# Pricing in Service Systems with Strategic Customers

Refik Güllü

Boğaziçi University
Industrial Engineering Department
Istanbul, Turkey

# Before the talk

- My professor's advice on queueing theory versus game theory
  - mathematical difficulty versus conceptual maturity
- How to learn new stuff?
  - teaching, writing a code, thesis supervision

# Before the talk

- A review from a personal perspective
- A great place to start reading:
  Rafael Hassin and Moshe Haviv, To queue or not the queue: equilibrium behavior in queueing systems, Springer, 2003.
  http://www.math.tau.ac.il/ hassin/book.html
- A follow up survey book, "Rational Queueing" to appear soon

# Outline

- A simple example of an "unobservable" queue
- Parameter uncertainty
- Effect of delay information
- Observable queues: residual service time
- Multiple customer types: identical price
- Multiple customer types: differentiation
- An inventory model

# A Framework Example

- A student cafeteria in a university
- University administration regulates the price
  - possibly in the form of a subsidy
- Students arrive according to a Poisson process
  - $\Lambda$ is the rate of potential students
- A single server with exponential rate $\mu > \Lambda$.

# A Framework Example

- Cafeteria is sort of far away from the main building
- Students decide eating there or not before observing the congestion
- There are other dining facilities on campus
  - Once a decision is made, it can not be changed
- Students are identical
  - with respect to their valuation of the service, value of time, and their behaviour towards risk
  - all are rational decision makers

# A Framework Example

- $R$ : the value of service (as judged by students)
- $c$ : the unit cost for waiting
- $p$ : fee for dining at the cafeteria
- The expected utility of a student from the service

$$R - p - cE[\text{sojourn time}]$$

- the system is at steady state

# A Framework Example: equilibrium behaviour

- As students are "identical", their equilibrium behaviour is expected to be the same
- Each student choose to enter the cafeteria with probability $q$
- Let $U(q_{tagged}, q_{others})$ be the utility of a tagged student when all the others behave with $q_{others}$
- Best response against $q_{others}$: $U(q', q_{others}) \geq U(q, q_{others})$ for all $q$.
- Symmetric Nash equilibrium: best response against itself

$$U(q_e, q_e) \geq U(q, q_e) \text{ for every } q$$

# A Framework Example: equilibrium behaviour

- $0 \leq q_e(p) \leq 1$ is the equilibrium probability of joining the cafeteria when the fee is $p$
- $\lambda_e(p) = \Lambda q_e(p) < \mu$
- For effective arrival rate $\lambda < \mu$

$$w(\lambda) = 1/(\mu - \lambda)$$

- 3 cases need to be examined

# Case 1

- Nobody else is joining and

$$p + cw(0) > R \implies R < p + c\frac{1}{\mu}$$

$$\mu < \frac{c}{R - p}$$

  - $q_e(p) = 0$
  - $\lambda_e(p) = 0$
- $w(\lambda_e(p)) = 1/\mu$

# Case 2

- If everybody else is joining and

$$p + cw(\Lambda) \leq R \implies R \geq p + c\frac{1}{\mu - \Lambda}$$

$$\mu \geq \Lambda + \frac{c}{R - p}$$

  - $q_e(p) = 1$
  - $\lambda_e(p) = \Lambda$
- $w(\lambda_e(p)) = 1/(\mu - \Lambda)$

# Case 3

- $p + cw(0) \leq R < p + cw(\Lambda)$

$$R = p + cw(\lambda_e(p))$$

  - $q_e(p) = \lambda_e(p)/\Lambda$
  - $w(\lambda_e(p)) = 1/(\mu - \lambda_e(p))$

$$\lambda_e(p) = \mu - \frac{c}{R - p}$$

# Administration's Problem: social optimization

- The administration cares about the overall performance
- Solves the following problem

$$\max_{0 \leq \lambda \leq \Lambda} \{\lambda(R - c\frac{1}{\mu - \lambda})\}$$

- strictly concave
- maximum occurs at

$$\lambda^* = \mu - \sqrt{\frac{c\mu}{R}}$$

- $\lambda^* \geq 0$ (by $R \geq c/\mu$)

# Administration's Problem: social optimization

- By considering the constraint $\lambda \leq \Lambda$

$$\lambda^* = \min\{\Lambda, \mu - \sqrt{\frac{c\mu}{R}}\}$$

- if $\Lambda \geq \mu - \sqrt{\frac{c\mu}{R}}$
  - optimal objective function value

$$(\sqrt{R\mu} - \sqrt{c})^2$$

- $w(\lambda^*) = \sqrt{\frac{R}{c\mu}}$

# Administration's Problem: social optimization

- if $\Lambda \leq \mu - \sqrt{\frac{c\mu}{R}}$
  - optimal objective function value

$$\Lambda(R - \frac{c}{\mu - \Lambda})$$

- $w(\lambda^*) = \frac{1}{\mu - \Lambda}$

- By assuming $R \geq c/\mu$

$$\lambda_e(0) = \mu - \frac{c}{R} \geq \mu - \sqrt{\frac{c\mu}{R}} = \lambda^*$$

- Individual optimization leads to longer queues than imposed by social optimization
- Admission fee can regulate this

# Revenue maximization

- Let $p_m$ be the admission fee charged for dining
- $p_m = R - cw(\lambda)$

$$\max_{0 \le \lambda \le \Lambda} p_m \lambda$$

- Same as the social optimization objective
- The socially optimal arrival rate can be induced by the fee

$$p_m = p^* = R - cw(\lambda^*) = R - \sqrt{\frac{cR}{\mu}}$$

- $\lambda_e(p^*) = \lambda^*$, profit $= (\sqrt{R\mu} - \sqrt{c})^2$

Rafael Hassin and Moshe Haviv, To queue or not the queue: equilibrium behavior in queueing systems, Springer, 2003. (Chapter 3)

Bell, C. E., Stidham, Jr., 1983, Individual versus Social Optimization in the Allocation of Customers to Alternate Servers, *Management Science*, 29, 831-839.

# Parameter Uncertainty

- The preceding model considered an "unobservable" system
- The queue length or the waiting times upon arrival are unobservable
- Need to be careful
  - Many things are known and/or intelligently computable: service rate, expected waiting time, service value, etc.
- These parameters are known with certainty

# Uncertainty in the Service Rate

- Suppose $\mu$, the service rate, can take two values

$$\mu = \begin{cases} \mu_1 & \text{with probability } \alpha \\ \mu_2 & \text{with probability } 1-\alpha \end{cases}$$

- $\mu_1 > \mu_2$
- Do students know the realised service rate?
  - No, they are "uninformed"
  - Yes, they are "informed", and the server charges either a different fee or the same fee

# Service Rate Uncertainty

- For the "uninformed" case
- $v = (R - p)/c$
- $v = \frac{\alpha}{\mu_1 - \lambda} + \frac{1-\alpha}{\mu_2 - \lambda}$ in equilibrium
- $v$ is a solution of a nonlinear equation

$$\Pi^{un} = \lambda(R - cv)$$

# Service Rate Uncertainty

- "informed" case with two prices
- $\lambda_i = \mu_i - \frac{c}{R-p_i}$, $p_i = R - \sqrt{\frac{cR}{\mu_i}}$

$$\Pi_2^{in} = \begin{cases} \alpha(\sqrt{R\mu_1} - \sqrt{c})^2 + (1-\alpha)(\sqrt{R\mu_2} - \sqrt{c})^2 & \text{if } R \geq \frac{c}{\mu_2} \\ \alpha(\sqrt{R\mu_1} - \sqrt{c})^2 & \text{if } \frac{c}{\mu_1} < R < \frac{c}{\mu_2} \\ 0 & \text{if } R \leq \frac{c}{\mu_1} \end{cases}$$

# Service Rate Uncertainty

- "informed" case with a single price
- $\lambda_i = \mu_i - \frac{c}{R-p}$ (rate of arrivals given the service rate)
- If $p$ is small enough to attract customers for both values of $\mu$
- average arrival rate for the single price $p$

$$\begin{aligned} \bar{\lambda} &= \alpha(\mu_1 - \frac{c}{R-p}) + (1-\alpha)(\mu_2 - \frac{c}{R-p}) \\ &= \bar{\mu} - \frac{c}{R-p} \end{aligned}$$

# Service Rate Uncertainty

- Maximizing $p\bar{\lambda} = p\bar{\mu} - \frac{pc}{R-p}$
- $(R-p)^2 = Rc/\bar{\mu}$

$$\Longrightarrow \; p = R - \sqrt{\frac{Rc}{\bar{\mu}}}$$

- Resulting profit

$$(\sqrt{R\bar{\mu}} - \sqrt{c})^2$$

# Service Rate Uncertainty

- But a higher price may be chosen: so that customers opt out when $\mu = \mu_2$.
- $p = R - \sqrt{\frac{Rc}{\mu_1}}$
- Resulting profit: $\alpha(\sqrt{R\mu_1} - \sqrt{c})^2$
- Two profit terms are equal for

$$\frac{R}{c} = \eta = \left( \frac{1 - \sqrt{\alpha}}{\sqrt{\bar{\mu}} - \sqrt{\mu_1 \alpha}} \right)^2$$

# Service Rate Uncertainty

$$\Pi_1^{in} = \begin{cases} (\sqrt{R\bar{\mu}} - \sqrt{c})^2 & \text{if } \frac{R}{c} \geq \eta \\ \alpha(\sqrt{R\mu_1} - \sqrt{c})^2 & \text{if } \frac{1}{\mu_1} \leq \frac{R}{c} \leq \eta \\ 0 & \text{if } \frac{R}{c} \leq \frac{1}{\mu_1} \end{cases}$$

# Parameters Uncertainty

- The service provider benefits from revealing the service rate, and from pricing accordingly

$$\Pi_2^{in} \geq \Pi_1^{in} \geq \Pi^{un}$$

- As the variability in service rate increases, $\Pi_2^{in}$ increases
  - The server provider may lose $(1-\alpha)$ fraction of the customers but extracts higher revenue from the remaining

# Parameters Uncertainty

Hassin, R., 2007, Information and Uncertainty in a Queuing System, *Probability in the Engineering and Informational Sciences*, 21, 361-380.

- waiting cost uncertainty
- service valuation uncertainty

# Delay Information

- So far: "unobservable" systems
- What if the customers are revealed information on the possible delay before they decide to join or not
  - "observable" systems
- Three levels of information

1. No information (same as before)
2. Partial information: how many customers are in front of me?
3. Full information: what is my exact waiting time?

# Delay Information

- M/M/1 type service system
- $W$: waiting time (in the queue)
- $\theta$: customer type parameter, a random variable, $\theta \in [0,1]$ with cdf $H$, pdf $h$
- $c(t)$: cost of waiting $t$ time units

$$U = R - \theta E[c(W)]$$

- Previously: $\theta \equiv 1$, $c(t) = ct$

# Delay Information

- Scale $R$ and $c(t)$ so that $R = 1$
- Assume that $c(0) > 1$
  - Customers with $\theta > 1/c(0)$ balk
- Scale $\Lambda$ (ignore them) to $\lambda$, and assume (new) $c(0) = 1$
- Customers with $\theta \approx 1$ are also attracted to join when $W = 0$.

$$U(\text{no waiting}) = 1 - \theta$$

# Delay Information: who stays in the system?

- Information $I$, a random variable
- Want: $U|I = 1 - \theta E_W[c(W)|I] \geq 0$
- Given information $I = i$, an arriving customer stays if

$$\theta \leq \theta_i = \frac{1}{E_W[c(W)|I = i]}$$

- $\Pr\{\text{stays}|I = i\} = H(\theta_i)$
- Fraction of customers who stay: $E_I[H(\theta_I)]$
- Throughput

$$\lambda E_I[H(\theta_I)]$$

# Average utility

- Define

$$J(\theta) = \frac{1}{\theta} \int_0^{\theta} H(x)dx$$

- Average utility

$$\begin{aligned} u &= E[U^+] = E_{\theta,I}[(1 - \theta E_W[c(W)|I])^+] \\ &= E_I \left[ \int_0^{\theta_I} (1 - xE_W[c(W)|I])h(x)dx \right] \\ &= E_I \left[ H(\theta_I) - (1/\theta_I) \int_0^{\theta_I} xh(x)dx \right] \\ &= E_I[J(\theta_I)] \end{aligned}$$

# Case 1: No information

- The equilibrium arrival rate:

$$\lambda^{NI} = \lambda H\left(\frac{1}{E[c(W^{NI})]}\right)$$

- $\rho^{NI} = \lambda^{NI}/\mu$
- $\widetilde{c}(s) = \int_0^\infty e^{-st}c(t)dt$ LST of $c(t)$.
- $\Pr\{W^{NI} > t\} = \rho^{NI}e^{-\mu(1-\rho^{NI})t}$

$$E[c(W^{NI})] = (1-\rho^{NI}) + \rho^{NI}\mu(1-\rho^{NI})\widetilde{c}(\mu(1-\rho^{NI}))$$

# Case 1: No information

- An example: Uniform customers with Linear Cost

$$c(t) = 1 + t$$

- $\widetilde{c}(s) = \frac{1}{s} + \frac{1}{s^2}$

$$\lambda^{NI} = \frac{\lambda}{1 + \rho^{NI}/(\mu(1-\rho^{NI}))}$$

$$\Longrightarrow \ (1-\mu)(\rho^{NI})^2 + (\mu+\lambda)\rho^{NI} - \lambda = 0$$

$$\Longrightarrow \ \rho^{NI} = \frac{-(\mu+\lambda) + \sqrt{(\lambda+\mu)^2 + 4\lambda(1-\mu)}}{2(1-\mu)}$$

- $\pi_n^{NI} = (1-\rho^{NI})(\rho^{NI})^n$

# Case 2: Partial information

- The service provider tells the arriving customer $N(t)$
- The customer computes $c_n = E[c(W)|N(t) = n]$
- Stays if $\theta \leq \theta_n = 1/c_n$
- Birth-death process with state dependent arrival rate $\lambda_n = \lambda H(\theta_n)$
- Steady-state probabilities

$$\begin{aligned}
\Theta_n &= \prod_{m=0}^{n-1} H(\theta_m), \quad \Theta = \sum_{n=1}^{\infty} \Theta_n (\lambda/\mu)^n \\
\pi_0^{PI} &= 1/(1+\Theta) \\
\pi_n^{PI} &= \Theta_n (\lambda/\mu)^n \pi_0^{PI}
\end{aligned}$$

# Case 2: Partial information

- An example: Uniform customers with Linear Cost

$$c(t) = 1 + t$$

- $\theta_m = \frac{1}{1+\frac{m}{\mu}}$, $\Theta_n = \prod_{m=0}^{n-1} \frac{1}{1+\frac{m}{\mu}}$
- $\Theta_n(\lambda/\mu)^n = \frac{\Gamma(\mu)\Gamma(\mu+n)}{\lambda^n}$

$$\pi_0^{PI} = \frac{1}{1+\gamma(\mu,\lambda)\lambda^{1-\mu}e^{\lambda}}, \quad \pi_n^{PI} = \pi_0^{PI}\frac{\Gamma(\mu)}{\Gamma(\mu+n)}\lambda^n$$

$$\Gamma(x) = \int_0^{\infty} t^{\mu-1}e^{-t}dt \text{ and } \gamma(\mu,\lambda) = \int_0^{\lambda} t^{\mu-1}e^{-t}dt$$

# Case 3: Full information

- The service provider tells the arriving customer workload $V(t) = v$
- Critical point: $\theta_v = 1/c(v)$
- Effective arrival rate $\lambda(v) = \lambda H(\theta_v)$
- $f(v)$: the pdf of the stationary workload $V$

under linear cost and uniform customers

$$\pi_0^{FI} = \frac{1}{1 + \lambda e^{\mu} \mu^{-(\lambda+1)} \gamma(\lambda + 1, \mu)}$$

$$f^{FI}(v) = \lambda \pi_0^{FI} (1 + v)^{\lambda} e^{-\mu v}, \quad v > 0$$

# Benefit of Information

- The service provider would like to have large throughput $\lambda E_I[H(\theta_I)]$
- Customers would like to have a large average utility $E_I[J(\theta_I)]$

$$E_I\left[\frac{1}{\theta_I}\int_0^{\theta_I} H(x)dx\right]$$

- What is the impact of information on these measures?
- Clearly the answer depends on $H(x)$.

# Benefit of Information

- There are cases where the service provider and the customers are aligned

$$H(x) = x^{\alpha}, \quad \alpha > 0$$

- $J(\theta) = \frac{1}{\theta}\int_0^{\theta} x^{\alpha} dx = \frac{1}{\alpha+1}\theta^{\alpha} = \frac{1}{\alpha+1}H(\theta)$
- Average utility $\sim$ Throughput
- More information is better for all parties

# Benefit of Information

- Consider constant $\theta$
- No information: throughput is $\lambda$ for sufficiently small $\lambda$
- Partial information: there is a threshold $n^*$ beyond which customers do not join
  - throughput $< \lambda$
- The service provider may hide information
- Essentially whether information beneficial to one party or the other depends on the shape of $H(\theta)$

# Benefit of Information

- If $H(1/x)$ is convex in $x \geq 1$, then more information benefits the service provider by increasing throughput
- If $J(H^{-1}(y))$ is convex on $[0,1]$, then more information benefits the customers by increasing the average utility.

$$H(x) = \frac{\gamma e^{-\gamma x}}{1 - e^{-\gamma}} \quad x \in [0,1]$$

$$\Longrightarrow \quad \gamma \leq 2 \ (h(x) \text{ does not decrease too rapidly})$$

# Benefit of Information

- If $H(1/x)$ is convex, then $\pi_0^{PI} \leq \pi_0^{NI}$

$$\begin{aligned} \pi_0^{NI} &= 1 - \frac{\lambda}{\mu} H\left(\frac{1}{E[c_{N^{NI}}]}\right) \\ \pi_0^{PI} &= 1 - \frac{\lambda}{\mu} E\left[H\left(\frac{1}{c_{N^{PI}}}\right)\right] \end{aligned}$$

- $\pi_0^{PI} > \pi_0^{NI} \implies N^{PI} \leq_{st} N^{NI} \implies E[c_{N^{PI}}] \leq E[c_{N^{NI}}]$

$$\implies H\left(\frac{1}{E[c_{N^{NI}}]}\right) \leq E\left[H\left(\frac{1}{c_{N^{PI}}}\right)\right]$$

by convexity of $H(1/x)$ and Jensen's inequality

- a contradiction

Guo, P., Zipkin, P., 2007, Analysis and Comparison of Queues with Different Levels of Delay Information, *Management Science*, 53, 962-970.

Guo, P., Zipkin, P., 2009, The Effects of the Availability of Waiting-time Information on a Balking Queue, *EJOR*, 198, 199-209.

# Observable Queues: Residual Service Time

- Consider an observable M/G/1 queue
- The arriving customer observes the queue length before joining
- If the service time is exponential
  - Customer joins if the number in the system is less than

$$\left\lfloor \frac{R}{cE[\text{Service Time}]} \right\rfloor$$

- With non-exponential service times
  - Residual service time matters

# Observable Queues: Residual Service Time

- A customer who observes $n$ customers upon arrival joins with probability $q_n$

$$(q_1, q_2, \ldots)$$

- The behaviour of others has an effect on the assessment of residual service time

$E[RST_n] = E[\text{residual ST when the arriving customer finds } n \text{ in the system}]$

- $E[RST_n] = f_n(q_1, q_2, \ldots, q_n)$ (a recursive expression)

$$E[RST_1] = \frac{E[ST]}{1 - \tilde{G}(\Lambda q_1)} - \frac{1}{\Lambda q_1}$$

  - Suppose deterministic service time and $q_1$ is high
  - information about the current service state

# Observable Queues: Residual Service Time

- Once $q_1, q_2, \ldots, q_{n-1}$ are known

$$\begin{aligned}
&\text{if } nE[ST] + f_n(q_1, \ldots, q_{n-1}, 1) \leq R/c && \Longrightarrow \quad q_n = 1 \\
&\text{if } nE[ST] + f_n(q_1, \ldots, q_{n-1}, 0) \geq R/c && \Longrightarrow \quad q_n = 0 \\
&\text{otherwise } nE[ST] + f_n(q_1, \ldots, q_{n-1}, q) = R/c && \Longrightarrow \quad q_n = q
\end{aligned}$$

- Intuition: $q_1 \geq q_2 \geq q_3 \geq \cdots$
- Which turns out to be wrong!

# Observable Queues: Residual Service Time

- $ST = 1$ with probability $\epsilon$ (small), $ST = 0$ with probability $1 - \epsilon$
- $E[RST_1] = \frac{1}{1-e^{-\Lambda q_1}} - \frac{1}{\Lambda q_1}$
- Solve $\epsilon + E[RST_1] = R/c$ to find $\Lambda q_1$.

$$\begin{aligned} \Lambda \leq \Lambda_1 &\implies q_1 = q_2 = 1 \\ \Lambda_1 < \Lambda \leq \Lambda_2 &\implies 0 < q_1 < 1, q_2 = 1 \\ \Lambda > \Lambda_2 &\implies 0 < q_1 < q_2 < 1 \end{aligned}$$

- The fact that there are two customers (one in service) means that the we are probably nearing the end of the current service time, and the service time of the next customer is very likely to be zero anyway

# Observable Queues: Residual Service Time

- If the service time distribution is of type "decreasing mean residual life" (DMRL)

  that is, $E[ST - t|ST > t]$ is monotone decreasing in $t$

- There is $n_e$ as the smallest integer satisfying

$$nE[S] + f_n(1, 1, 1, \ldots, 1) \geq R/c$$

- $q_e \in [0, 1)$ satisfying

$$n_e E[S] + f_n(1, 1, 1, \ldots, q_e) = R/c$$

$$q_n = \begin{cases} 1 & n < n_e \\ q_e & n = n_e \\ 0 & n > n_e \end{cases}$$

# Observable Queues: Residual Service Time

- Intuition: waiting time difference between states $n$ and $n+1$

$$(n+1)E[ST]+E[RST_{n+1}]-nE[ST]-E[RST_n] > E[ST]-E[RST_n] \geq 0$$

  - Waiting times are increasing in $n$
  - There exist at most one $n$ with mixed strategy
  - $[RST_n]$ is increasing in $q$ for $(q_1, \ldots, q_{n-1}, q)$
  - Hence $q_e$ is unique.
- "Avoiding the crowd" versus "following the crowd"

Naor, P., 1969, The Regulation of Queue Size by Levying Tolls, *Econometrica*, 37, 15-24.

Kerner, Y., 2011, Equilibrium Joining Probabilities for an M/G/1 queue, *Games and Economic Behavior*, 71, 521-526.

Haviv, M., Kerner, Y., 2007, On Balking from an Empty Queue, *Queueing Systems*, 55, 239-249.

Kerner, Y., 2008, The Conditional Distribution of the Residual Service Time in the Mn/G/1 Queue, *Stochastic Models*, 24, 364-375.

Manou, A., Economou, A., Karaesmen, F., 2014, Strategic Customers in a Transportation Station: When Is It Optimal to Wait?, *Operations Research*, 62, 910-925.

# Multiple Customer Classes-identical price

- Things get complicated with multiple customer classes
- Consider two classes of customers with M/G/1 type service facility
- $R_1$, $R_2$, $c_1$, $c_2$, $\Lambda_1$, $\Lambda_2$
- Suppose that the customers are either
  - indistinguishable to the service provider
  - or price discrimination is not possible
- Service times are identically distributed
- Customers are treated in a FCFS manner
- They can not observe the queue length

# Multiple Customer Classes-identical price

- $\lambda = \lambda_1 + \lambda_2$, the equilibrium arrival rate
- $z_i = R_i - (c_i/\mu)$, $i = 1, 2$

$$u_i(\lambda, p) = R_i - p - c_i(w_Q(\lambda) + (1/\mu)) = z_i - p - c_i w_Q(\lambda)$$

- $z_1 \geq z_2$, and $p \leq z_i$ (otherwise does not join)
- $u_i(\lambda, p)$ is strictly decreasing and concave in $\lambda$

since $w_Q(\lambda)$ is convex increasing

# Multiple Customer Classes-identical price

- For any price $p$ and $0 \leq \lambda < \mu$,
  - $c_1 \leq c_2 \implies u_1(\lambda, p) \geq u_2(\lambda, p)$
  - If $c_1 > c_2$, there is a critical value $\tilde{\lambda}$ so that

$$\lambda > \tilde{\lambda} \implies u_1(\lambda, p) < u_2(\lambda, p)$$

- As system gets more congested, the one with higher sensitivity to delay hurts more

# Multiple Customer Classes-identical price

- By solving $u_i(\lambda, p) = 0$ for $\lambda$ and $p$

$$\lambda_i(p) = \frac{2\mu^2(z_i - p)}{2\mu(z_i - p) + c_i(1 + cv^2)}$$

$$p_i(\lambda) = z_i - \frac{\lambda c_i(1 + cv^2)}{2\mu(\mu - \lambda)}$$

- maximum arrival rate for $i$ for a given price $p$
- maximum price $i$ is willing to pay for total arrival rate $\lambda$

# Multiple Customer Classes-identical price

- How do customers behave for a given price $p$?
  - Depends on $c_1$ and $c_2$
- If $c_1 \leq c_2$, then $\Lambda_1$ and $\lambda_2(p)$ are compared
- $\lambda_1(p) \geq \lambda_2(p)$

$$\Lambda_1 \geq \lambda_2(p) \implies (q_e^1, q_e^2) = (\min\{1, \lambda_1(p)/\Lambda_1\}, 0)$$

$$\Lambda_1 < \lambda_2(p) \implies (q_e^1, q_e^2) = (1, \min\{1, (\lambda_2(p) - \Lambda_1)/\Lambda_2\})$$

# Multiple Customer Classes-identical price

- Suppose $\Lambda_1 \geq \lambda_1(p)$ ( $\Longrightarrow \geq \lambda_2(p)$)
- First: Class-1 enters with rate $\lambda_1(p)$

$$u_2(\lambda_1(p), p) \leq u_1(\lambda_1(p), p) = 0$$

$$\Longrightarrow \; q_e^2 = 0$$

- Next: Class-2 does not enter the system

$$u_1(\lambda, p) > 0 \iff \lambda < \lambda_1(p)$$

$u_1(\lambda, p)$ is decreasing in $\lambda$

$$\Longrightarrow \; q_e^1 = \lambda_1(p)/\Lambda_1$$

# Multiple Customer Classes-identical price

- How the server provider sets the price?
- $\max_p \lambda(p)p$
- If $c_1 \leq c_2$, then depends on the market size of Type-1

$$\Lambda_1 \text{ "large"} \implies p_e = \max\{p_1^*, p_1(\Lambda_1)\}$$

$$\text{and } (q_e^1, q_e^2) = (\min\{1, \lambda_1(p_1^*)/\Lambda_1\}, 0)$$

$$p_1^* = \arg\max p\lambda_1(p)$$

$$p_1^* = z_1 - \frac{\sqrt{c_1(1+cv^2)(c_1(1+cv^2)+2\mu z_1)} - c_1(1+cv^2)}{2\mu}$$

Zhou, W., Chao, X., Gong, X., 2014, Optimal Uniform Pricing Strategy of a Service Firm when Facing Two Classes of Customers, *POM*, 23, 676-688.

# Multiple Customer Types-differentiation

- $M$ customer types
- $R_i$, $c_i$, $\Lambda_i$
- Some questions:
  - what will be the "control policy"
  - what will be the price/delay menu?
  - what is the information structure: who knows what?
- Incentive compatibility

$$p_i + c_i w_i \leq p_j + c_i w_j \quad j \neq i$$

- individual rationality

$$R_i \geq p_i + c_i w_i$$

# Multiple Customer Types-Revenue Maximization Model

$$
\begin{aligned}
\max_{p,W,u} \quad & \textstyle\sum_{i=1}^{M} p_i \lambda_i & & \\
\lambda_i \quad & = \quad & & \Lambda_i \Pr\{R_i \geq p_i + c_i w_i\} \ \ \forall\, i \\
p_i + c_i w_i \quad & \leq \quad & & p_j + c_i w_j \ \ \ j \neq i \\
\sum_{i=1}^{M} \lambda_i \quad & < \quad & & \mu
\end{aligned}
$$

# Multiple Customer Types-General Results

- If the service provider observes the types of the customers
  - Set priorities with a work conserving discipline ($c\mu$ rule)
- If the service provider does not observe the types
  - Set priorities with strategic delays
  - Work conserving discipline may not be optimal
  - Delay cost minimization is not the dominant criterion
  - Strategic delay (for low priority items) deters high priority customers purchasing a low priority menu
  - This accomplishes incentive compatibility

# Multiple Customer Types-An example (Allon, 2010)

- M/M/1 queue with $\mu = 1$
- $\lambda_1 = 0.2$, $R_1 = 100$, $c_1 = 20$
- $\lambda_2 = 0.3$, $R_2 = 30$, $c_2 = 4$
- Under $c\mu$ rule, $w_1 = 1/(\mu - \lambda_1) = 1.25$,
  $w_2 = 1/(\mu(1-\rho_1)(1-\rho_1-\rho_2)) = 2.5$
  - $p_1 = R_1 - c_1 w_1 = 100 - 20(1.25) = 75$
  - $p_2 = R_2 - c_2 w_2 = 30 - 4(2.5) = 20$
  - Revenue: $0.2(75) + 0.3(20) = 21$
- Overall delay cost=$2.5(4) + 1.25(20) = 35$
- Not IC: $75 + 20(1.25) > 20 + 20(2.5) = 70$

# Multiple Customer Types-An example

- Highest IC price under $c\mu$:

$$p_1 = p_2 + c_1 w_2 - c_1 w_1 = 20 + 20(2.5) - 20(1.25) = 45$$

- with revenue=$45(0.2) + 20(0.3) = 15$
- If one can have $w_2 \leftarrow 2.5 + 1 = 3.5$
- Price for Type-2 becomes: $p_2 = 30 - (3.5)4 = 16$
- IC Type 1 price

$$p_1 = 16 + 20(3.5) - 20(1.25) = 61$$

- Revenue=$61(0.2) + 16(0.3) = 17.$
- Overall delay cost=$3.5(4) + 1.25(20) = 39$

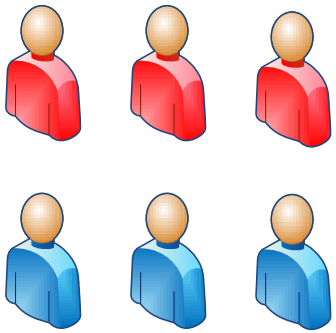
# A manufacturing and service system

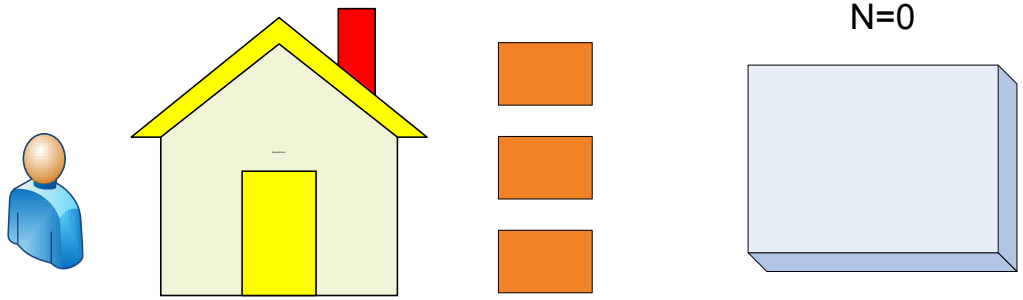
- A common part is used for service activity of heterogenous customers
  - AudiA4 and VW Passat use the same engine, transmission and some other features
  - Design for after-sales-service
- Customers have different sensitivity for waiting and service valuation
- Service provider keeps a common spare parts inventory
- Operates with a base stock policy (base stock level $y$)
- Parts are replenished through a finite capacity system
  - M/M/1 but can be generalized

# A manufacturing and service system

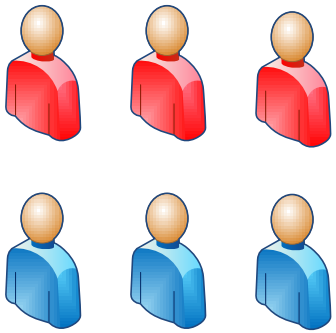
- Whenever there is on-hand stock, customer demand is satisfied irrespective of the type
- If on-hand stock is zero, customers have to wait
- Non-preemptive priorities
  - A customer is tagged with an incoming part (irrespective of the type)
- $y =$ net inventory $+$ outstanding parts $(N)$

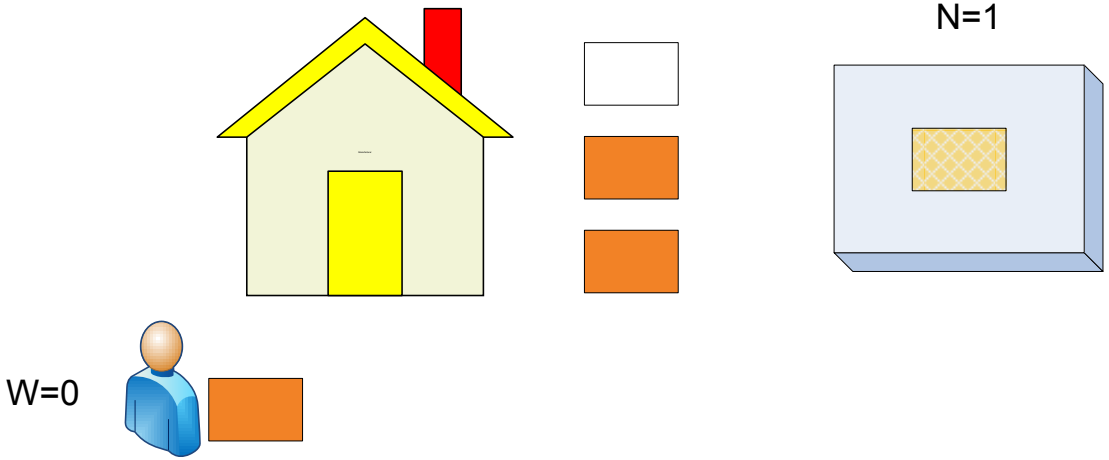
# An Illustration



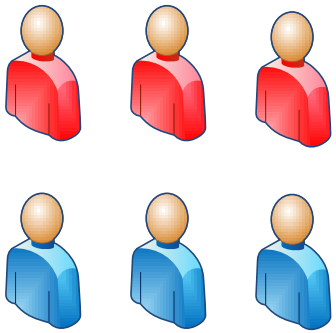


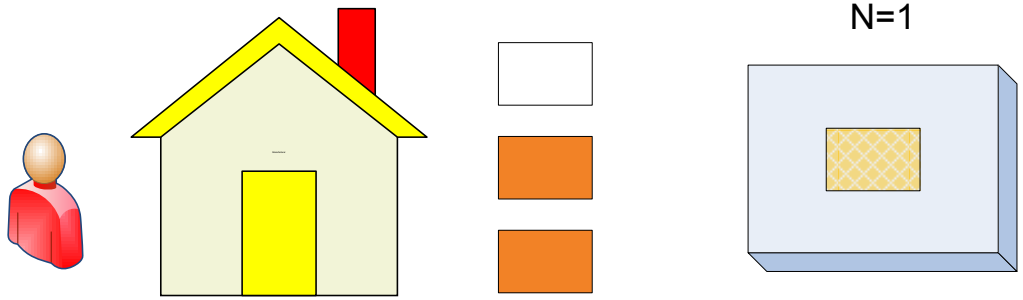
# An Illustration



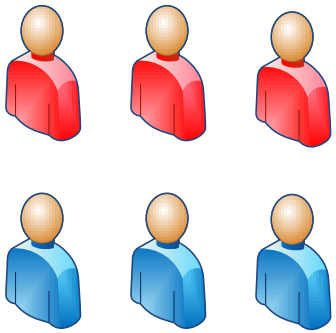


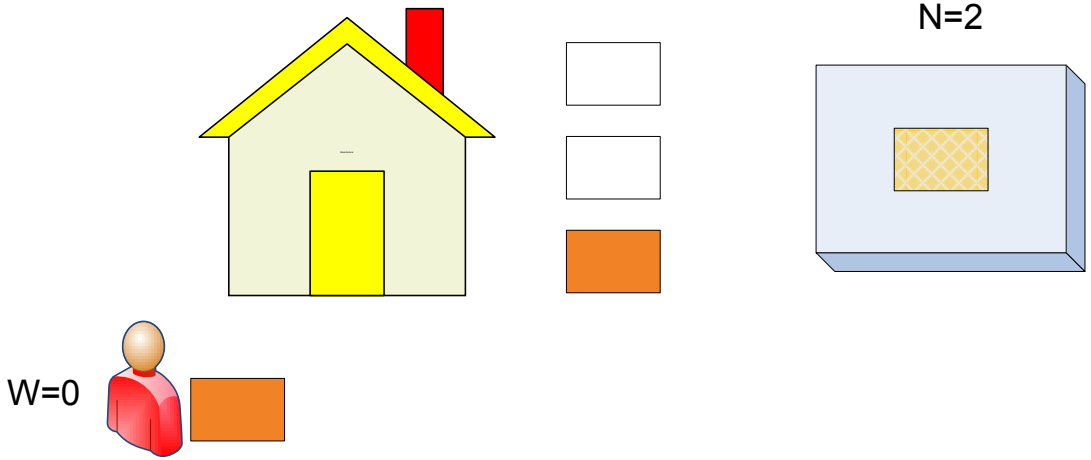
# An Illustration



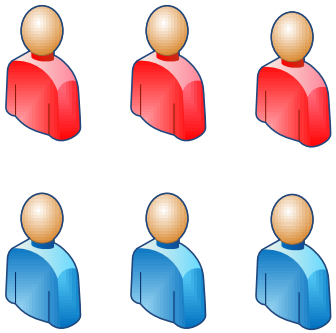


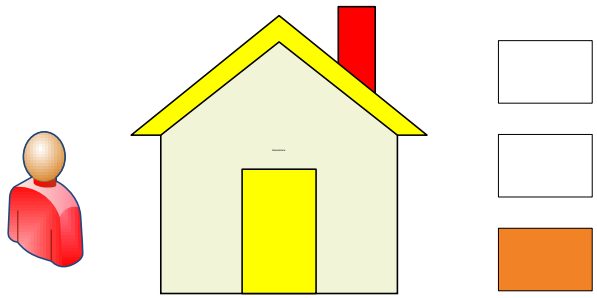
# An Illustration

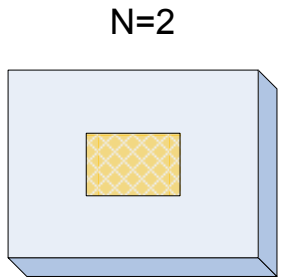




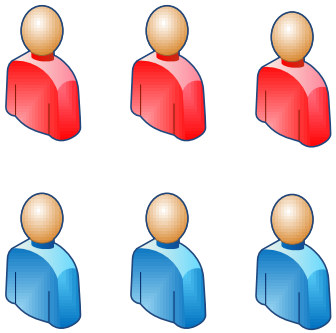
# An Illustration

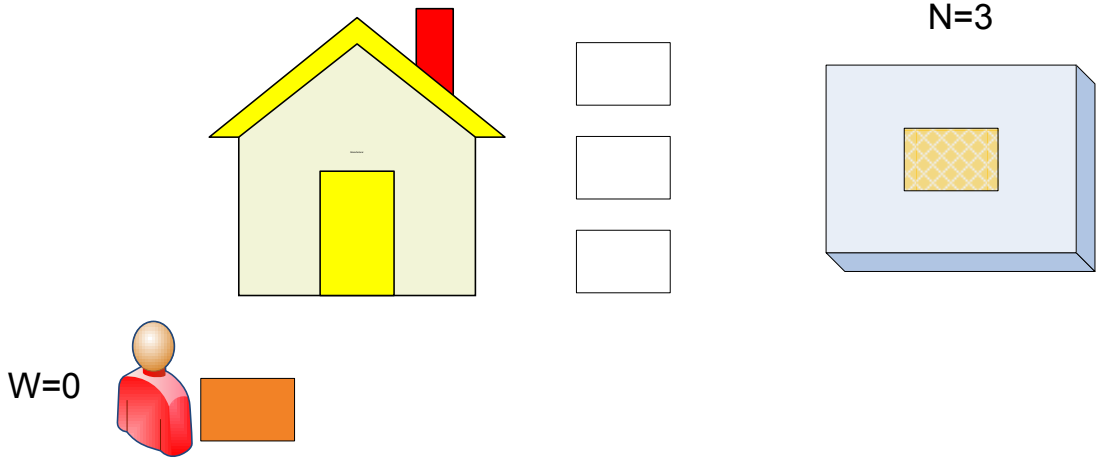




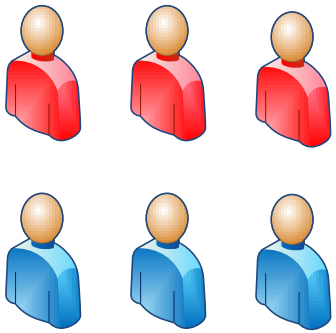


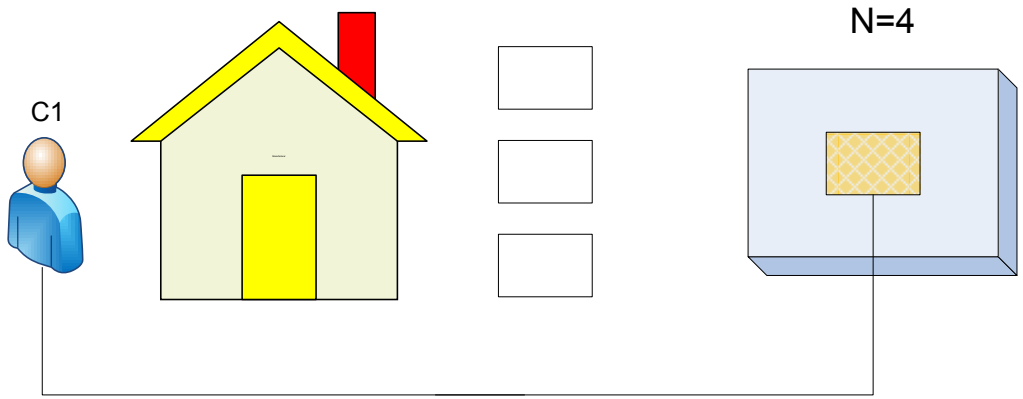
# An Illustration



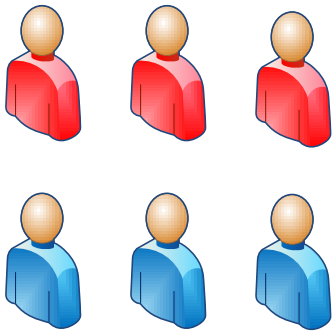


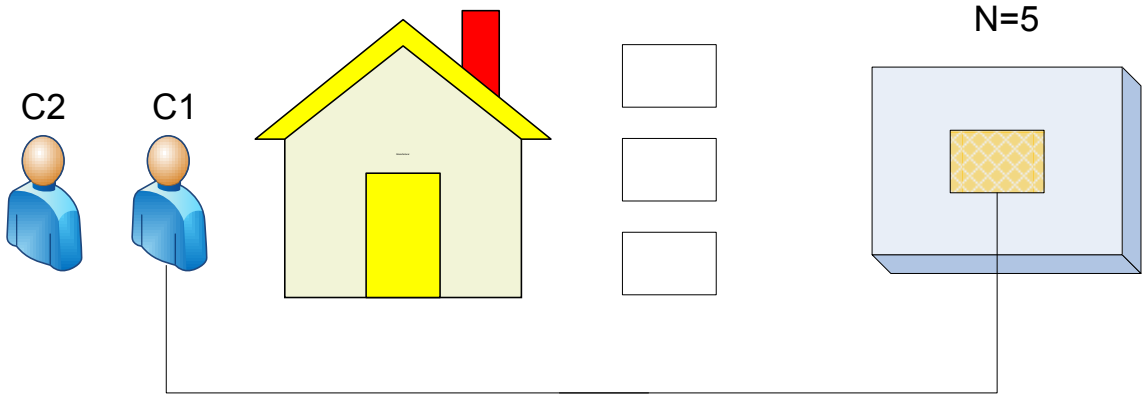
# An Illustration



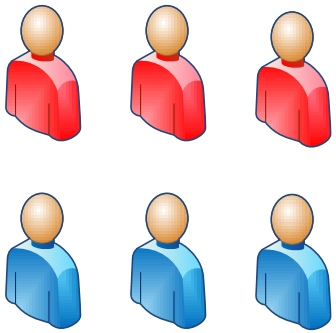


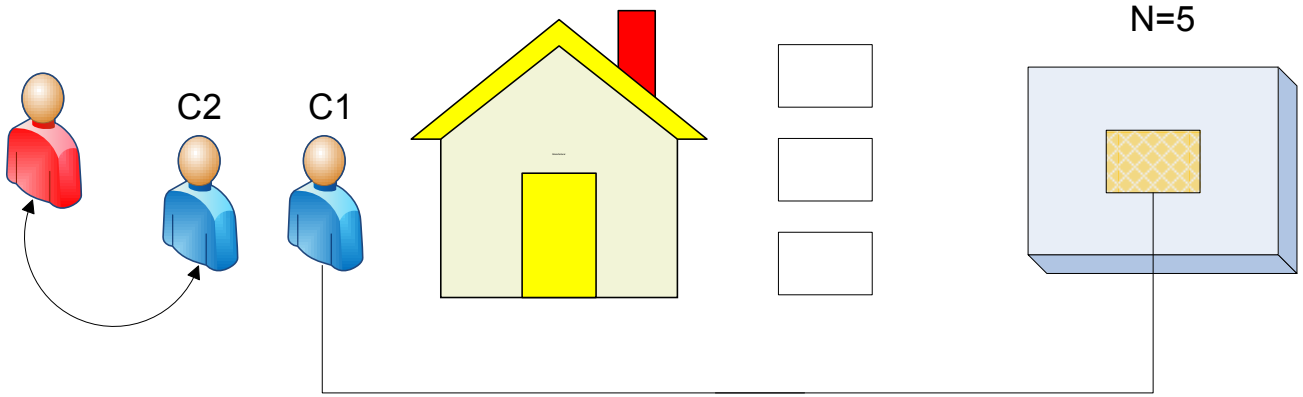
# An Illustration



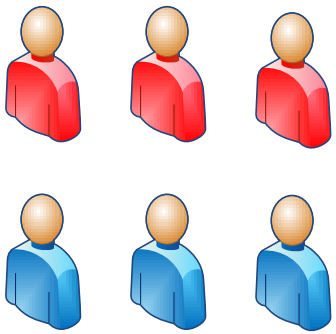


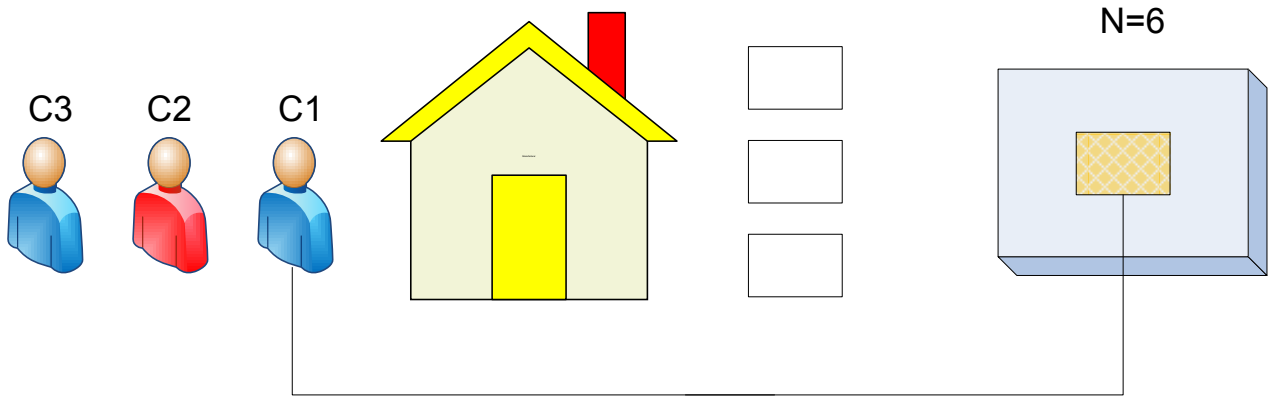
# An Illustration



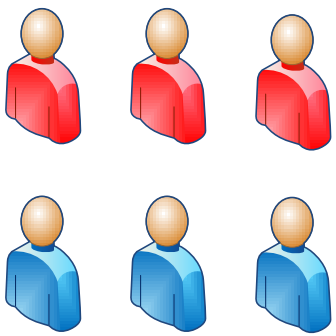


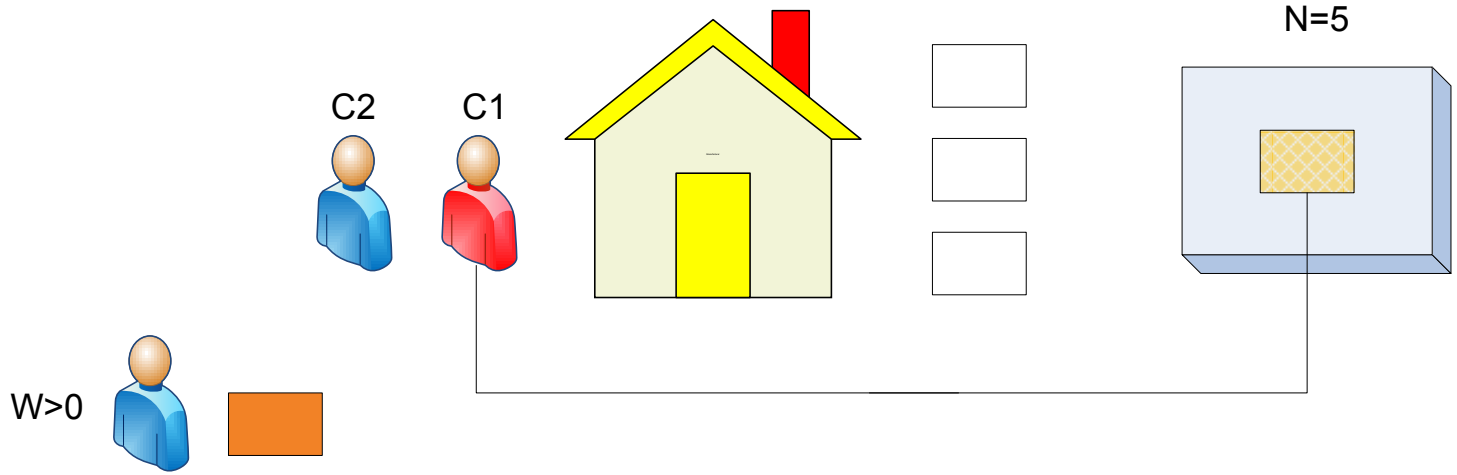
# An Illustration





# An Illustration





- The service provider determines
  - base stock level $y$
  - price menu $p = (p_1, \ldots, p_M)$
  - the priority scheme
- Customers react by arriving with $\lambda = (\lambda_1, \ldots, \lambda_M)$

$$R_i = p_i + c_i E[\text{waiting time}]$$

$$i = 1, 2, \ldots, M$$

- The distribution of the outstanding parts
  - M/M/1: $\Pr\{N = k\} = \rho^k(1-\rho)$
  - $\rho = \sum_{i=1}^{M} \lambda_i/\mu$
- By PASTA property

$$\begin{aligned} E[\text{waiting with } y \geq 0] &= \Pr\{N \geq y\} E[\text{waiting in a standard queue}] \\ &= \rho^y E[\text{waiting in a standard queue}] \end{aligned}$$

- The service provider's problem

$$\max_{y,\lambda,u} \left\{ \sum_{i=1}^{M} \lambda_i R_i - \rho^y \sum_{i=1}^{M} c_i \lambda_i w_i - hE[(y-N)^+] \right\}$$

- $u = (u(1), u(2), \ldots, u(M))$ the priority order

# Main Results

- Restricted to work-conserving disciplines:
  - $c\mu$ rule is optimal: $c_1 \geq c_2 \geq \cdots c_M$
- Given $\lambda$, optimal base stock level:

$$y^*(\lambda) = \min\left\{y \geq 0 : \rho^{y+1} \leq \frac{h}{h + vH(\lambda)}\right\}$$

$$v = (1-\rho)/\rho, \qquad H(\lambda) = \sum_{i=1}^{M} c_i \lambda_i E[\text{waiting}(\lambda)]$$

- Prices given by the first order conditions are incentive compatible

# A reduction:

- If $R_i \geq R_j$ and $c_i \leq c_j$, then Type-i dominates Type-j
- $\lambda_j^* = 0$ $(p_j^* = R_j)$
- $c_i = c$, $R_k = \max\{R_i\}$ or $R_i = R$, $c_k = \min\{c_i\}$

$$\max_{y \geq 0, \rho \in [0,1)} \left\{ R\mu\rho - c\rho^y \frac{\rho}{1-\rho} - h(y - \frac{\rho}{1-\rho}(1-\rho^y)) \right\}$$

$$y(\rho) = \min\{y \geq 0 : \rho^{y+1} \leq h/(c+h)\}$$

# A continuous approximation: $\Pr\{N \geq y\} \approx e^{-vy}$

| Variable | Value | $\partial/\partial h$ | $\partial/\partial c$ |
|---|---|---|---|
| $\lambda$ | $\mu - \sqrt{\mu K / R}$ | $\downarrow$ | $\downarrow$ |
| $y$ | $\left(\sqrt{R\mu K} - K\right) / h$ | $\downarrow$ | $\uparrow, \downarrow$ |
| $E[W]$ | $\sqrt{R / (\mu K)}$ | $\downarrow$ | $\downarrow$ |
| $p$ | $R - c\sqrt{R / (\mu K)}$ | $\uparrow$ | $\downarrow$ |
| $K$ | $h \log(1 + c/h)$ | | |

- $K = h \log(1 + c/h) \to c$ as $h \to \infty$
- Compare $p$ above with

$$p = R - \sqrt{\frac{cR}{\mu}} = R - c\sqrt{\frac{R}{\mu c}}$$

- Optimal profit: $(\sqrt{R\mu} - \sqrt{K})^2$

$$K < c \implies (\text{profit with } y > 0) \geq (\text{profit with } y = 0)$$

- Attracts more demand (with smaller price) and achieves a higher total profit.

Mendelson, H., Wang, S., 1990, Optimal Incentive-Compatible Priority Pricing for the M/M/1 Queue, *Operations Research*, 38, 870-883.

Afèche, P., 2013, Incentive-Compatible Revenue Management in Queueing Systems: Optimal Strategic Delay, *MSOM*, 15, 423-443.

Allon, G., 2010, Pricing and Scheduling Decisions, *Wiley Encyclopedia of Operations Research and Management Science edited by James J. Cochran*, Wiley.

Maglaras, C., Yao, J., Zeevi, A., 2013, Optimal Price and Delay Differentiation in Queueing Systems, *Working Paper*.

Guler, G., Bilgic, T., Gullu, R., 2014, Joint Inventory and Pricing Decisions when Customers are Delay Sensitive, *to appear in IJPE*.

# Some less explored topics

- Alternative cost and reward structures
  - "willingness to wait": $E[(W - WtoW)^+]$
- Correlation of $R$ and $c$
- Competition and cooperation among service providers
- Distribution free bounds

$$\max_{\lambda} \min_{f_R} \lambda E[(R - cw(\lambda))^+]$$

- Estimation errors in parameters
- Bounded rationality (Huang, Allon and Bassamboo, 2014)