On Pooled versus Dedicated Service in the Presence of Triage Errors

Yonit Barron, Opher Baron

2024 Sandra Rotman Centre for Health Sector Strategy

Thx: Ren Yi

Access to Care- Long Lines, Long Waits

FRASER INSTITUTE

Waiting [is] a defining characteristic of Canadian health care. … A median waiting time of **27.7 weeks** [between referral and treatment in 2023] , 27.4 weeks [in 2022], 9.3 weeks [1993].[1]

U.S.News & WORLD REPORT

The median wait time far **exceeds two hours** in some states. The rate of … visits has increased significantly… admitted patients in the nation's capital wait a median of **286 minutes** for their room in the hospital.

CNN Money

The average time to see (any) dermatologist is 72 days in Boston, 56 days in Minneapolis, and only 14 days in San Diego.

How long will you wait to see a doctor?

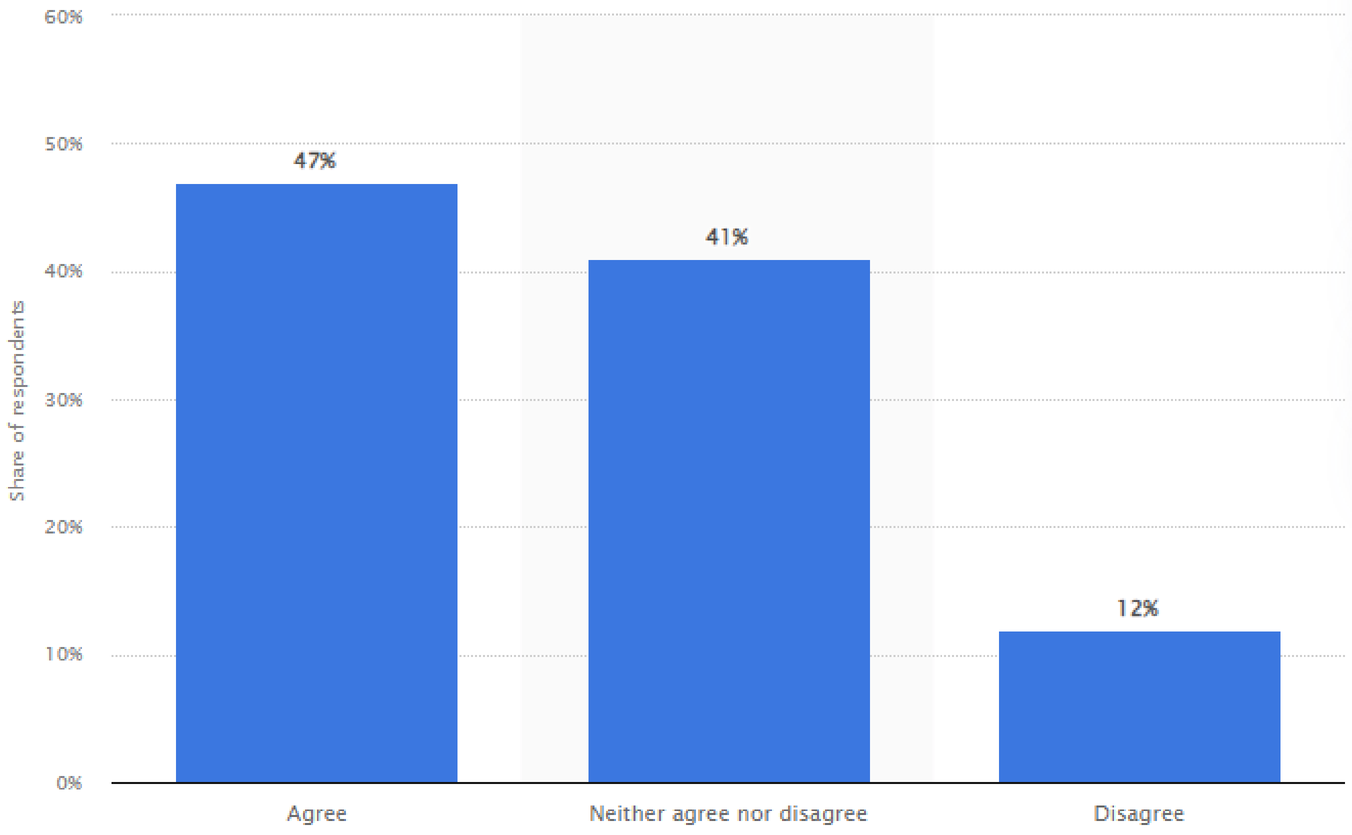
Here's the average wait time for a new patient to see a doctor for a non-emergency issue.

	Wait Time What does this show me?	Expected Length of Stay What does this show me?	Status What does this show me?
Vancouver & Area			
Vancouver General Hospital Patients of ages 17 and older seen	06:24	07:15	
St. Paul's Hospital Patients of all ages seen	00:40	05:12	
Mount Saint Joseph Hospital Patients of all ages seen		Currently closed	
UBC Hospital (UBCH) Patients of all ages seen		Currently closed	
City Centre Urgent & Primary Care Centre Patients of all ages seen UPCC is for mild to moderate illness		Currently closed	
REACH Urgent and Primary Care Centre Patients of all ages seen (lab & x-ray offsite) UPCC is for mild to moderate illness		Currently closed	
Northeast Urgent and Primary Care Centre Patients of all ages seen (lab & x-ray offsite) UPCC is for mild to moderate illness		Currently closed	
Southeast Urgent and Primary Care Centre Patients of all ages seen (lab & x-ray offsite) UPCC is for mild to moderate illness		Currently closed	
BC Children's Hospital Patients seen up to age 16	07:38	06:35	

Health, Pharma & Medtech › State of Health

PREMIUM +

Share of individuals perceiving the waiting times to get an appointment with doctors as too long in Japan in 2018



© Statista 2024

Additional Information

Show source

DOWNLOAD

PDF + XLS + PNG + PPT +

Source

→ Show sources information

→ Show publisher information

→ Use Ask Statista Research Service

Release date

July 2018

Region

Japan

Survey time period

May 25 to June 8, 2018

Number of respondents

Approx. 1,000

Age group

16-64 years

Access to Medical Care in Emergency Departments in America, Asia, and, well, Globally

- EDs serve different types of patients
- Triage: emergency ($\leq$3%), acute(~20%), non acute (~80%)
- Wait targets for patients of different acuity level
  - CTAS: TPIA for each patients' type, measures of LOS
  - Chinese govt. guideline on triage motivate EDs to provide a high level of service to patients
- Fundamental queueing insight: pooling is effective [Smith & Whitt, 1981]
  - Ignores multiple customer types with different targets and priorities
  - Pooling may not be helpful in EDs [Song et al 2019]
- Importantly, in EDs this insight ignores triage and **triage errors**
  - Nurses from 4 Swiss hospitals triage only 59.6% of the patients correctly [Jordi, et al 2015]
  - For elderly patients, 117 out of 519 cases were assigned to a lower type [Grossmann 2012]

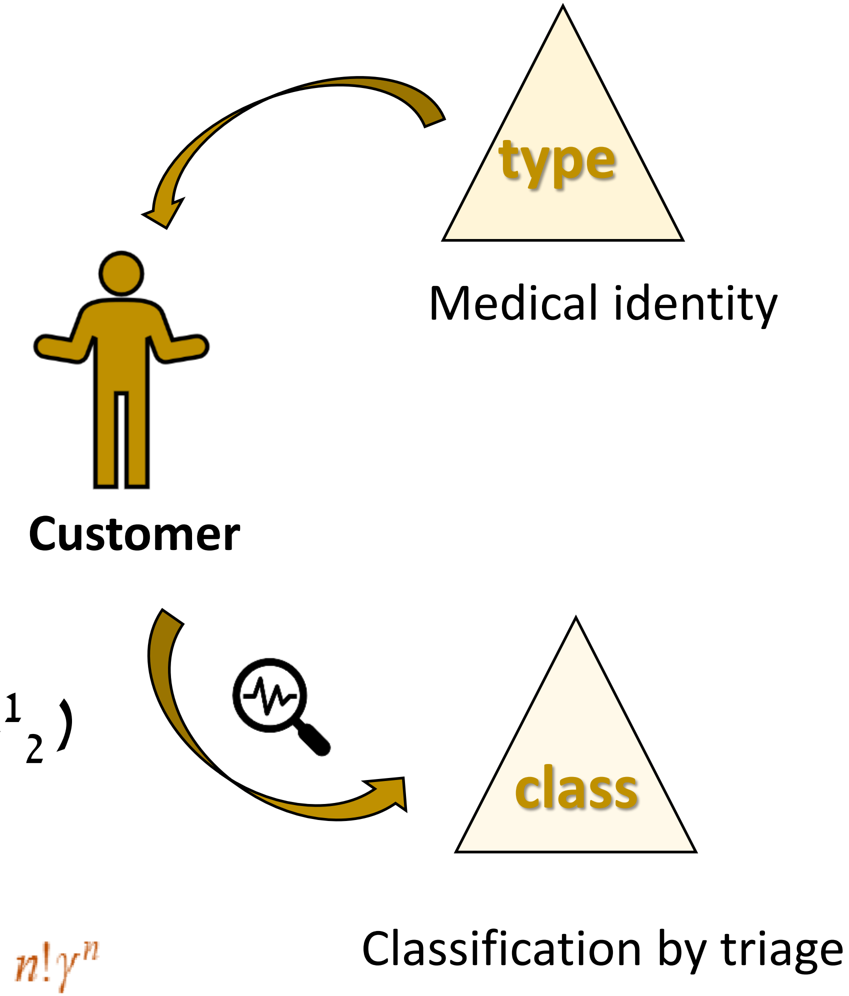
Model Description

- ED system
- Acute (type 1), non-acute (type 2) patients
- Poisson arrival, rate λ_i, $i=1,2$.
- Workload brought $S_i \sim exp(1/\gamma_i)$
- Triage: p_{ij} = type **i** is classified as type **j**
- System's moments (total workload $L^1 = L^1{}_1 + L^1{}_2$)

  $L^r{}_j := p_{1j}\lambda_1 (\gamma_1)^r + p_{2j}\lambda_2 (\gamma_2)^r$
- Capacity μ_i, $i=0$ (pooled),1,2.

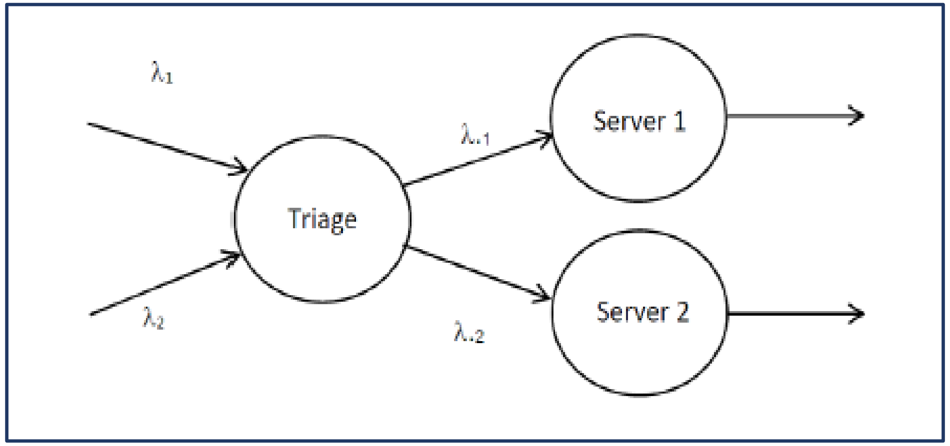
$Y \sim exp(\mu/\gamma),$

$$E[Y^n] = \frac{E[S^n]}{\mu^n} = \frac{n!\gamma^n}{\mu^n}.$$



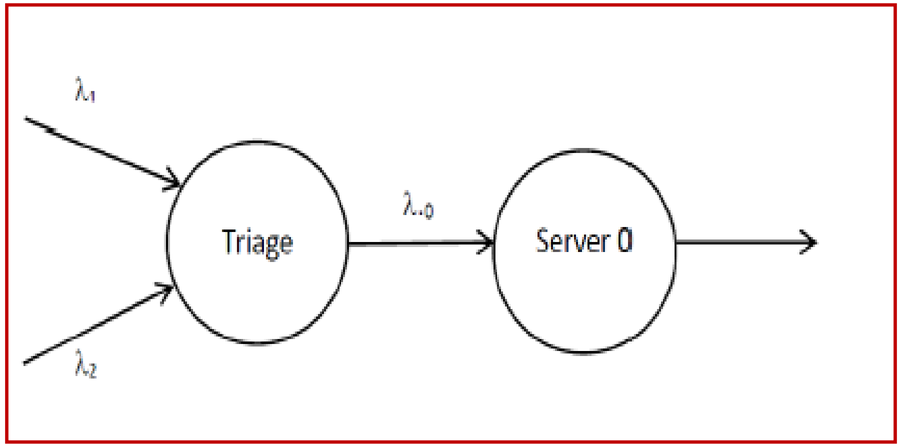
Two M/M/1 like models:

Dedicated system



FCFS

Pooled system



FCFS

Priority (PR)

(non-preemptive)

Objective functions and SL Constraints

w_i - target for expected waiting time of type i customer

W_i - (realized) expected waiting time of type i customer

c_j - capacity cost for one unit of workload of server j

Dedicated Cost

$$\min C_d = c^1\mu^1 + c^2\mu^2$$
$$s.t.\ W_i \le w_i, \forall i,$$

Pooled Cost FCFS

$$\min C_p = c^0\mu^0$$
$$s.t.\ W \le w_1.$$

Pooled Cost PR

$$\min C_p = c^0\mu^0$$
$$s.t.\ W_1 \le w_1.$$
$$W_2 \le w_2$$

Our *study*

- Deriving the optimal capacities μ_0, $\mu_{1,}$ μ_2 *under SL constrains*
- *Compare dedicated system vs. pooled systems (FCFS and PR policies)*
- The impact of triage errors

Three points of view:

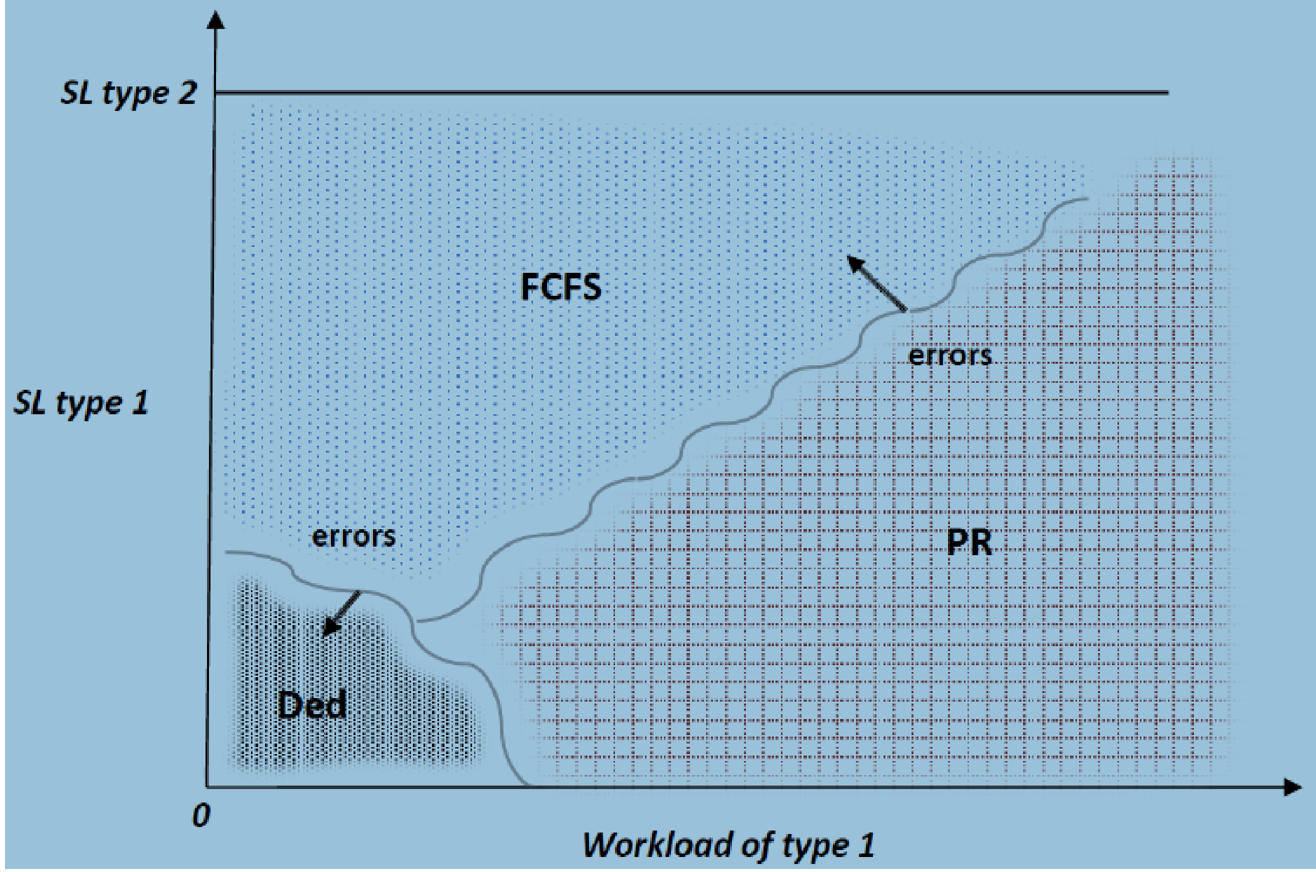
I. Servers' point of view (capacities, utilizations, waiting time at server)

II. Customers' point of view (waiting time observed by customers)

III. System's point of view (total cost)

- Many comparisons…

Qualitative Areas where Each Policy is Cost Minimizing System's POV



Triage errors significantly impact the optimal dedicated system, slightly impact the optimal PR pooled system, and have no impact on the FCFS pooled system => arrows

The dedicated system w. errors may be better jointly on all 3 POV:

1. Servers: lower utilization (one capacity increases)
2. Customers: are served faster (one type strictly faster)
3. System: costs are lower

Analysis with no triage errors

Optimal Dedicated system

Safety stock of servers 1 and 2

$$
\begin{aligned}
C_d^* &= c_1\left(\frac{\lambda_1\gamma_1}{2}+\sqrt{\frac{\lambda_1\gamma_1^2}{w_1}+\left(\frac{\lambda_1\gamma_1}{2}\right)^2}\right)+c_2\left(\frac{\lambda_2\gamma_2}{2}+\sqrt{\frac{\lambda_2\gamma_2^2}{w_2}+\left(\frac{\lambda_2\gamma_2}{2}\right)^2}\right)\\
&= c_1\left(\lambda_1\gamma_1+\frac{\lambda_1\gamma_1^2}{w_1\left(\sqrt{\frac{\lambda_1\gamma_1^2}{w_1}+\left(\frac{\lambda_1\gamma_1}{2}\right)^2}+\frac{\lambda_1\gamma_1}{2}\right)}\right)+c_2\left(\lambda_2\gamma_2+\frac{\lambda_2\gamma_2^2}{w_2\left(\sqrt{\frac{\lambda_2\gamma_2^2}{w_2}+\left(\frac{\lambda_2\gamma_2}{2}\right)^2}+\frac{\lambda_2\gamma_2}{2}\right)}\right)
\end{aligned}
$$

Analysis with no triage errors

Optimal Pooled FCFS

$$C^*_{FCFS} = c_0\left(\frac{L}{2} + \sqrt{\frac{L^{(2)}}{w_1} + \frac{L^2}{4}}\right) = c_0\left(L + \frac{L^{(2)}}{w_1\left(\frac{L}{2} + \sqrt{\frac{L^{(2)}}{w_1} + \frac{L^2}{4}}\right)}\right)$$

Safety stock

$W_1 = W_2 = W_{FCFS} \leq w_1$

Fixed w_1 $\longrightarrow$ C^*_{FCFS} independent of w_2

Analysis with no triage errors

Optimal Pooled PR

Type 1 dominates

$$C^*_{PR} = \begin{cases} (i)\ c_0\left(\frac{\lambda_1\gamma_1}{2}+\sqrt{\frac{L^{(2)}}{w_1}+\left(\frac{\lambda_1\gamma_1}{2}\right)^2}\right) = c_0\left(\lambda_1\gamma_1 + \boxed{\frac{L^{(2)}}{w_1\left(\sqrt{\frac{L^{(2)}}{w_1}+\left(\frac{\lambda_1\gamma_1}{2}\right)^2}+\frac{\lambda_1\gamma_1}{2}\right)}}\right), & w_1 \le w_1^*, \\ (ii)\ c_0\left(\lambda_1\gamma_1+\frac{\lambda_2\gamma_2}{2}+\sqrt{\frac{L^{(2)}}{w_2}+\left(\frac{\lambda_2\gamma_2}{2}\right)^2}\right) = c_0\left(L + \boxed{\frac{L^{(2)}}{w_2\left(\sqrt{\frac{L^{(2)}}{w_2}+\left(\frac{\lambda_2\gamma_2}{2}\right)^2}+\frac{\lambda_2\gamma_2}{2}\right)}}\right), & w_1 > w_1^*, \end{cases}$$

Both customers impact

where

$$w_1^* = \frac{(d-1)^2 L^{(2)}}{d(\lambda_1\gamma_1 + d\lambda_2\gamma_2)L}.$$

$d=w_1/w_2$, ($d=1 \Rightarrow w^*_1=0$)

Systems' comparisons-no triage errors

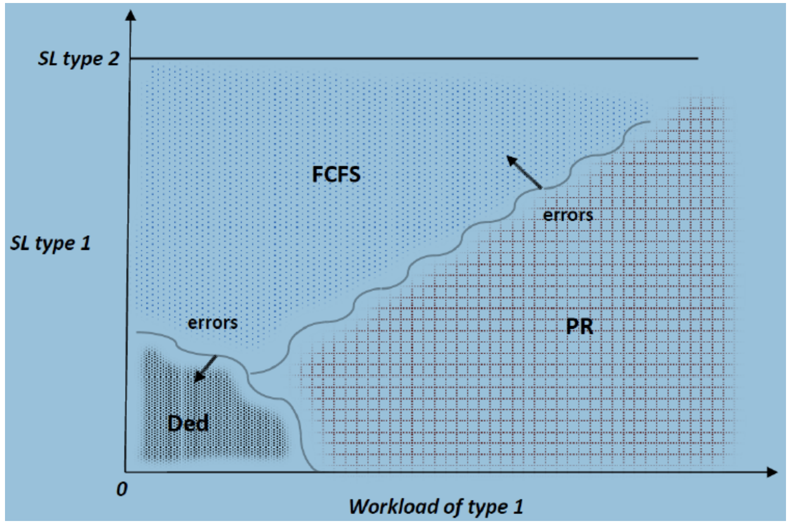
Conclusions:

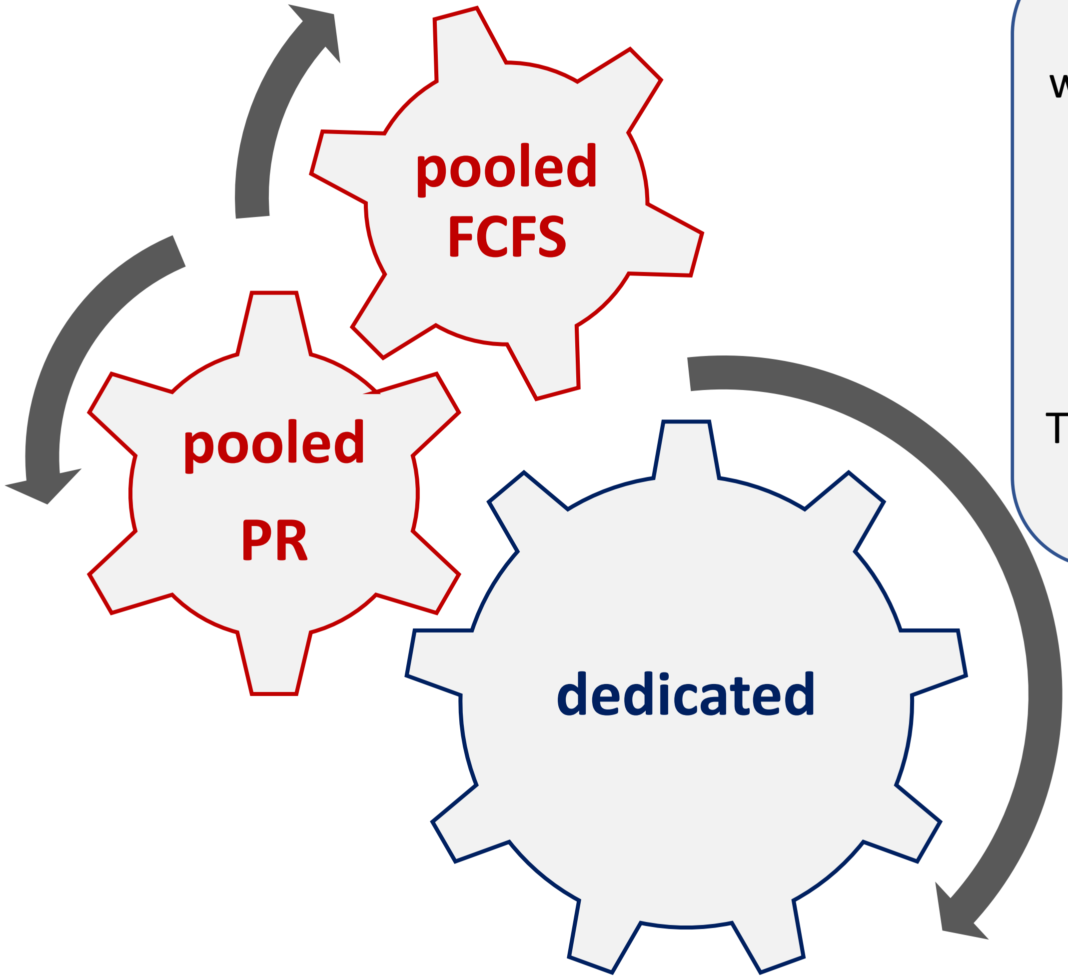
$w_1 = w_2$ ⟹ *Pooled FCFS*

$w_1 << w_2$ ⟹ *Pooled PR*

$w_1 \to \infty$ ⟹ Similar

or $w_1 = 0$ ⟹ Similar





w_1 is strict (low)

workload of type 1 small

μ_0 increases

The benefit from pooling

decreases

Analysis With Triage Errors

Better or worse?

Triage errors highlight **two** effects:

- Servers' workload patters
- Customers service pattern

$L_j = \mathbf{p_{1j}} \lambda_1 \gamma_1 + \mathbf{p_{2j}} \lambda_2 \gamma_2$

Workload of **server *j***
(dedicated systems)

$W_i = \mathbf{p_{ii}} W_{.i} + \mathbf{p_{ij}} W_{.j}$

Service pattern => waits of customer **type *i***

SL is for **Realized** waits

Analysis With Triage Errors

Optimal Dedicated system

$$\min_{\mu_1,\mu_2} \{C_d = c_1\mu_1 + c_2\mu_2\}$$

$s.t.$

$$W_i = p_{ii}\frac{L_i^{(2)}}{(\mu_i - L_i)\mu_i} + p_{ij}\frac{L_j^{(2)}}{(\mu_j - L_j)\mu_j} \le w_i, \qquad i = 1,2$$

$$\mu_j > L_j, \qquad j = 1,2.$$

$$\min_{W_{\cdot 1},W_{\cdot 2}} C_d = c_i\left(\frac{L_i}{2} + \sqrt{\left(\frac{L_i}{2}\right)^2 + \frac{L_i^{(2)}}{W_{\cdot i}}}\right) + c_j\left(\frac{L_j}{2} + \sqrt{\left(\frac{L_j}{2}\right)^2 + \frac{L_j^{(2)}}{W_{\cdot j}}}\right)$$

$s.t.$

$(a)\ W_i = p_{ii}W_{\cdot i} + p_{ij}W_{\cdot j} \le w_i,$

$(b)\ W_j = p_{ji}W_{\cdot i} + p_{jj}W_{\cdot j} \le w_j.$

W_i, W_j decision variables

The optimal waits have

at least one tight SL

$W_i = w_i$

or

$W_j = w_j$

Solution steps (dedicated system…)

✓Transform the problem to a single decision variable w

✓Under some constraints, the objective function $C^*_d(w)$ is convex in w

⬇

$C^*_d(w)$ has a unique minimizer $\bar{w}_i$

Proposition:

(a) If $\bar{w}_i \geq \hat{W}(w_i)$ for both $i = 1$ and $i = 2$, the optimal solution is $W^*_{.i} = \hat{W}(w_i), i = 1, 2.$

$W^*_i = w_i$,$W^*_j = w_j$

(b) Otherwise,

$$W^*_{.i} = \bar{w}_i, W^*_{.j} = \frac{w_j - p_{ji}\bar{w}_i}{p_{jj}}.$$

$W^*_i < w_i$,$W^*_j = w_j$

Conclusions (dedicated system)

The interplay between the changes of the ***workload*** & ***service patterns*** effects the performance of the dedicated system

The system w. errors may be better **jointly** on all 3 POV:

1. Servers: lower utilization (one capacity increases)
2. Customers: are served faster (one type strictly faster)
3. System: costs are lower

Pondering

Intuitively, this may occur when the cheaper server becomes busier due to triage errors. But, this happens more generally!

Analysis With Triage Errors

Pooled FCFS: no analysis ☺

Pooled PR: (no closed form)

There exist a unique $\mu_i > L$ that solves $W_i = w_i$, $i=1,2$. The minimized cost for the PR w. triage errors is

$C_{PR} = c_o \mu_o^*$, where $\mu_o^* = max\{\mu_i\}$.

Some insights (pooled system...)

Comparing FCFS and PR pooled systems w. errors:

(i) When $p_{ii} = 0.5$ for i=1,2 (and all other parameters are the same), FCFS and PR have the same capacity and cost, i.e., $\mu_0^* = \mu_{FCFS}{}^*$;

(ii) ow., FCFS may perform strictly better only when the SL of type 2 customers is tight under PR, i.e., when under PR $W_2^* = w_2$ holds.

Numerical observations:

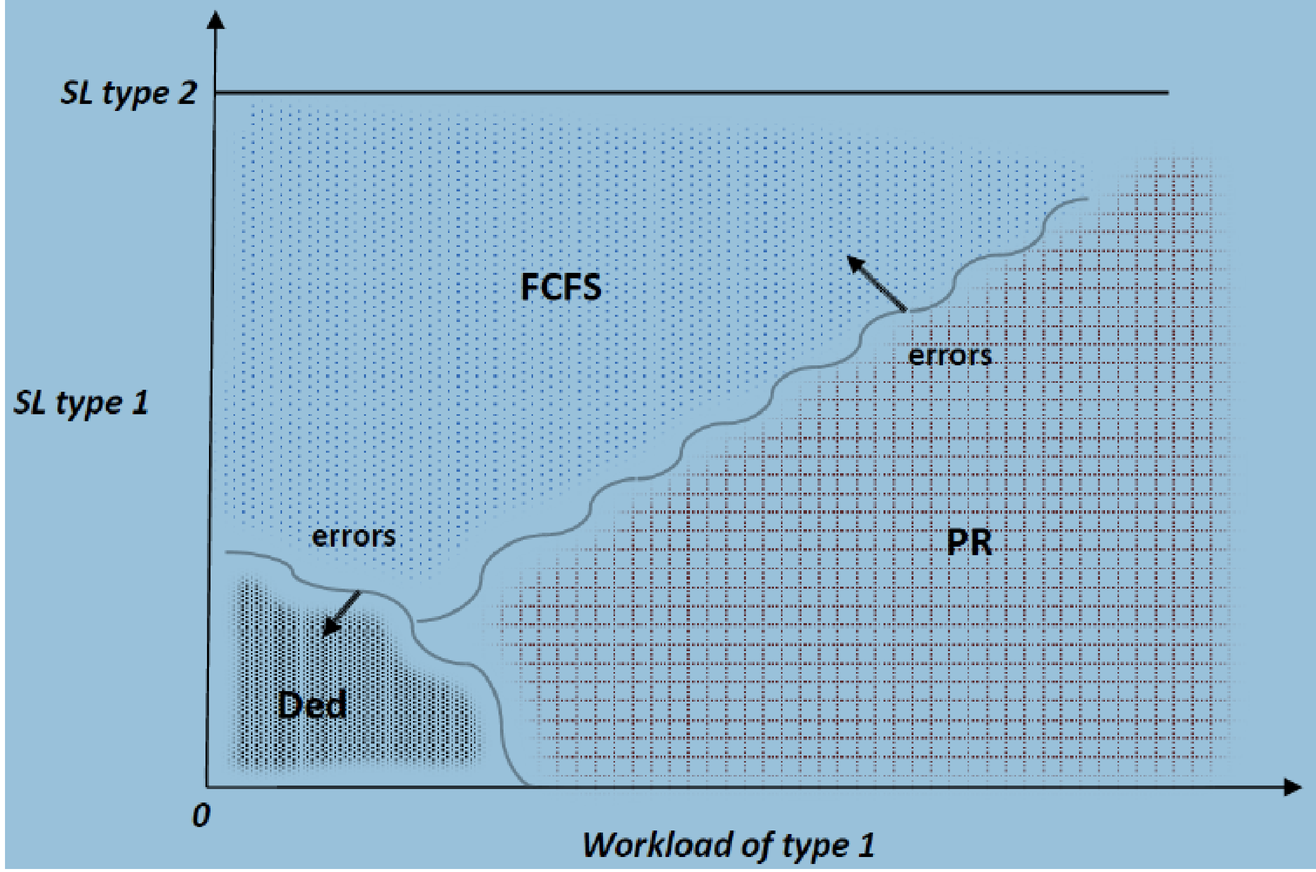
1. FCFS is better than PR when the proportion of type 1 customers is large

Intuition: Increasing the workload of type 1, increases the delay of type 2. To meet their SL, the PR server increases its capacity...

2. FCFS is better than PR when SL are identical, PR is better when these differ

Intuition: under PR, low priority customers face longer waits=>higher capacity (&cost)

Thx!



Triage errors significantly impact the optimal dedicated system, slightly impact the optimal PR pooled system, and have no impact on the FCFS pooled system => arrows

The dedicated system w. errors may be better **jointly** on all 3 POV:

1. Servers: lower utilization (one capacity increases)
2. Customers: are served faster (one type strictly faster)
3. System: costs are lower