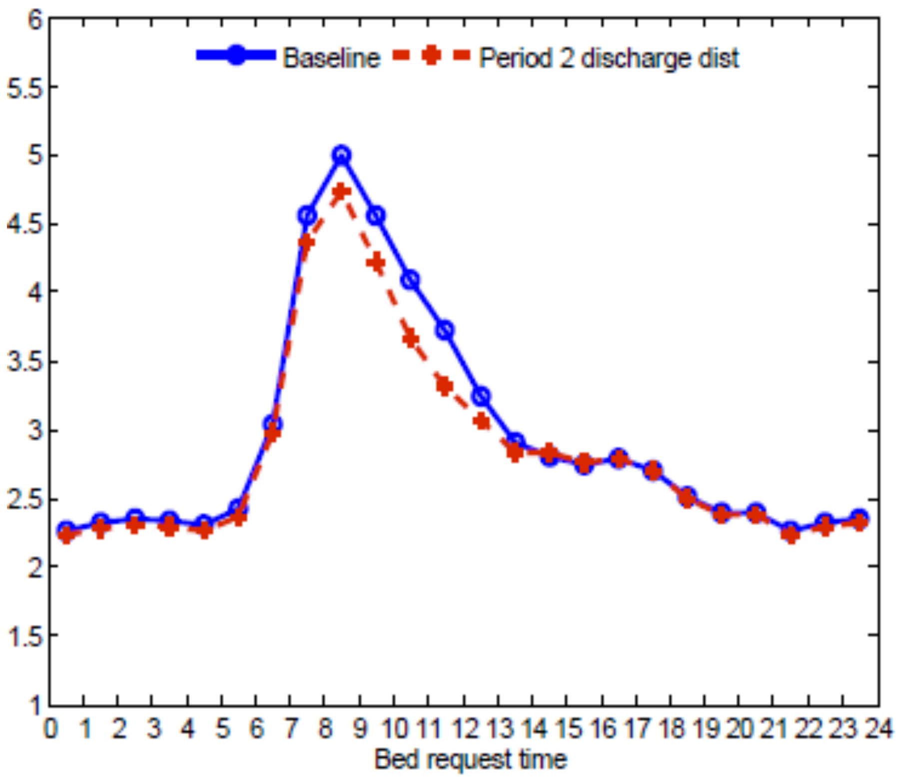
- Impact of discharge policy
- Steady-state, time-of-day mean waiting time

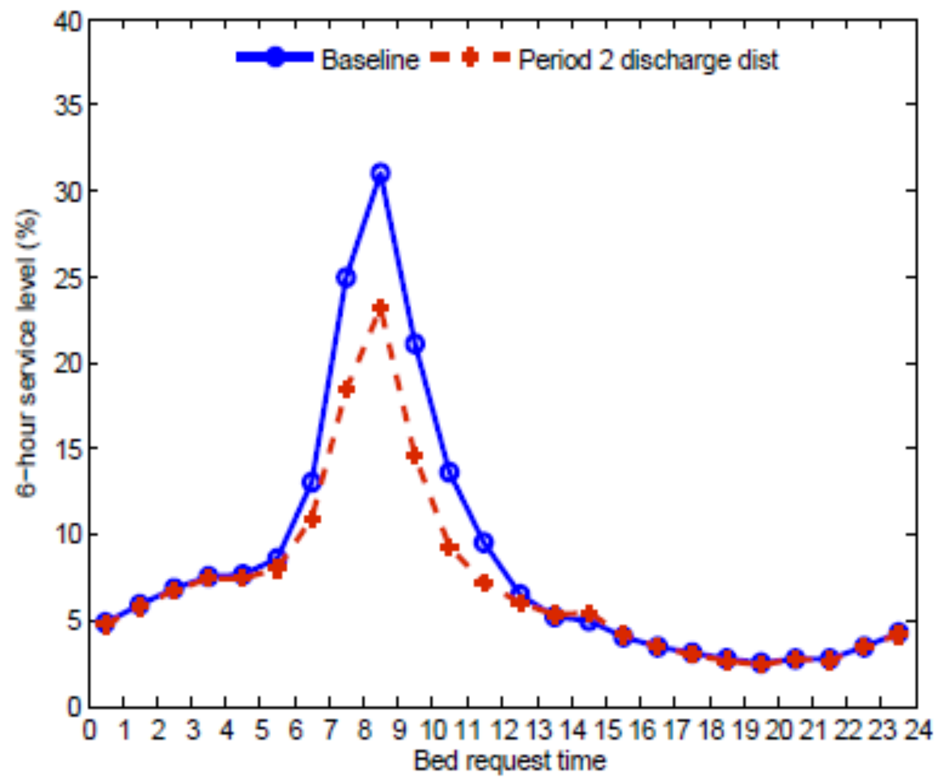


Simulation results

- Simulation shows NUH early discharge policy has little improvement

  (a) hourly avg. waiting time (b) 6-hour service level

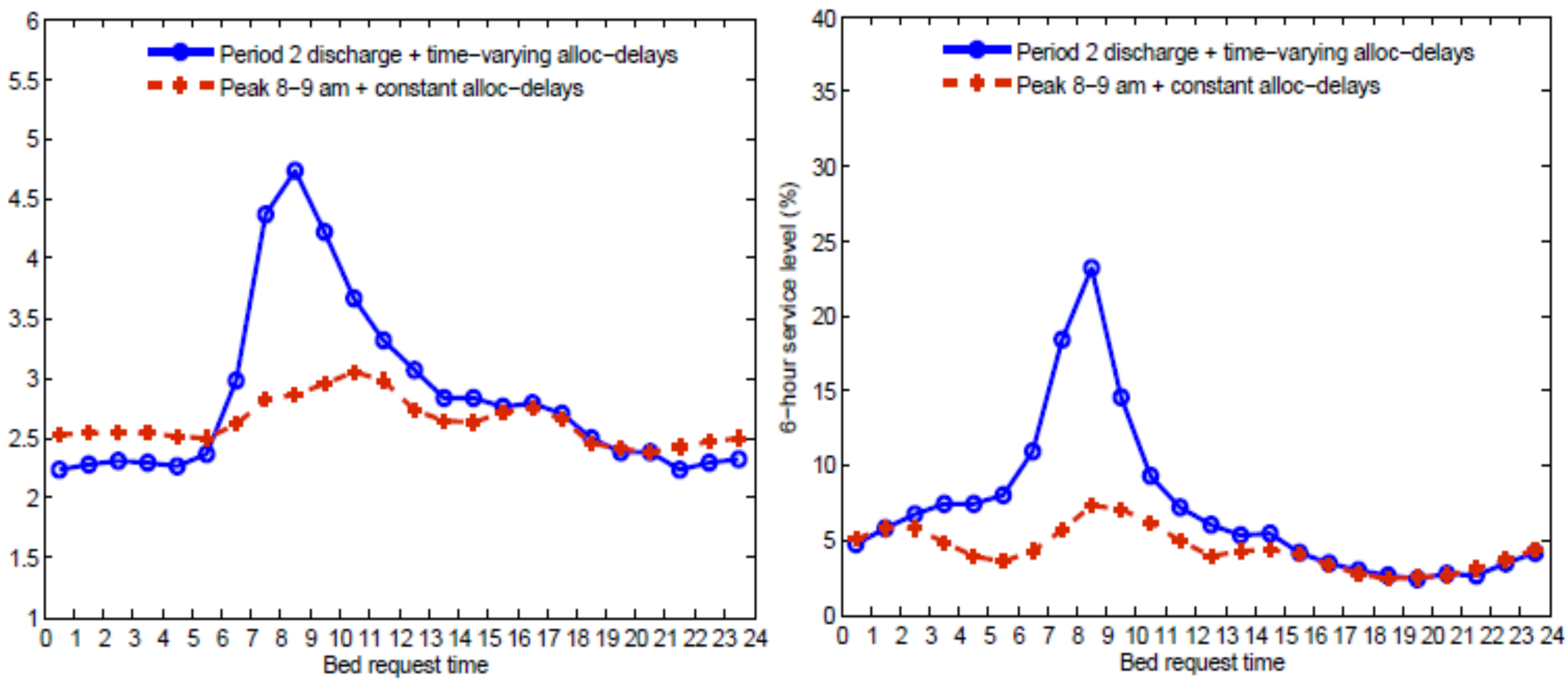




Aggressive early discharge + smooth allocation delay

- 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?