YEQT 2021

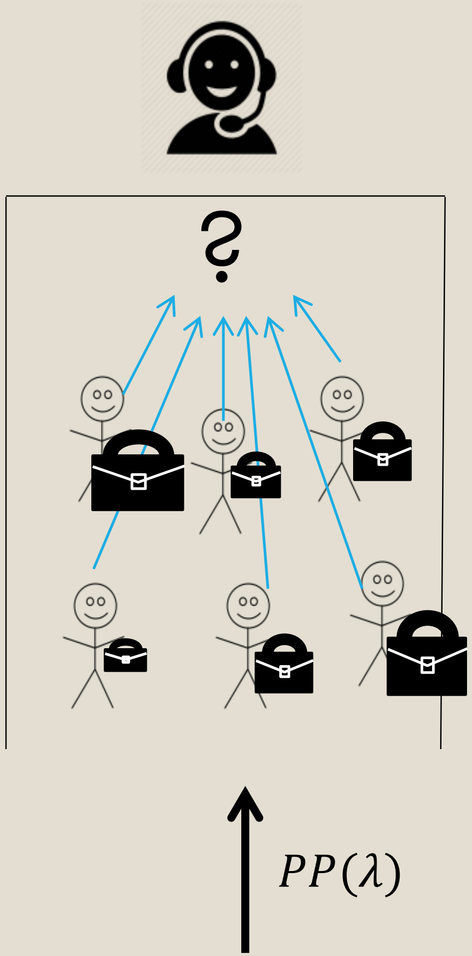
WHEN DOES THE GITTINS POLICY HAVE OPTIMAL RESPONSE TIME TAIL?

Ziv Scully (CMU) & **Lucas van Kreveld (UvA)**

The M/G/1-queue

Objective: Choose service policy to minimize the response time T (sojourn time)

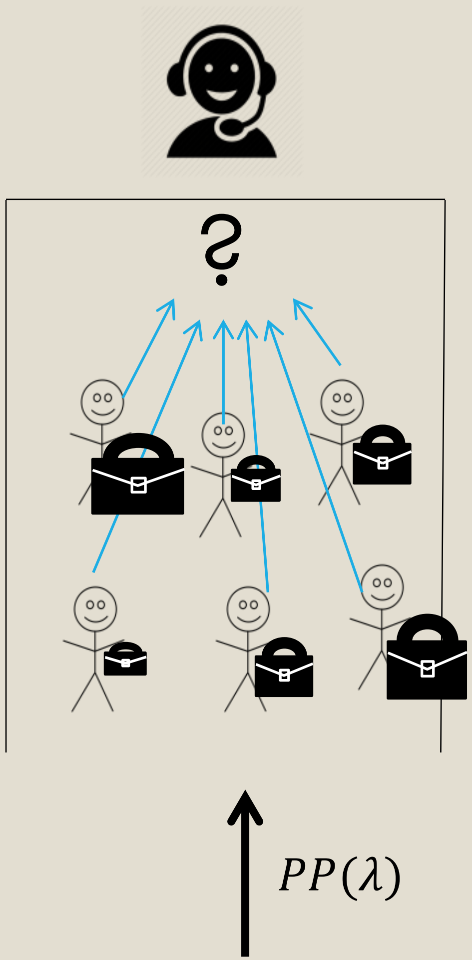
- Minimize <u>mean</u> response time $\mathbb{E}(T)$
- Minimize response time <u>tail</u> $\mathbb{P}(T > t)$
  - Focus on $t \to \infty$

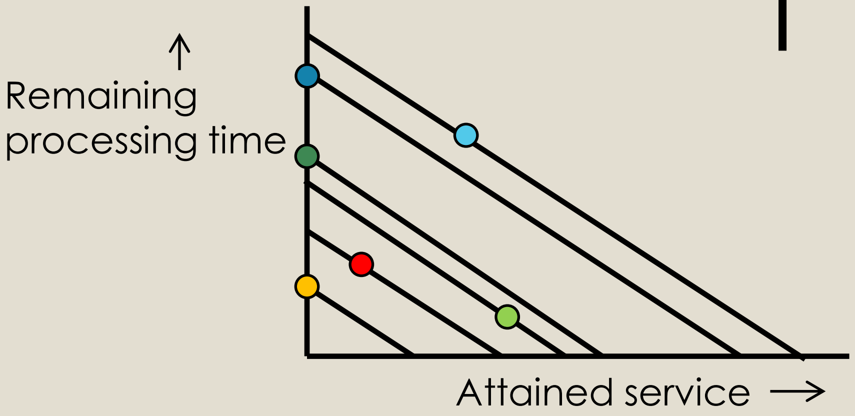


Size-aware policies

- Idea: serve the job that is closest to finishing
- Shortest Remaining Processing Time policy (SRPT)

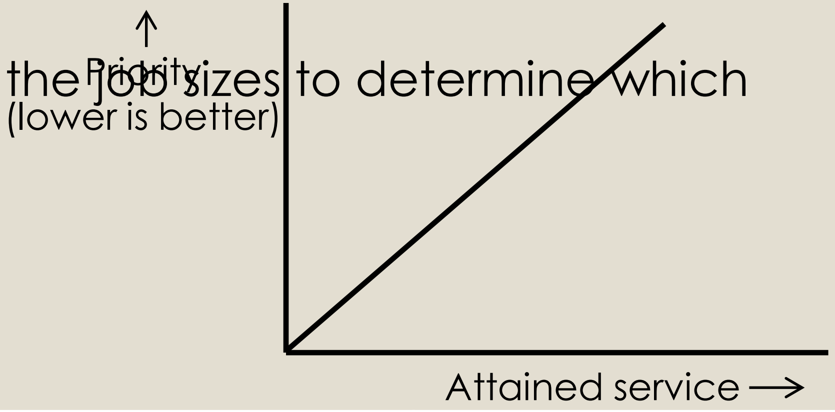
- SRPT
  - Mean response time: optimal
  - (Asymptotic) response time tail:
    "Optimal" under heavy-tailed job sizes
    "Pessimal" under light-tailed job sizes

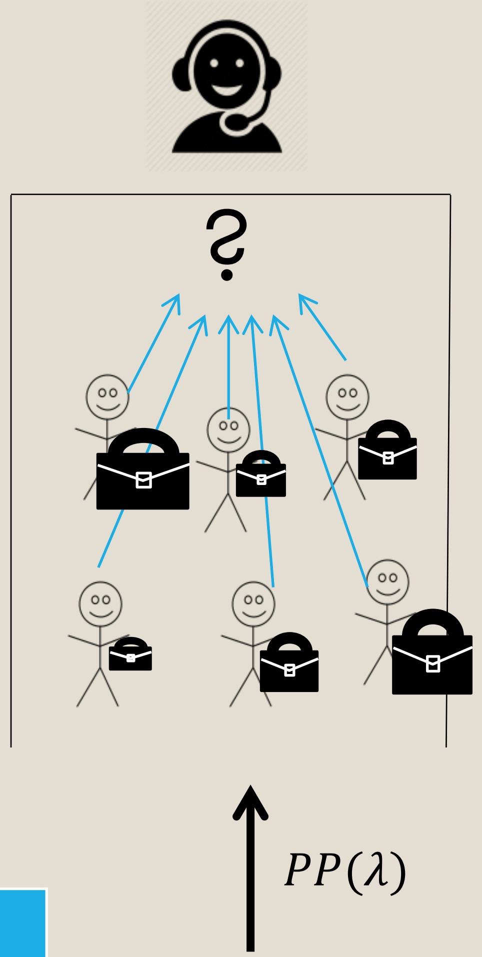




Size-blind policies

- Foreground-Background (FB): serves the job that has received the least amount of service
- Gittins policy: uses distribution of the job sizes to determine which job is "most likely to finish soon"





	$\mathbb{E}(T)$	$\lim_{t\to\infty} \mathbb{P}(T > t)$ for heavy-tailed job sizes	$\lim_{t\to\infty} \mathbb{P}(T > t)$ for light-tailed job sizes
FCFS	Generally not optimal	Pessimal	Optimal
FB	Generally not optimal	Optimal	Pessimal
Gittins	Optimal	Optimal assuming condition on hazard rate	Optimal if Gittins=FCFS, pessimal if Gittins=FB

SOAP-class of service policies

- Age = attained service so far
- SOAP-scheduling: Based on its age a, assign to each job a rank $r(a)$, and serve the job with smallest rank
- A SOAP-policy is characterized by its rank function

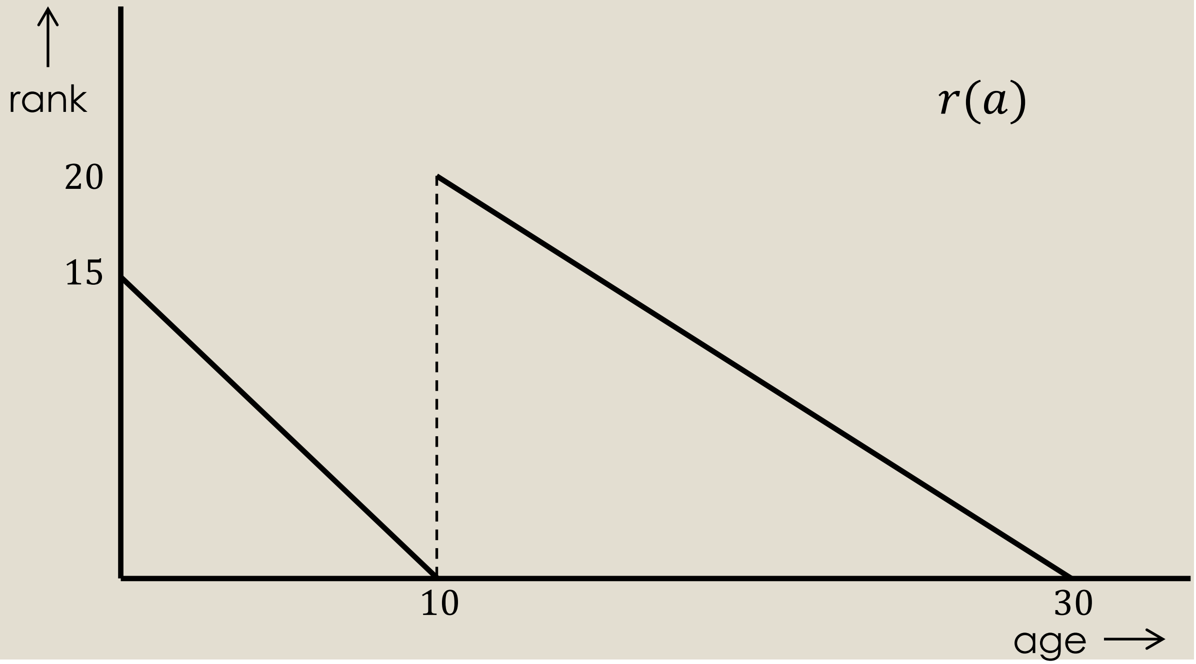
Gittins policy

X = job size

- The Gittins policy is the SOAP-policy with rank function

$$r(a) = \inf_{b>a} \frac{\int_a^b \mathbb{P}(X > x)dx}{\mathbb{P}(X \in (a,b])}$$

- Job size distribution $X \sim \begin{cases} 10 & w.p. \ \frac{2}{3} \\ 30 & w.p. \ \frac{1}{3} \end{cases}$



Outline of results

- Heavy-tailed case:
  - A SOAP-policy is tail-optimal if its rank function satisfies a certain condition
  - The Gittins policy satisfies this condition
- Light-tailed case:
  - The tail behavior of a SOAP-policy is determined by the age of the global maximum of the rank function
  - The Gittins policy can be tail-optimal, tail-pessimal or in between
- Tail-pessimality of Gittins can sometimes be remedied by tweaking the rank function

	$\mathbb{E}(T)$	$\lim_{t\to\infty} \mathbb{P}(T > t)$ for heavy-tailed job sizes	$\lim_{t\to\infty} \mathbb{P}(T > t)$ for light-tailed job sizes
FCFS	Generally not optimal	Pessimal	Optimal
FB	Generally not optimal	Optimal	Pessimal
Gittins	Optimal	Optimal assuming condition on hazard rate	Optimal if Gittins=FCFS, pessimal if Gittins=FB

Tail characterizations

- General goal: minimize $\mathbb{P}(T > x)$ for large x

T = response time
X = job size

Heavy-tailed job sizes

- A non-negative random variable X is heavy-tailed if $\mathbb{P}(X > x) = \Theta(x^{-\alpha})$ for some $\alpha > 1$
- A service policy π is tail-optimal if $\mathbb{P}(T_\pi > x) = \Theta(x^{-\alpha})$

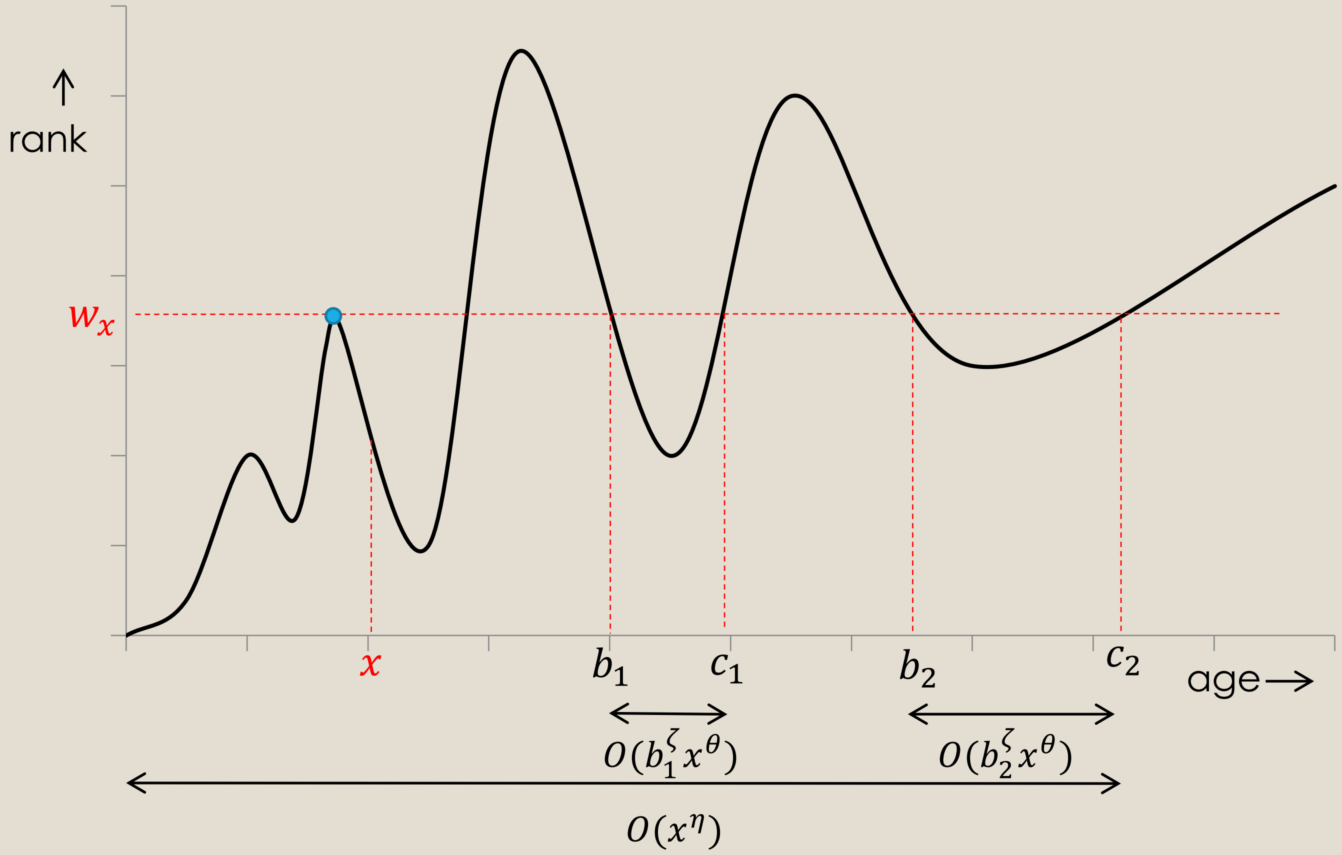
Light-tailed job sizes

- A non-negative random variable X is light-tailed if $\mathbb{E}(e^{\gamma X}) < \infty$ for some $\gamma > 0$
- With $d(T) = \lim_{x\to\infty} -\frac{1}{x}\log \mathbb{P}(T > x)$, a service policy π is

  tail-optimal if π maximizes $d(T_\pi)$
  tail-pessimal if π minimizes $d(T_\pi)$
  tail-intermediate otherwise

Heavy-tail results

- Let $\mathbb{P}(X > x) = \Theta(x^{-\alpha})$ for some $\alpha > 1$
- Let w_x be the highest rank of a size x job
- Condition: There exist $\zeta, \theta, \eta \in [0, \infty]$ such that for any job size x and for any interval (b, c) with rank below w_x with $b \geq x$, we have
  - $c - b = O(b^{\zeta} x^{\theta})$
  - $c = O(x^{\eta})$



Theorem

A SOAP-policy is tail-optimal if

$$\zeta + (\theta - 1)^+ - \frac{(1-\theta)^+}{\eta} < \frac{\alpha - 1}{\alpha}$$

Theorem

$$\zeta_{Gittins} = 0 \text{ and } \theta_{Gittins} = 1$$

Light-tail results

Suppose there exists $\gamma > 0$ such that $\mathbb{E}(e^{\gamma X}) < \infty$

Let $a^* = \inf\{a \geq 0 \mid r(a) \geq r(b) \ \ \forall b \in \mathbb{R}_+\}$ be the age at which the rank function has its (first) global maximum

Theorem

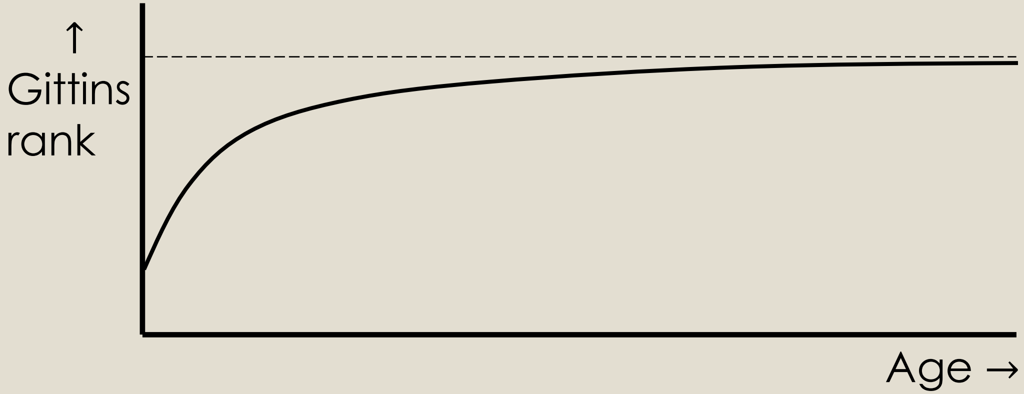
A SOAP-policy is

- tail-optimal if $a^* = 0$ → FCFS
- tail-intermediate if $0 < a^* < \infty$
- tail-pessimal if $a^* = \infty$

Theorem

Gittins policy can be tail-optimal, tail intermediate or tail-pessimal

$$X \sim \begin{cases} Exp(1) & w.p.\ \frac{1}{2} \\ Exp\left(\frac{1}{100}\right) & w.p.\ \frac{1}{2} \end{cases}$$



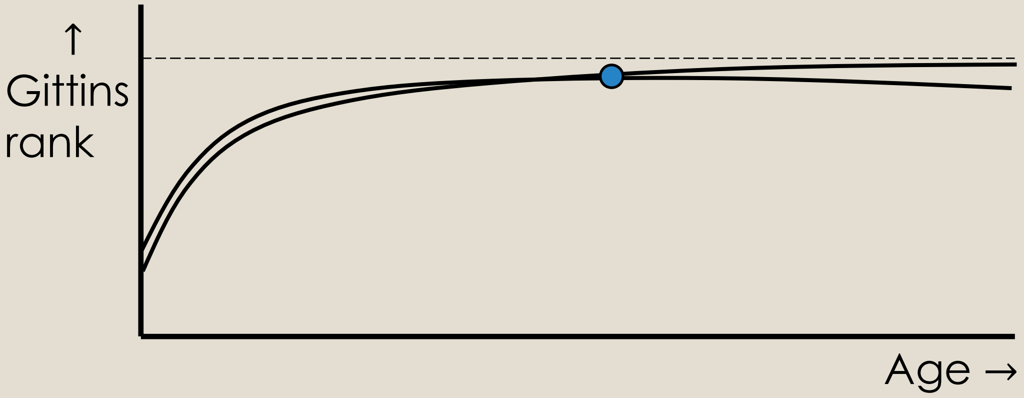
Light-tail results

Theorem

Let π be a SOAP-policy such that $\frac{1}{q} \leq \frac{r_\pi(a)}{r_{Gittins}(a)} \leq q$ for all $a \geq 0$. Then $E[T_\pi] \leq q^2 E[T_{Gittins}]$

Theorem

Suppose a job's Gittins rank is uniformly bounded at all ages. Then for all $\varepsilon > 0$, there exists a tail-intermediate policy that has within a factor $1 + \epsilon$ of optimal mean response time.



Conclusions

	$\mathbb{E}(T)$	$\lim_{t\to\infty} \mathbb{P}(T > t)$ for heavy-tailed job sizes	$\lim_{t\to\infty} \mathbb{P}(T > t)$ for light-tailed job sizes
FCFS	Generally not optimal	Pessimal	Optimal
FB	Generally not optimal	Optimal	Pessimal
Gittins	Optimal	Optimal assuming condition on hazard rate Optimal	Optimal, if Gittins=FCFS, pessimal if Gittins=FB Intermediate pessimal

- Under some conditions, an approximation of Gittins policy guarantees intermediate tail with almost optimal mean response time