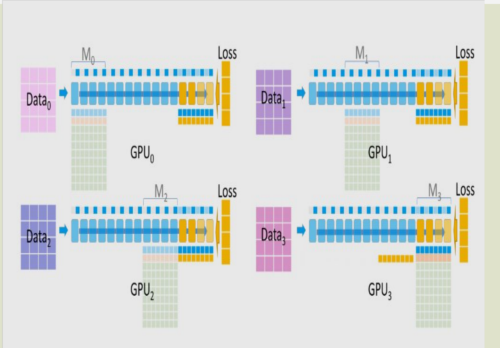
The Power of Two in Large Service-Marketplaces

Parimal Parag
Dheeraj Narasimha
Srinivas Nomula
Srinivas Shakkottai

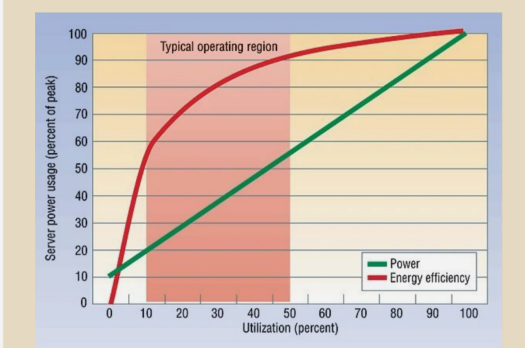
Electrical Communication Engineering,
Indian Institute of Science

ece.iisc.ac.in

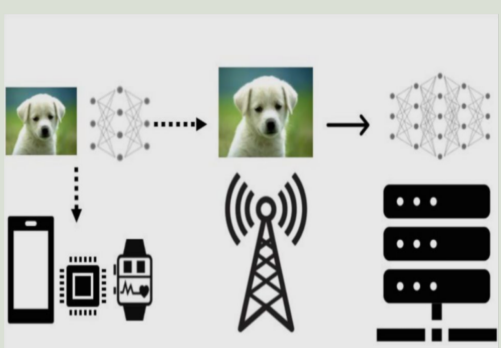
Distributed Systems



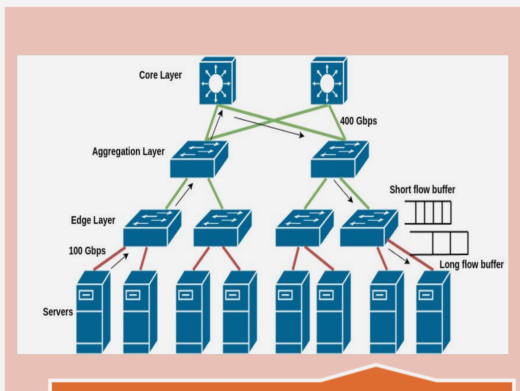
Distributed ML



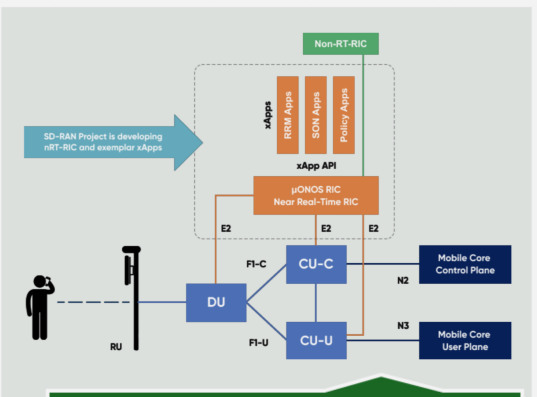
Sustainable computing



Heterogeneous computing and Hierarchical inference



DCN-optimized transport



Edge computing in private 5G/6G

Acknowledgements






CENTRE FOR
NETWORKED INTELLIGENCE
Indian Institute of Science

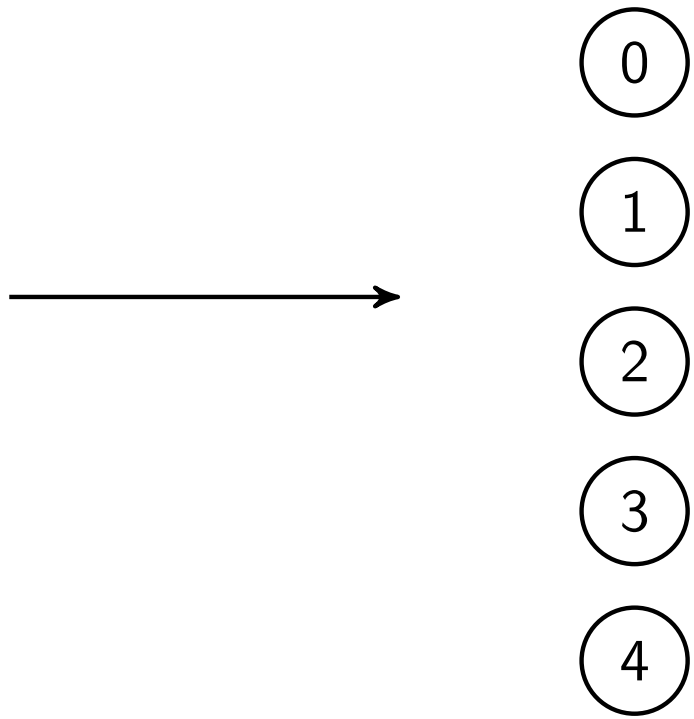
IBM

Qualcomm

NRF National Research Foundation

intel

Problem setup



Questions

Objective: Maximize revenue

- **Routing:** How to route arriving tasks?
- **Pricing:** How to price the service?

State-of-the-art

Revenue maximizing dynamic pricing

- For a single server queue
  - Random valuation: [Naor, 1969] [1], [Borgs et al, 2011] [2]
  - Arbitrary valuation: [Ashok et al, 2023], [3]
- Multiple servers with no queues and random valuation
  - Centralized routing and pricing:[Ashok et al, 2022] [4]
  - Our work: power-of-2 routing and rational pricing

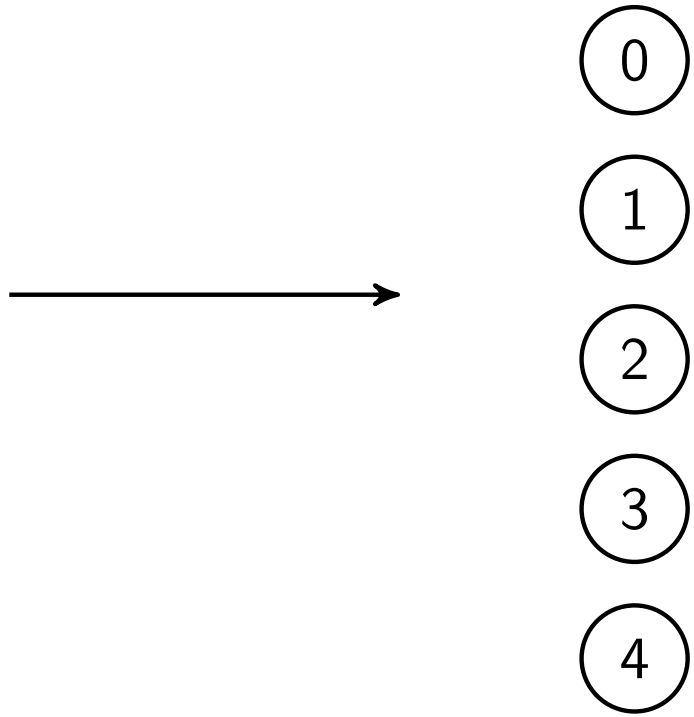
[1] P. Naor, "The regulation of queue size by levying tolls," Econometrica, vol. 37, no. 1, pp. 15–24, Jan. 1969.

[2] C. Borgs et al, "The optimal admission threshold in observable queues with state dependent pricing," Probability in the Engineering and Informational Sciences, vol. 28, no. 1, p. 101–119, 2014.

[3] Ashok et al., "Optimal pricing in a single server system," ACM Trans. Model. Perform. Eval. Comput. Syst., vol. 8, no. 4, pp. 1–32, Dec. 2023.

[4] Ashok et al, "Optimal pricing in multi server systems," Performance Evaluation, vol. 154, p. 102282, 2022.

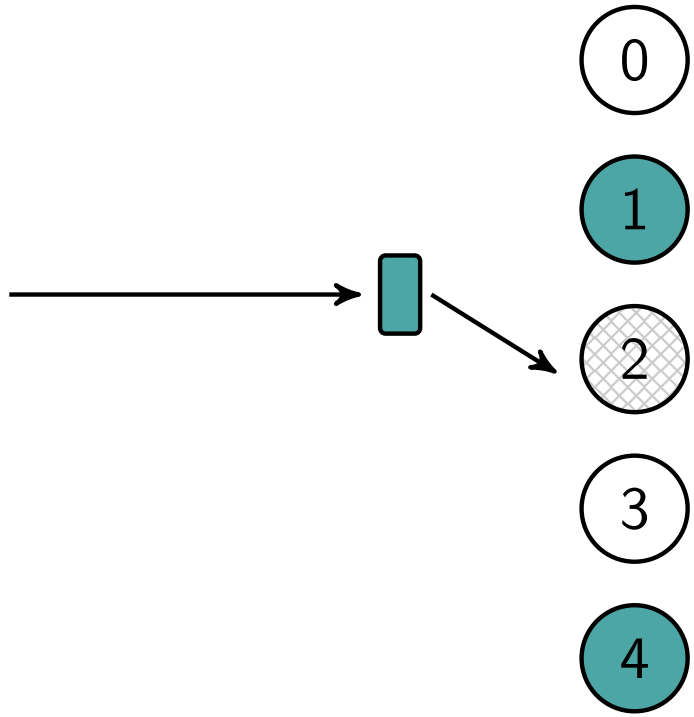
System model



N-server loss system

- Random *i.i.d.* unit mean exponential service times
- Poisson arrivals of rate $N\lambda$
- Server n is busy or idle denoted $X_n(t)$
- Random *i.i.d.* valuation with distribution G for each task

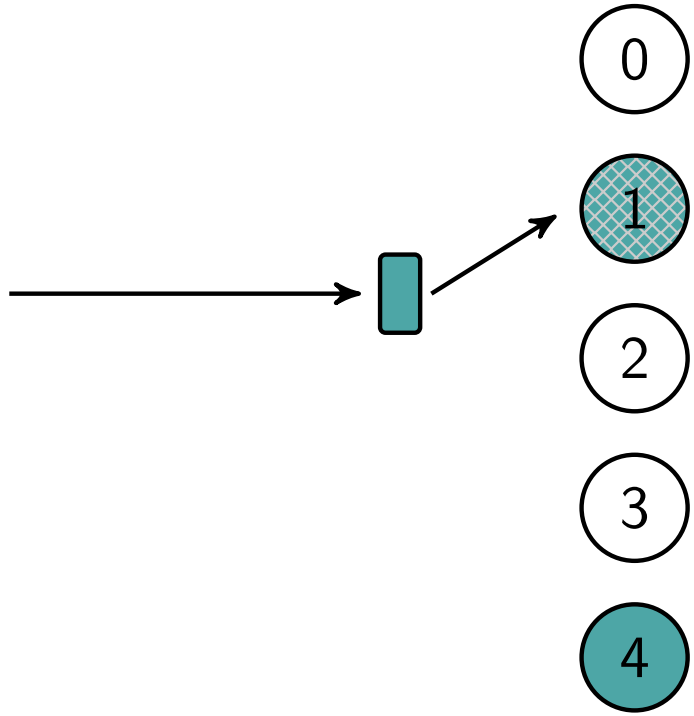
Deterministic routing D_1



Join an empty server

- Requires state information from all servers
- Loss only when all servers are busy
- Revenue if price less than valuation

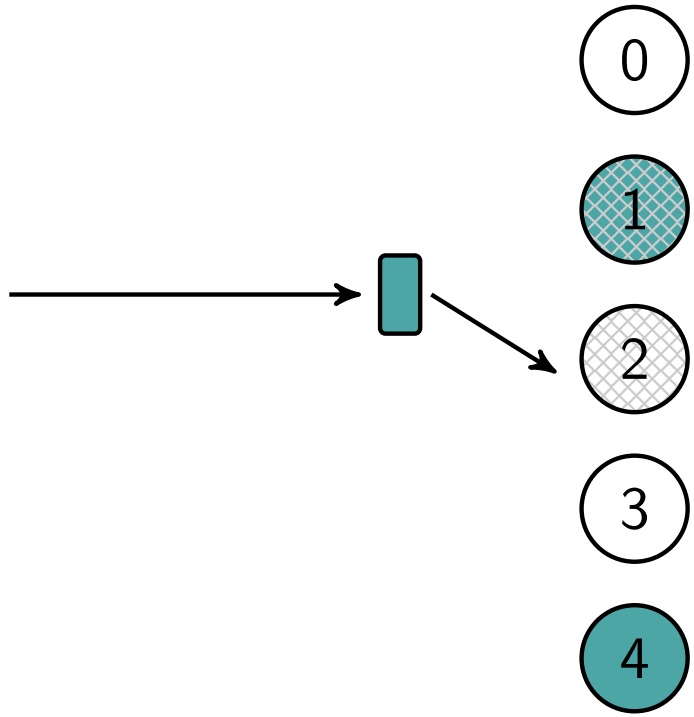
Random routing R_1



Join a random server

- Requires no server state feedback
- Loss when a busy server is selected
- No revenue can be generated

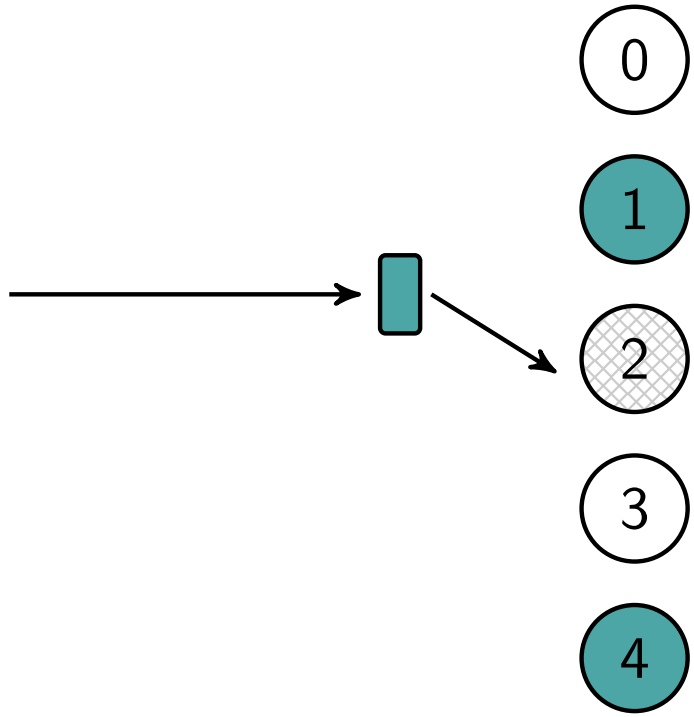
Power-of-two routing R_2



Join one of two randomly selected servers

- Requires server state feedback from two servers at each arrival
- Loss when both busy servers are selected
- No revenue if both servers are busy

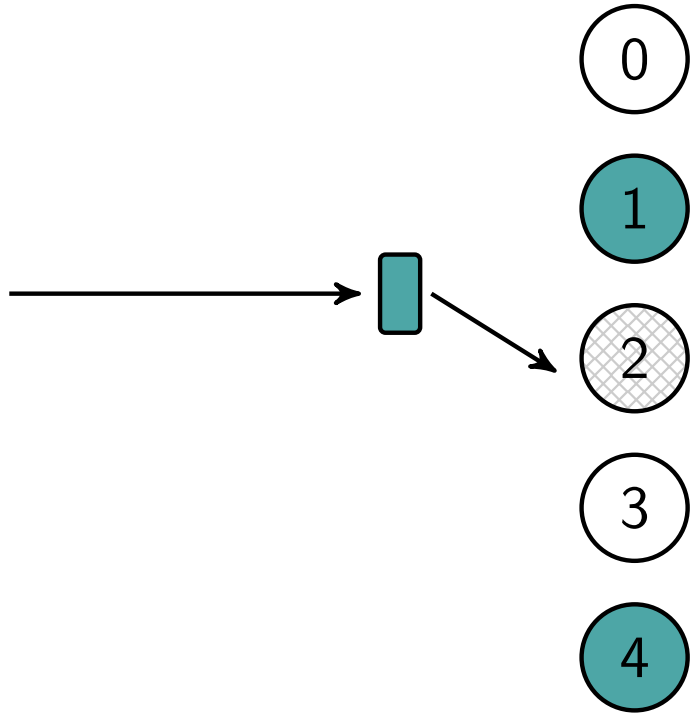
Pricing



Centralized and deterministic

- Centrally decided for all the servers
- Decided by individual servers
- Deterministic versus random

Centralized pricing for deterministic routing

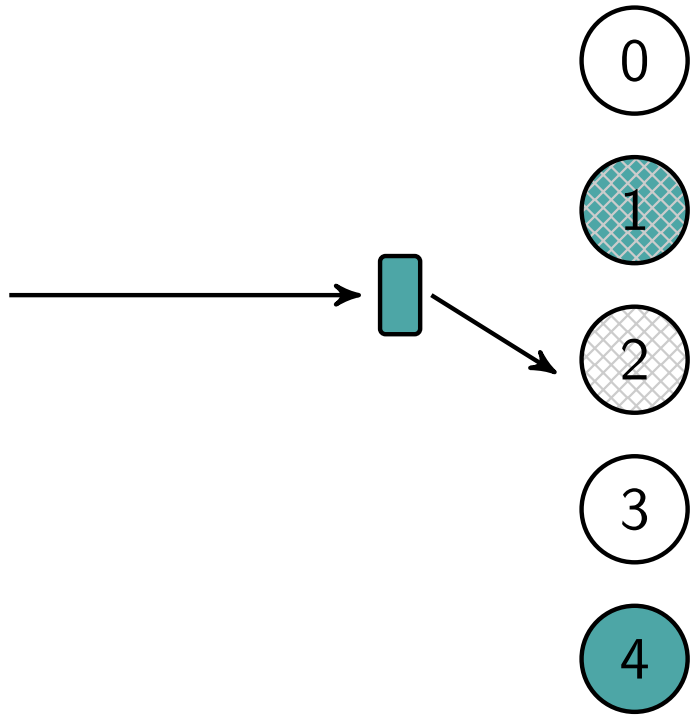


D_1C: State dependent pricing

- Revenue maximizing price given # busy servers [5]
- For large N state independent pricing maximizes revenue
- For price P at all servers, effective arrival rate $N\lambda\bar{G}(P)$
- For uniform pricing revenue rate per server is $\lambda P\bar{G}(P)$

[5] Ashok et al., "Optimal pricing in multi server systems," Performance Evaluation, vol. 154, p. 102282, 2022.

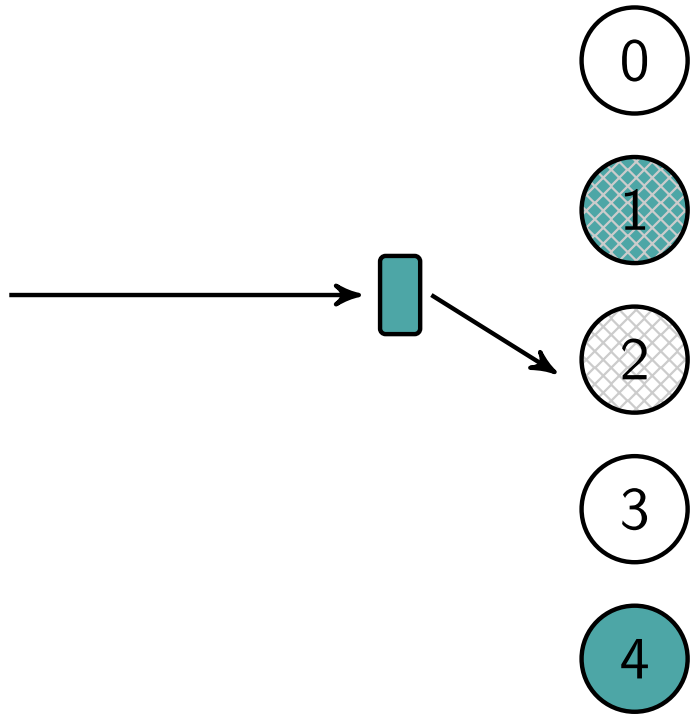
Decentralized pricing for power-of-2 routing



$\mathrm{R_2G}$: Mean-field game

- Task joins the idle server with lower price if lower than value
- Each server picks its own price based on the empirical average of busy servers

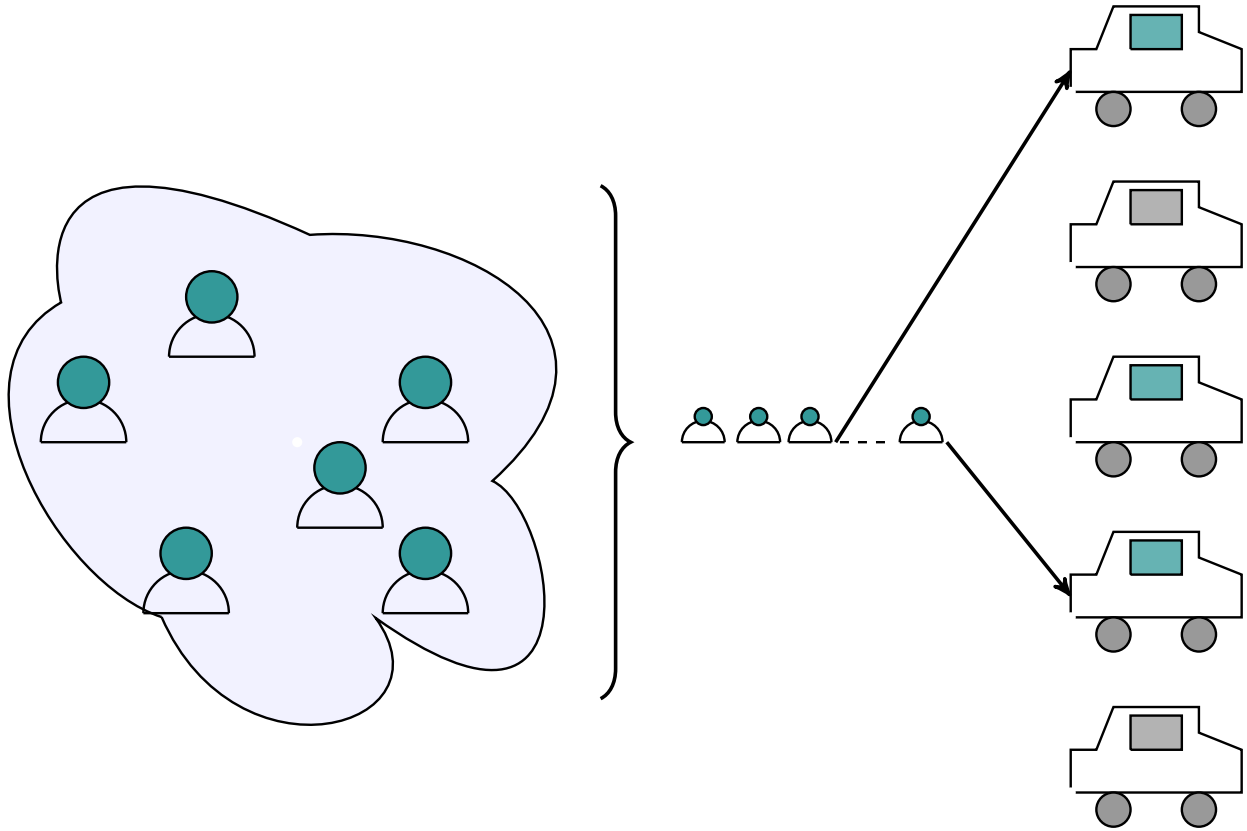
Problem Statement



$\mathrm{R_2G}$: Mean-field game

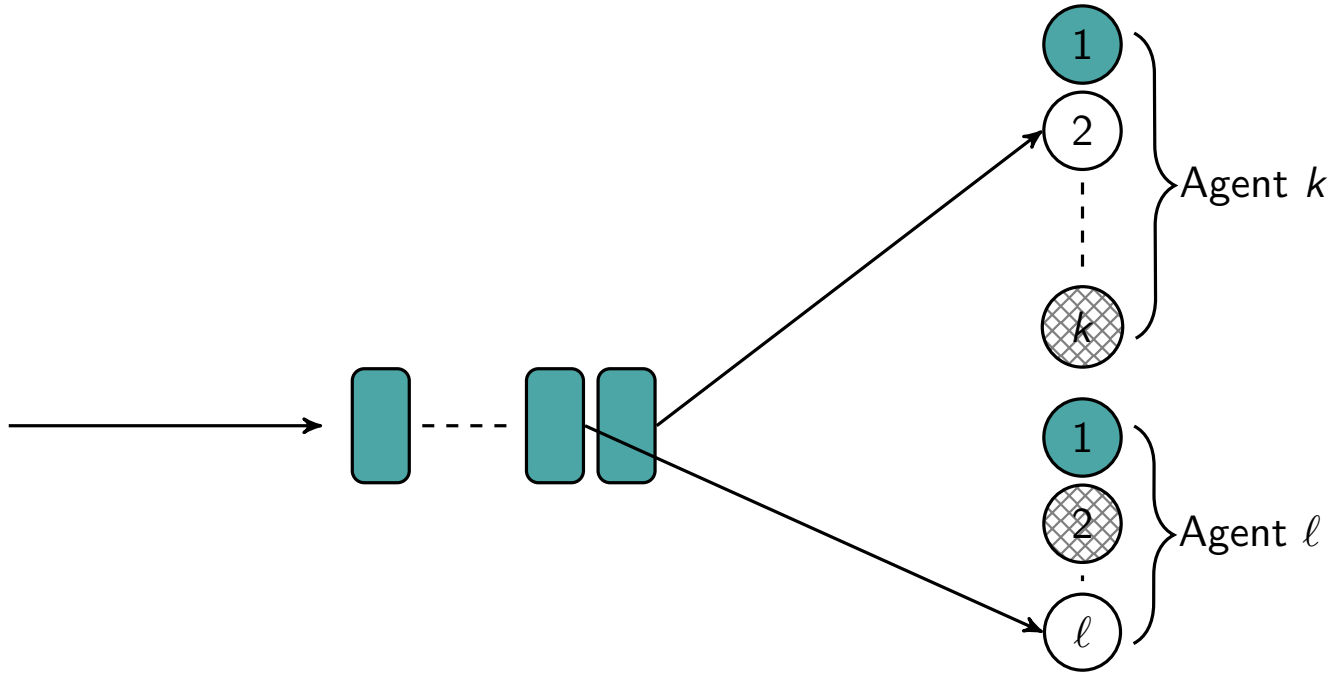
- Is there mean-field game equilibrium for this problem?
- Find the revenue rate under the mean-field game equilibrium

Ride sharing and on demand services



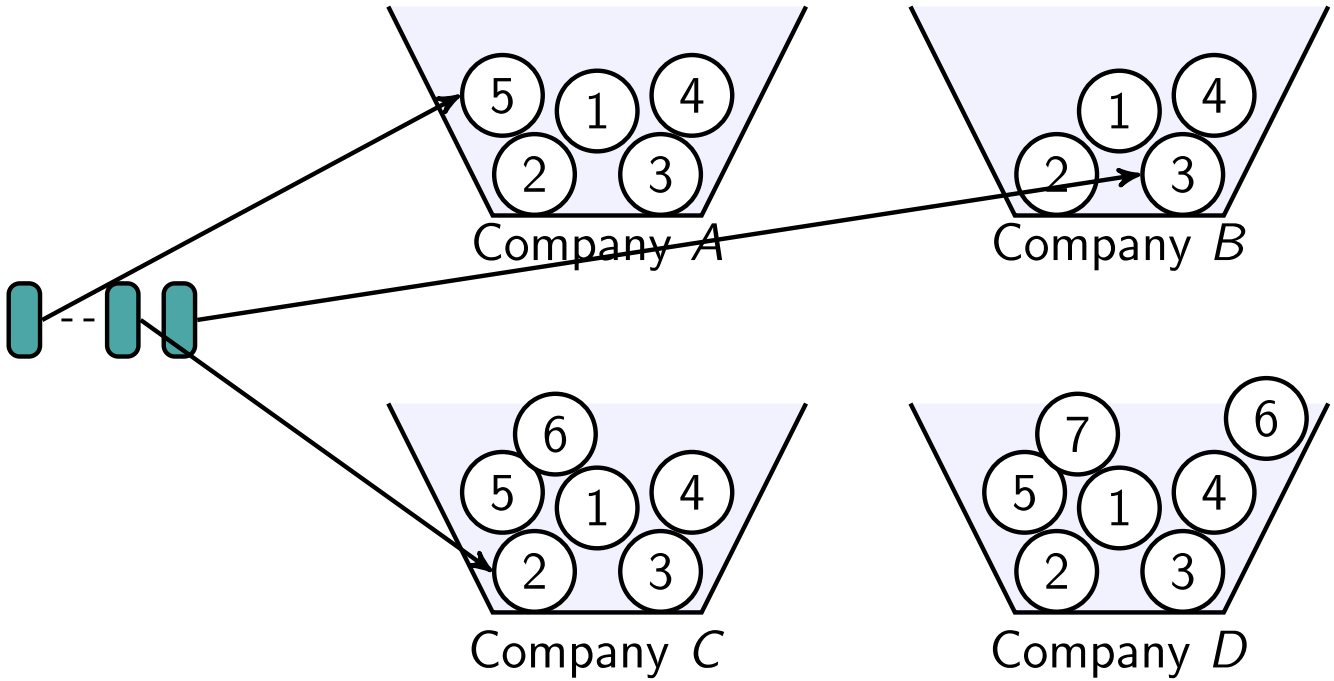
- Ride-hailing platforms like Uber and Lyft use dynamic pricing to match drivers with riders based on demand
- The two-server matching principle is similar to *two drivers competing for a ride* based on price and availability.

Online Cloud Marketplaces



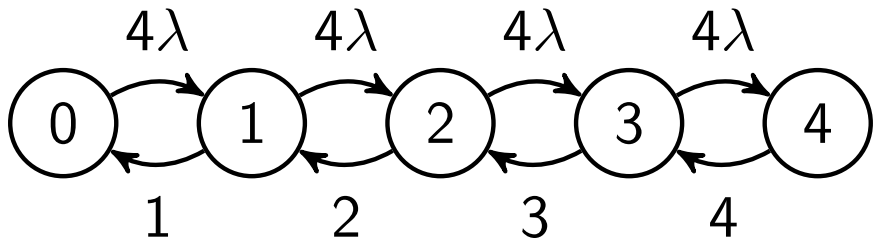
- Google cloud and AWS marketplace allow independent cloud service providers to list their services
- Multiple providers compete for customer jobs, similar to the two-server price competition model

Online stock marketing



- Each conglomerate has a list of stocks whose prices vary
- We assume that these variations follow a specific distribution

Deterministic routing D_1

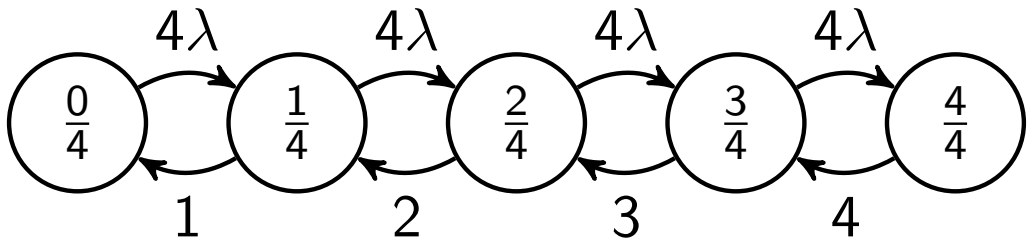


Number of busy servers $\sum_{n=1}^{N} X_n(t)$

- Evolve as a continuous time Markov chain with

$$Q_{x,x-1} = x, \qquad Q_{x,x+1} = N\lambda$$

Deterministic routing D_1



Fraction of busy servers $Z(t) \triangleq \frac{1}{N}\sum_{n=1}^{N} X_n(t)$

- Evolve as a continuous time Markov chain with

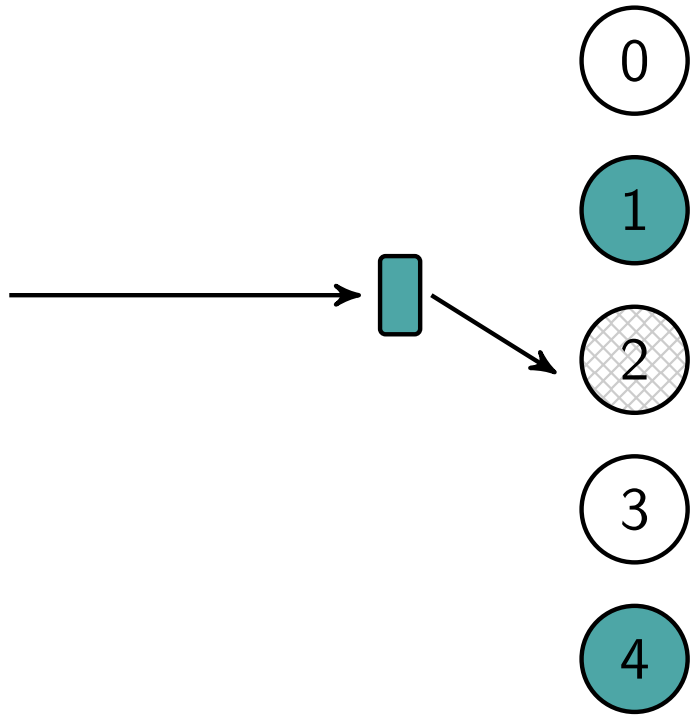
$$Q_{z,z-\frac{1}{N}} = Nz, \qquad Q_{z,z+\frac{1}{N}} = N\lambda$$

- Mean rate of change of fraction of busy servers is

$$f(z) \triangleq \sum_{w} Q_{z,w}(w-z) = \lambda - z$$

- Mean-field limit $\frac{dz}{dt} \approx f(z) = \lambda - z$
  - If $\lambda < 1$, then stationary fraction $z^* = \lambda$
  - If $\lambda > 1$, then stationary fraction $z^* = 1$

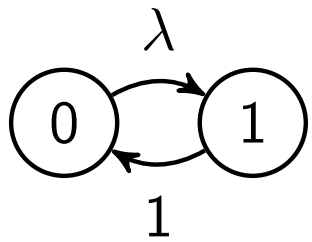
Centralized pricing for deterministic routing



$\mathrm{D}_1\mathrm{C}$ and uniform pricing

- Effective arrival rate $\lambda \bar{G}(P)$ for common price P
- If $\lambda \bar{G}(P) < 1$, then revenue rate is $\lambda P \bar{G}(P)$
- If $\lambda \bar{G}(P) > 1$, then revenue rate is P maximized at $\bar{G}^{-1}(1/\lambda)$

Random routing R_1

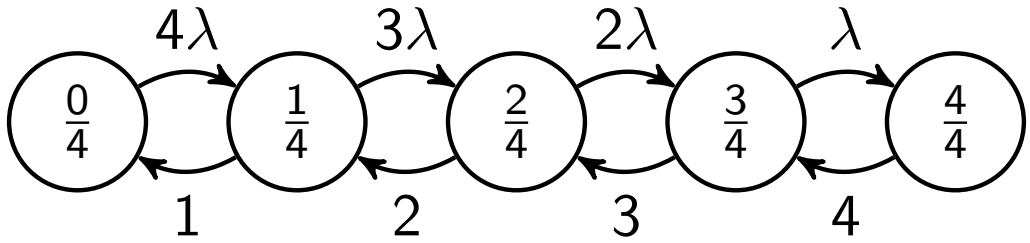


Number of busy servers

- Each server evolves independently as a continuous time Markov chain with

$$Q_{1,0} = 1, \qquad Q_{0,1} = \lambda$$

Random routing R_1



Fraction of busy servers

- Evolve as a continuous time Markov chain with

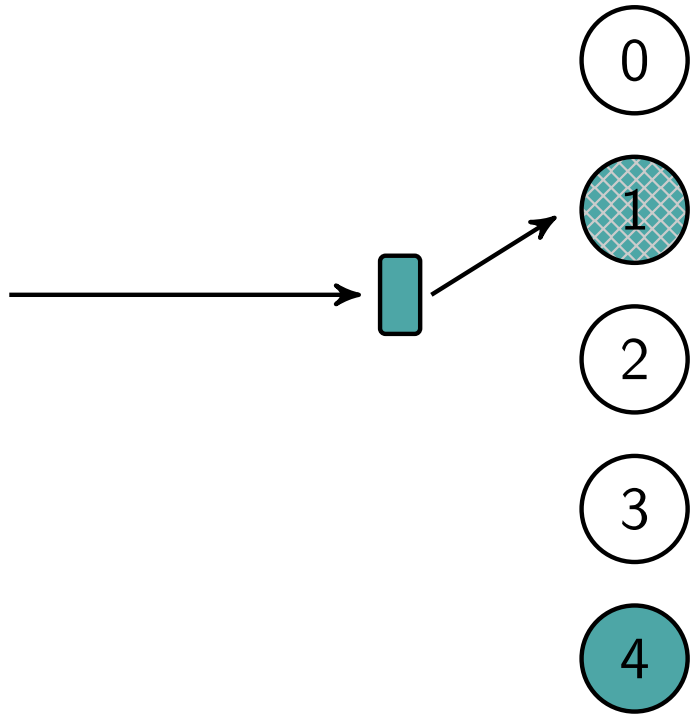
$$Q_{z,z-\frac{1}{N}} = Nz, \qquad Q_{z,z+\frac{1}{N}} = N\lambda(1-z)$$

- Mean rate of change of fraction of busy servers is

$$f(z) \triangleq \sum_{w} Q_{z,w}(w-z) = \lambda(1-z) - z$$

- Mean-field limit $\frac{dz}{dt} \approx f(z) = \lambda(1-z) - z$
  - Stationary fraction $z^* = \frac{\lambda}{1+\lambda}$

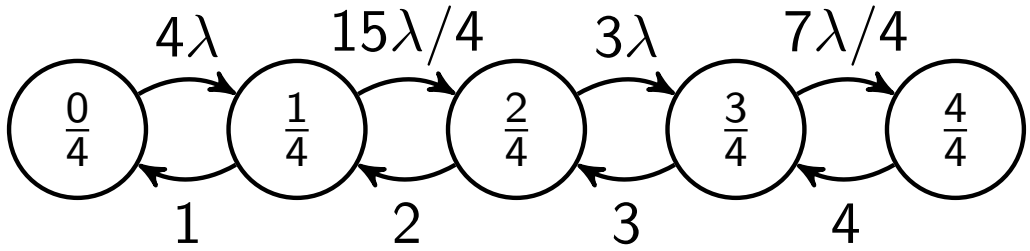
Centralized pricing for random routing



$\mathrm{R}_1\mathrm{C}$ and uniform pricing

- Effective arrival rate $\lambda \bar{G}(P)$ for common price P
- Stationary fraction $z^* = \frac{\lambda \bar{G}(P)}{1+\lambda \bar{G}(P)}$
- Revenue rate is $\lambda(1 - z^*)P\bar{G}(P)$

Power of two routing R_2



Fraction of busy servers

- Evolve as a continuous time Markov chain with

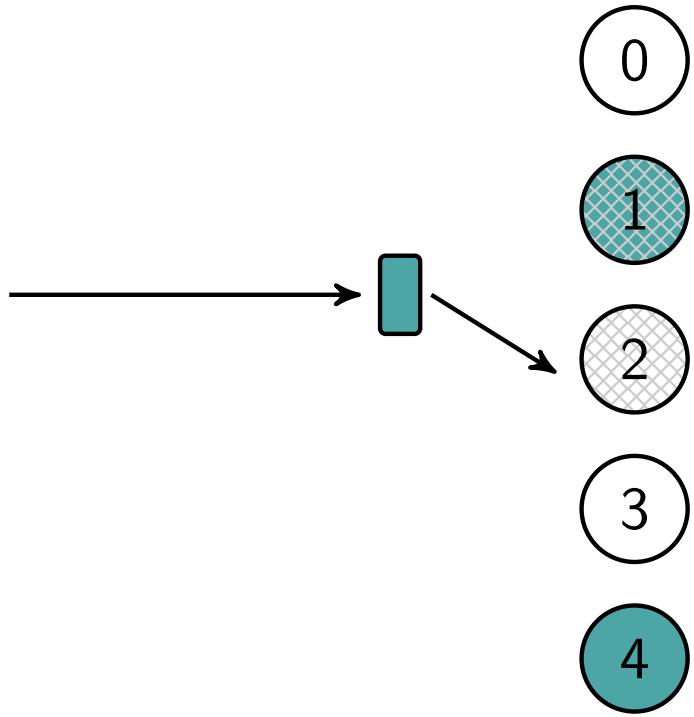
$$Q_{z,z-\frac{1}{N}} = Nz, \qquad Q_{z,z+\frac{1}{N}} = N\lambda(1-z^2)$$

- Mean rate of change of fraction of busy servers is

$$f(z) \triangleq \sum_w Q_{z,w}(w-z) = \lambda(1-z^2) - z$$

- Mean-field limit $\frac{dz}{dt} \approx f(z) = \lambda(1-z^2) - z$
  - Stationary fraction $z^* = -\frac{1}{2\lambda} + \sqrt{1+\frac{1}{4\lambda^2}}$

Centralized pricing for power of two routing



$\mathrm{R_2C}$ and uniform pricing

- Effective arrival rate $\lambda \bar{G}(P)$ for common price P
- Stationary fraction $z^* = -\frac{1}{2\lambda \bar{G}(P)} + \sqrt{1 + \frac{1}{4\lambda^2 \bar{G}(P)^2}}$
- Revenue rate is $\lambda(1 - z^{*2}) P \bar{G}(P)$

Mean-field game

Approach

- Valuation distribution is exponential with rate v
- Servers in $[N]$ follow same pricing, *i.i.d.* exponential price with rate d_1
- Fraction of busy servers $Z_t^N \triangleq \frac{1}{N}\sum_{n=1}^{N} X_{t,n}$
- Find mean-field limit $z^*(d_1) = \lim_{N\to\infty}\lim_{t\to\infty} Z_t^N$ of the fraction of busy servers
- Tag server 0 that has exponential price with rate d_0
- Find revenue rate of server 0 given d_1
- Choose best response rate $d_0^*(d_1)$ that maximizes revenue rate of server 0
- Is there a mean field game equilibrium?
- What is the per server revenue rate at this equilibrium?

System evolution

Admission indicators

- For kth arrival: task valuation V_k, price $P_{k,n}$ at server n
- Admission indicators

$$\eta_{k,10} \triangleq \mathbb{1}_{\{V_k > P_{k,0}\}}, \qquad \eta_{k,20} \triangleq \mathbb{1}_{\{V_k > P_{k,0}, P_{k,0} < P_{k,n}\}},$$
$$\zeta_{k,1} \triangleq \mathbb{1}_{\{V_k > P_{k,n}\}}, \qquad \zeta_{k,2} \triangleq \mathbb{1}_{\{V_k > P_{k,n} \wedge P_{k,m}\}}.$$

- Admission probabilities

$$q_1 \triangleq \mathbb{E}\eta_{k,10}, \qquad q_{20} \triangleq \mathbb{E}\eta_{k,20}, \qquad p_1 \triangleq \mathbb{E}\zeta_{k,1}, \qquad p_2 \triangleq \mathbb{E}\zeta_{k,2}.$$

Evolution

Selection indicator for tagged server 0 by the kth task

$$\xi_k^N = \mathbb{1}_{\{0 \in I_k\}} \bar{X}_{A_k,0} \sum_{n=1}^{N} \mathbb{1}_{\{n \in I_k\}} \left(X_{A_k,n} \eta_{k,10} + \bar{X}_{A_k,n} \eta_{k,20} \right).$$

System evolution

Generator matrix

The process $(X_{t,0}, Z_t^N)$ is a CTMC with the generator matrix Q^N defined as

$$Q^N_{(x,z),(y,w)} = \begin{cases} Nz, & w = z - \frac{1}{N}, y = x \\ \lambda\bar{z}(2p_1(x + Nz) + 2\bar{x}q_{21} + p_2(N\bar{z} - 1)), & w = z + \frac{1}{N}, y = x \\ x, & w = z, y = x - 1, \\ 2\lambda\bar{x}(zq_1 + \bar{z}q_{20}), & w = z, y = x + 1. \end{cases}$$

Mckean-Vlasov equation

Consider an autonomous dynamic system $\dot{z} = h(z)$, where

$$h(z) \triangleq \lim_{N\to\infty} \sum_{y,w} Q^N_{(x,z),(y,w)}(w - z) = \lambda\bar{z}(2zp_1 + \bar{z}p_2) - z.$$

Limiting fraction of busy servers

Let $\alpha \triangleq \frac{\nu}{d_1}$ and $x \triangleq \frac{1}{2}\left(\alpha + \frac{(1+\alpha)(2+\alpha)}{2\lambda}\right)$, then the unique rest point z^* such that $h(z^*) = 0$ is

$$z^* \triangleq -x + \sqrt{1 + \alpha + x^2}.$$

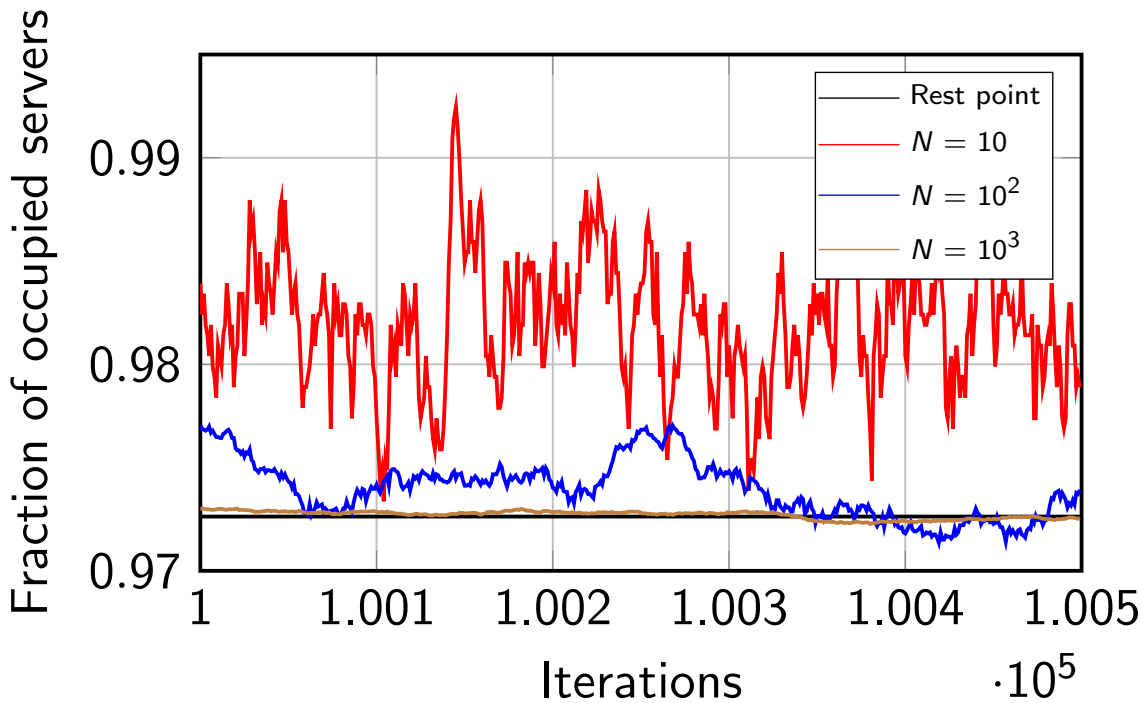
Our Contributions

- Calculated the deterministic occupancy z^* of the sub-system using Mckean-Vlasov equation
- Derived the tagged server's limiting revenue expression as a function of z^*, price and value rates
- Designed an algorithm that plays a game between the agents to choose the optimum price parameter which maximizes their revenue
- Derived the numerical results for mean price, limiting revenue and throughput of ours' as well as the state-of-art techniques and compared them

Mean-field convergence

Mean-field convergence

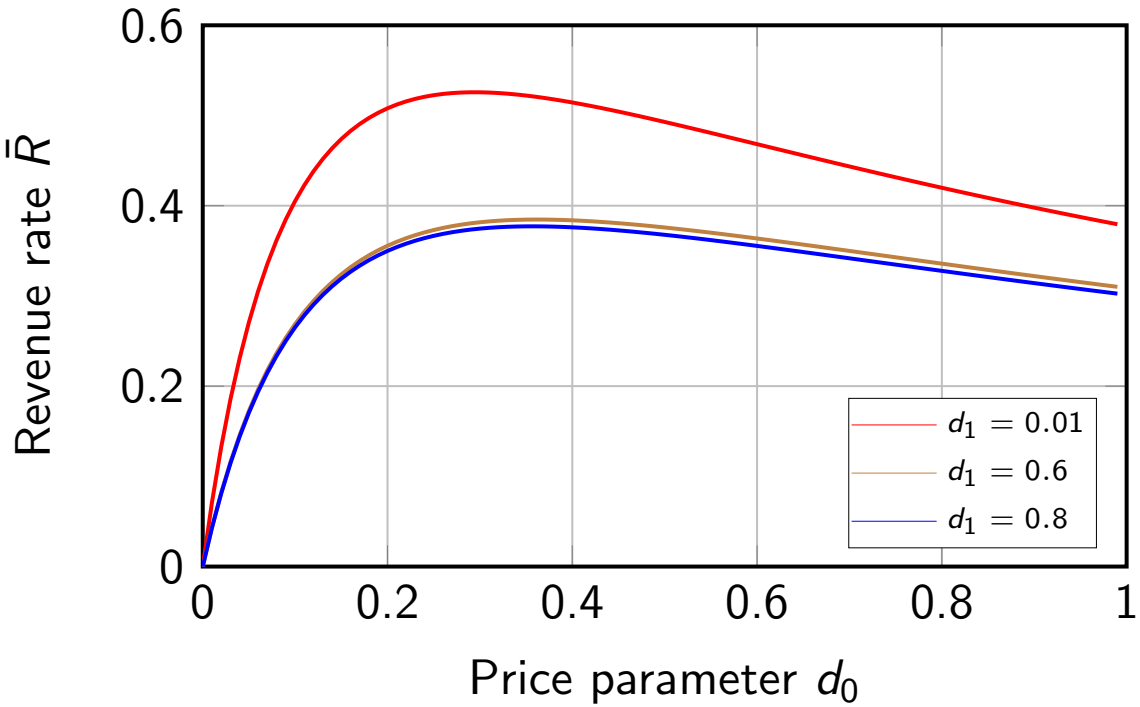
The stationary fraction of busy servers Z_∞^N converges in the mean-square sense to unique rest point z^* of mean-field model with rate $\frac{1}{N}$. That is, $\lim_{N\to\infty}\lim_{t\to\infty}\mathbb{E}\left|Z_t^N - z^*\right|^2 = O\left(\frac{1}{N}\right)$



Tagged server revenue

Limiting revenue rate at tagged server 0

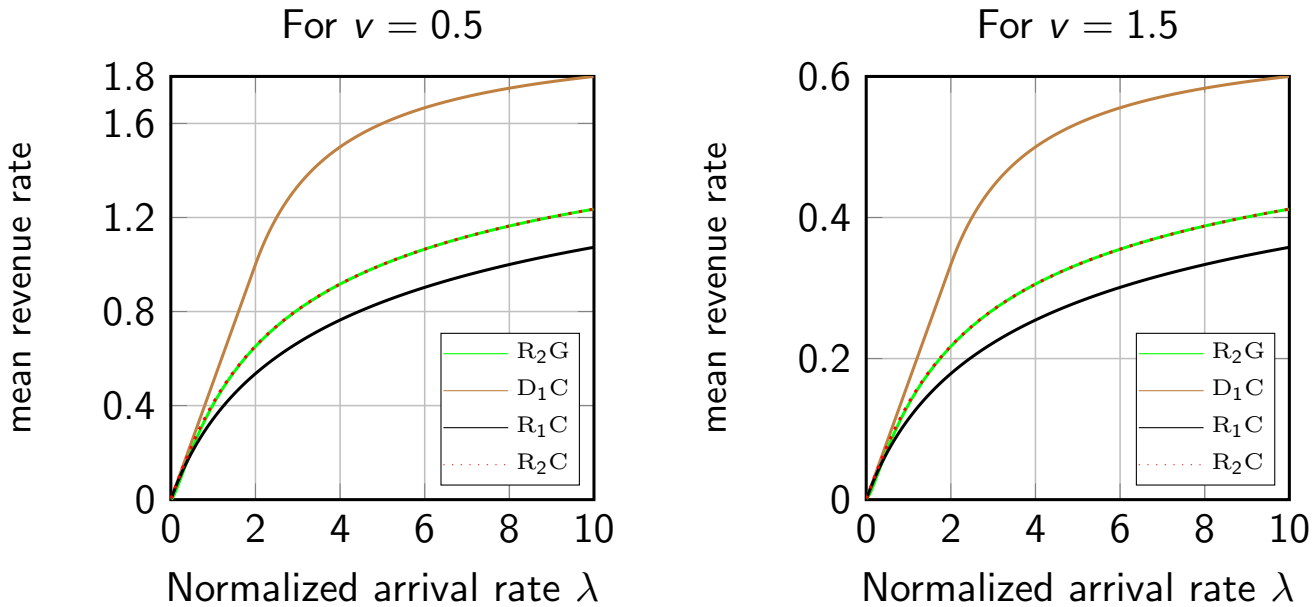
$$\bar{R} \triangleq \lim_{N\to\infty} \lim_{K\to\infty} \frac{1}{A_K} \sum_{k=0}^{K} P_{k,0} \xi_k^N = \frac{z^* q_1^2 + (1-z^*) q_{20}^2}{d_0 (\frac{1}{2\lambda} + z^* q_1 + (1-z^*) q_{20})}$$



- We can show $d_0 \mapsto z^* \mapsto d_0^*$ is composition of continuous maps
- There exists a fixed-point which is the mean-field game equilibrium

Performance comparison

Mean revenue rate



- $\mathrm{D_1C}$ has the best revenue rate at the cost of highest server feedback
- $\mathrm{R_2G}$ has same performance as $\mathrm{R_2C}$ without coordinated pricing
- $\mathrm{R_1C}$ has the worst performance since it is completely agnostic of system state

References

- P. Naor, "The regulation of queue size by levying tolls," Econometrica, vol. 37, no. 1, pp. 15–24, Jan. 1969.
- C. Borgs et al, "The optimal admission threshold in observable queues with state dependent pricing," Probability in the Engineering and Informational Sciences, vol. 28, no. 1, p. 101–119, 2014.
- A. Krishnan K.S., C. K. Singh, S. T. Maguluri, and P. Parag, "Optimal pricing in multi server systems," Performance Evaluation, vol. 154, p. 102282, 2022.
- L. Ying, "On the approximation error of mean-field models," in ACM SIGMETRICS Inter. Conf. Meas. Model. Comp. Sci., Jun. 2016, pp. 285–297.
- A. Krishnan K.S., C. K. Singh, S. T. Maguluri, and P. Parag, "Optimal pricing in a single server system," ACM Trans. Model. Perform. Eval. Comput. Syst., vol. 8, no. 4, pp. 1–32, Dec. 2023.
- N. D. Vvedenskaya, R. L. Dobrushin, and F. I. Karpelevich, "Queueing system with selection of the shortest of two queues: An asymptotic approach," Prob. Info. Transmission, vol. 32, no. 1, pp. 15–27, 1996.
- D. Narasimha, S. Nomula, S. Shakkottai, and P. Parag, "The Power of Two in Large Service-Marketplaces", in IEEE Conf. Comp. Commun. (INFOCOM), 2025, pp. 1–10.