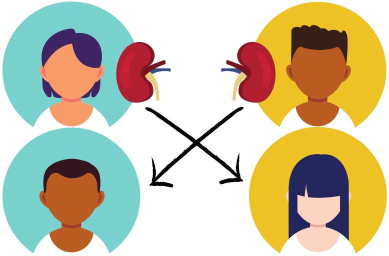
The Cost of Impatience in Dynamic Matching:

Scaling Laws and Operating Regimes

Angela Kohlenberg and Itai Gurvich

BLOOD ALLOCATION

PAIRED KIDNEY EXCHANGE



- HOW DOES IMPATIENCE IMPACT MATCH RATE?

  THE RELATIONSHIP BETWEEN PARAMETERS

- NEED TO CONSIDER: TIME TO ABANDON & TIME TO MATCH

How Does Impatience Impact Match Rate?

1. Classify settings based on how impatience impacts match loss

   [Operating Regimes]

2. Identify key determinants of match loss from impatience

   [Scaling Laws]

Related Literature (A Sample): Matching and Single-Sided Queues with Abandonment

Exact Results

- Transient and steady-state performance of a two-sided queue (Conolly et al. 2002, Afèche et al. 2014, Diamant and Baron 2019)
- Performance analysis / stability region under specific policy (Castro et al. 2020b, Zubeldia et al. 2022)

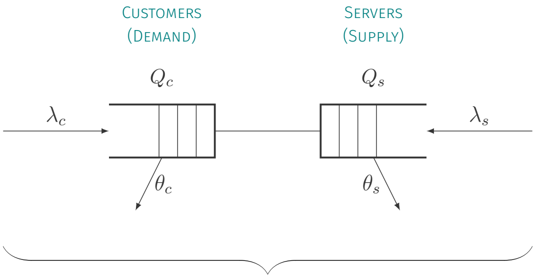
All parameters are equally important

Approximations (mostly heavy traffic)

- Performance analysis and control of a two-sided queue (Liu et al. 2015, Büke and Chen 2017, Chen and Hu 2020, Aveklouris et al. 2023)
- Performance analysis and control of a single-sided queue (Ward and Glynn 2003, Lee and Ward 2019)
- Control policies for a matching network (Aveklouris et al. 2021, Collina et al. 2020, Aouad and Saritaç 2022, Castro et al. 2020a, Wang et al. 2022)

Imposes specific relationship on parameters

Our Focus: simple model, universal results (in parameters) & general characterization



MATCH RATE WITHOUT ABANDONMENT (UPPER BOUND):

$$\min\{\lambda_c, \lambda_s\} = \lambda_c$$

LABEL: $\lambda_c \leq \lambda_s$

UTILIZATION (MEASURE OF EXCESS CAPACITY):

$$\rho = \lambda_c / \lambda_s$$

ACTUAL MATCH RATE:

$$d = \lim_{t \uparrow \infty} \frac{1}{t}\mathbb{E}[D(t)]$$

ARRIVALS = MATCHES + ABANDONMENTS:

$$\lambda_c = d + \theta_c \mathbb{E}[Q_c]$$

Definition (Cost-of-Impatience, CoI).

$$\text{CoI} = \text{No-abandonment match rate } (\lambda_c) - \text{Actual match rate } (d) = \theta_c \mathbb{E}[Q_c] = \theta_s \mathbb{E}[Q_s] - (\lambda_s - \lambda_c)$$

1. CLASSIFY SETTINGS based on how impatience impacts match loss

   [OPERATING REGIMES]

2. IDENTIFY KEY DETERMINANTS of match loss from impatience

   [SCALING LAWS]

Operating Regimes: A Classification by Relative Impatience

- **Patient vs. Impatient:** measure of **mean patience** relative to **excess capacity** $(\rho = \lambda_c / \lambda_s)$
- **The CoI behaves different in each operating regime**

Operating Regimes: A Classification by Relative Impatience

Patient-Customer

Impatient-Customer

Impatient-Server

$$\frac{1}{1-\rho} \geq \sqrt{\frac{\lambda_s}{\theta_s}} \longleftrightarrow \lambda_s - \lambda_c \leq \sqrt{\theta_s \lambda_s}$$

Server Impatience

$$\frac{\rho}{1-\rho} \leq \sqrt{\frac{\lambda_c}{\theta_c}}$$

$$\frac{\rho}{1-\rho} \geq \sqrt{\frac{\lambda_c}{\theta_c}}$$

Intuition:

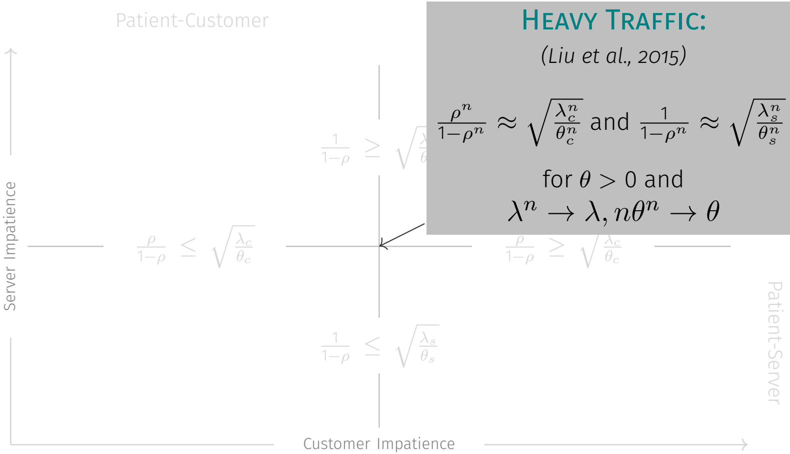
Customers & enough servers will wait longer than the time to match

$$\frac{1}{1-\rho} \leq \sqrt{\frac{\lambda_s}{\theta_s}}$$

Patient-Server

Customer Impatience

- **Patient vs. Impatient:** measure of **mean patience** relative to **excess capacity** $(\rho = \lambda_c/\lambda_s)$
- **The CoI behaves different in each operating regime**



Heavy Traffic:

(Liu et al., 2015)

$$\frac{\rho^n}{1-\rho^n} \approx \sqrt{\frac{\lambda_c^n}{\theta_c^n}} \text{ and } \frac{1}{1-\rho^n} \approx \sqrt{\frac{\lambda_s^n}{\theta_s^n}}$$

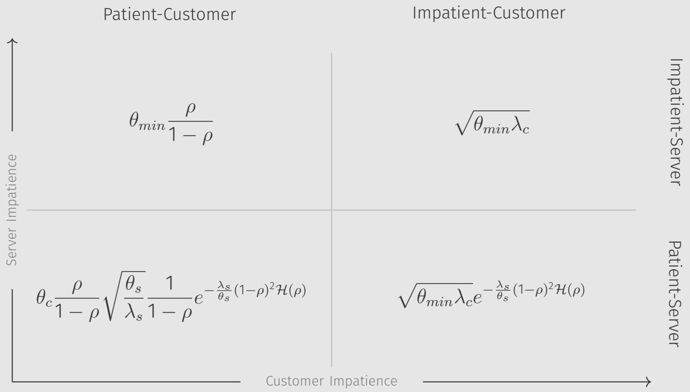
for $\theta > 0$ and

$$\lambda^n \to \lambda, n\theta^n \to \theta$$

- **Patient vs. Impatient:** measure of **mean patience** relative to **excess capacity** $(\rho = \lambda_c/\lambda_s)$
- **The CoI behaves different in each operating regime**

Key Determinants of Match Loss from Impatience: Sneak Peak

Theorem (Cost-of-Impatience Scaling by Operating Regime).



- **Scaling law, $\mathcal{S} \sim$ CoI:** Characterize how the CoI changes as a function of parameters

How Does Impatience Impact Match Rate?

1. Classify settings based on how impatience impacts match loss

   [Operating Regimes]

2. Identify key determinants of match loss from impatience

   [Scaling Laws]

Scaling Law Definition: A Universal Approximation

Definition (Scaling Law).

$$\mathsf{CoI} = \lambda_c - d \sim \mathcal{S}$$

when, **for "all" parameter combinations,**

$$\frac{1}{\Gamma} \leq \frac{\mathsf{CoI}(\lambda, \theta)}{\mathcal{S}(\lambda, \theta)} \leq \Gamma$$

for some function $\mathcal{S}(\lambda, \theta)$ and a constant $\Gamma \geq 1$ that does not depend on $\lambda = (\lambda_c, \lambda_s), \theta = (\theta_c, \theta_s)$. Recall that d denotes the actual match rate.

- **Scaling Law:** Characterize how the CoI changes as a function of parameters
- **Goal:** Identify the **relative importance** of each parameter on match loss

Theorem (Universal Cost-of-Impatience Scaling).

CoI $\sim \mathcal{S}$

$$= \theta_c \min\left\{\frac{\rho}{1-\rho}, \sqrt{\frac{\lambda_c}{\theta_c}}\right\}\left(1 + \left[1 + \frac{\rho}{1-\rho}\frac{\theta_c}{\lambda_c}\min\left\{\frac{\rho}{1-\rho}, \sqrt{\frac{\lambda_c}{\theta_c}}\right\}\right]\sqrt{\frac{\lambda_s}{\theta_s}}(1-\rho)e^{\frac{\lambda_s}{\theta_s}(1-\rho)^2\mathcal{H}(\rho)}\right)^{-1}$$

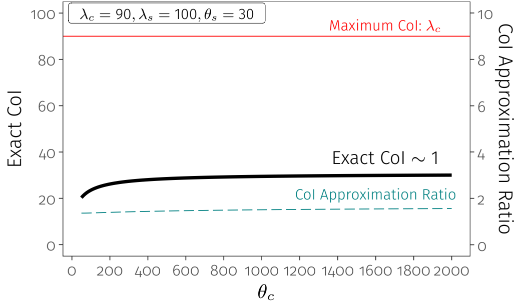
where $\rho = \lambda_c/\lambda_s$ and $\mathcal{H}(\rho) = \sum_{n=1}^{\infty} \frac{1}{n(n+1)}(1-\rho)^{n-1}$.

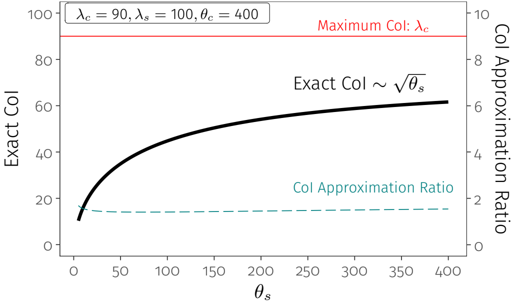
Exact CoI: CoI $= \theta_c \sum_{n=1}^{\infty} n \prod_{i=1}^{n} \frac{\lambda_c}{\lambda_s + i\theta_c} \left(1 + \sum_{n=1}^{\infty}\prod_{i=1}^{n}\frac{\lambda_c + i\theta_s}{\lambda_s} + \sum_{n=1}^{\infty}\prod_{i=1}^{n}\frac{\lambda_c}{\lambda_s + i\theta_c}\right)^{-1}$

Proof concept: **coupling arguments** and **expansions of explicit expressions** to upper- and lower-bound steady-state distributions

Universal CoI Scaling Law: An Example

Impatient-Customer, Impatient-Server Regime:





- **Scaling Law, $\mathcal{S}$:** Characterize how the CoI changes as a function of parameters
- **CoI Approximation Ratio** = $\mathcal{S}$ / Exact CoI

Theorem (Cost-of-Impatience Scaling by Operating Regime).

Patient-Customer

Impatient-Customer

$$\theta_{min}\frac{\rho}{1-\rho}$$

$$\sqrt{\theta_{min}\lambda_c}$$

Impatient-Server

Patient-Server

Server Impatience

Customer Impatience

"Winner-Take-All" Competition between Excess Capacity and Impatience

$$\mathbb{E}[Q_c | Q_s = 0] \sim \min\left\{\frac{\rho}{1-\rho}, \sqrt{\frac{\lambda_c}{\theta_c}}\right\}$$

Excess Capacity "Wins": M/M/1 Queue

Ignores Abandonment

Impatience "Wins": M/M/1+M Queue

Critically-Loaded, Ignores Excess Capacity

- In heavy traffic, $M/M/1 + M \sim M/M/1$ when $\sqrt{\theta} \ll 1 - \rho$ (Ward and Glynn, 2003)
- Only excess capacity OR impatience matters - not both
- Patient customers: CoI sensitive to small changes in excess capacity, ρ

Ability to Accumulate "Inventory" of Servers

Theorem (Cost-of-Impatience Scaling by Operating Regime).

Patient-Customer

Impatient-Customer

$$\theta_{min}\frac{\rho}{1-\rho}$$

$$\sqrt{\theta_{min}\lambda_c}$$

Server Impatience

Impatient-Server

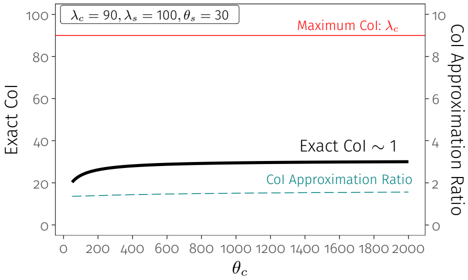
Patient-Server

Customer Impatience

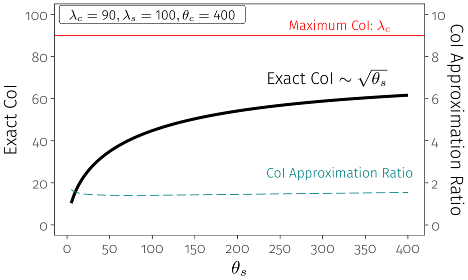
- **Patient Customers: Not just reduction to single-sided customer queue**

Only the Most Patient Type Matters ($\min\{\theta_c, \theta_s\}$)

Maximum Abandonment Rate (θ_c):



Minimum Abandonment Rate (θ_s):



- It only matters that one type is "patient enough"
- Lower match loss $\implies$ Focus on most patient type
- Waiting servers offset match loss

Theorem (Cost-of-Impatience Scaling by Operating Regime).

Patient-Customer | Impatient-Customer

Server Impatience

Impatient-Server:

$\longrightarrow$ $\sqrt{\theta_{min}\lambda_c}$

Patient-Server:

$\longrightarrow$ $\sqrt{\theta_{min}\lambda_c}\,e^{-\frac{\lambda_s}{\theta_s}(1-\rho)^2\mathcal{H}(\rho)}$

Customer Impatience

What is the optimal capacity level?

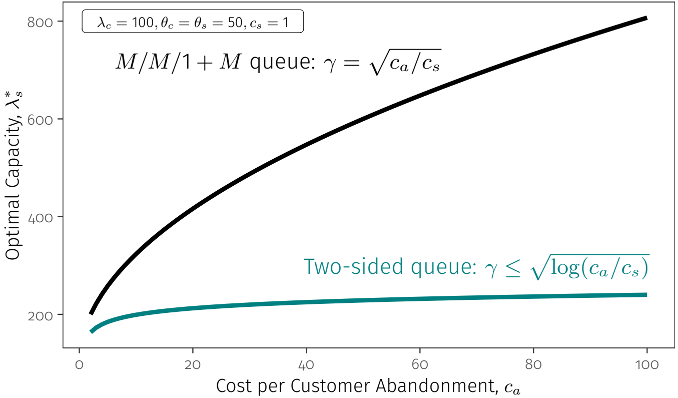
Optimal Capacity: Logarithmic Scaling in Abandonment Cost

Lemma (Optimal Capacity Scaling).

The optimal safety capacity, $\lambda_s^* - \lambda_c$, that balances abandonment and capacity costs, c_a and c_s, is:

$$\lambda_s^* - \lambda_c \sim \gamma\sqrt{\theta_{min}\lambda_c}$$

where
$\lambda_s^* = \text{argmin}_{\lambda_s \geq \lambda_c} \{c_a\theta_c\mathbb{E}[Q_c] + c_s\lambda_s\}.$



- Ability to hold inventory of servers $\Longrightarrow$ slower scaling of capacity

SUPPLY

TYPE O- (UNIVERSAL DONOR)

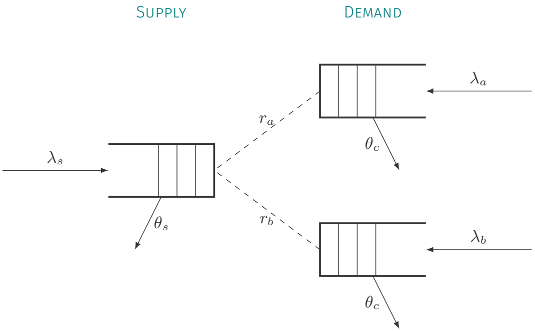
DEMAND

TYPE O- (HIGH REWARD)

TYPE AB+ (LOW REWARD)

- WHAT IS THE OPTIMAL TIMING OF MATCHES?

The Simplest Matching Model Involving a Match Decision



- **Goal:** Identify nearly optimal **policies** (specific policy for any combination of parameters)
- **Scaling Laws:** Lower bound on match loss
- Our Focus: simple model, universal results (in parameters) & general characterization

Thank you!

References

Afèche, P., Diamant, A., and Milner, J. (2014). Double-sided batch queues with abandonment: Modeling crossing networks. *Operations Research*, 62(5):1179–1201.

Aouad, A. and Saritaç, Ö. (2022). Dynamic Stochastic Matching under Limited Time. *Operations Research*, 70(4):2349–2383.

Aveklouris, A., DeValve, L., and Ward, A. R. (2021). Matching Impatient and Heterogeneous Demand and Supply.

Aveklouris, A., Puha, A. L., and Ward, A. R. (2023). A fluid approximation for a matching model with general reneging distributions. *Queueing Systems*.

Büke, B. and Chen, H. (2017). Fluid and diffusion approximations of probabilistic matching systems. *Queueing Systems*, 86:1–33.

Castro, F., Frazier, P., Ma, H., Nazerzadeh, H., and Yan, C. (2020a). Matching queues, flexibility and incentives. Available at SSRN.

Castro, F., Nazerzadeh, H., and Yan, C. (2020b). Matching queues with reneging: a product form solution. *Queueing Systems*, 96(3-4):359–385.

Chen, Y. and Hu, M. (2020). Pricing and matching with forward-looking buyers and sellers. *Manufacturing & Service Operations Management*, 22(4):717–734.

Collina, N., Immorlica, N., Leyton-Brown, K., Lucier, B., and Newman, N. (2020). Dynamic Weighted Matching with Heterogeneous Arrival and Departure Rates. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 12495 LNCS.

Conolly, B., Parthasarathy, P., and Selvaraju, N. (2002). Double-ended queues with impatience. *Computers & Operations Research*, 29(14):2053–2072.

Diamant, A. and Baron, O. (2019). Double-sided matching queues: Priority and impatient customers. *Operations Research Letters*, 47(3):219–224.

Lee, C. and Ward, A. R. (2019). Pricing and capacity sizing of a service facility: Customer abandonment effects. *Production and Operations Management*, 28(8):2031–2043.

Liu, X., Gong, Q., and Kulkarni, V. G. (2015). Diffusion Models for Double-Ended Queues with Renewal Arrival Processes. *Stochastic Systems*, 5(1):1–61.

Wang, G., Zhang, H., and Zhang, J. (2022). On-demand ride-matching in a spatial model with abandonment and cancellation. *Operations Research*, page forthcoming.

Ward, A. R. and Glynn, P. W. (2003). A Diffusion Approximation for a Markovian Queue with Reneging. *Queueing Systems*, 43(1-2).

Zubeldia, M., Jhunjhunwala, P. R., and Maguluri, S. T. (2022). Matching queues with abandonments in quantum switches: Stability and throughput analysis. arXiv preprint arXiv:2209.12324.