Queues, Tolls and Welfare

Tejas Bodas [1] A J Ganesh [2] D. Manjunath [1]

[1]E.E. Dept, IIT Bombay

[2]School of Mathematics, University of Bristol

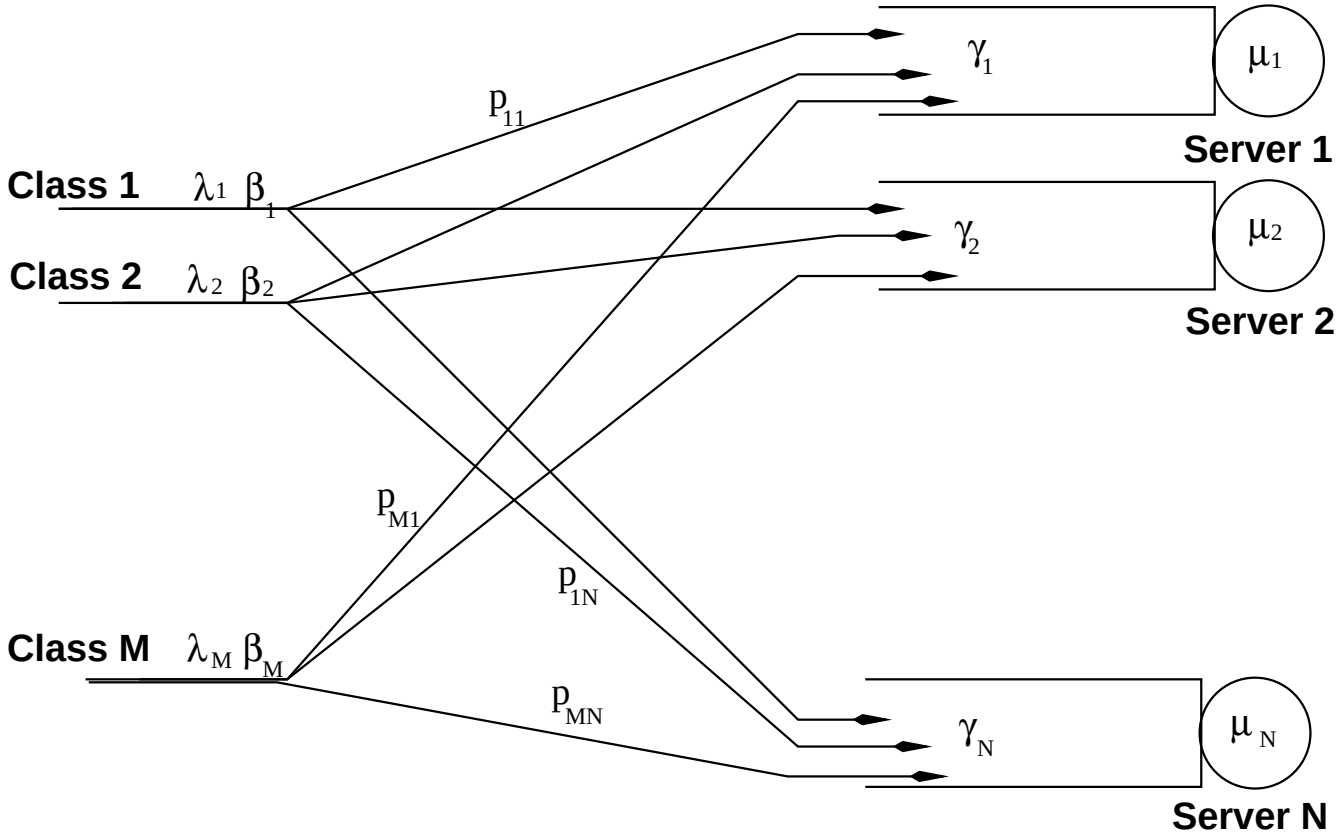
5th November, 2014

Motivation

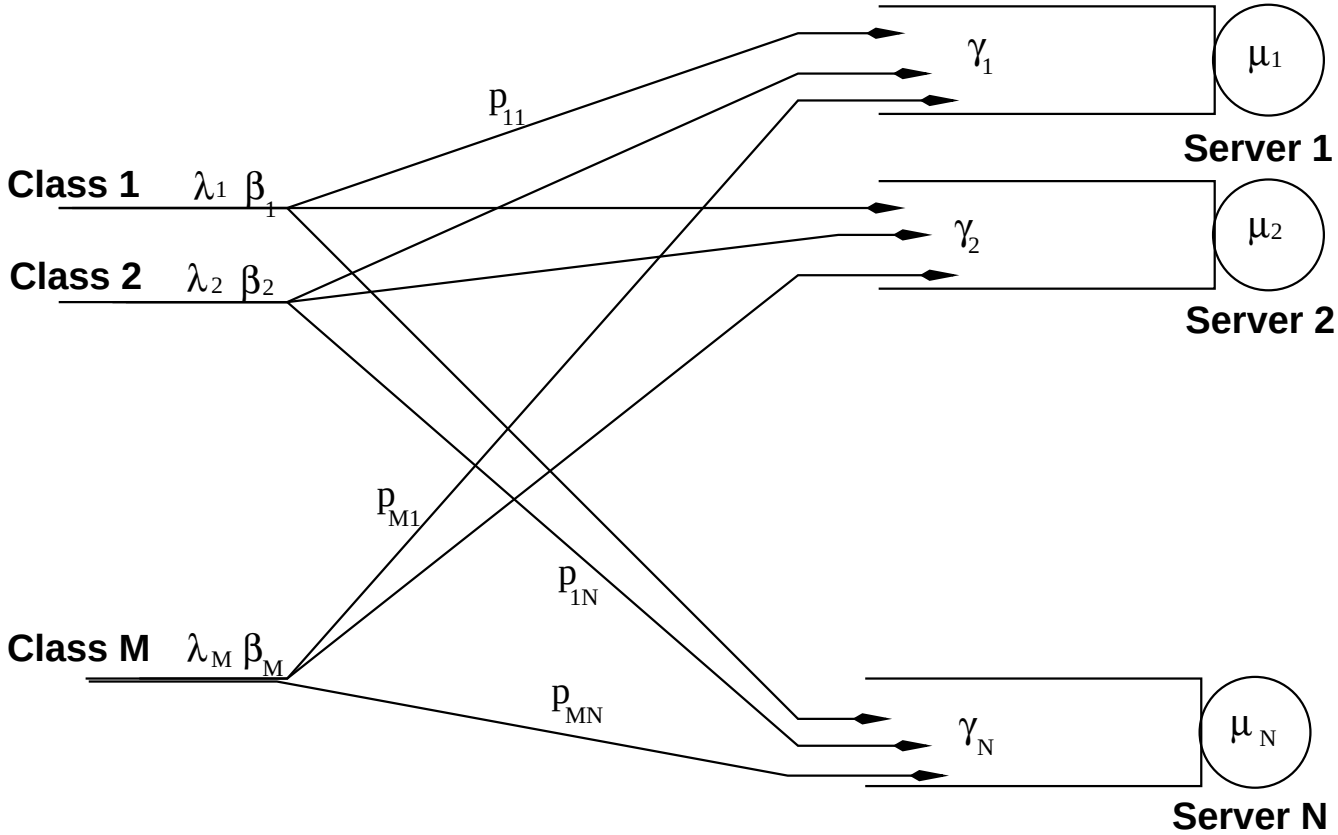
- Service systems typically handle **heterogenous customers**.
- Some examples from our daily lives ...
  1. Public transport
     - Different aversions to crowding.
  2. Road transport
     - Different sensitivities to congestion, delay, and ride quality.
  3. The Internet.
     - Different loss and delay requirements for different traffic streams, e.g., games, voice, video, data.
- Efficient routing of heterogeneous customers in service systems is a very relevant problem.

A formal description of the setting

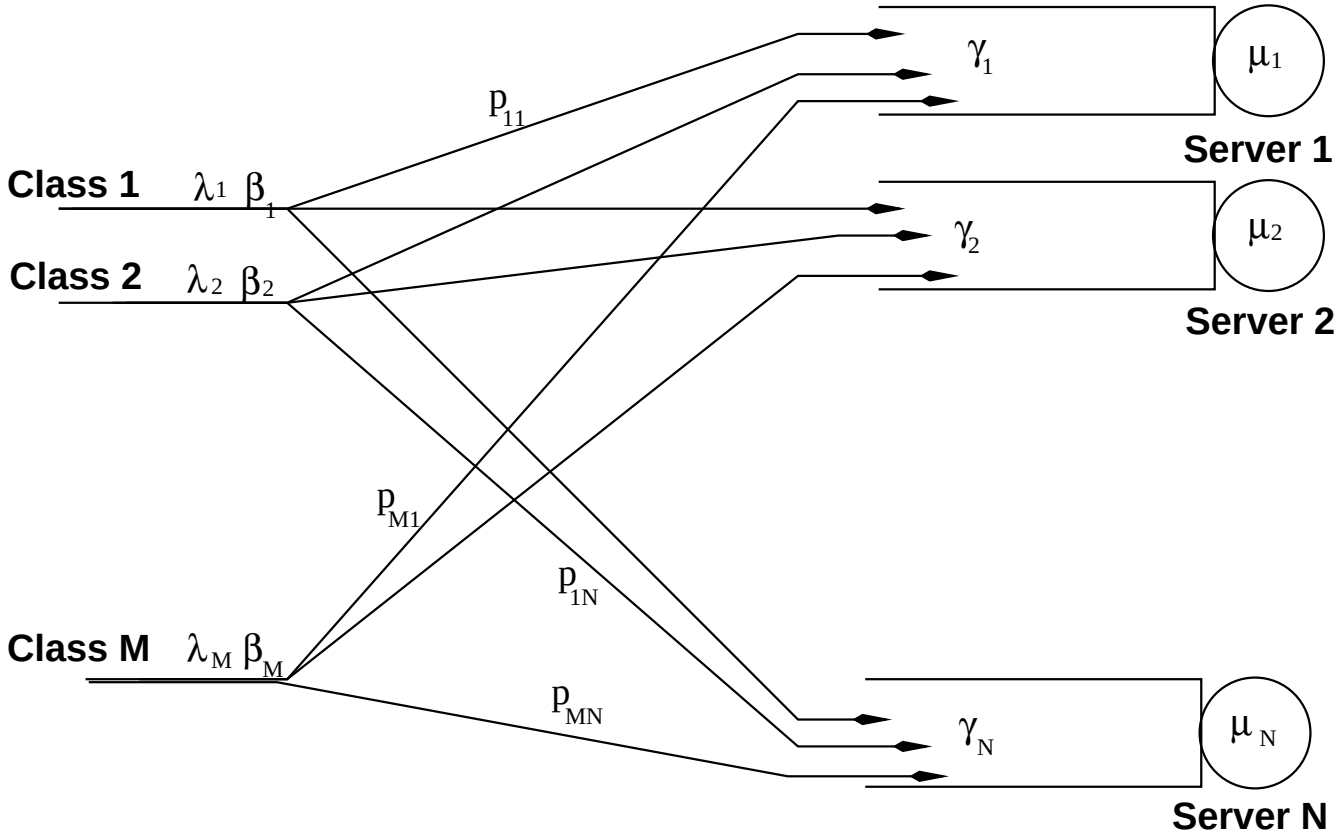
- Heterogeneous parallel queues each with its own server.
- Heterogeneous customer classes
  - Each class has a different valuation of its delay function.
- Decision: *Oblivious* routing of customers to different queues.
- Two kinds of routing schemes are of interest.
  - Social welfare maximisation (will be made precise).
  - Selfish or individually optimal decisions.



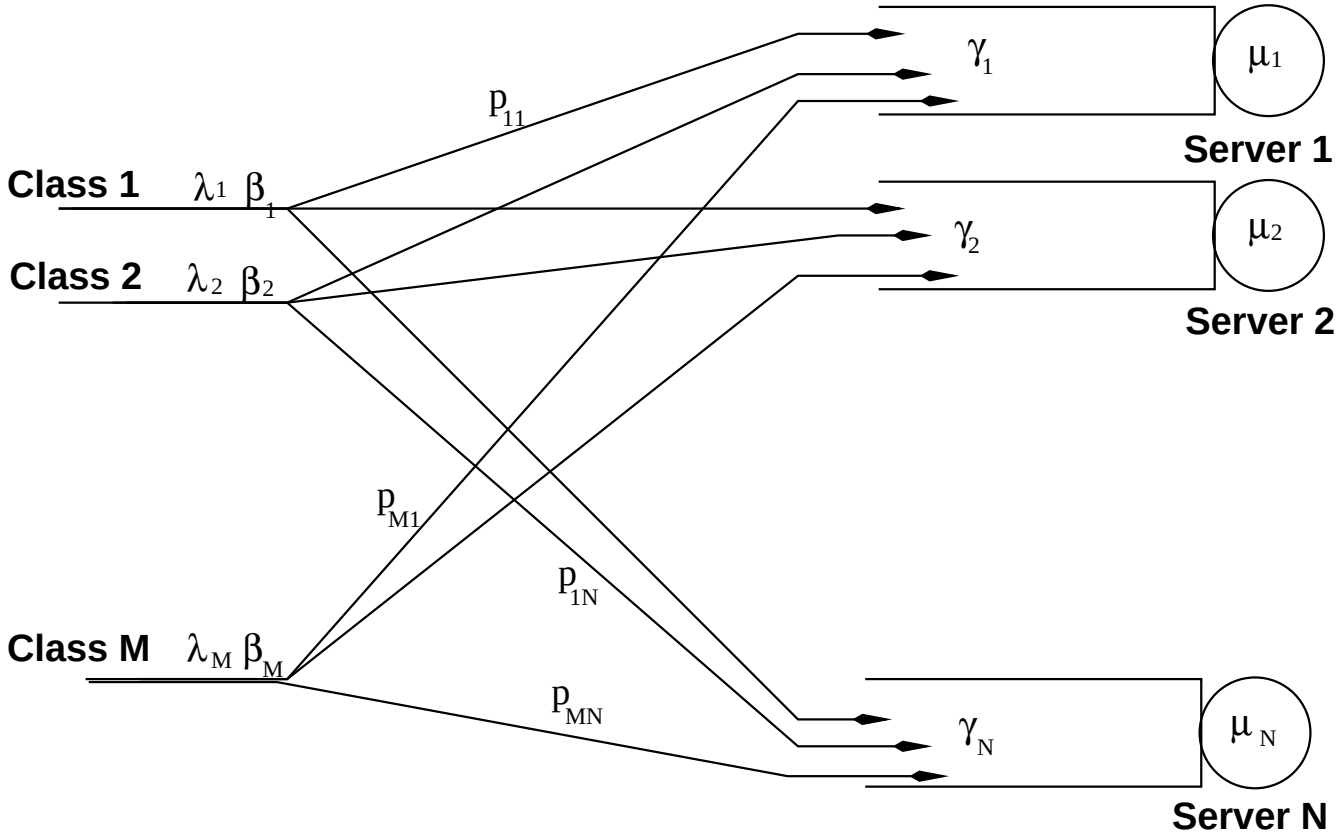
- M classes of customers and N queues.
- Class m arrives according to a Poisson process of rate λ_m.
- Each queue has its own service discipline that is arbitrary and does not depend on class or queue length.



- Classes have the same service time distribution for a job.
- β_m (delay sensitivity) is cost of waiting per unit time of a class m job.
- WLOG, $\beta_m > \beta_{m+1}$

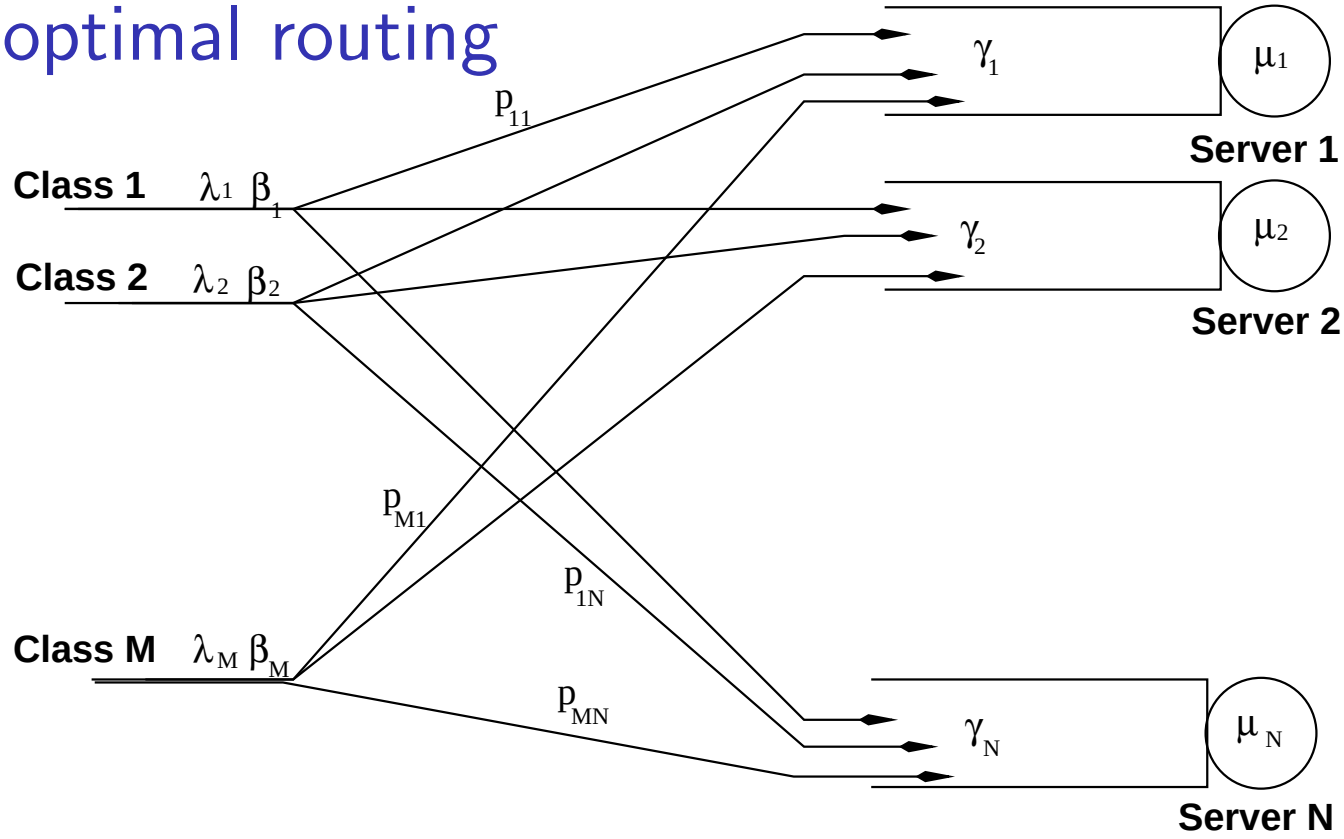


- Oblivious routing
  - Routing decision does not depend on the current queue length, previous arrival times, previous routing decisions etc.
  - p_{mn} is probability that class m customer is routed to queue n.
  - $P = [[p_{mn}]]$ is a stochastic matrix.
  - $\gamma_n = \sum_{m=1}^{M} p_{mn}\lambda_m$ is the offered traffic to queue n.



- $D_n(\gamma_n)$ is the "delay function" in Queue n and it depends only on the total traffic γ_n. We assume that D_n is
  - increasing and continuous and
  - differentiable in the interior of its domain (arrival rates for which D_n is finite) with a strictly positive derivative.
- D_n is fairly generic.

Socially optimal routing



- Social cost

$$U(P) = \sum_m \lambda_m \left[\beta_m \left(\sum_n p_{mn} D_n(\gamma_n) \right) \right]$$

Socially optimal routing

- The social welfare maximization problem is

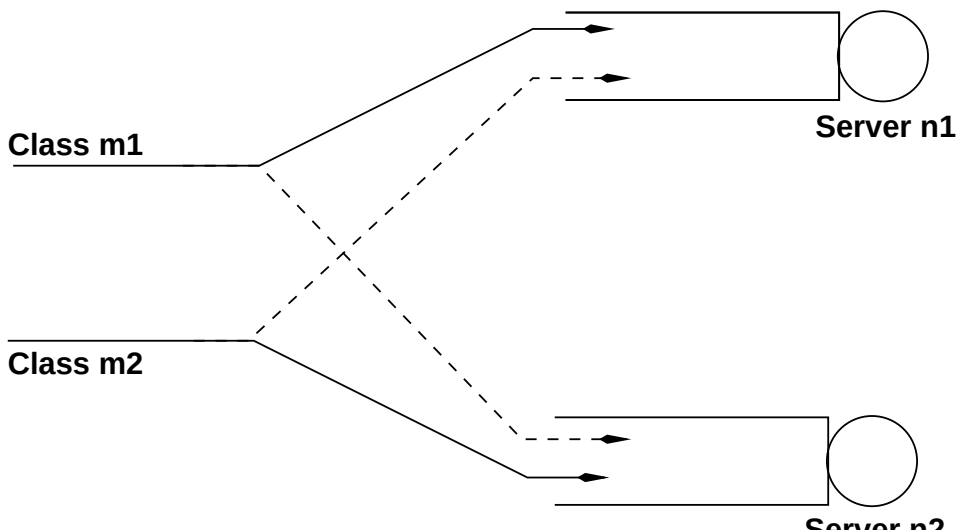
$$\min_{P} U(P) = \sum_{i=1}^{M} \sum_{j=1}^{N} \beta_i \lambda_i p_{ij} D_j(\gamma_j), \tag{1}$$

- $U(P)$ is a continuous function.
- The minimum is taken over all right stochastic matrices P which is a convex set.
- There is a matrix P^* achieving the infimum.
- However $U(P)$ need not be a convex function. This can be proved for the case of $M/M/1$ servers.

Characterizing P^*

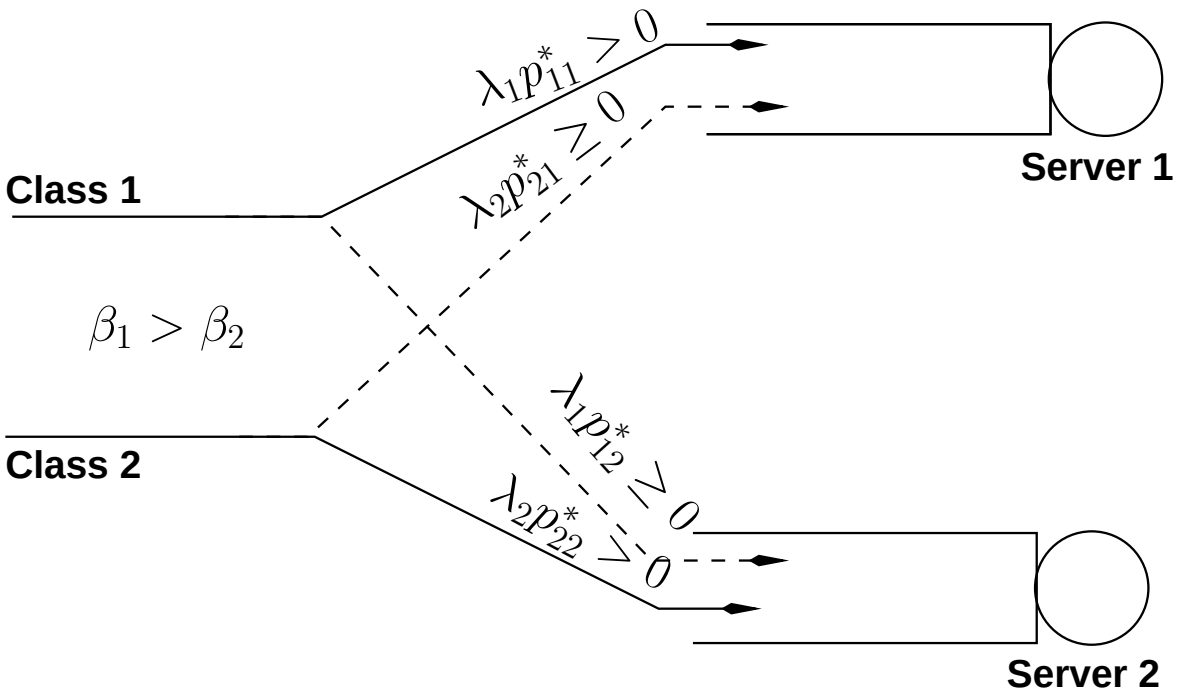
Theorem 1

- Let P^* achieve social optimality.
- Let γ^* be the arrival rates corresponding to P^*.
- Consider two customer classes $m_1 < m_2$, so that $\beta_{m_1} > \beta_{m_2}$.
  - Suppose n_1 and n_2 are distinct queues such that m_1 uses n_1 and m_2 uses n_2, i.e., $p^*_{m_1 n_1} > 0$ and $p^*_{m_2 n_2} > 0$.
  - Then $D_{n_1}(\gamma^*_{n_1}) < D_{n_2}(\gamma^*_{n_2})$.



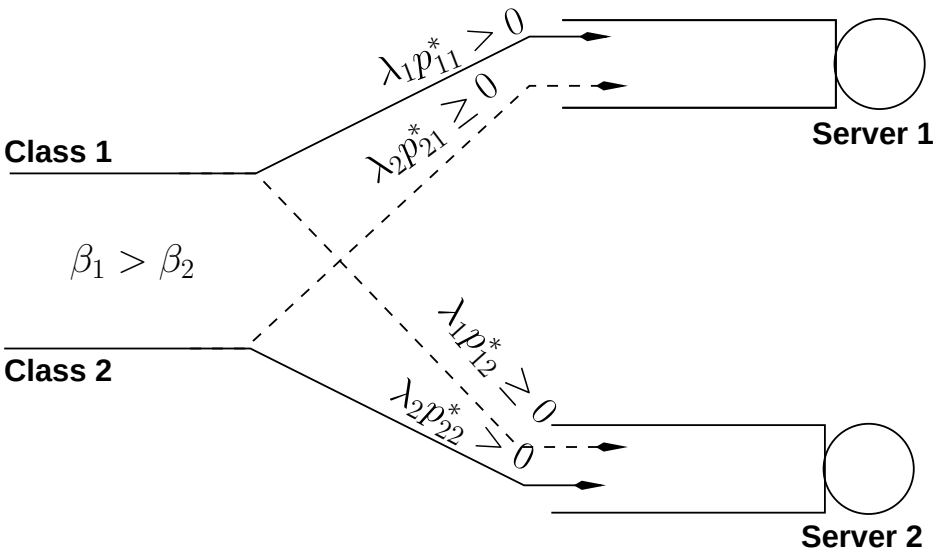
Proof by contradiction.

- Consider two queues 1 and 2 and two classes 1 and 2.

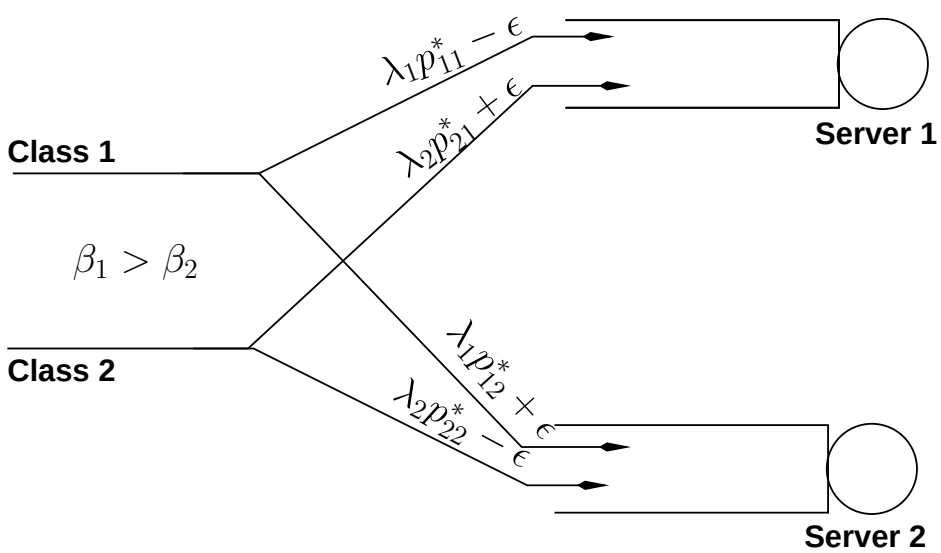


$$D_1(\gamma_1^*) < D_2(\gamma_2^*)$$

First suppose $D_1^* > D_2^*$.



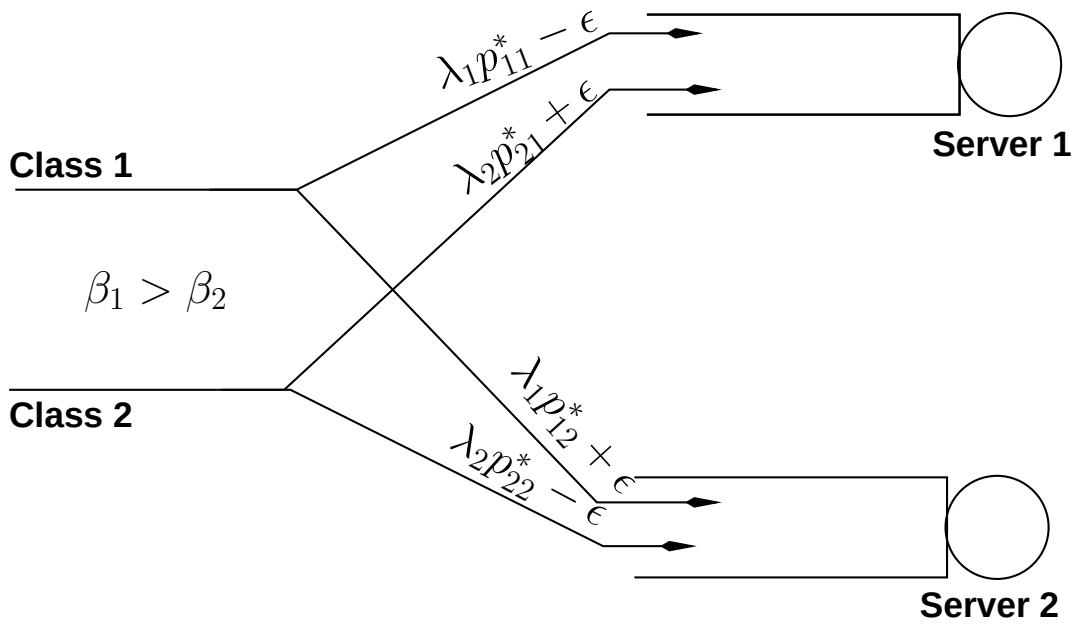
Allocation P^*



Allocation P

- For sufficiently small $\epsilon > 0$ find p_{11}, p_{12}, p_{21}, p_{22}, such that

$$\lambda_1 p_{11} = \lambda_1 p_{11}^* - \epsilon, \qquad \lambda_1 p_{12} = \lambda_1 p_{12}^* + \epsilon,$$
$$\lambda_2 p_{22} = \lambda_2 p_{22}^* - \epsilon, \qquad \lambda_2 p_{21} = \lambda_2 p_{21}^* + \epsilon.$$



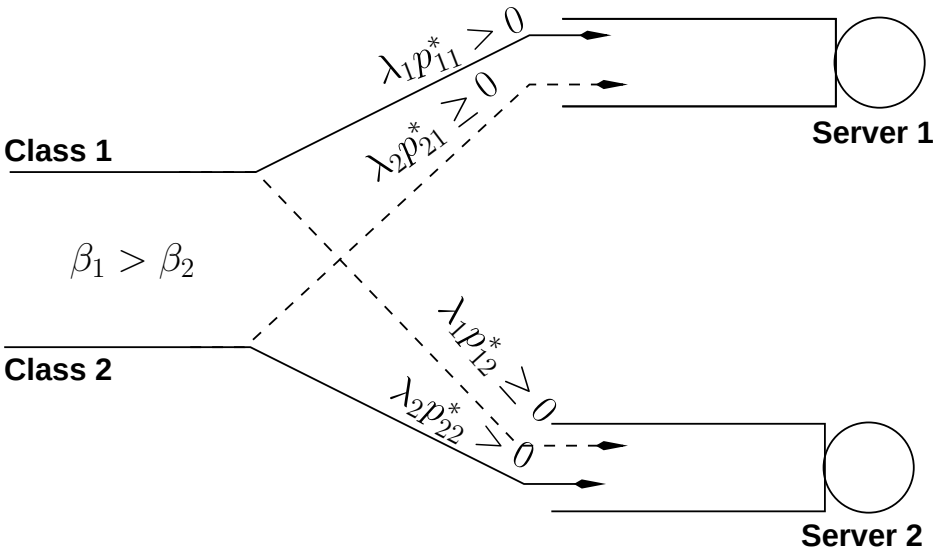
$$D_1(\gamma_1^*) > D_2(\gamma_2^*)$$

- γ remains unchanged. Hence D_n are the same as before. But

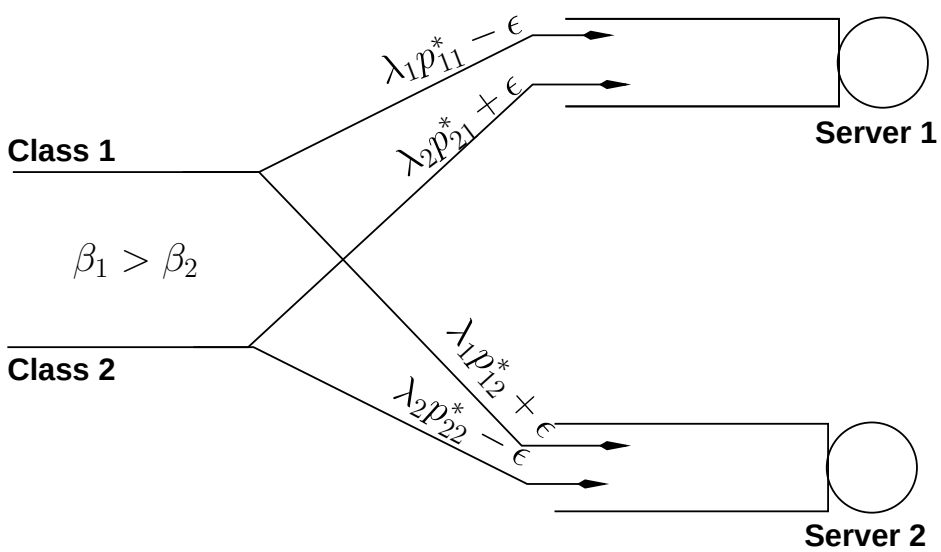
$$U(P) - U(P^*) = \epsilon(\beta_1 - \beta_2)(D_2^* - D_1^*).$$

- Since $\beta_1 > \beta_2$ and it is assumed that $D_1^* > D_2^*$, this implies that $U(P) < U(P^*)$ giving us the contradiction and proving $D_1^* \leq D_2^*$.

Suppose $D_1^* = D_2^*$.



Allocation P^*



Allocation P

- Define the matrices $P^\alpha = \alpha P + (1-\alpha)P^*$ for $\alpha \in [0,1]$.
- But total flow rates $\gamma^\alpha = \gamma^*$.
- Further, $U(P^\alpha) = U(P^*)$.

Showing $D_1^* \neq D_2^*$ (contd.)

- Apply the KKT conditions for optimality at P^α, where p_{11}, p_{12}, p_{21} and p_{22} are all strictly between 0 and 1. The KKT conditions imply that

$$\frac{\partial U(P^\alpha)}{\partial p_{11}} = \frac{\partial U(P^\alpha)}{\partial p_{12}}, \quad \frac{\partial U(P^\alpha)}{\partial p_{21}} = \frac{\partial U(P^\alpha)}{\partial p_{22}}.$$

- Using the definitions of U and γ_j, we can rewrite the first equality above as

$$\beta_1 \lambda_1 D_1^\alpha + \beta_1 \lambda_1^2 p_{11}^\alpha (D_1^\alpha)' = \beta_1 \lambda_1 D_2^\alpha + \beta_1 \lambda_1^2 p_{12}^\alpha (D_2^\alpha)',$$

- But $D_1^* = D_2^*$ by assumption, and γ^α coincides with γ^* for all α in $[0, 1]$ by construction.

Showing $D_1^* \neq D_2^*$ (contd.)

- Hence, we obtain that

$$p_{11}^{\alpha}(D_1^*)' = p_{12}^{\alpha}(D_2^*)',$$

  for all $\alpha \in (0, 1)$.
- This is impossible because the D' are non-zero by assumption, one of p_{11}^{α} and p_{12}^{α} is an increasing function of α while the other is a decreasing function. □

Structure P^*

Corollary

Suppose P^* solves the optimization problem (1), and let γ_j^* denote the resulting flow rates. Consider a re-ordering of the queues such that $D_1(\gamma_1^*) \leq D_2(\gamma_2^*) \leq \ldots \leq D_N(\gamma_N^*)$. Then, there exist numbers $n_1, \ldots, n_M$, with $1 \leq n_1 \leq n_2 \leq \ldots n_M \leq N$, such that

$$p_{ij} \begin{cases} > 0, & \text{if } j \in \{n_{i-1}+1, \ldots n_i - 1\}, \\ = 0, & \text{if } j \notin \{n_{i-1}, \ldots n_i\}. \end{cases}$$

Moreover, for each n_i, either p_{i,n_i} or p_{i+1,n_i} or both are strictly positive. [a]

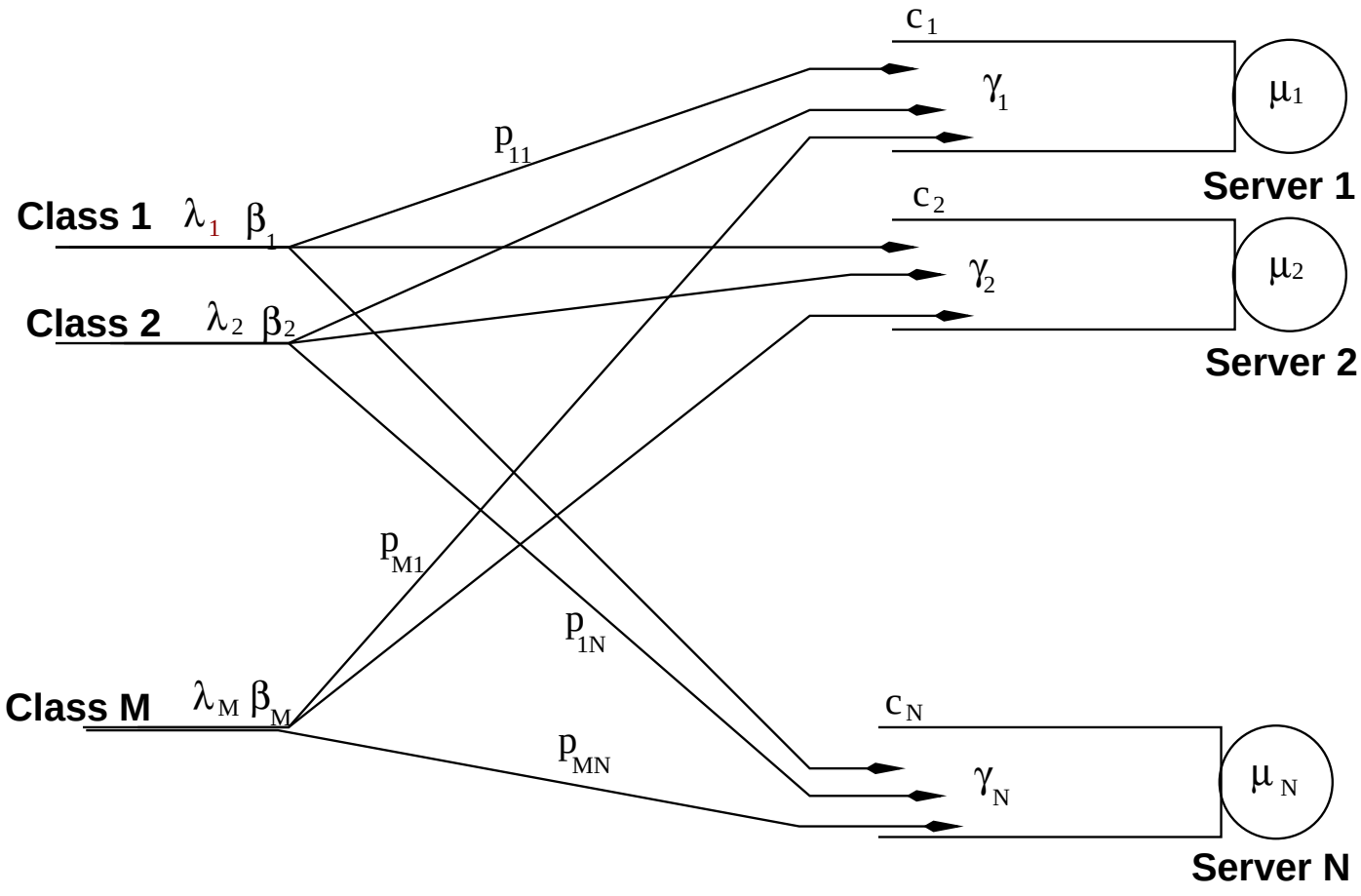
---

[a] We use the convention that the set $\{a, a+1, \ldots, b\}$ is assumed empty if $a > b$.

$$\mathbf{P}^* = \begin{array}{c} \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{array} \begin{array}{c} \begin{array}{ccccc} D_1 < & D_2 < & D_3 < & D_4 < & D_5 \end{array} \\ \begin{pmatrix} \times & \times & 0 & 0 & 0 \\ 0 & \times & 0 & 0 & 0 \\ 0 & \times & \times & 0 & 0 \\ 0 & 0 & \times & \times & 0 \\ 0 & 0 & 0 & 0 & \times \end{pmatrix} \end{array}$$

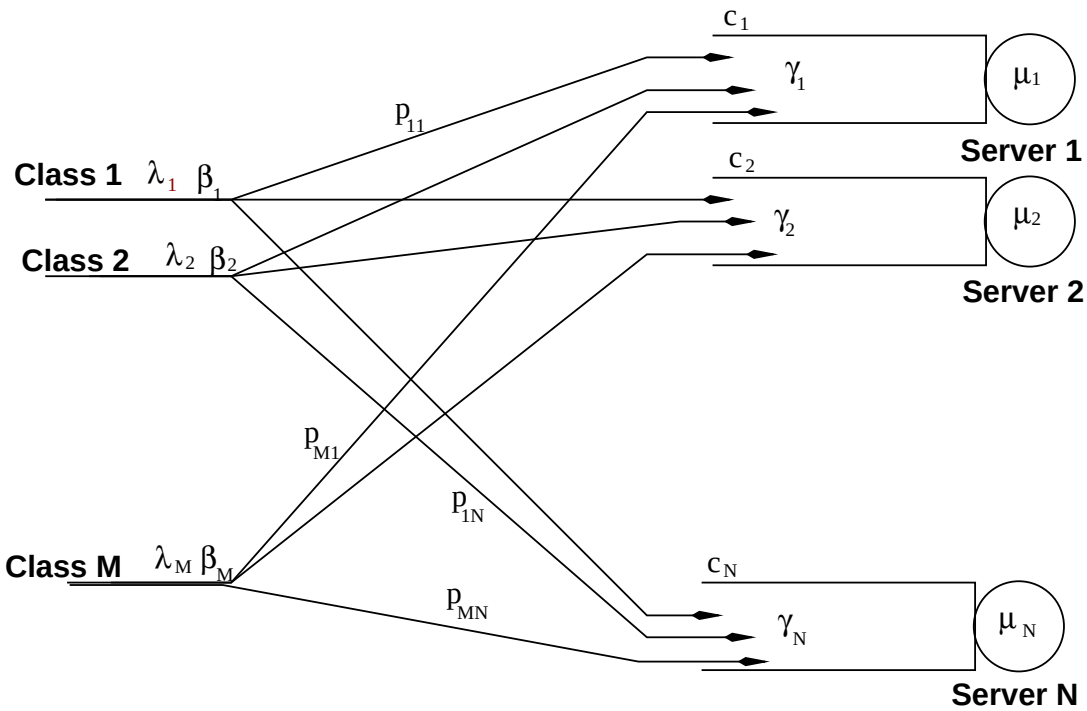
- Each class uses a nearly dedicated set of queues with a possible overlap only at the boundaries of the sets.
- More than two classes can use the same queue.
- It is possible that there are some queues that are not used.

Admission prices and selfish routing



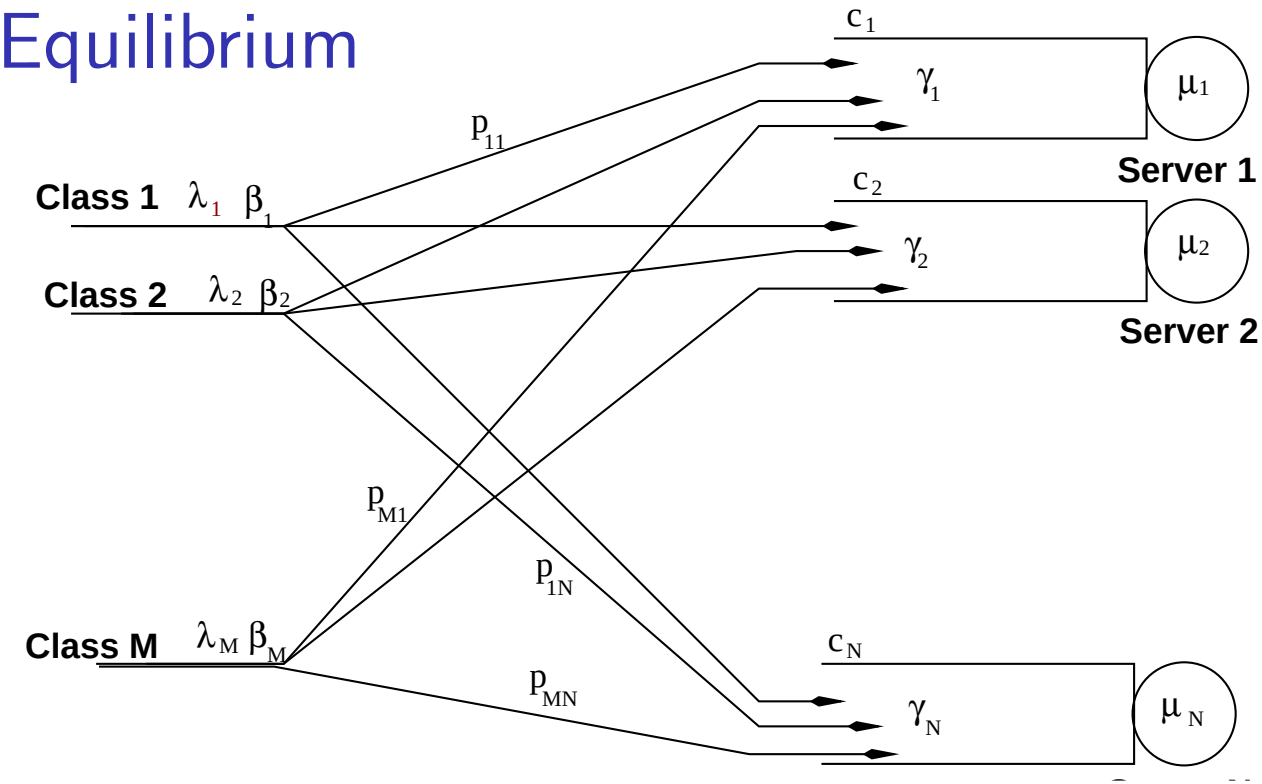
- Queue n charges an admission price c_n.
- No balking, i.e., all arriving customers must join a queue.

Admission prices and selfish routing



- Each class m customer seeks to minimise its expected cost.
- Cost for a class m customer at queue n is $c_n + \beta_m D_n(\gamma_n)$.
- Traffic rates γ_n are determined by the customer choices.

Wardrop Equilibrium

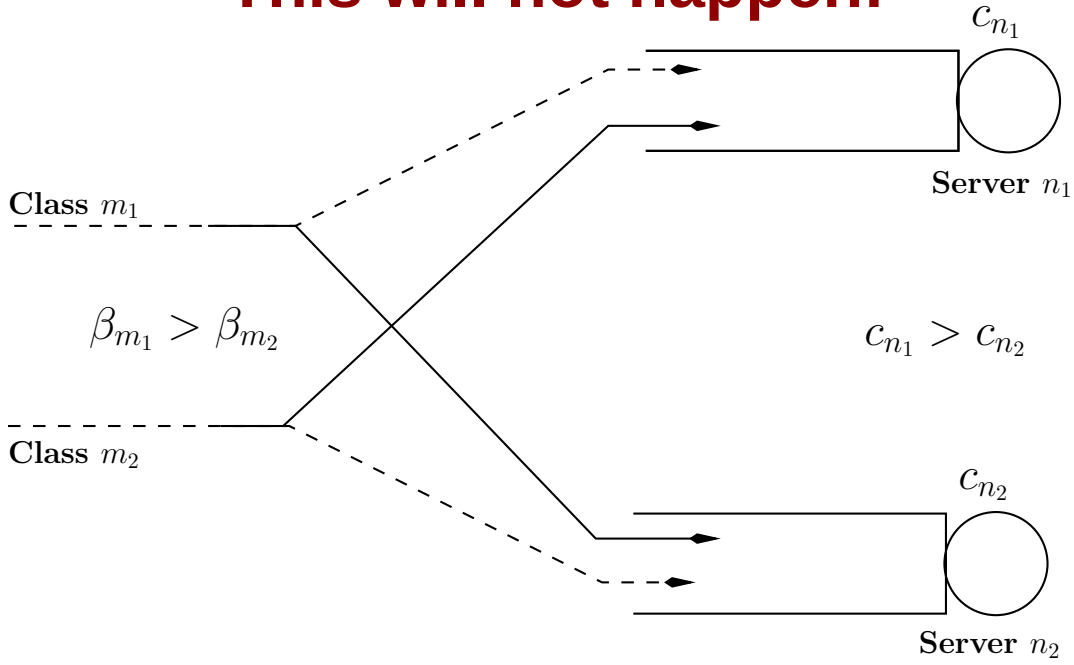


Definition

If class m used queue n, then at Wardrop equilibrium, for all k

$$c_n + \beta_m D_n(\gamma_n) \leq c_k + \beta_m D_k(\gamma_k).$$

This will not happen!



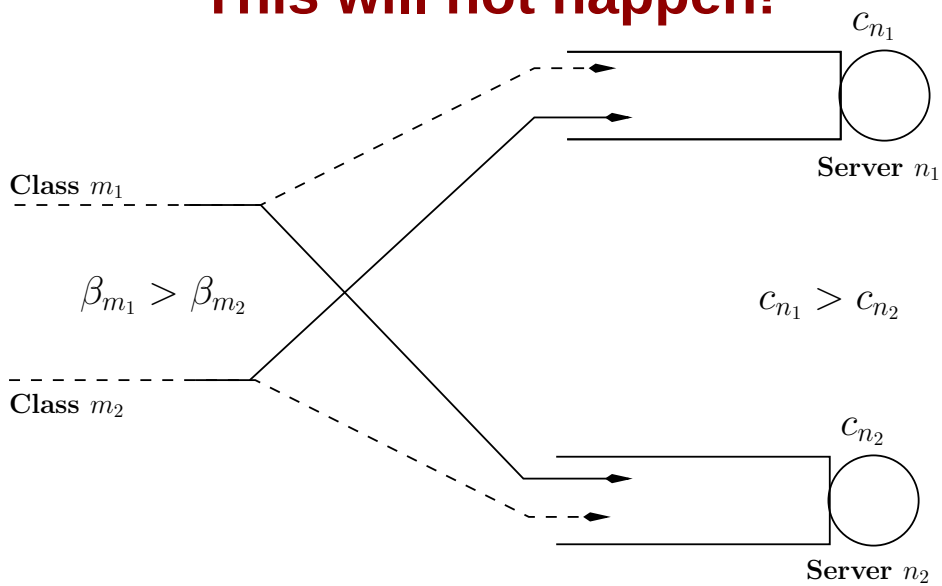
Theorem 2

- Consider two customer classes $m_1 < m_2$, with $\beta_{m_1} > \beta_{m_2}$,
- two queues $n_1 < n_2$, so that $c_{n_1} > c_{n_2}$.
- There is no Wardrop equilibrium P^W in which
  - class m_1 uses queue n_2 while
  - class m_2 simultaneously uses queue n_1,

  i.e., $p^W_{m_1,n_2} > 0$ and $p^W_{m_2,n_1} > 0$.

This will not happen!

Proof by contradiction



- Suppose $p_{12}^W > 0$ and $p_{21}^W > 0$. We then have

$$c_2 + \beta_1 D_2^W \leq c_1 + \beta_1 D_1^W, \quad c_1 + \beta_2 D_1^W \leq c_2 + \beta_2 D_2^W.$$

- Rearranging these, we get

$$\beta_1(D_2^W - D_1^W) \leq c_1 - c_2 \leq \beta_2(D_2^W - D_1^W).$$

- Since $c_1 > c_2$ the second inequality implies $(D_2^W - D_1^W) > 0$.
- Both equalities cannot be true at the same time with $(D_2^W - D_1^W) > 0$ because $\beta_1 < \beta_2$.

Structure of P^W

Corollary

Suppose P^W is a Wardrop equilibrium. There exist numbers $n_1, \ldots, n_M$, with $1 \leq n_1 \leq n_2 \leq \ldots n_M \leq N$, such that

$$p_{ij}^W = 0, \ \textit{if } j \notin \{n_{i-1}, \ldots n_i\}.$$

$$P^W = \begin{array}{c} \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{array} \begin{array}{c} \begin{array}{ccccc} c_1 > & c_2 > & c_3 > & c_4 > & c_5 \end{array} \\ \begin{pmatrix} \times & \times & 0 & 0 & 0 \\ 0 & \times & 0 & 0 & 0 \\ 0 & \times & \times & 0 & 0 \\ 0 & 0 & \times & \times & 0 \\ 0 & 0 & 0 & 0 & \times \end{pmatrix} \end{array}$$

Mechanism design

- Can prices be set to achieve optimal allocation under Wadrop equilibrium?
- Yes! A constructive algorithm is obvious.
- Interestingly, there is an alternate interpretation.
- Let P^* be the social welfare maximising routing matrix and γ^* the corresponding arrival rates into the queues.

Mechanism design

- A marginal increase of traffic at queue n increases the cost of each customer that uses the queue by $D_n'(\gamma_n^*)$.
- This imposes additional cost of $\beta_m D_n'(\gamma_n^*)$ on classs m customers entering queue n.
- And there are $\lambda_m p_{mn}$ of these customers per unit time.
- Thus total congestion externality caused by a marginal increase in offered load to queue n is

$$c_n = \sum_{m=1}^{M} \beta_m \lambda_m p_{mn} D_n'(\gamma_n^*)$$

Mechanism design (contd.)

- This is also called **Pigouvian tax,** a tax that is applied to a market activity (joining of queue n) that causes negative externality (added delay costs to other customers).

Theorem 3

Let P^* be a routing matrix solving the social welfare optimization problem. Let the admission prices $c_1, c_2, \ldots, c_N$ at queues $1, 2, \ldots, N$ be set according to

$$c_n = \sum_{m=1}^{M} \beta_m \lambda_m p_{mn} D_n'(\gamma_n^*)$$

Then P^* is a Wardrop equilibrium of the resulting game.

Open problems

- Computation
  - $U(P)$ is not convex and hence obtaining P^* is not easy.
  - If ordering in terms of $D(\cdot)$ is known, P^* is easy.
  - But optimal ordering is not obvious.

Open problems

- Computation
  - $U(P)$ is not convex and hence obtaining P^* is not easy.
  - If ordering in terms of $D(\cdot)$ is known, P^* is easy.
  - But optimal ordering is not obvious.
- Mechanism design
  - Need information about λ and β to obtain the optimal allocation and also to set prices.
  - λ can be measured and β can possible be elicited by varying admission charges.

Open problems

- Computation
  - $U(P)$ is not convex and hence obtaining P^* is not easy.
  - If ordering in terms of $D(\cdot)$ is known, P^* is easy.
  - But optimal ordering is not obvious.
- Mechanism design
  - Need information about λ and β to obtain the optimal allocation and also to set prices.
  - λ can be measured and β can possible be elicited by varying admission charges.
- Market structure
  - What happens if service provider sets prices to maximise revenue?
  - Does the price equilibrium always exist ?

Thank you for your attention.