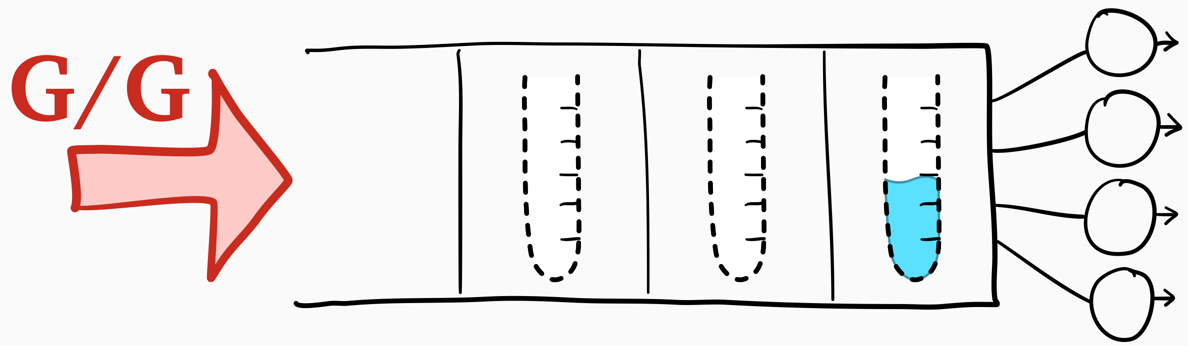
Performance of the

Gittins Policy *in the* **G/G/1 & G/G/*k*,**

With and Without Setup Times

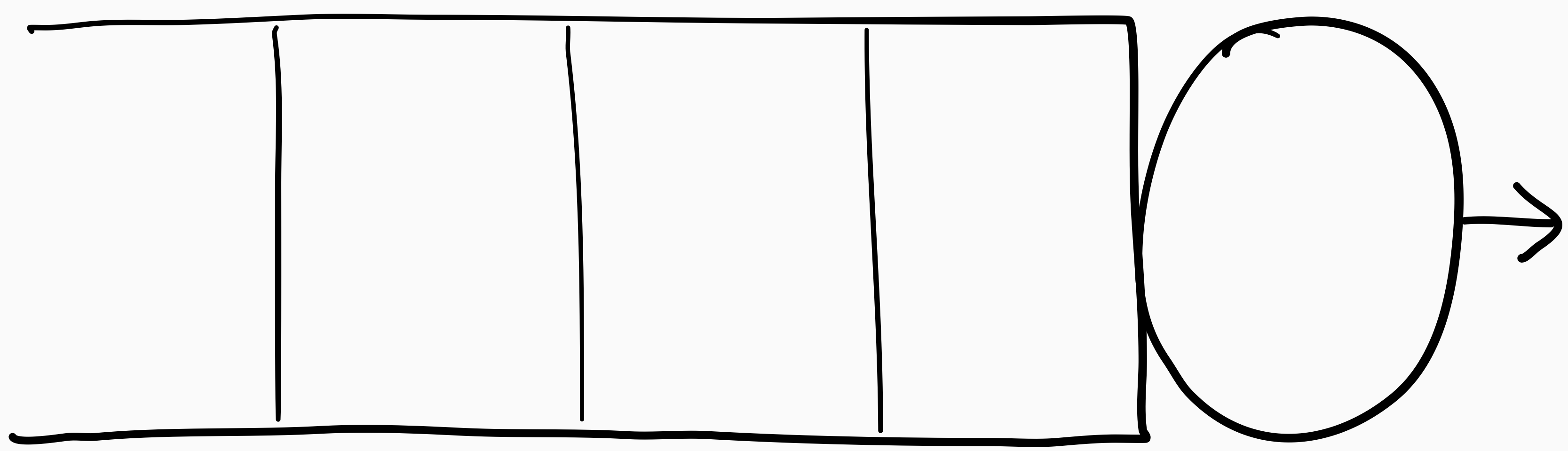
Yige Hong
Carnegie Mellon University

Ziv Scully
Cornell University

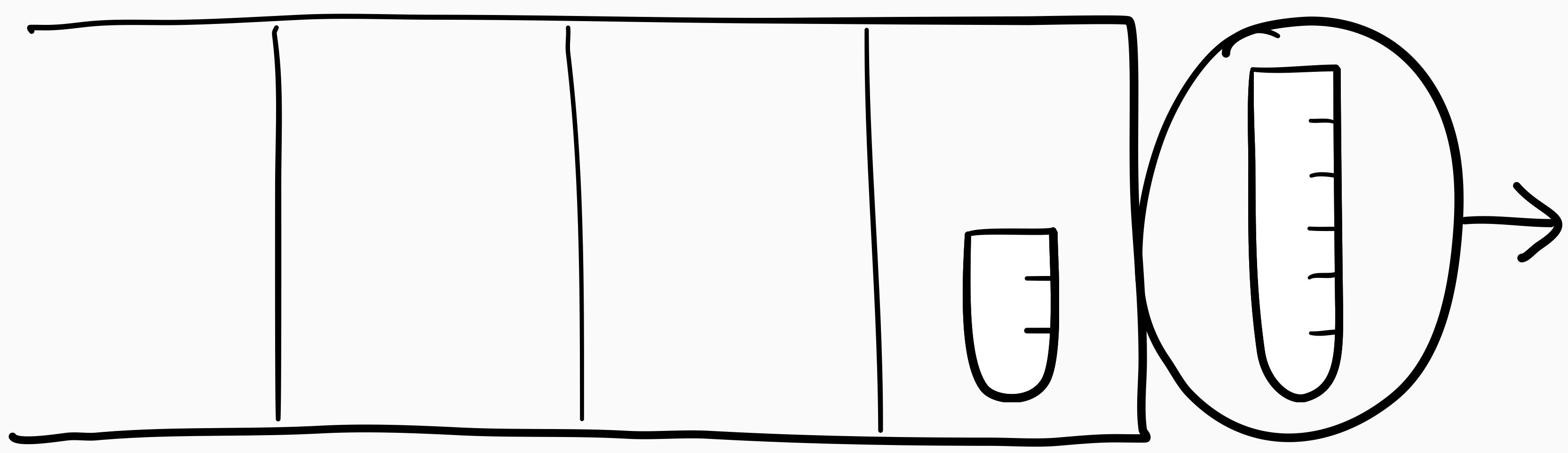


How should we schedule jobs to minimize delay?

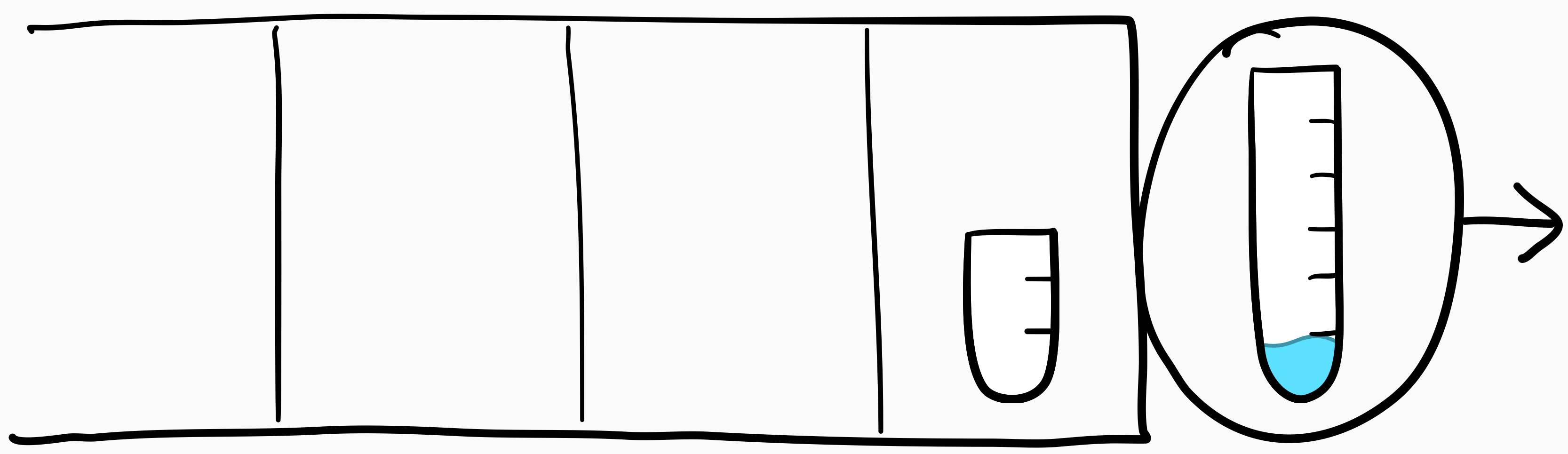
How should we schedule jobs to minimize delay?



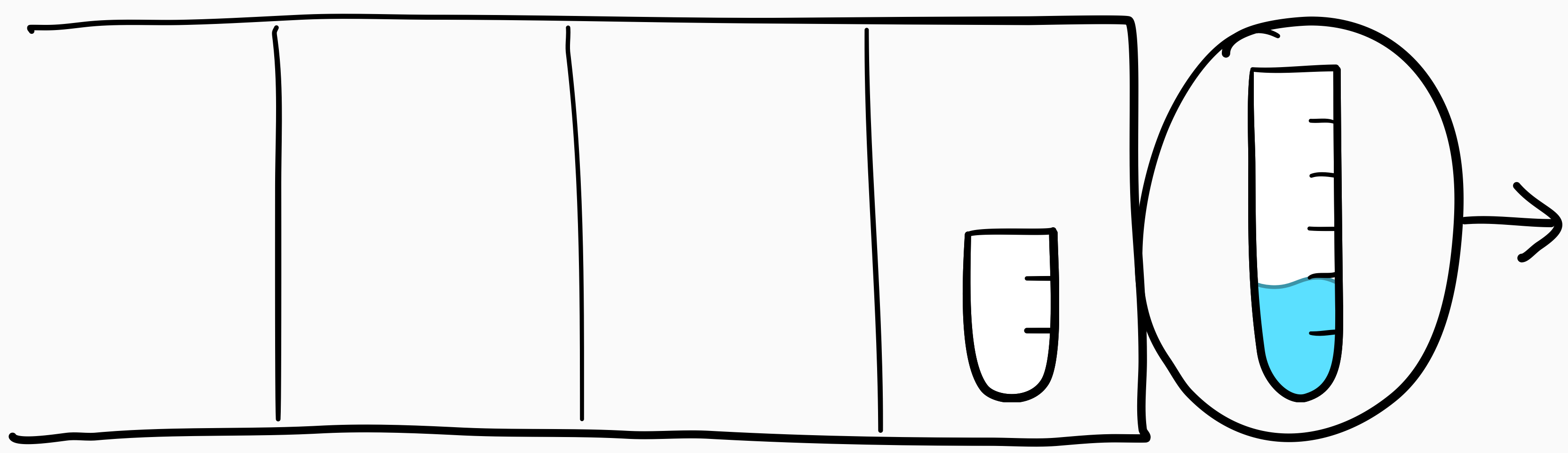
How should we schedule jobs to minimize delay?



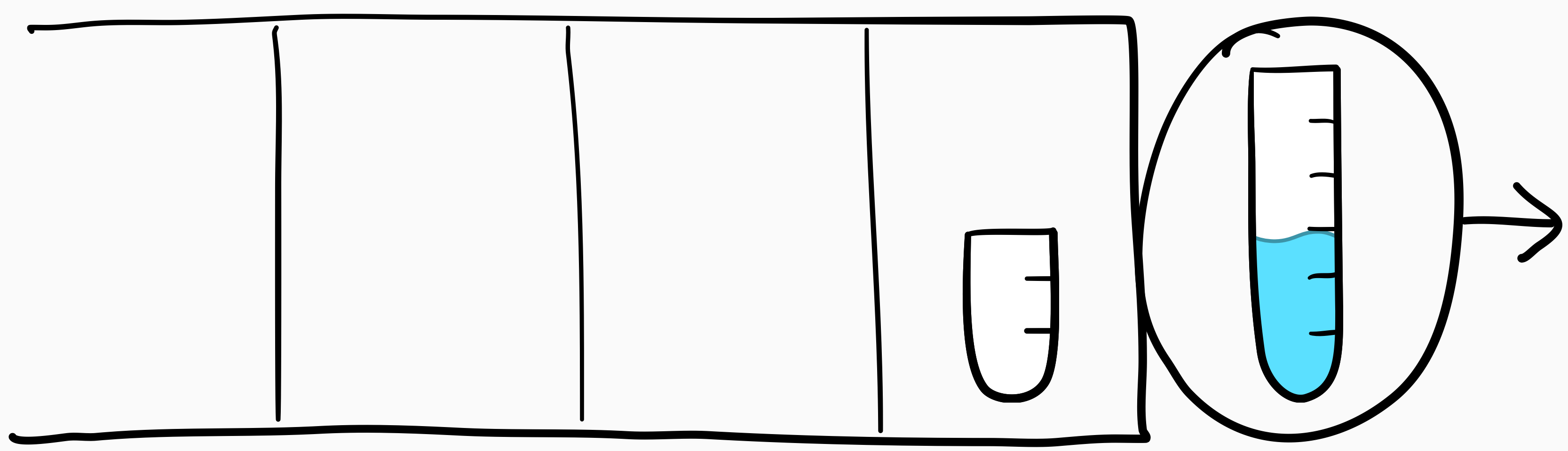
How should we schedule jobs to minimize delay?



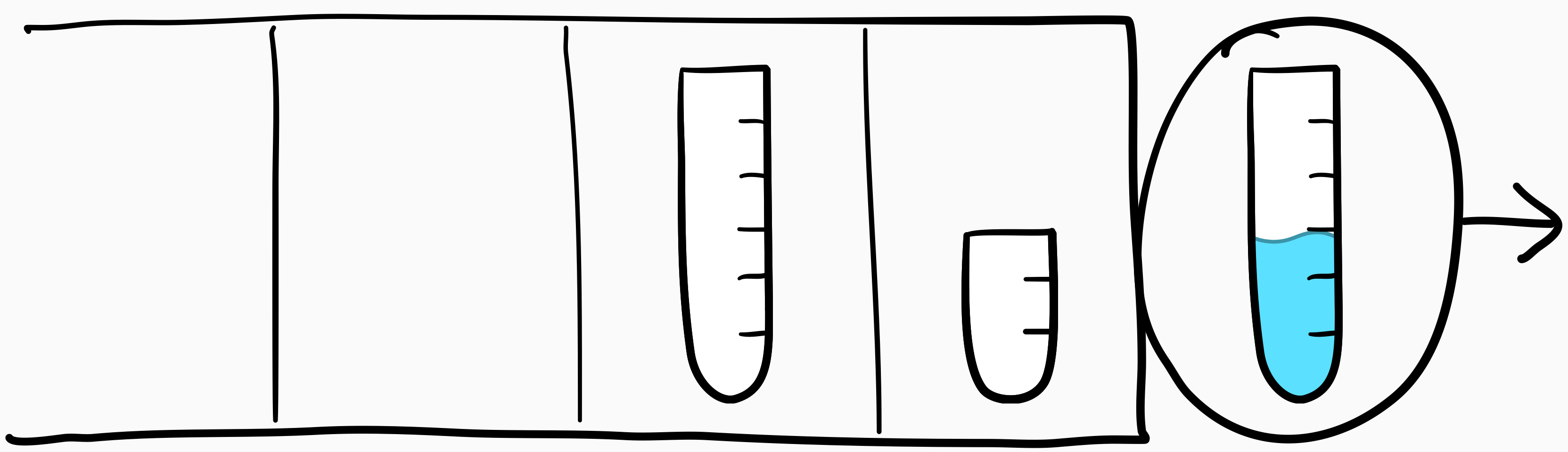
How should we schedule jobs to minimize delay?



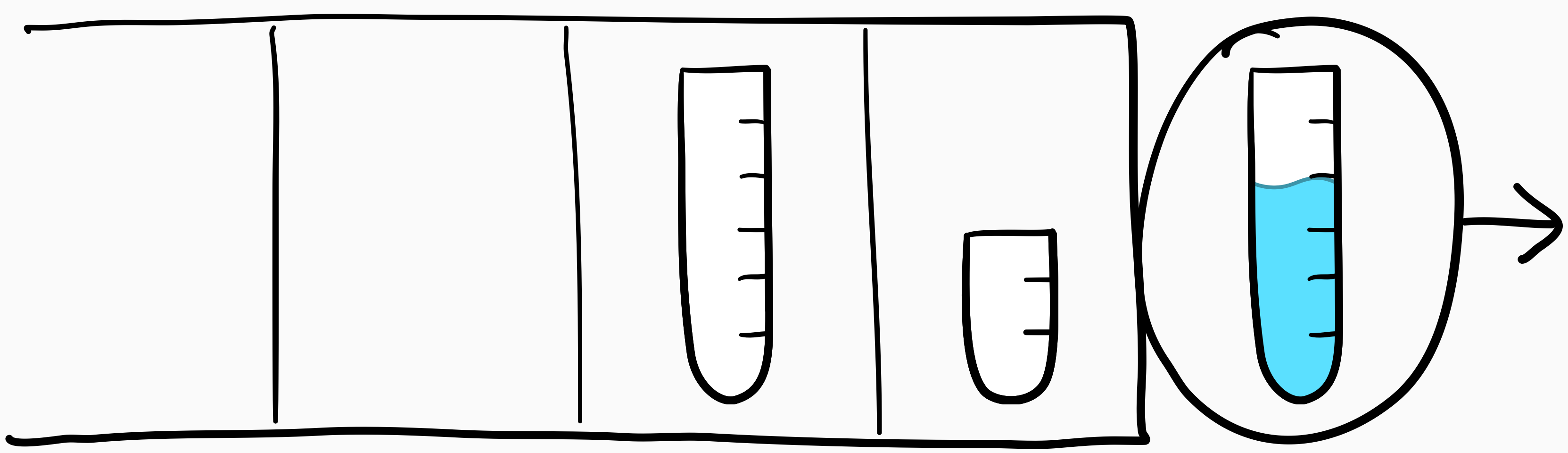
How should we schedule jobs to minimize delay?



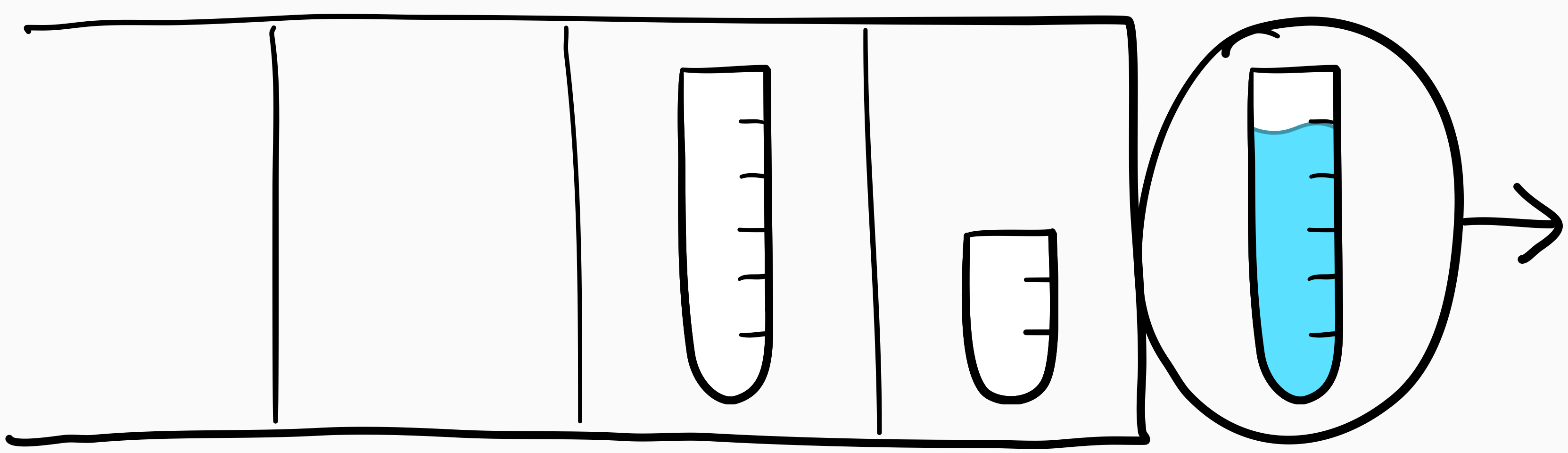
How should we schedule jobs to minimize delay?



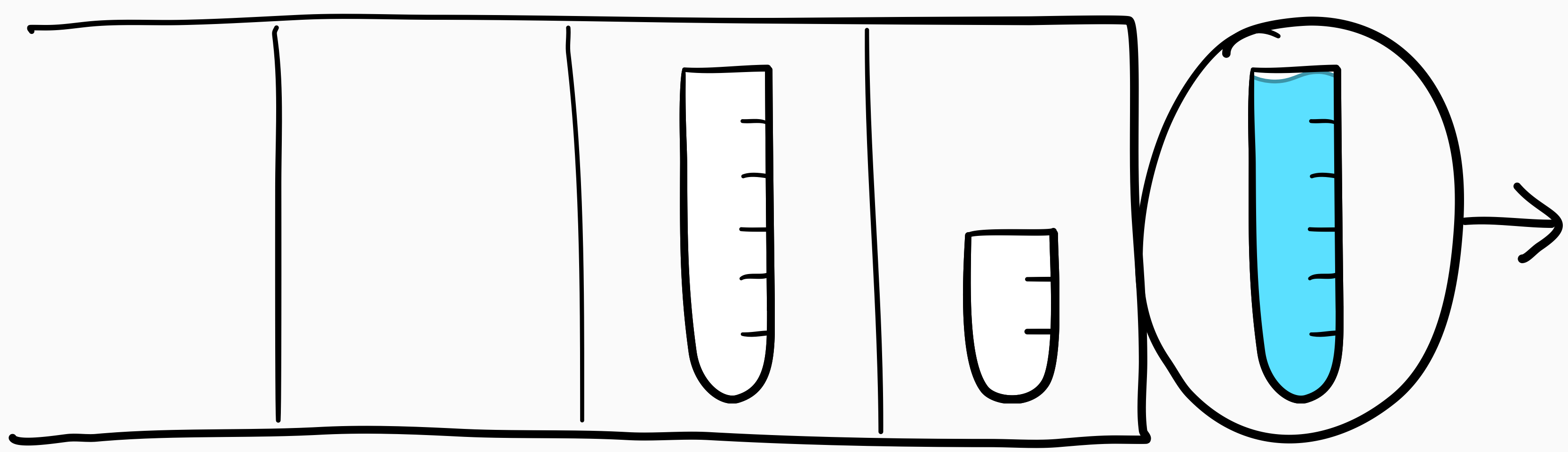
How should we schedule jobs to minimize delay?



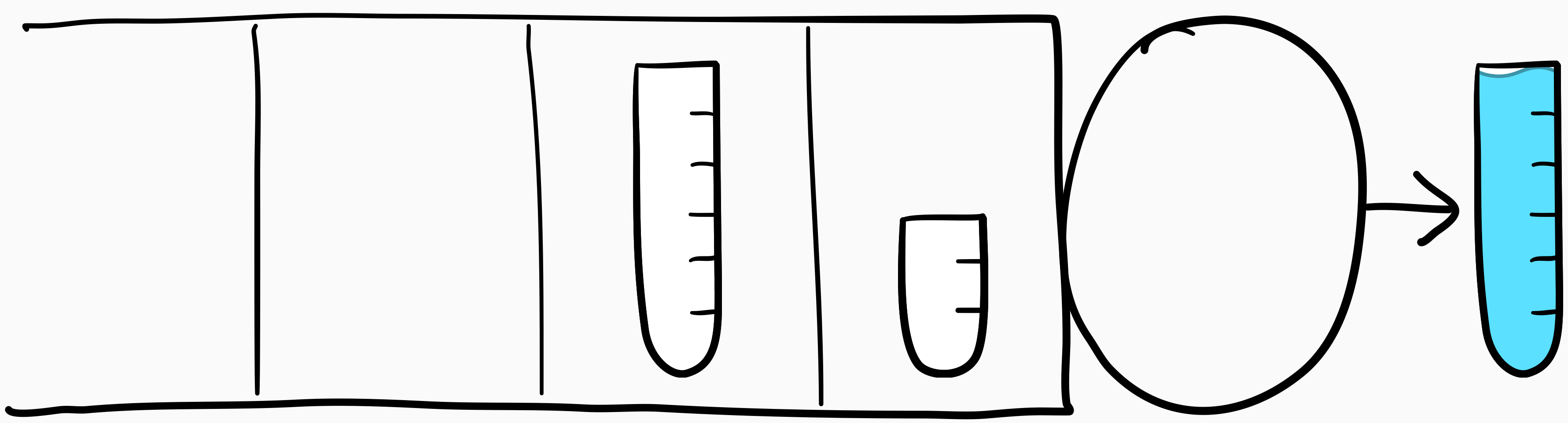
How should we schedule jobs to minimize delay?



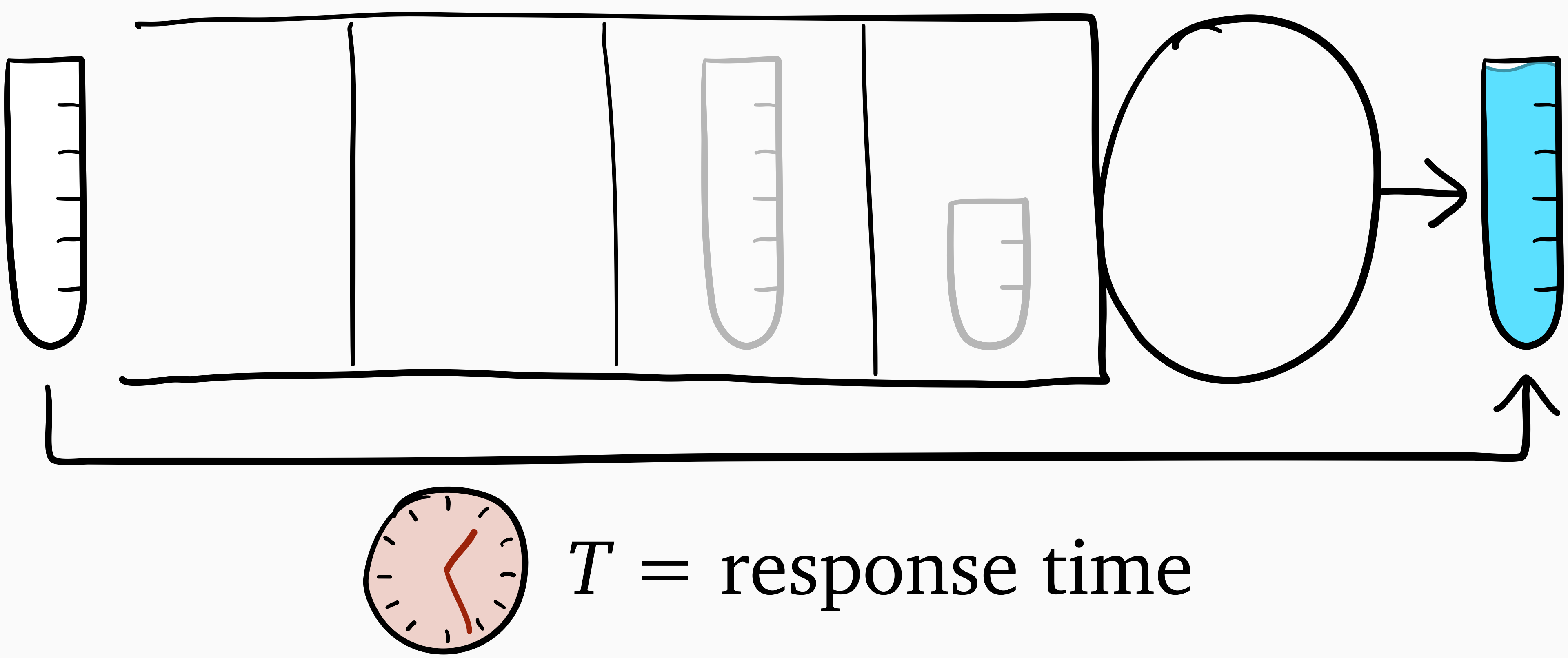
How should we schedule jobs to minimize delay?



How should we schedule jobs to minimize delay?



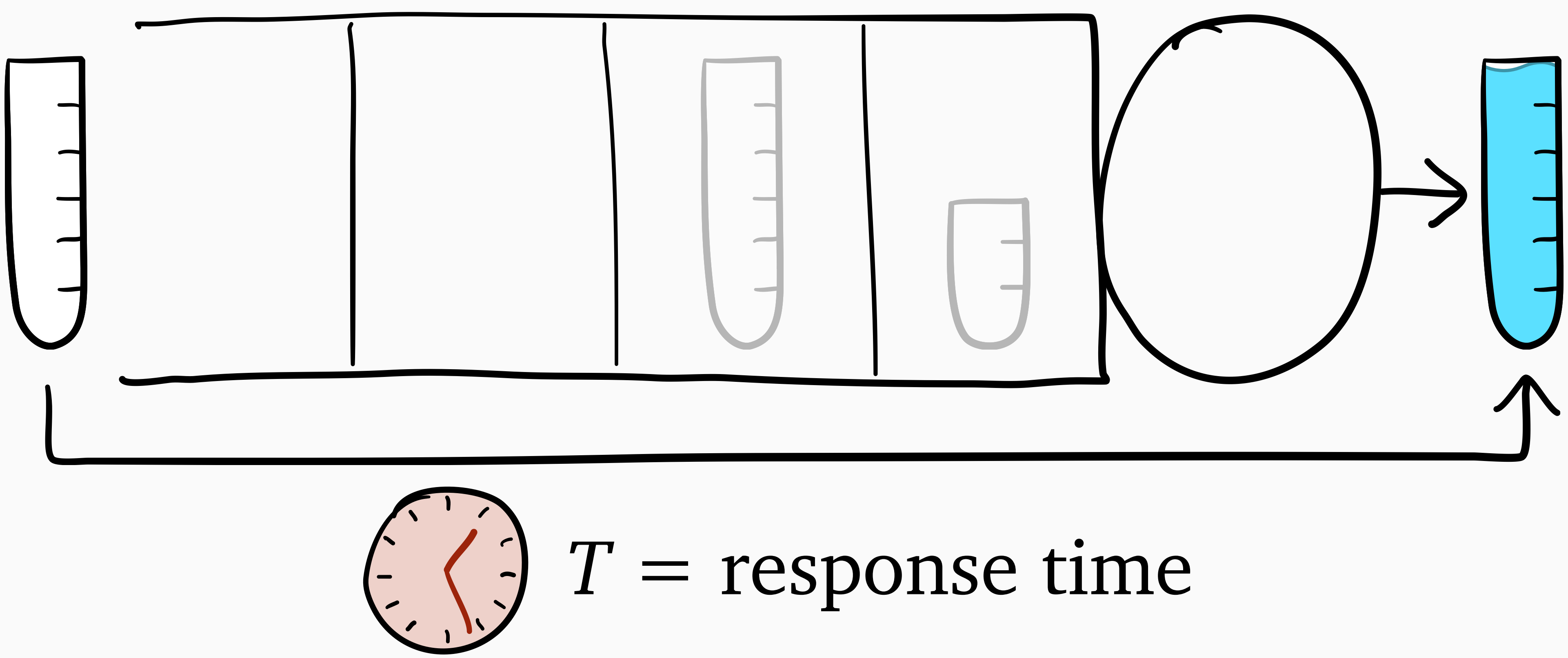
How should we schedule jobs to minimize delay?



T = response time

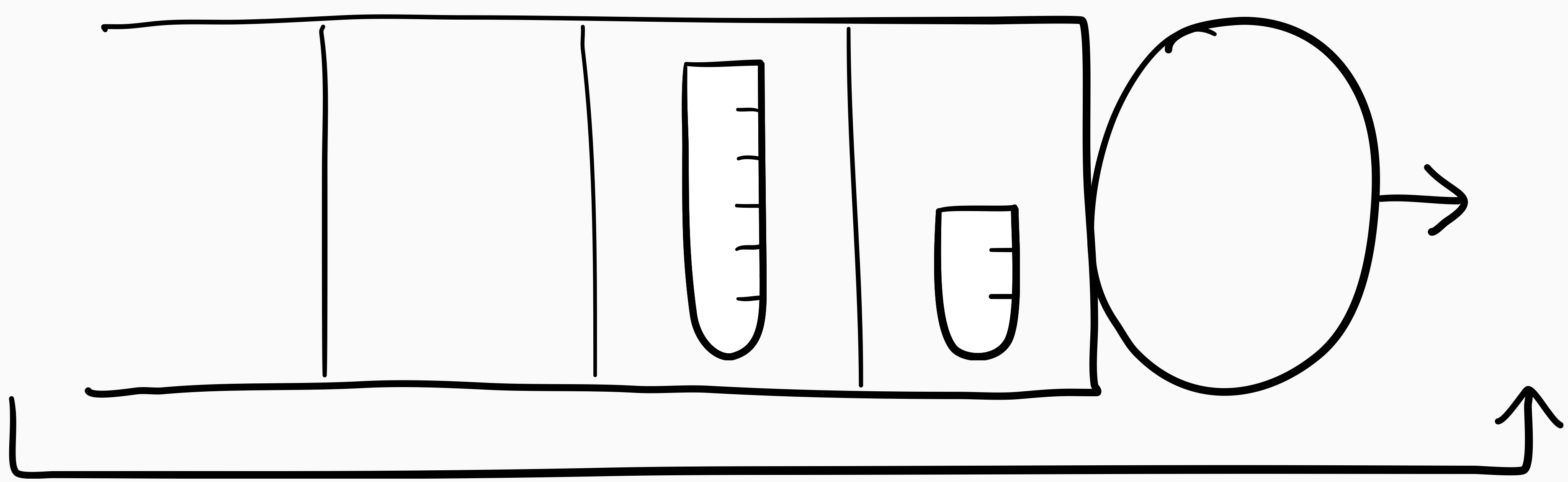
How should we schedule jobs to minimize delay?

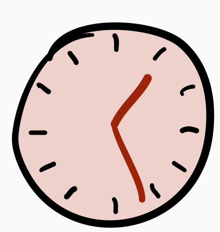
$\mathbf{E}[T]$



How should we schedule jobs to minimize delay?

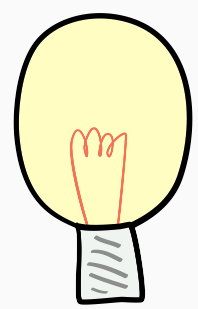
$\mathbf{E}[T]$



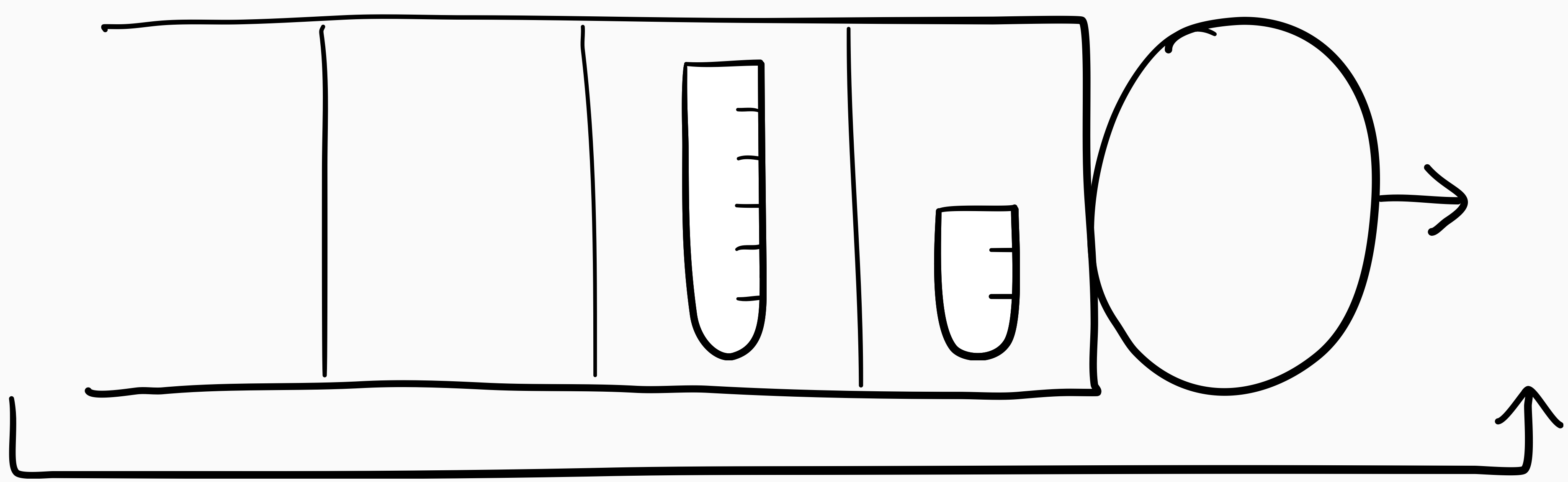
 T = response time

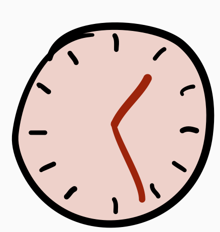
How should we schedule jobs to minimize delay?

$\mathbf{E}[T]$



Serve short jobs before long jobs

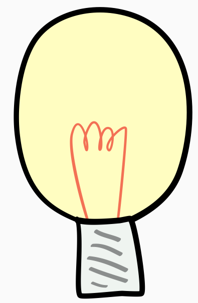




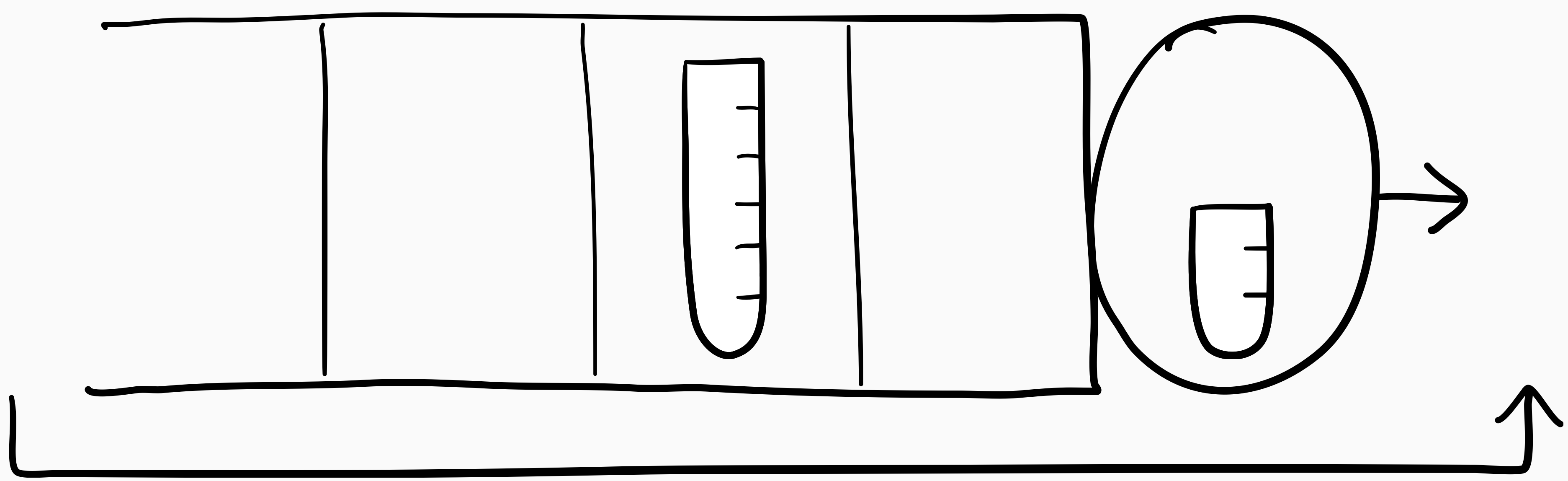
T = response time

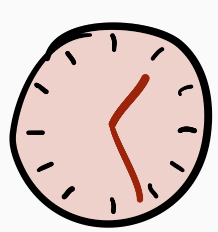
How should we schedule jobs to minimize delay?

$\mathbf{E}[T]$



Serve short jobs before long jobs

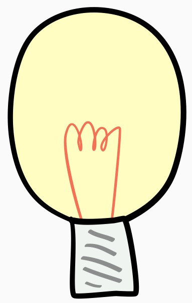




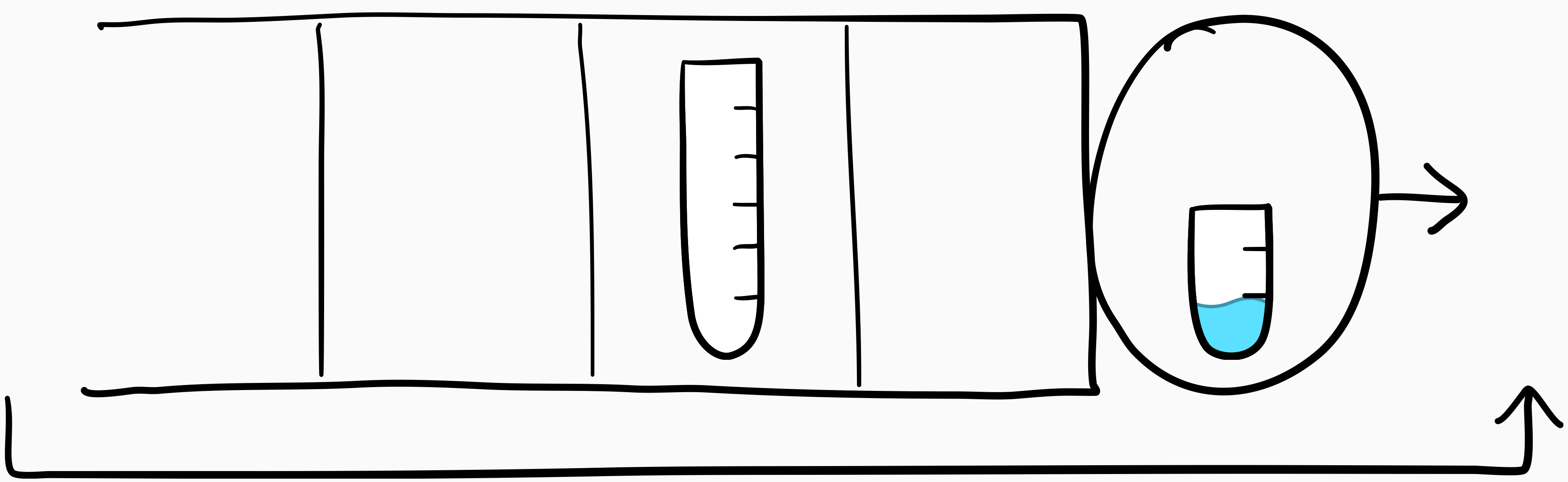
T = response time

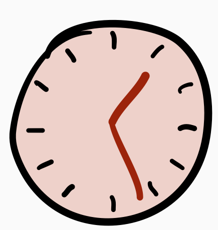
How should we schedule jobs to minimize delay?

$\mathbf{E}[T]$



Serve short jobs
before long jobs

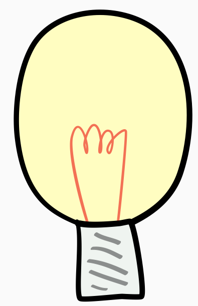




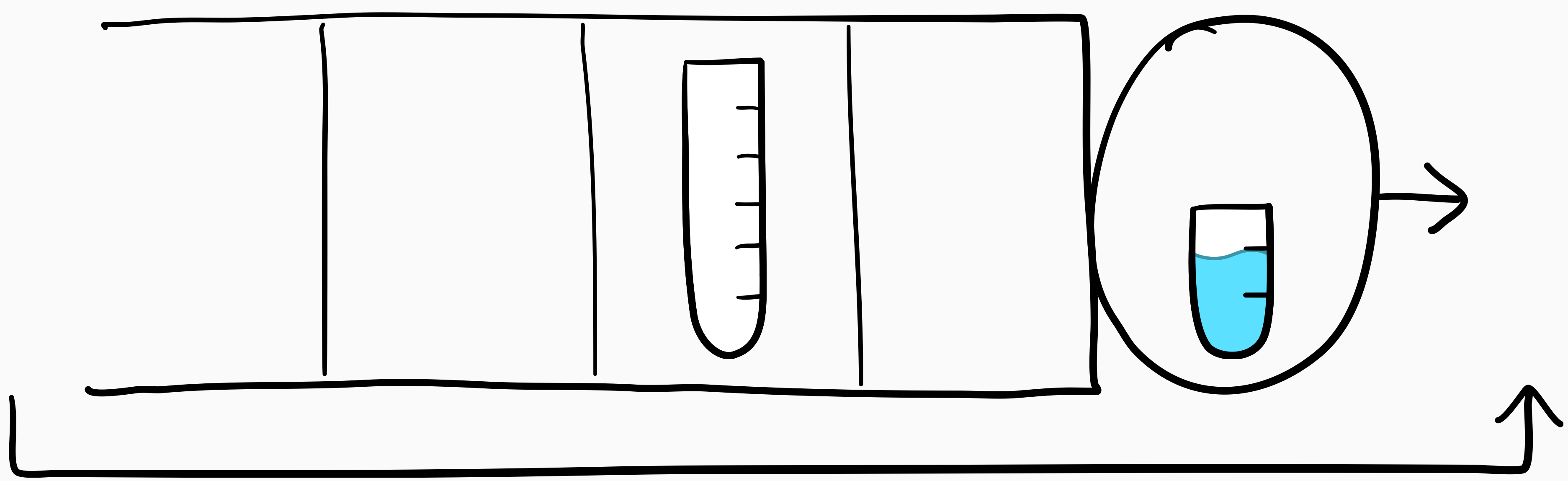
T = response time

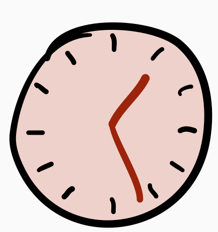
How should we schedule jobs to minimize delay?

$\mathbf{E}[T]$



Serve short jobs before long jobs

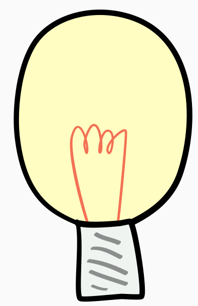




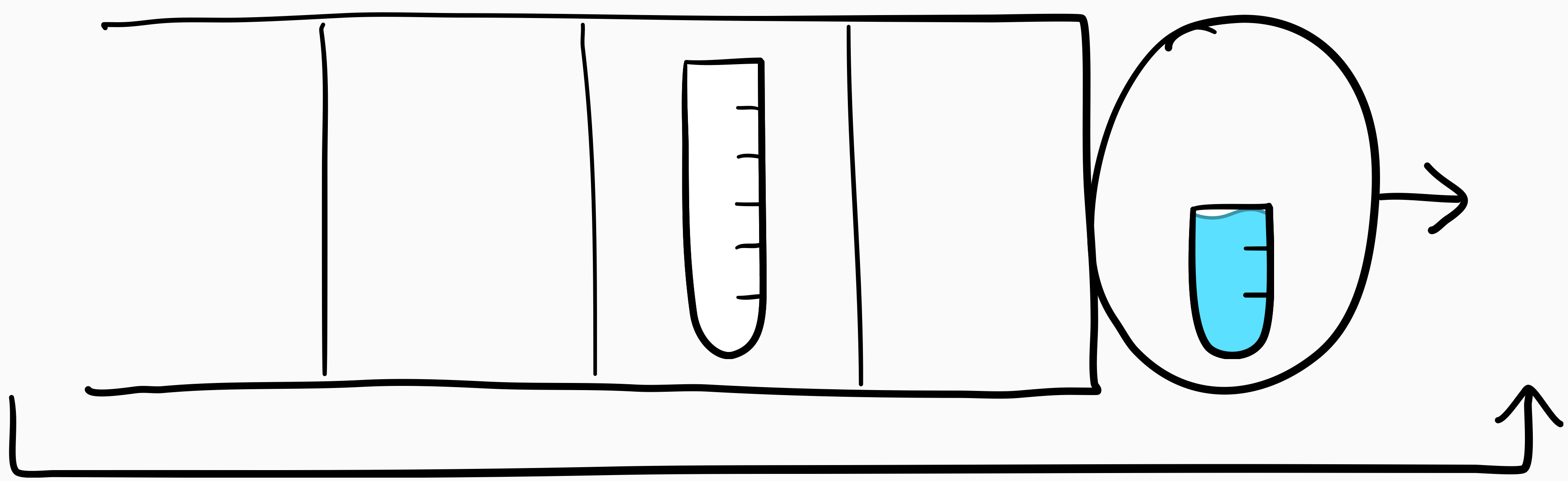
T = response time

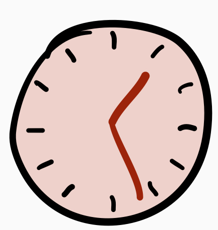
How should we schedule jobs to minimize delay?

$\mathbf{E}[T]$



Serve short jobs before long jobs

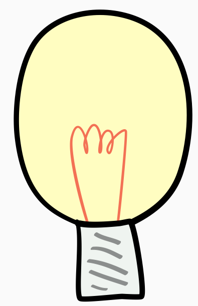




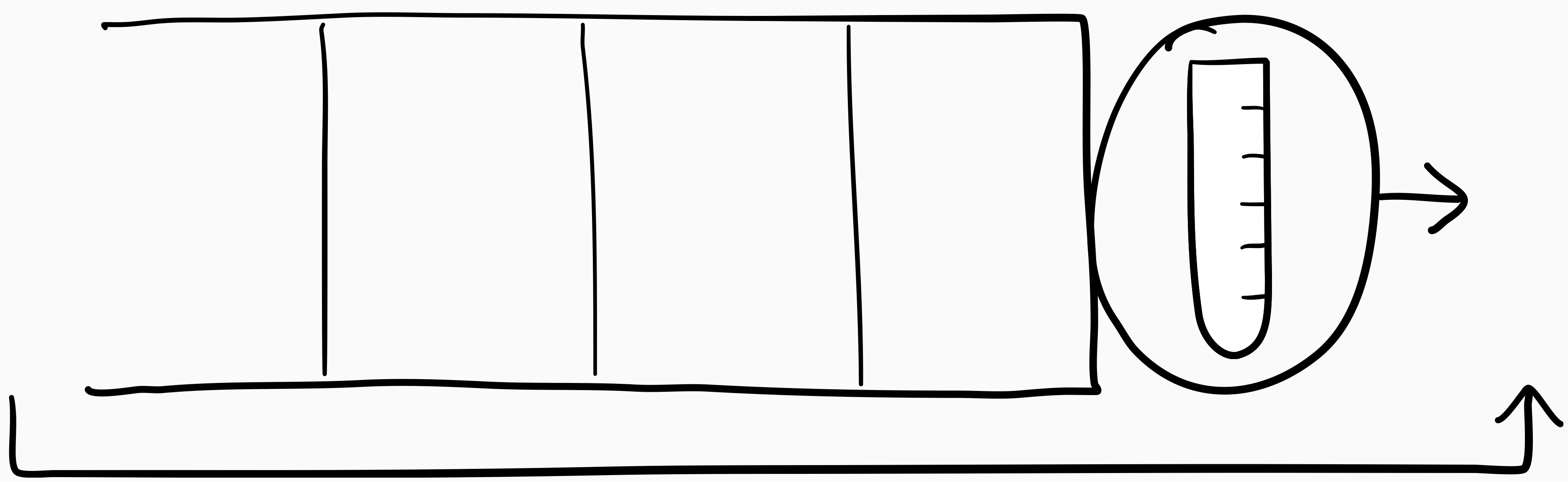
T = response time

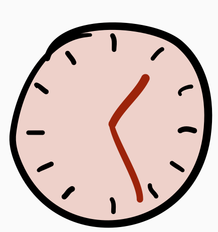
How should we schedule jobs to minimize delay?

$\mathbf{E}[T]$



Serve short jobs
before long jobs

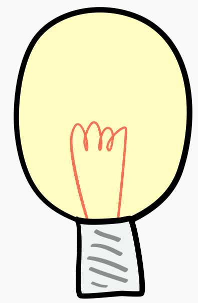




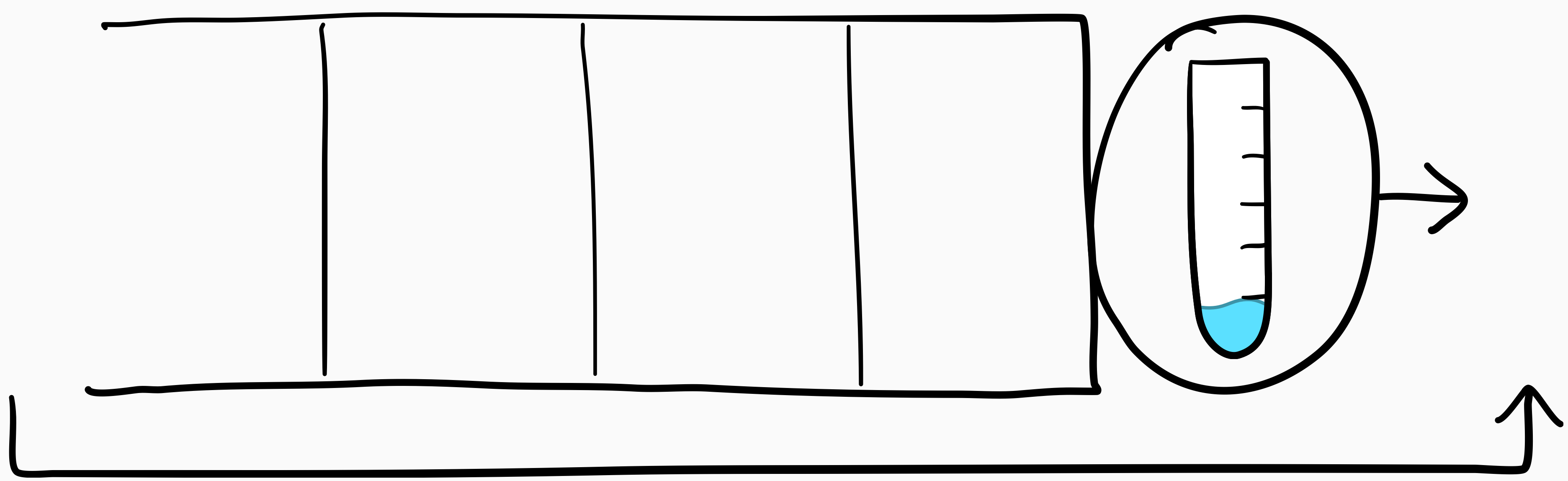
T = response time

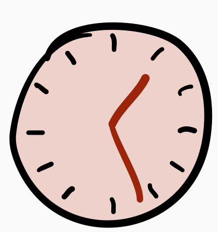
How should we schedule jobs to minimize delay?

$\mathbf{E}[T]$



Serve short jobs before long jobs

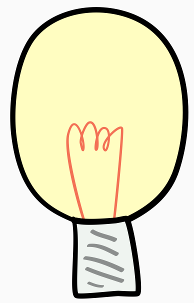




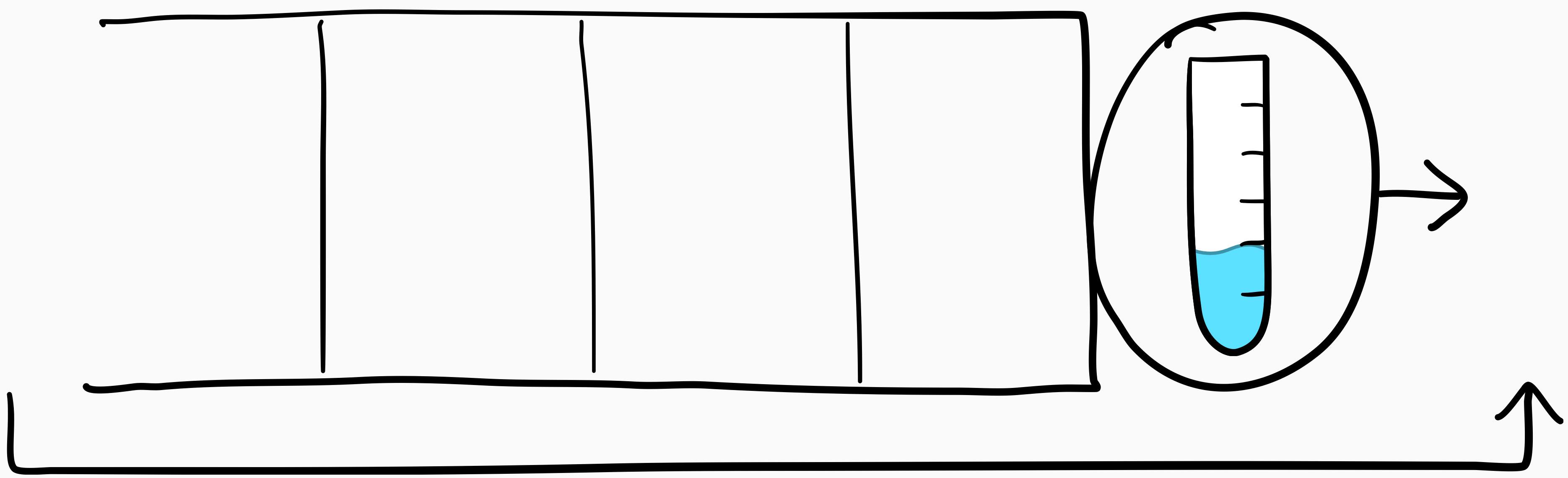
T = response time

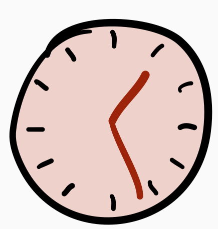
How should we schedule jobs to minimize delay?

$\mathbf{E}[T]$



Serve short jobs
before long jobs

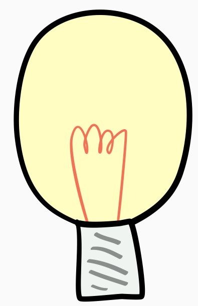




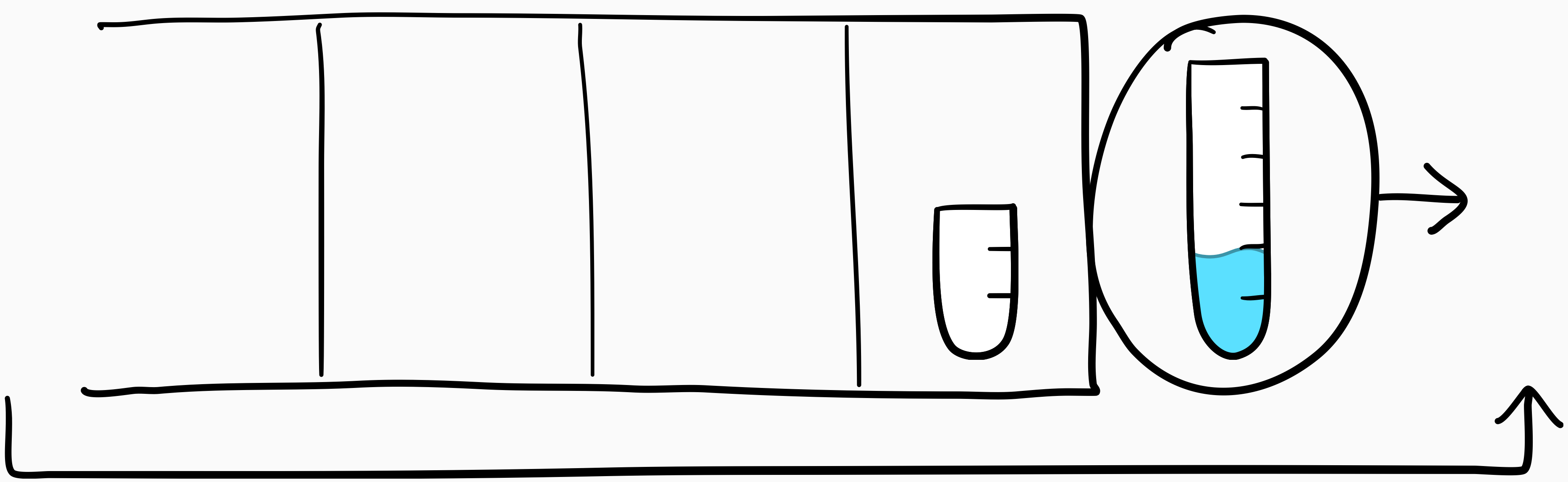
T = response time

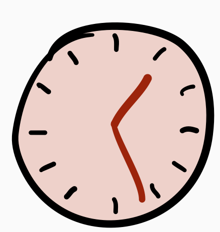
How should we schedule jobs to minimize delay?

$\mathbf{E}[T]$



Serve short jobs
before long jobs

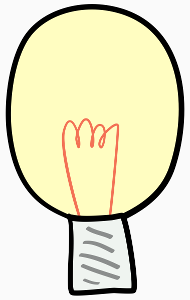




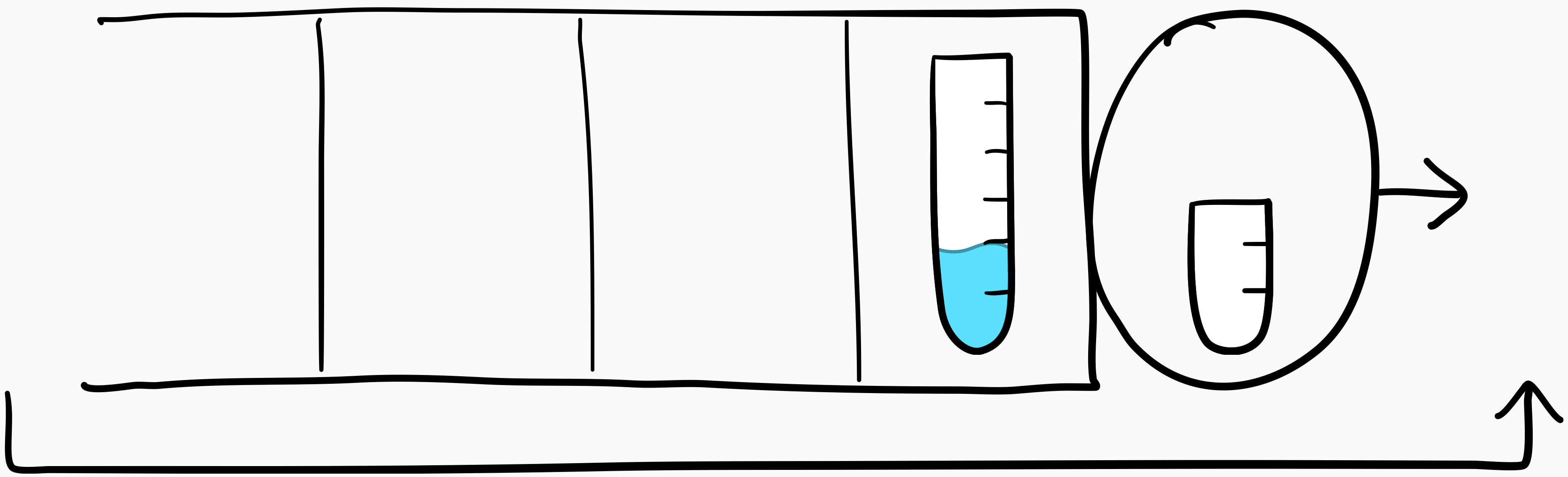
T = response time

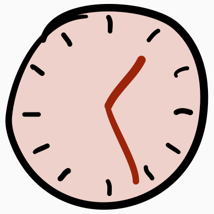
How should we schedule jobs to minimize delay?

$\mathbf{E}[T]$



Serve short jobs before long jobs

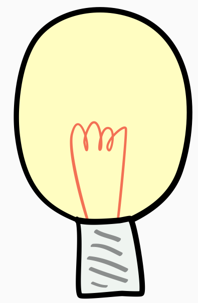




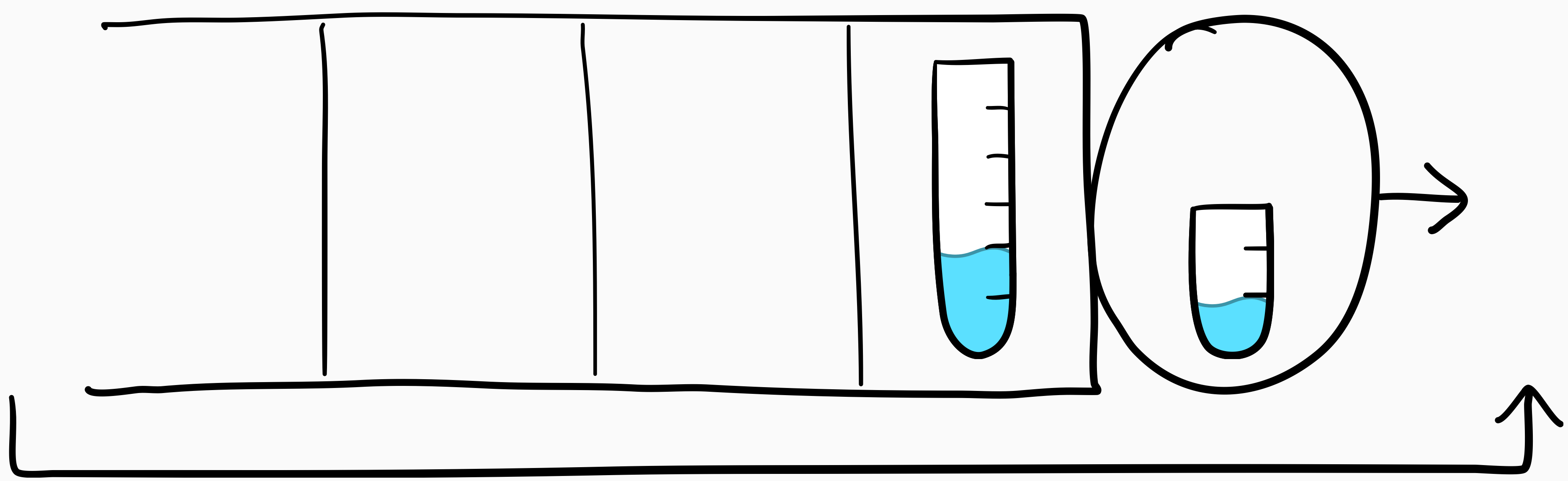
T = response time

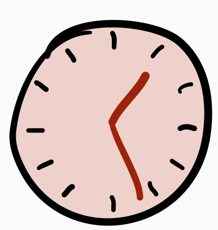
How should we schedule jobs to minimize delay?

$\mathbf{E}[T]$



Serve short jobs before long jobs

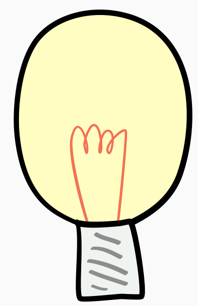




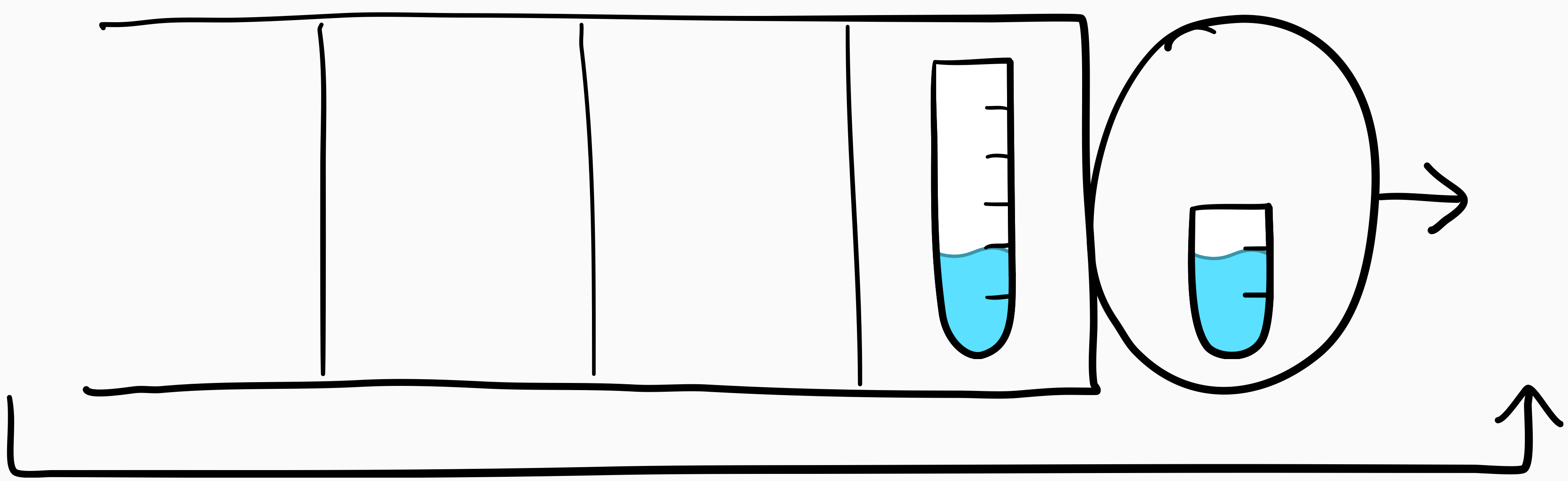
T = response time

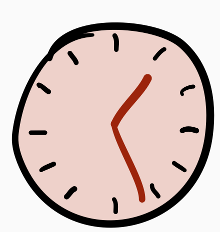
How should we schedule jobs to minimize delay?

$\mathbf{E}[T]$



Serve short jobs before long jobs



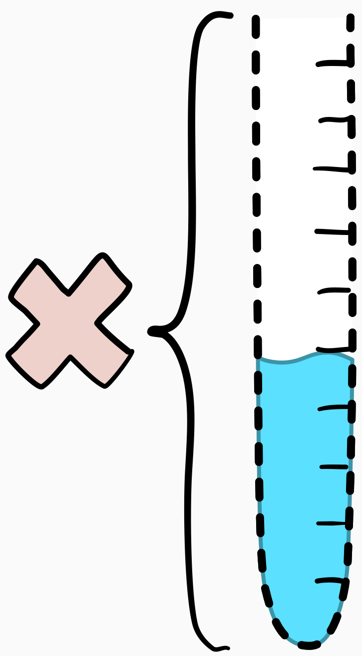


T = response time

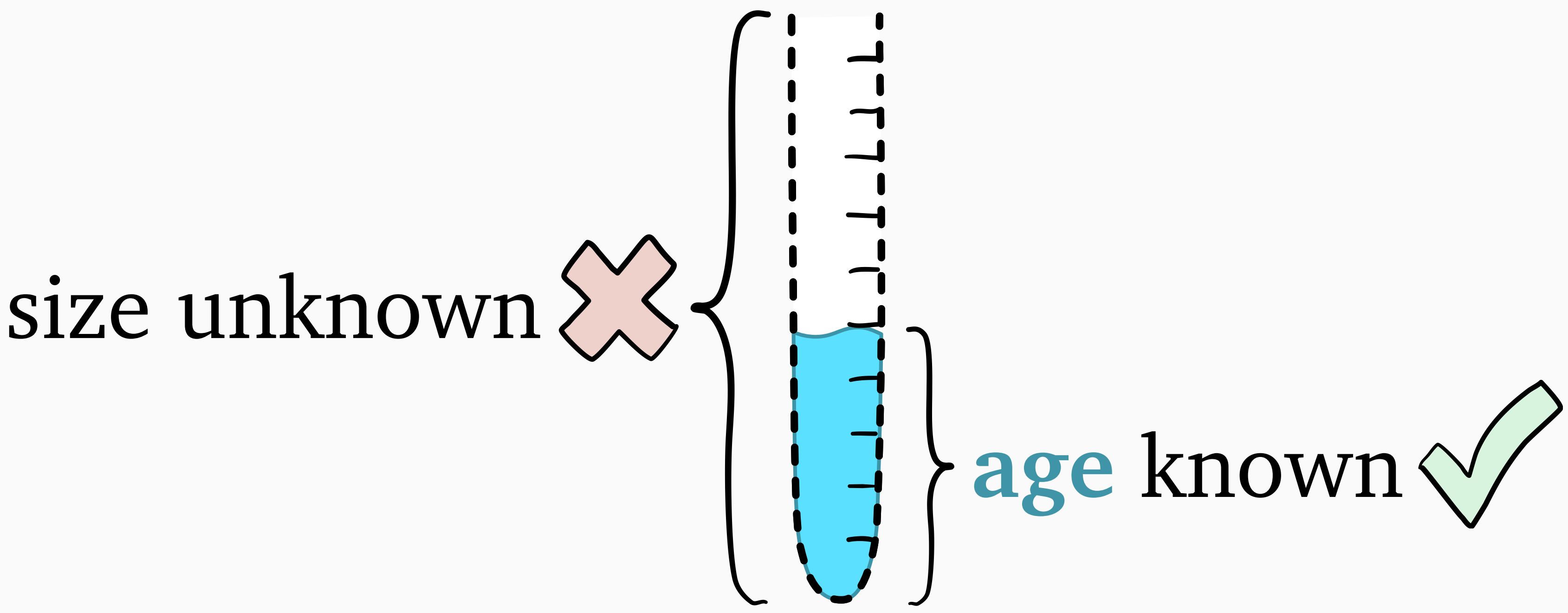
Scheduling with **unknown** sizes

Scheduling with **unknown** sizes

size unknown

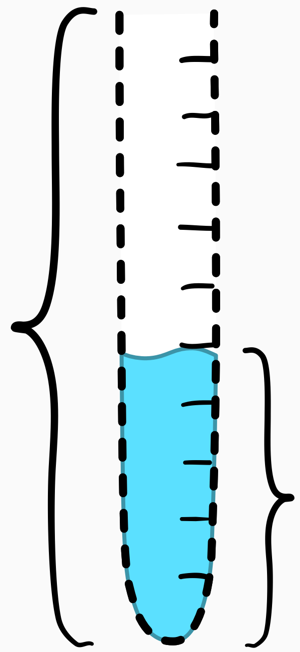


Scheduling with **unknown** sizes

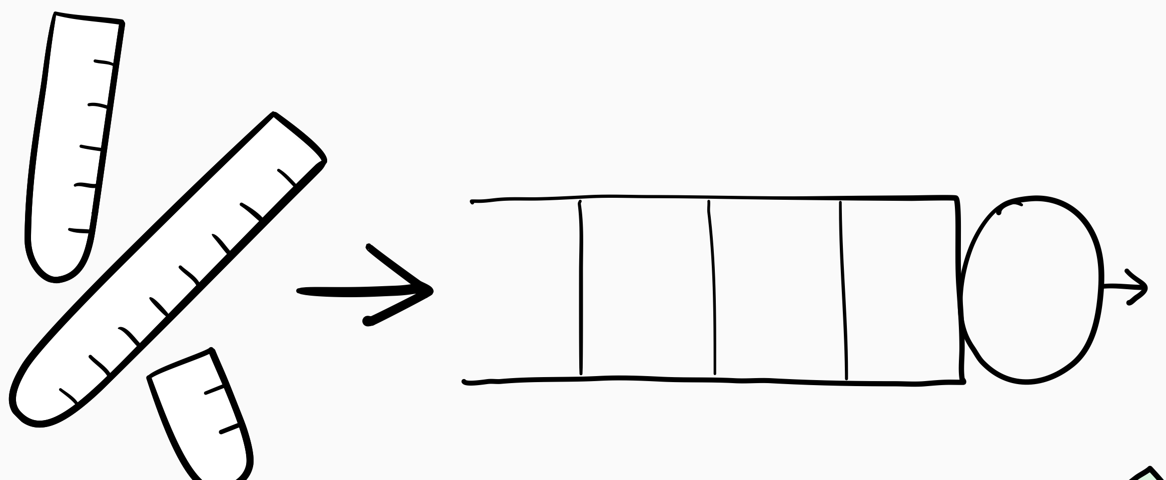


Scheduling with **unknown** sizes

size unknown



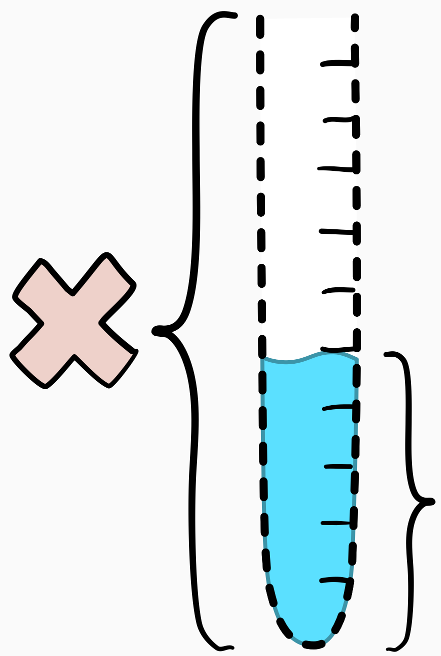
age known



size distribution S known

Scheduling with unknown sizes

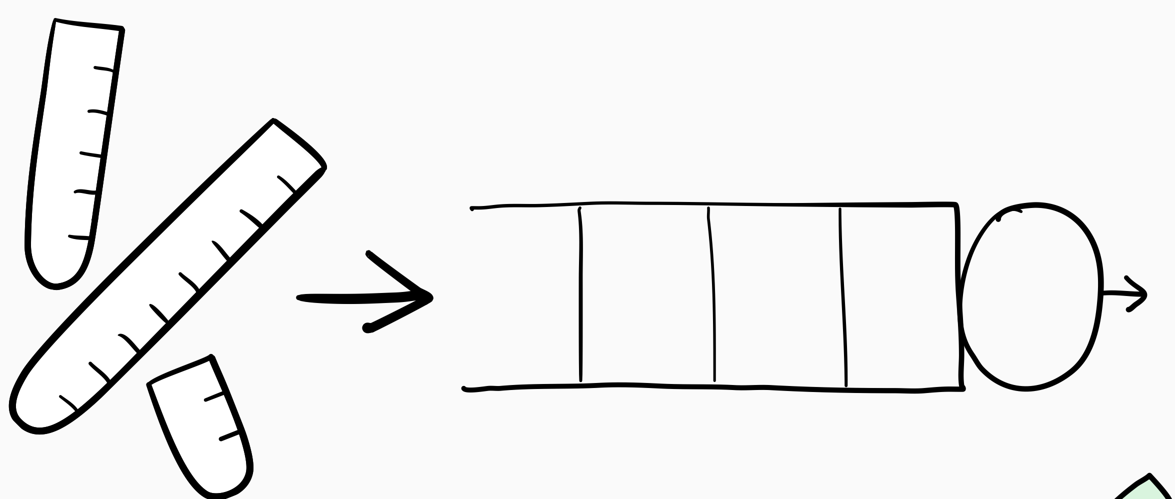
size unknown



age known

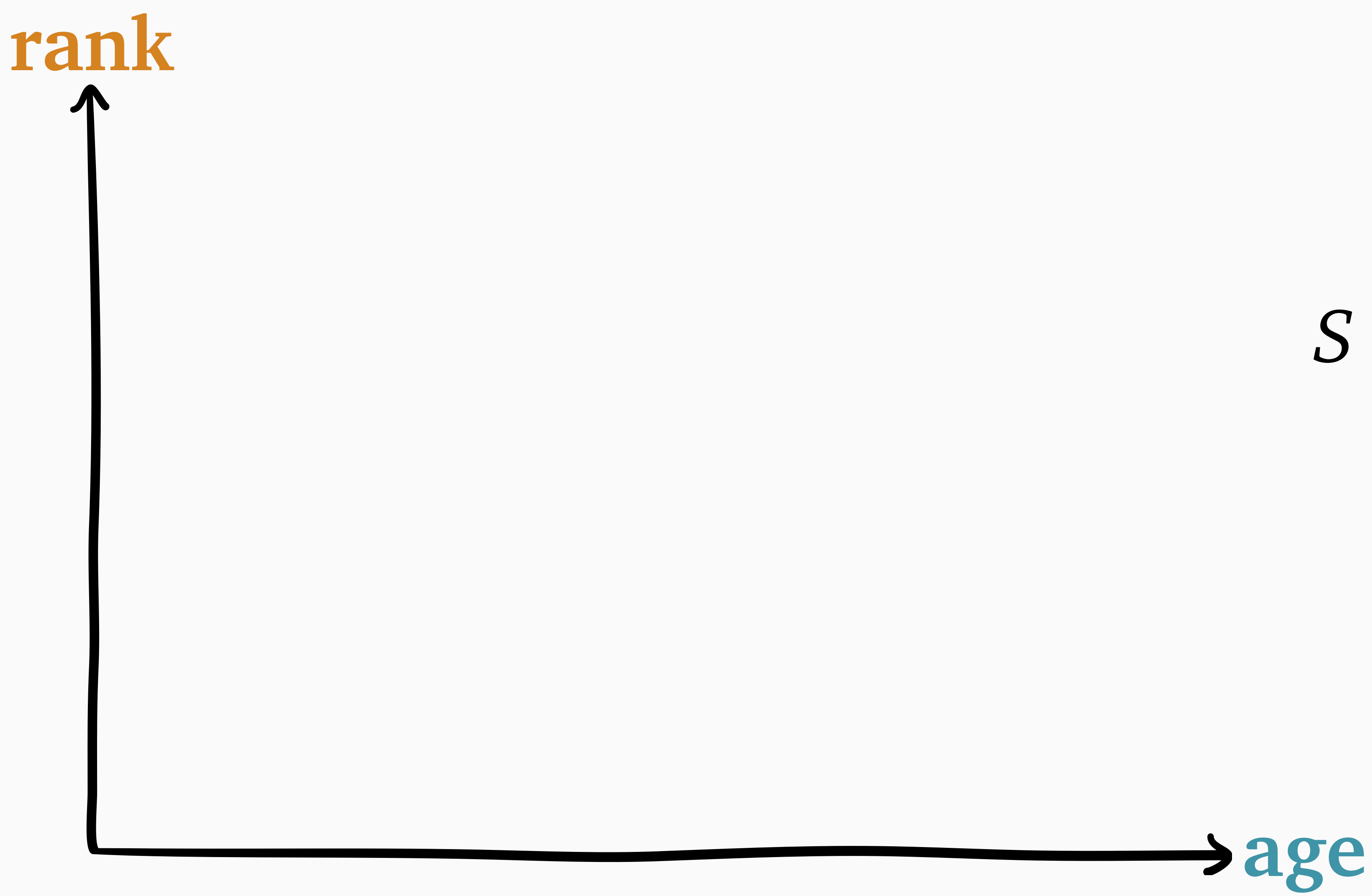
Example:

$$S = \begin{cases} 1 & \text{w.p. } \frac{1}{3} \\ 6 & \text{w.p. } \frac{1}{3} \\ 14 & \text{w.p. } \frac{1}{3} \end{cases}$$



size distribution S known

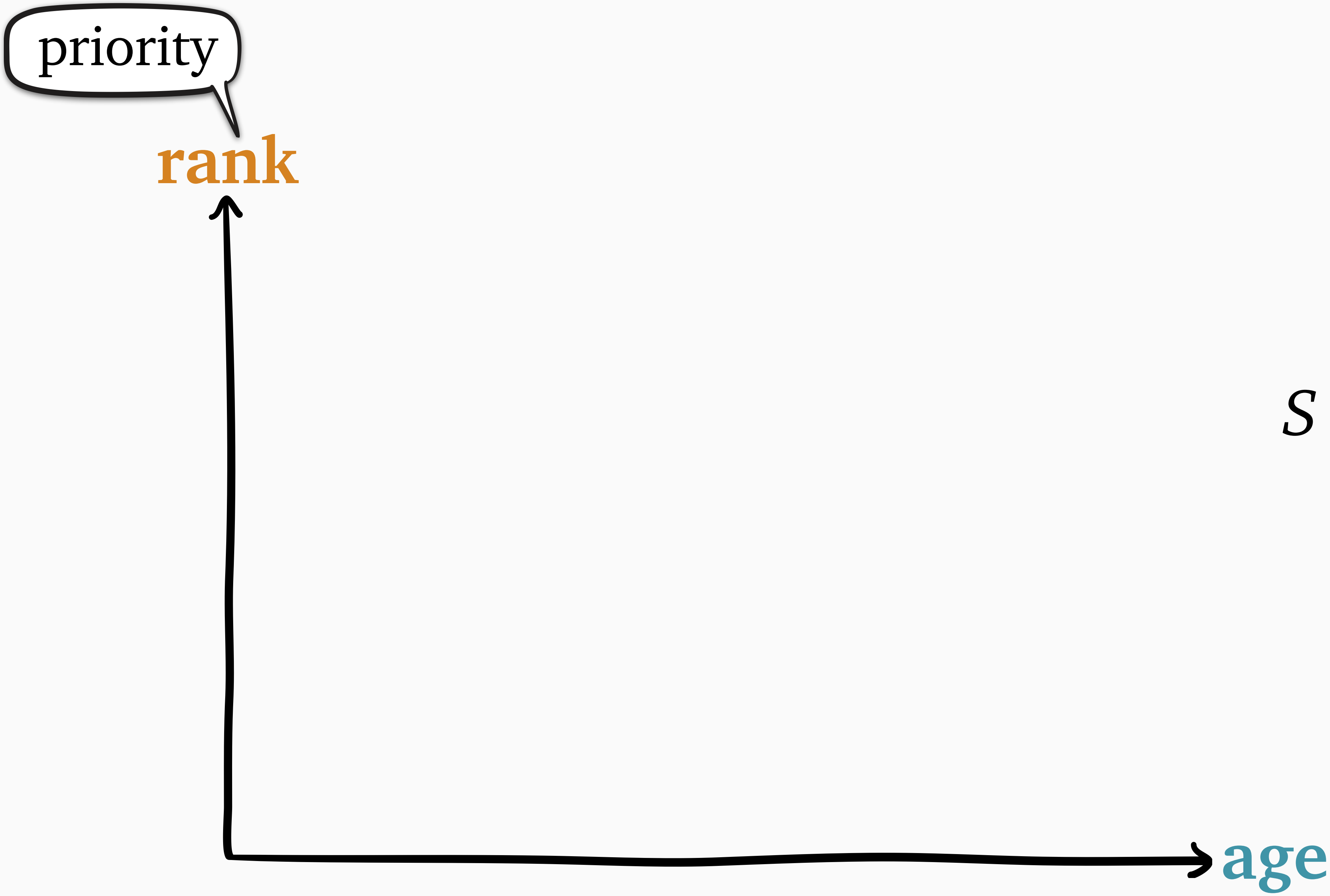
Scheduling with unknown sizes

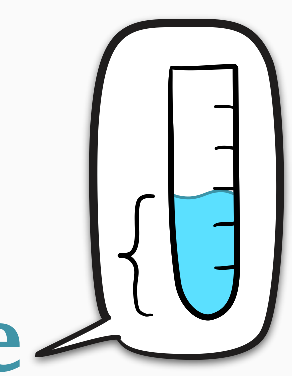


Example:

$$S = \begin{cases} 1 & \text{w.p. } \frac{1}{3} \\ 6 & \text{w.p. } \frac{1}{3} \\ 14 & \text{w.p. } \frac{1}{3} \end{cases}$$

Scheduling with unknown sizes

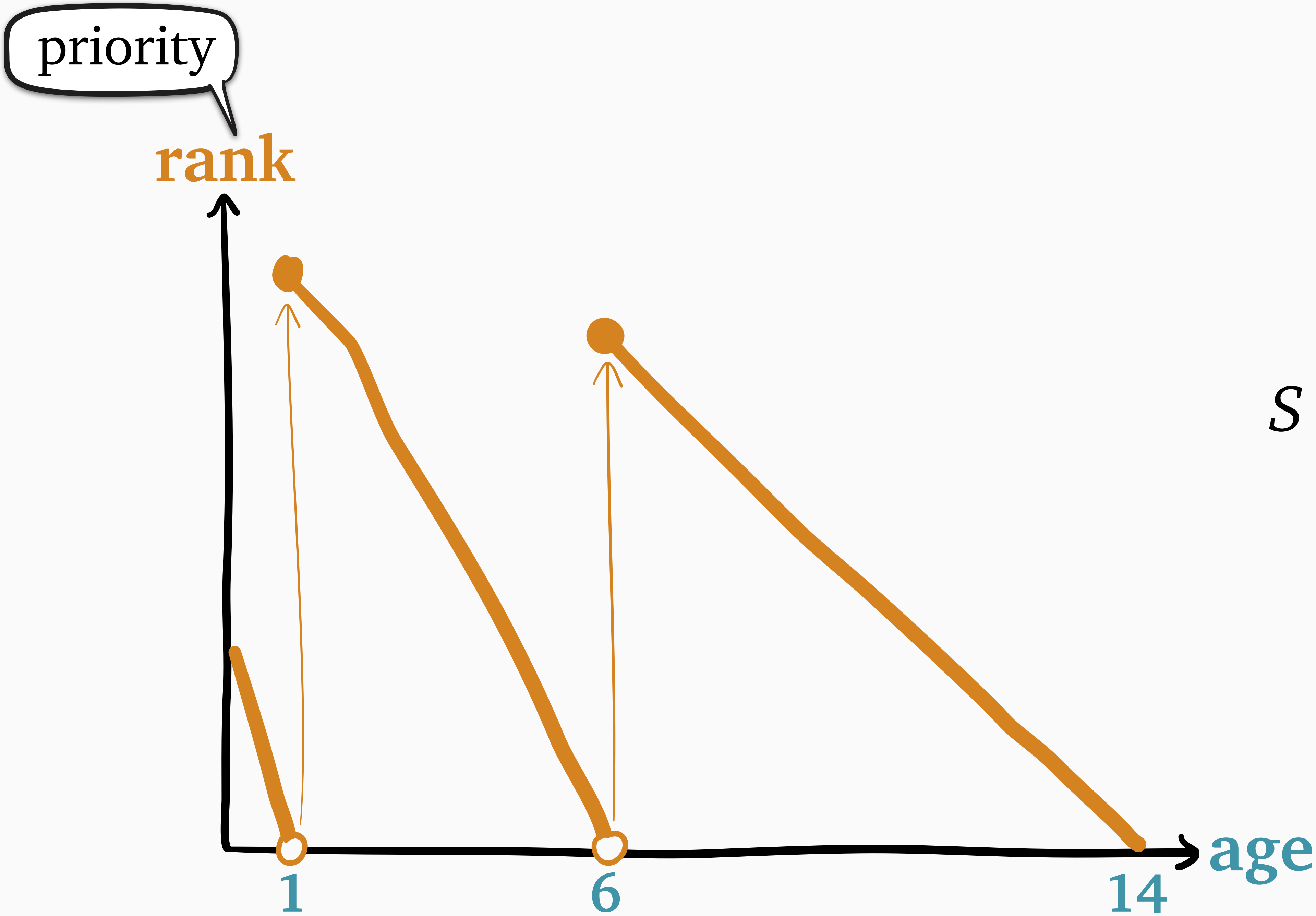




Example:

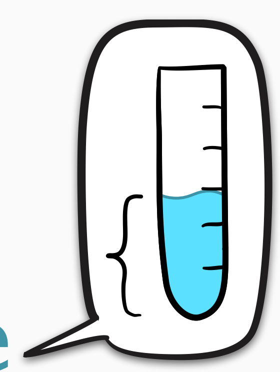
$$S = \begin{cases} 1 & \text{w.p. } \frac{1}{3} \\ 6 & \text{w.p. } \frac{1}{3} \\ 14 & \text{w.p. } \frac{1}{3} \end{cases}$$

Scheduling with **unknown** sizes



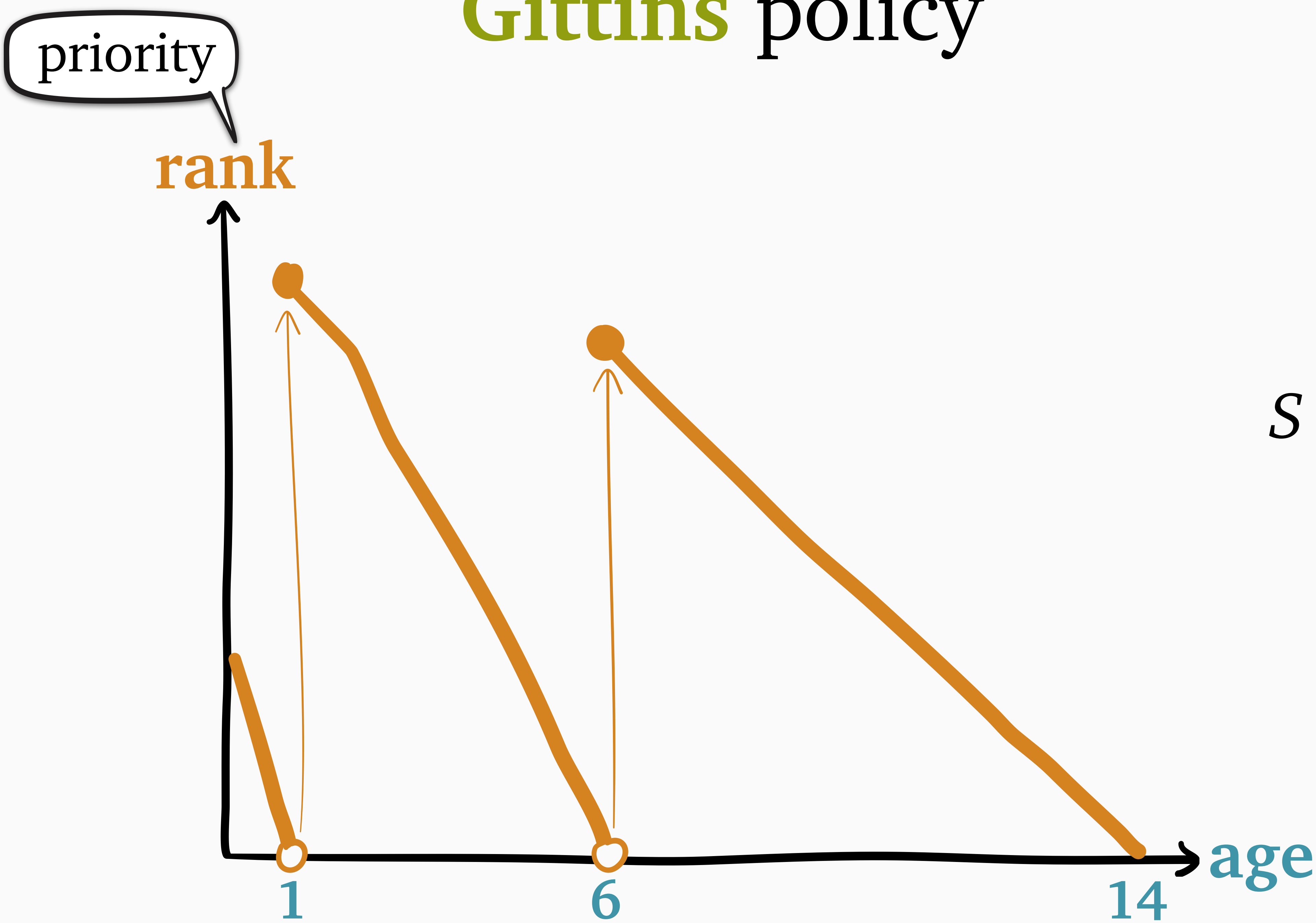
Example:

$$S = \begin{cases} 1 & \text{w.p. } \frac{1}{3} \\ 6 & \text{w.p. } \frac{1}{3} \\ 14 & \text{w.p. } \frac{1}{3} \end{cases}$$



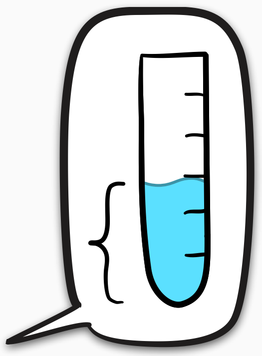
Scheduling with unknown sizes

Gittins policy



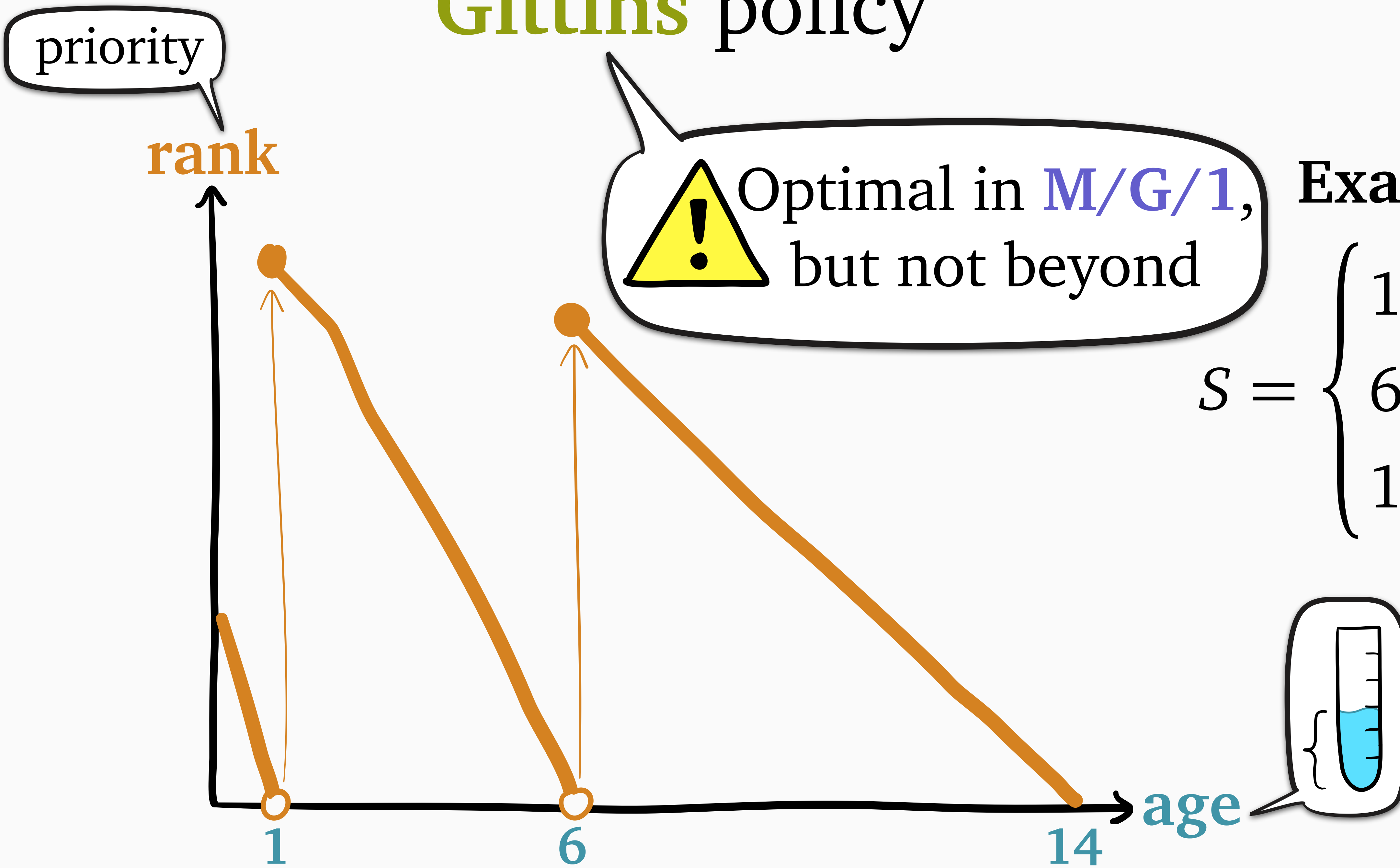
Example:

$$S = \begin{cases} 1 & \text{w.p. } \frac{1}{3} \\ 6 & \text{w.p. } \frac{1}{3} \\ 14 & \text{w.p. } \frac{1}{3} \end{cases}$$



Scheduling with **unknown** sizes

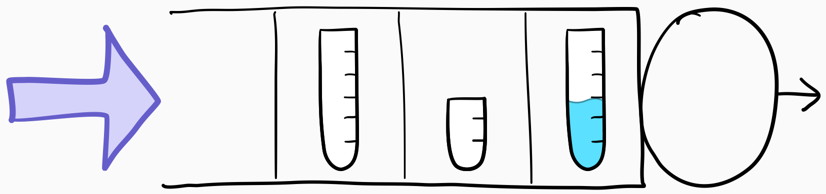
Gittins policy



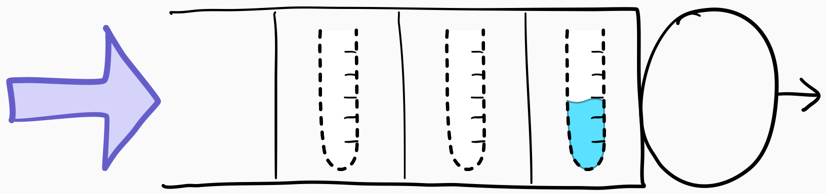
Example:

$$S = \begin{cases} 1 & \text{w.p. } \frac{1}{3} \\ 6 & \text{w.p. } \frac{1}{3} \\ 14 & \text{w.p. } \frac{1}{3} \end{cases}$$

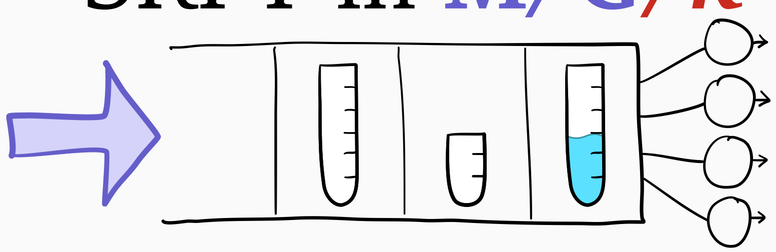
SRPT in M/G/1



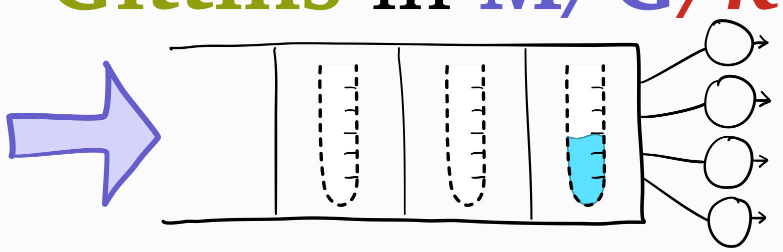
Gittins in M/G/1



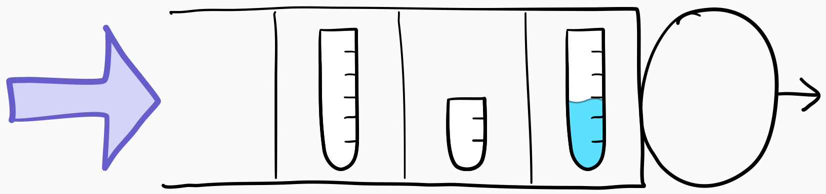
SRPT in M/G/*k*



Gittins in M/G/*k*

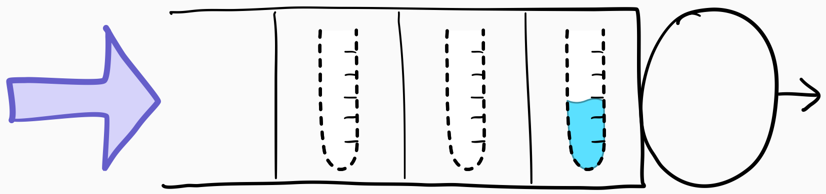


SRPT in M/G/1

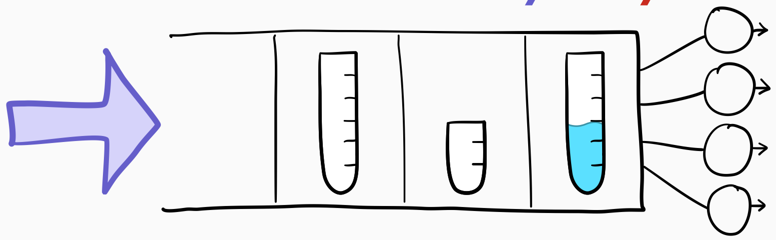


1966: compute $\mathbf{E}[T]$ (Schrage & Miller)

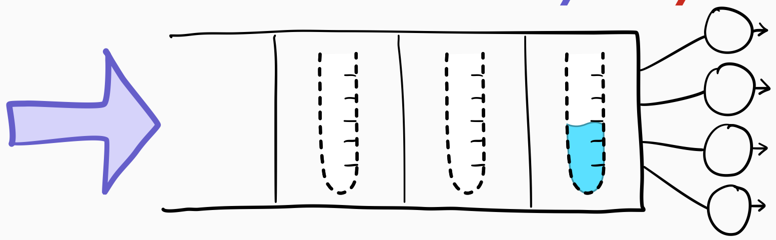
Gittins in M/G/1



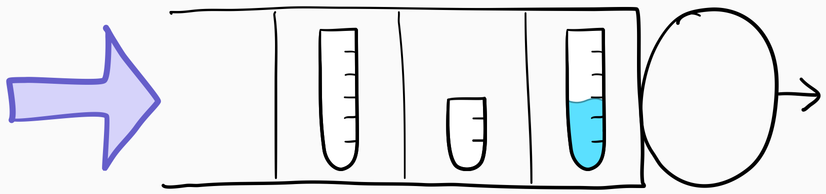
SRPT in M/G/k



Gittins in M/G/k



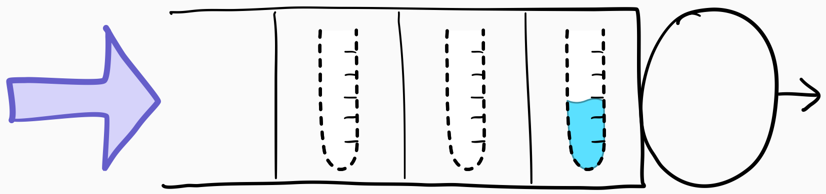
SRPT in M/G/1



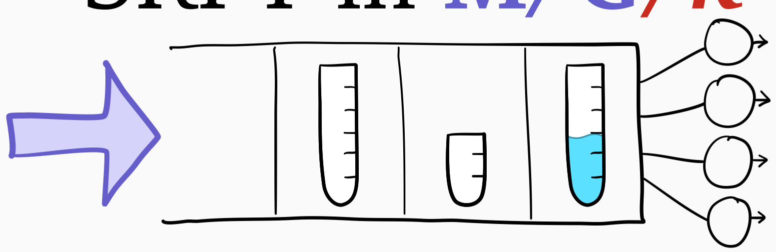
1966: compute $\mathbf{E}[T]$ (Schrage & Miller)

1968: optimal (Schrage)

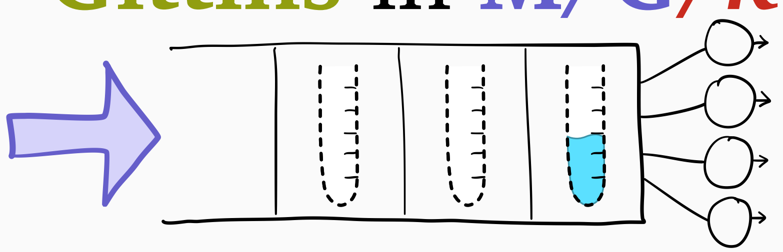
Gittins in M/G/1



SRPT in M/G/k



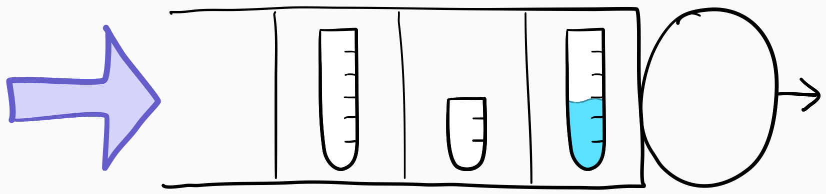
Gittins in M/G/k



SRPT in M/G/1

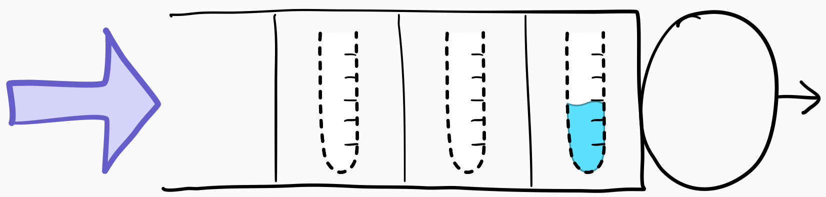
1966: compute $\mathbf{E}[T]$ (Schrage & Miller)

1968: optimal (Schrage)

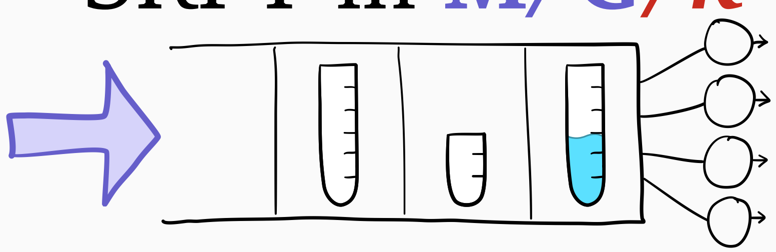


Gittins in M/G/1

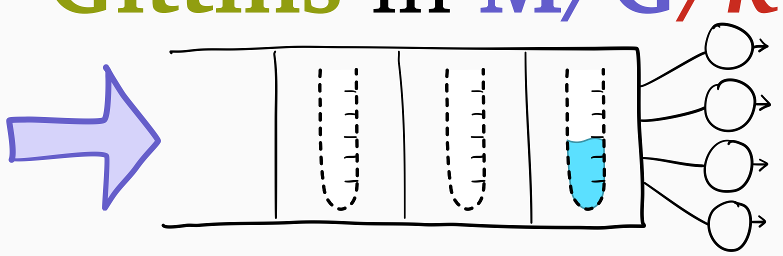
1970s: optimal (Sevcik; von Olivier; Gittins)



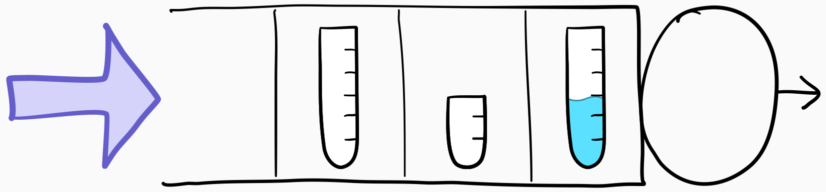
SRPT in M/G/k



Gittins in M/G/k



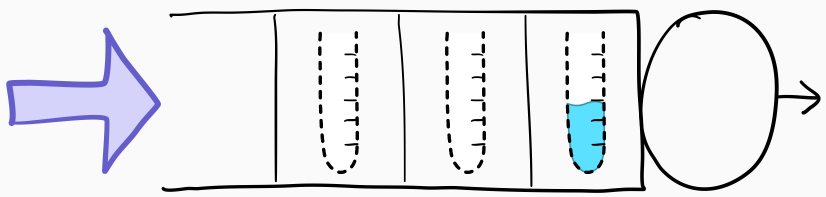
SRPT in M/G/1



1966: compute **E**[T] (Schrage & Miller)

1968: optimal (Schrage)

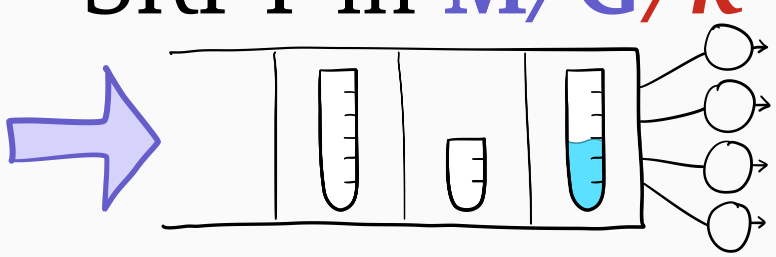
Gittins in M/G/1



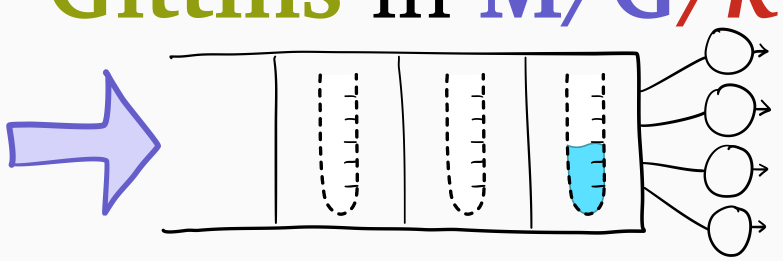
1970s: optimal (Sevcik; von Olivier; Gittins)

2005: **E**[T] in M/M/1 (Whittle)

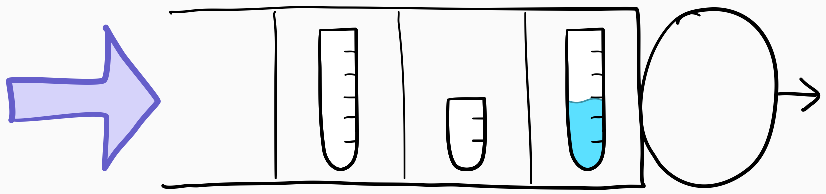
SRPT in M/G/k



Gittins in M/G/k



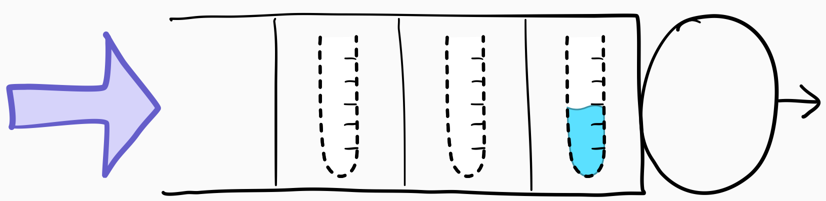
SRPT in M/G/1



1966: compute $\mathbf{E}[T]$ (Schrage & Miller)

1968: optimal (Schrage)

Gittins in M/G/1

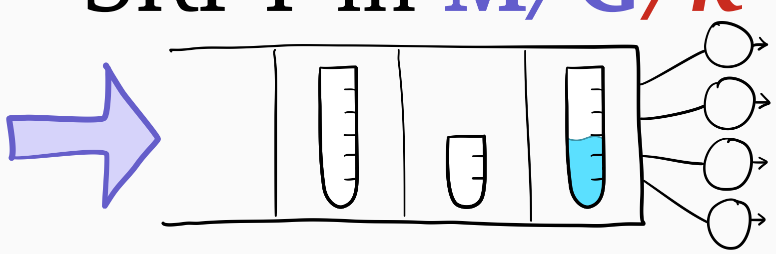


1970s: optimal (Sevcik; von Olivier; Gittins)

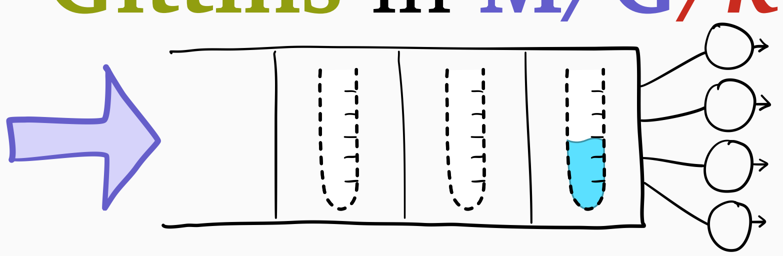
2005: $\mathbf{E}[T]$ in M/M/1 (Whittle)

2018: $\mathbf{E}[T]$ in M/G/1 (Scully et al.)

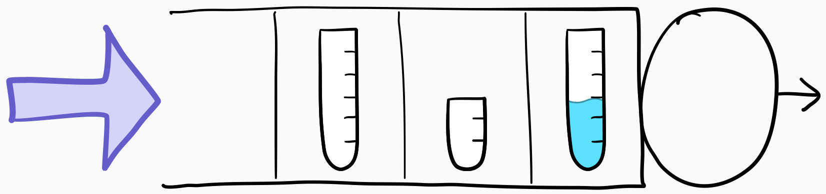
SRPT in M/G/k



Gittins in M/G/k



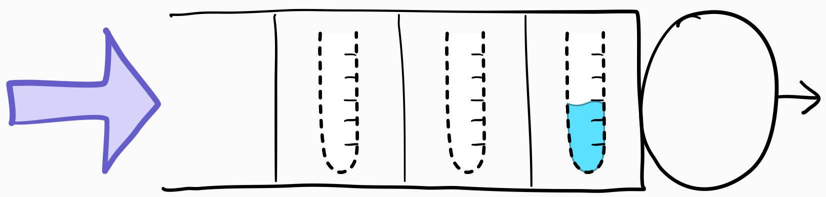
SRTP in M/G/1



1966: compute $\mathbf{E}[T]$ (Schrage & Miller)

1968: optimal (Schrage)

Gittins in M/G/1

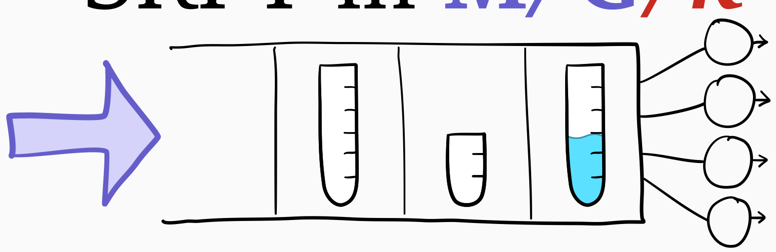


1970s: optimal (Sevcik; von Olivier; Gittins)

2005: $\mathbf{E}[T]$ in M/M/1 (Whittle)

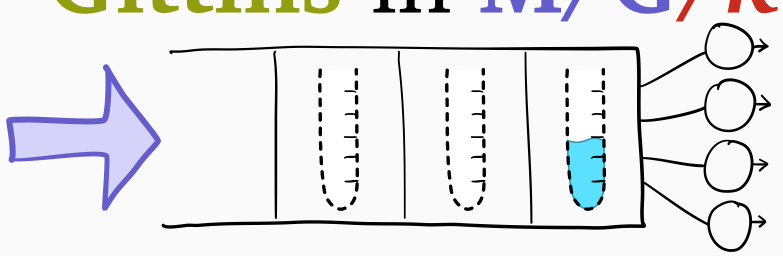
2018: $\mathbf{E}[T]$ in M/G/1 (Scully et al.)

SRPT in M/G/*k*

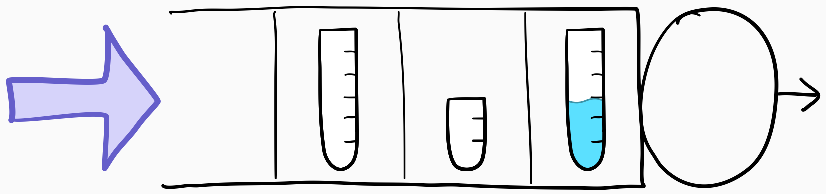


2018: $\mathbf{E}[T]$ bound, "near-optimal" (Grosof et al.)

Gittins in M/G/*k*



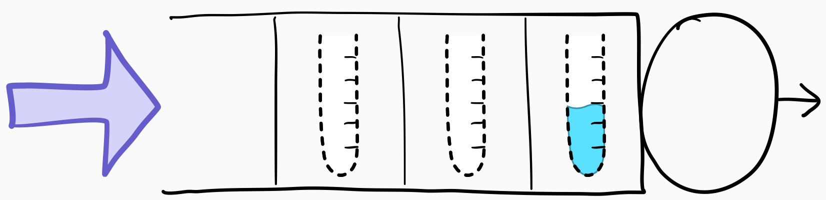
SRPT in M/G/1



1966: compute **E**[T] (Schrage & Miller)

1968: optimal (Schrage)

Gittins in M/G/1

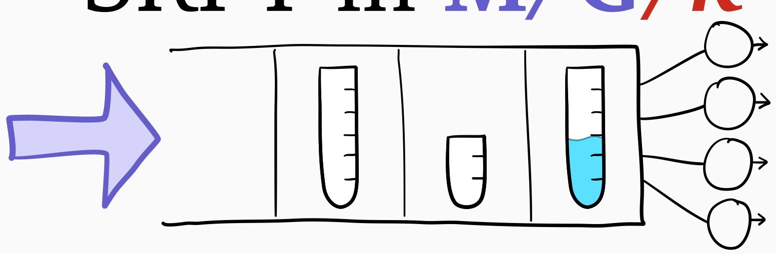


1970s: optimal (Sevcik; von Olivier; Gittins)

2005: **E**[T] in M/M/1 (Whittle)

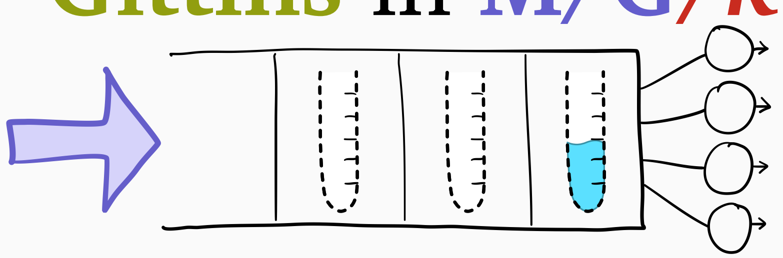
2018: **E**[T] in M/G/1 (Scully et al.)

SRPT in M/G/k



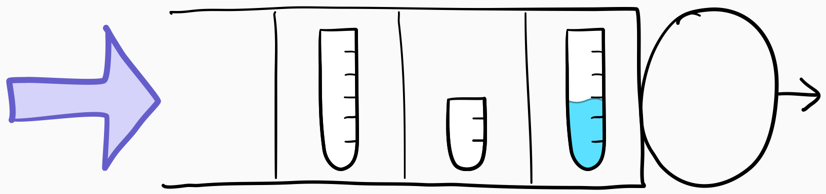
2018: **E**[T] bound, “near-optimal” (Grosof et al.)

Gittins in M/G/k



2001: M/M/k (Glazebrook & Niño-Mora)

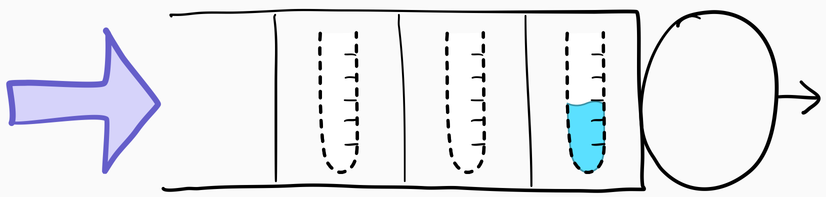
SRPT in M/G/1



1966: compute $\mathbf{E}[T]$ (Schrage & Miller)

1968: optimal (Schrage)

Gittins in M/G/1

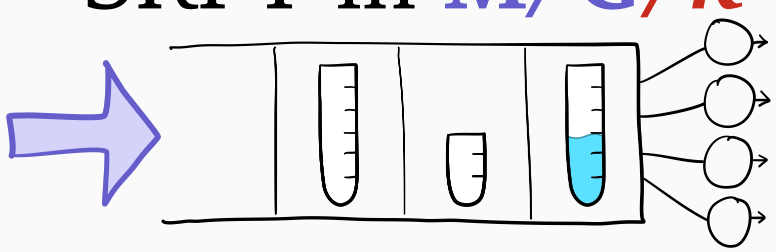


1970s: optimal (Sevcik; von Olivier; Gittins)

2005: $\mathbf{E}[T]$ in M/M/1 (Whittle)

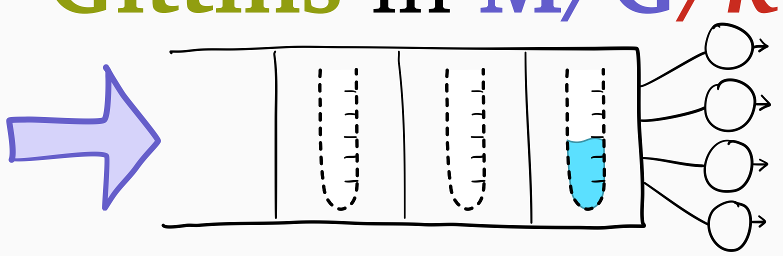
2018: $\mathbf{E}[T]$ in M/G/1 (Scully et al.)

SRPT in M/G/k



2018: $\mathbf{E}[T]$ bound, “near-optimal” (Grosof et al.)

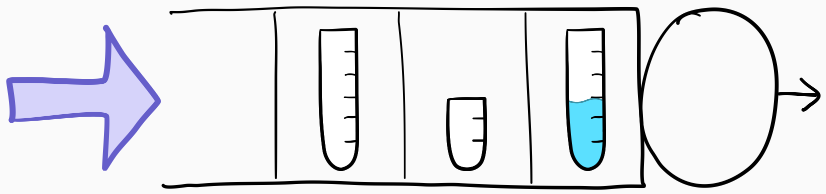
Gittins in M/G/k



2001: M/M/k (Glazebrook & Niño-Mora)

2003: nonpreemptive M/G/k (Glazebrook)

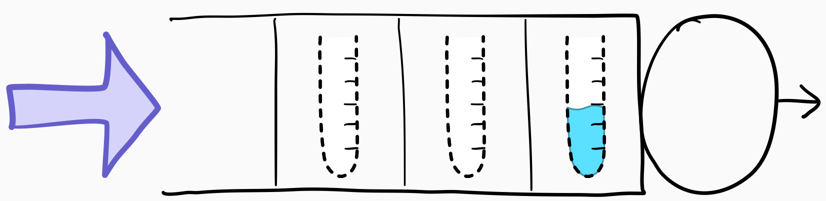
SRPT in M/G/1



1966: compute **E**[T] (Schrage & Miller)

1968: optimal (Schrage)

Gittins in M/G/1

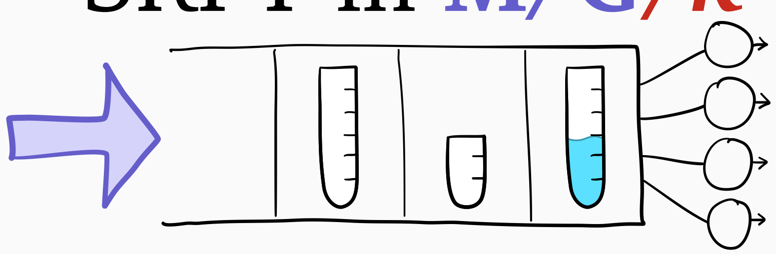


1970s: optimal (Sevcik; von Olivier; Gittins)

2005: **E**[T] in M/M/1 (Whittle)

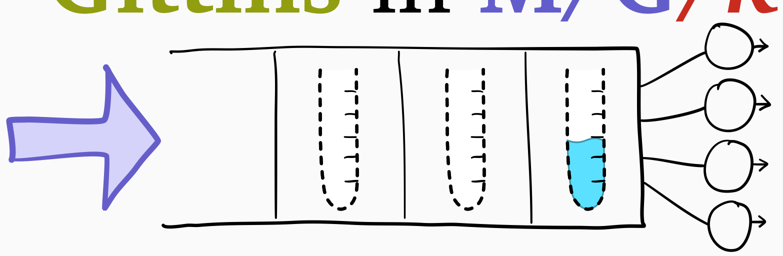
2018: **E**[T] in M/G/1 (Scully et al.)

SRPT in M/G/k



2018: **E**[T] bound, “near-optimal” (Grosof et al.)

Gittins in M/G/k

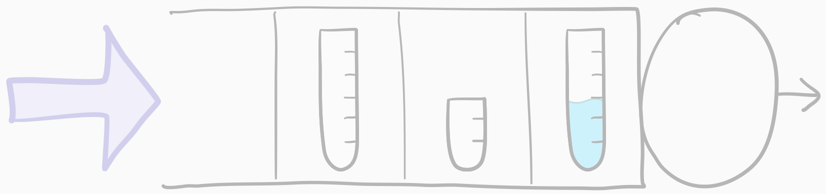


2001: M/M/k (Glazebrook & Niño-Mora)

2003: nonpreemptive M/G/k (Glazebrook)

2020: preemptive M/G/k (Scully et al.)

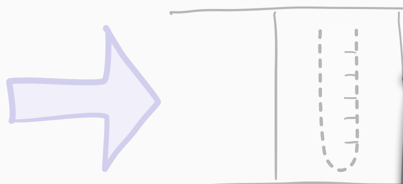
SRPT in M/G/1



1966: compute $\mathbf{E}[T]$ (Schrage & Miller)

1968: optimal (Schrage)

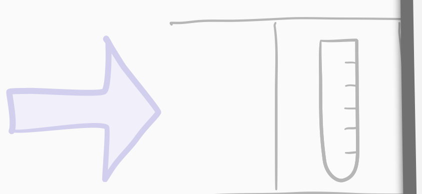
Gittins in M/G/1



1970s: optimal (Sevcik; von Olivier; Gittins)

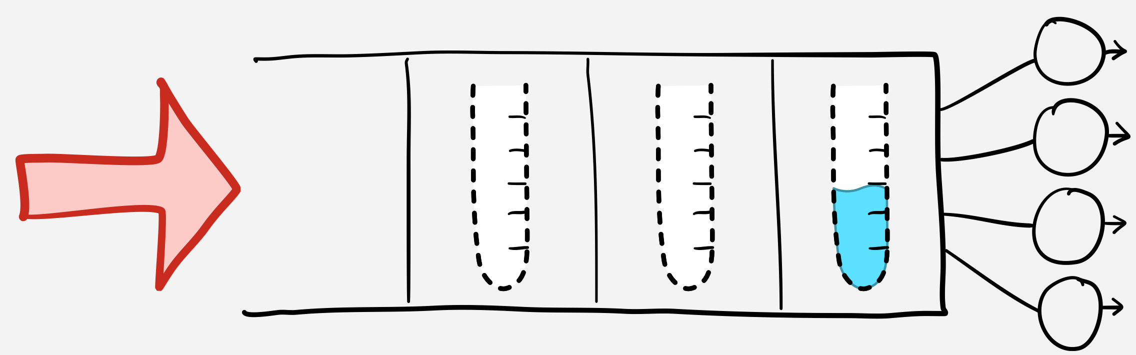
al.)

SRPT in

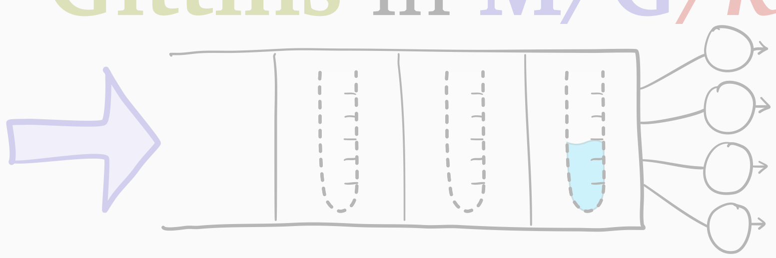


nal” (Grosof et al.)

Our work: G/G arrivals



Gittins in M/G/k



2001: M/M/k (Glazebrook & Niño-Mora)

2003: nonpreemptive M/G/k (Glazebrook)

2020: preemptive M/G/k (Scully et al.)

1966: compute $\mathbf{E}[T]$ (Schrage & Miller)

1968: optimal (Schrage)

1970s: optimal (Sevcik; von Olivier; Gittins)

2005: $\mathbf{E}[T]$ in M/M/1 (Whittle)

2018: $\mathbf{E}[T]$ in M/G/1 (Scully et al.)

2018: $\mathbf{E}[T]$ bound, "near-optimal" (Grosof et al.)

2001: M/M/k (Glazebrook & Niño-Mora)

2003: nonpreemptive M/G/k (Glazebrook)

2020: preemptive M/G/k (Scully et al.)

Strategy 0
just optimality

1966: compute $\mathbf{E}[T]$ (Schrage & Miller)

1968: optimal (Schrage)

1970s: optimal (Sevcik; von Olivier; Gittins)

2005: $\mathbf{E}[T]$ in M/M/1 (Whittle)

2018: $\mathbf{E}[T]$ in M/G/1 (Scully et al.)

2018: $\mathbf{E}[T]$ bound, “near-optimal” (Grosof et al.)

2001: M/M/k (Glazebrook & Niño-Mora)

2003: nonpreemptive M/G/k (Glazebrook)

2020: preemptive M/G/k (Scully et al.)

Strategy 0
just optimality

1966: compute $\mathbf{E}[T]$ (Schrage & Miller)

1968: optimal (Schrage)

Strategy 1
work in system

1970s: optimal (Sevcik; von Olivier; Gittins)

2005: $\mathbf{E}[T]$ in M/M/1 (Whittle)

2018: $\mathbf{E}[T]$ in M/G/1 (Scully et al.)

2018: $\mathbf{E}[T]$ bound, “near-optimal” (Grosof et al.)

2001: M/M/k (Glazebrook & Niño-Mora)

2003: nonpreemptive M/G/k (Glazebrook)

2020: preemptive M/G/k (Scully et al.)

Strategy 0
just optimality

1966: compute $\mathbf{E}[T]$ (Schrage & Miller)
1968: optimal (Schrage)

Strategy 1
work in system

1970s: optimal (Sevcik; von Olivier; Gittins)
2005: $\mathbf{E}[T]$ in M/M/1 (Whittle)
2018: $\mathbf{E}[T]$ in M/G/1 (Scully et al.)

Strategy 2
states of all jobs
(**Gittins**-specific)

2018: $\mathbf{E}[T]$ bound, "near-optimal" (Grosof et al.)

2001: M/M/k (Glazebrook & Niño-Mora)
2003: nonpreemptive M/G/k (Glazebrook)
2020: preemptive M/G/k (Scully et al.)

Strategy 0
just optimality

1966: compute $\mathbf{E}[T]$ (Schrage & Miller)

1968: optimal (Schrage)

Strategy 1
work in system

1970s: optimal (Sevcik; von Olivier; Gittins)

2005: $\mathbf{E}[T]$ in M/M/1 (Whittle)

2018: $\mathbf{E}[T]$ in M/G/1 (Scully et al.)

Strategy 2
states of all jobs
(**Gittins**-specific)

2018: $\mathbf{E}[T]$ bound, "near-optimal" (Grosof et al.)

Strategy 3
work in system
(**Gittins**-specific)

2001: M/M/k (Glazebrook & Niño-Mora)

2003: nonpreemptive M/G/k (Glazebrook)

2020: preemptive M/G/k (Scully et al.)

Strategy 0
just optimality

1966: compute $\mathbf{E}[T]$ (Schrage & Miller)
1968: optimal (Schrage)

Strategy 1
work in system

1970s: optimal (Sevcik; von Olivier; Gittins)
2005: $\mathbf{E}[T]$ in M/M/1 (Whittle)

Strate
states of
(**Gittins**-s

l” (Grosof et al.)

Our work:
Strategy 3 for **G/G** arrivals

Strategy 3
work in system
(**Gittins**-specific)

2001: M/M/k (Glazebrook & Niño-Mora)
2003: nonpreemptive M/G/k (Glazebrook)
2020: preemptive M/G/k (Scully et al.)

Our result

$$\mathbf{E}[T_{\mathbf{Gittins}}] - \mathbf{E}[T_{\mathrm{opt}}] \leq \ell_{\mathrm{multi}} + \ell_{\mathrm{G/G}} + \ell_{\mathrm{setup}}$$

Our result

$$\mathbf{E}[T_{\text{Gittins}}] - \mathbf{E}[T_{\text{opt}}] \le \ell_{\text{multi}} + \ell_{\text{G/G}} + \ell_{\text{setup}}$$

$$\ell_{\text{multi}} = 4\mathbf{E}[S] \log \frac{1}{1-\rho}$$

λ = arrival rate

S = job size

ρ = load = $\lambda \mathbf{E}[S]$

Our result

$$\mathbf{E}[T_{\text{Gittins}}] - \mathbf{E}[T_{\text{opt}}] \le \ell_{\text{multi}} + \ell_{\text{G/G}} + \ell_{\text{setup}}$$

$$\ell_{\text{multi}} = 4\mathbf{E}[S] \log \frac{1}{1-\rho}$$

$$\ell_{\text{G/G}} = \lambda(A_{\max} - A_{\min}) = O(1)$$

$\lambda =$ arrival rate

$S =$ job size

$\rho =$ load $= \lambda \mathbf{E}[S]$

$R =$ residual interarrival

$\mathbf{E}[R \mid \text{past}] \in [A_{\min}, A_{\max}]$

Our result

$$\mathbf{E}[T_{\text{Gittins}}] - \mathbf{E}[T_{\text{opt}}] \le \ell_{\text{multi}} + \ell_{\text{G/G}} + \ell_{\text{setup}}$$

$$\ell_{\text{multi}} = 4\mathbf{E}[S] \log \frac{1}{1-\rho}$$

$$\ell_{\text{G/G}} = \lambda(A_{\max} - A_{\min}) = O(1)$$

$$\ell_{\text{setup}} = O(1)$$

$\lambda =$ arrival rate

$S =$ job size

$\rho =$ load $= \lambda \mathbf{E}[S]$

$R =$ residual interarrival

$\mathbf{E}[R \mid \text{past}] \in [A_{\min}, A_{\max}]$

Our result

$$\mathbf{E}[T_{\text{Gittins}}] - \mathbf{E}[T_{\text{opt}}] \le \ell_{\text{multi}} + \ell_{\text{G/G}} + \ell_{\text{setup}}$$

$$\ell_{\text{multi}} = 4\mathbf{E}[S]\log\frac{1}{1-\rho}$$

$$\ell_{\text{G/G}} = \lambda(A_{\max} - A_{\min}) = O(1)$$

$$\ell_{\text{setup}} = O(1)$$

λ = arrival rate

S = job size

ρ = load = $\lambda\mathbf{E}[S]$

R = residual interarrival

$\mathbf{E}[R \mid \text{past}] \in [A_{\min}, A_{\max}]$

Gittins is optimal
in heavy traffic

Our result

$$\mathbf{E}[T_{\text{Gittins}}] - \mathbf{E}[T_{\text{opt}}] \le \ell_{\text{multi}} + \ell_{G/G} + \ell_{\text{setup}}$$

$\ell_{\text{multi}} =$ 1 rate

e

$\ell_{G/G} =$ $\lambda \mathbf{E}[S]$

l interarrival

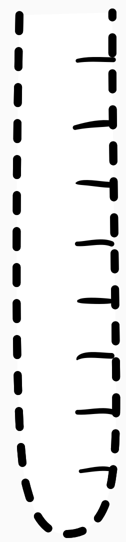
$\ell_{\text{setup}} =$ $A_{\max}]$

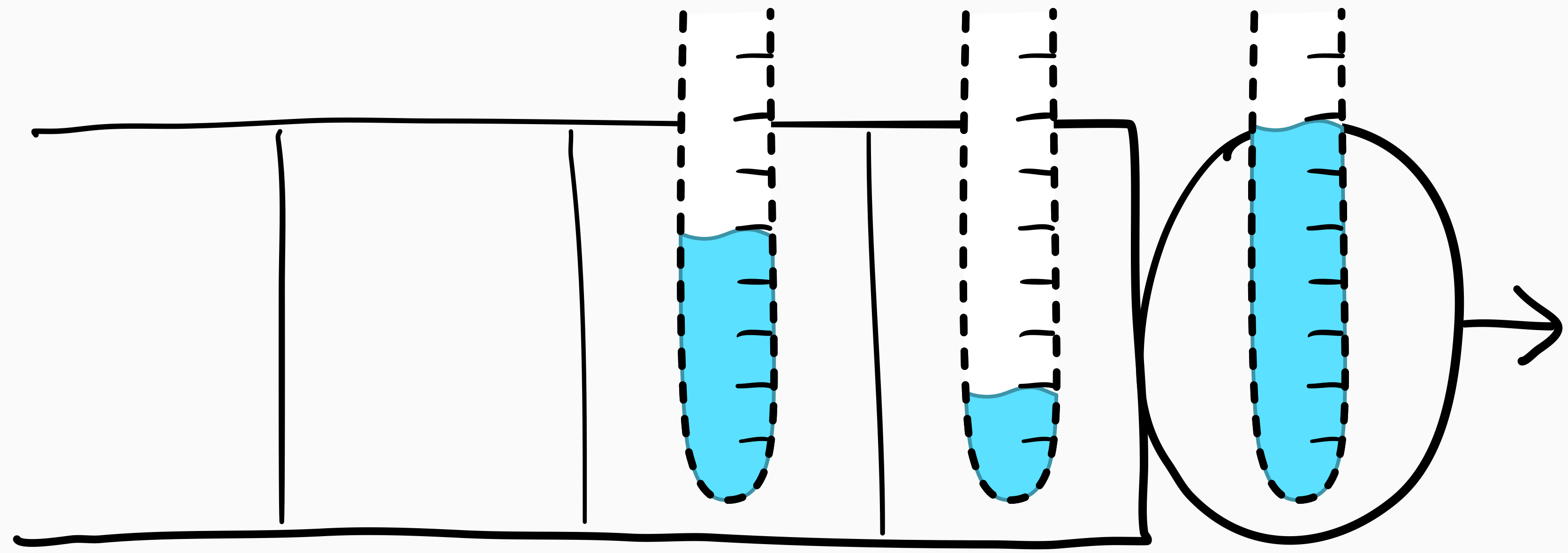
How does **Strategy 3** work?

How do we handle **G/G** arrivals?

Gittins is optimal
in heavy traffic

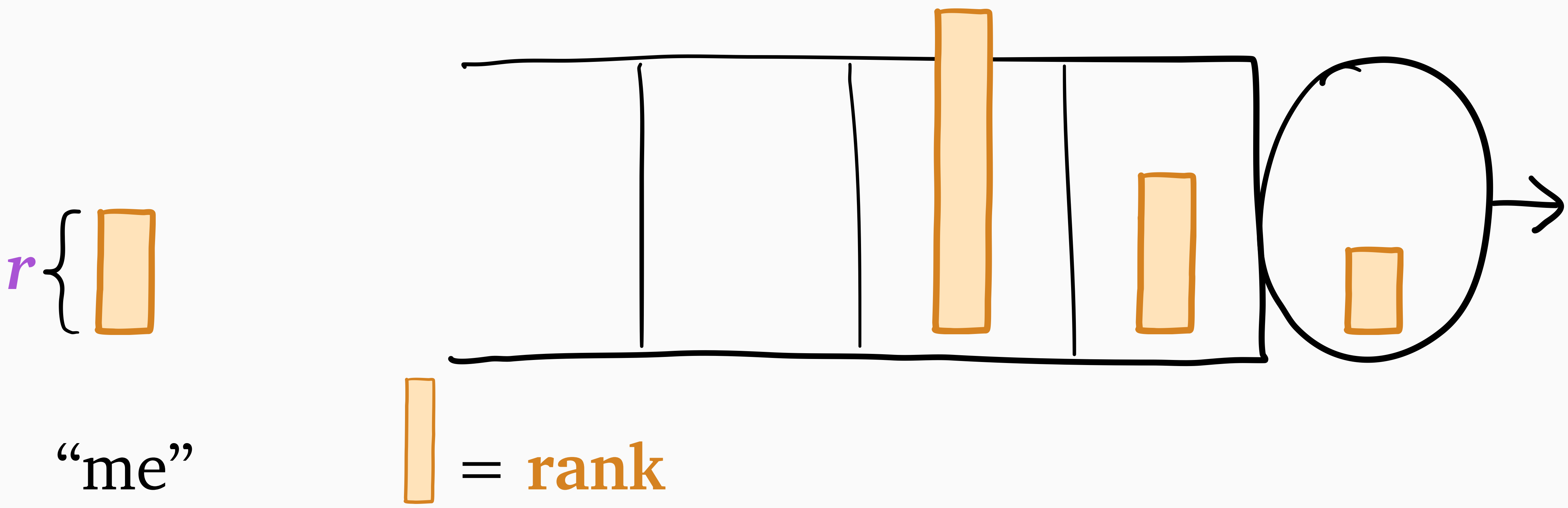
Key quantity: *r*-work



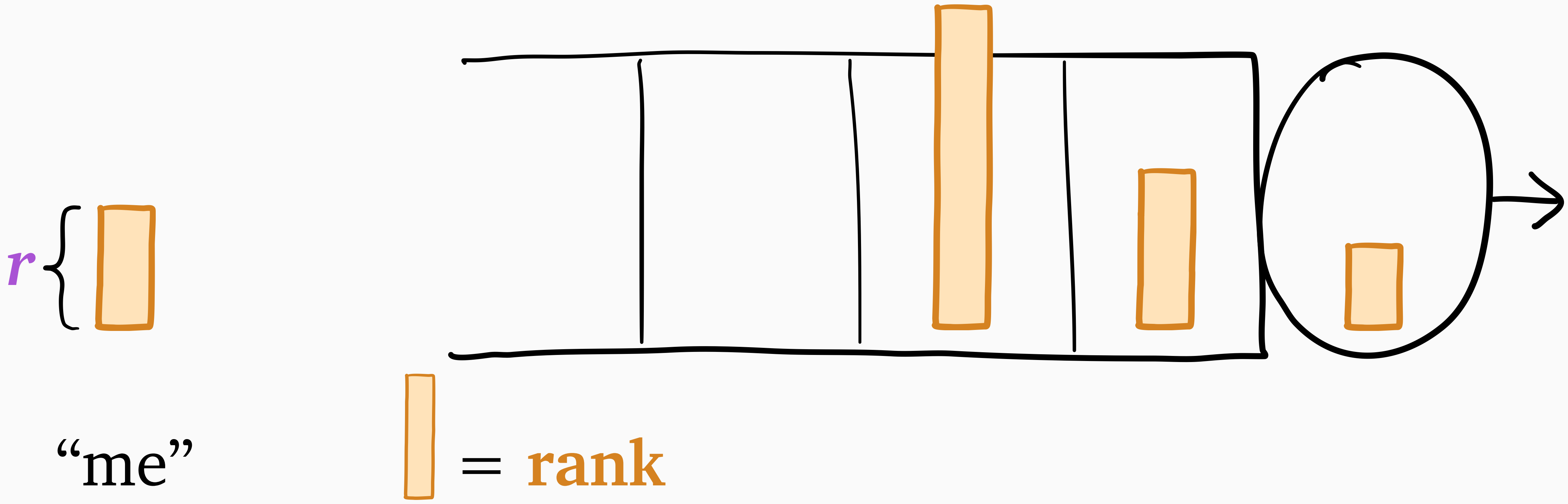


"me"

Key quantity: *r*-work

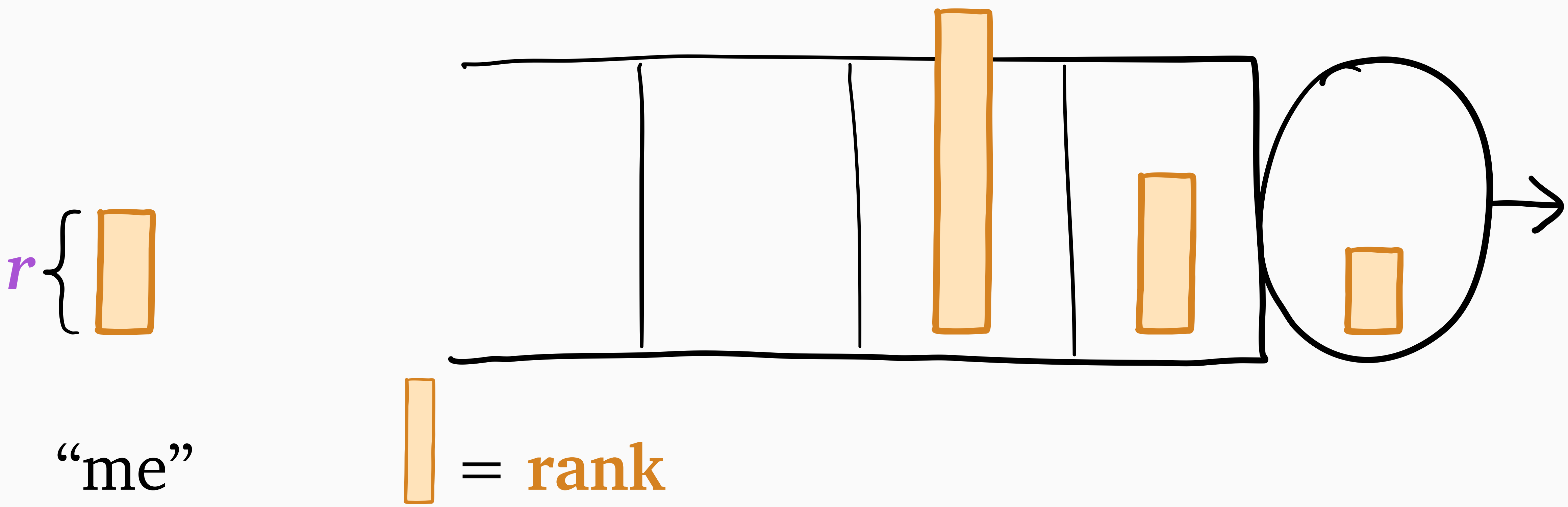


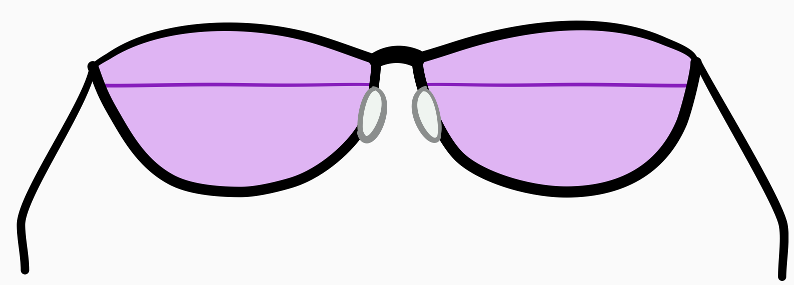
Key quantity: r-work



r-work: $W(r)$ = work relevant to job of **rank r**

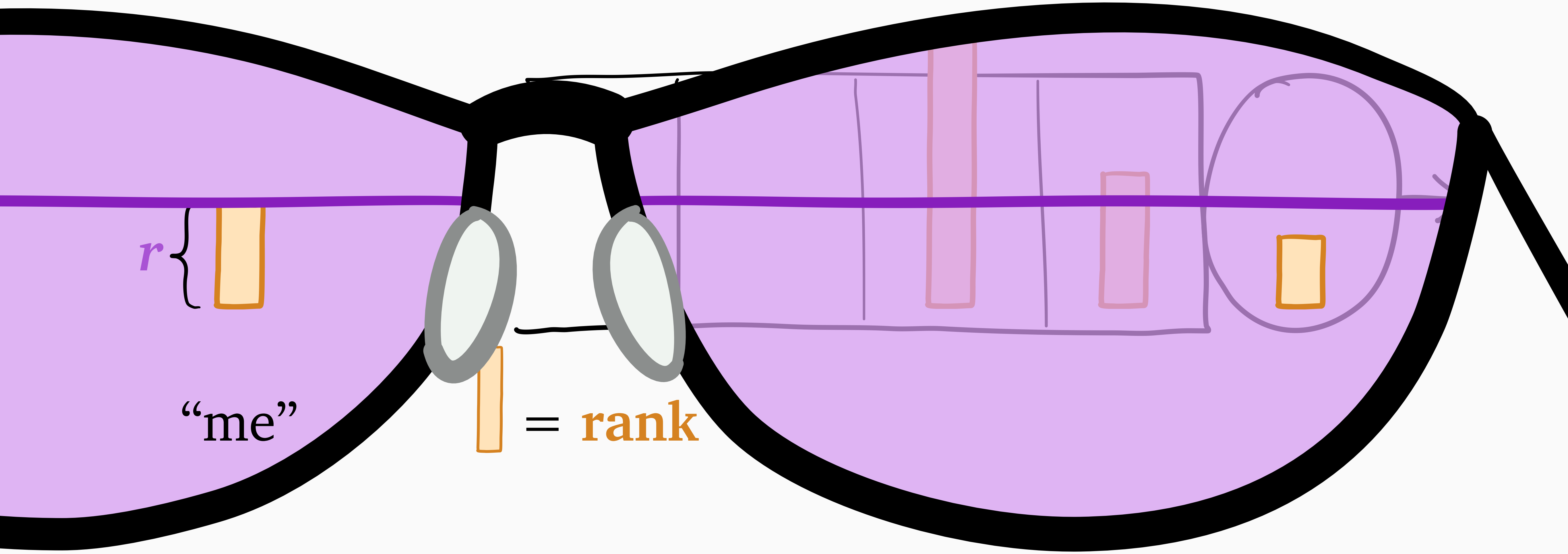
Key quantity: r-work





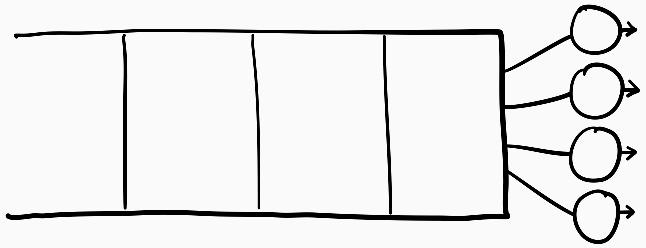
r-work: $W(r)$ = work relevant to job of **rank** r

Key quantity: *r*-work

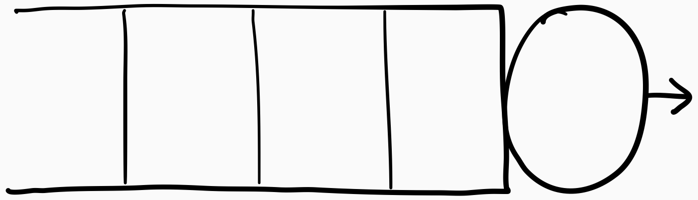


r*-work:** $W(r)$ = work relevant to job of **rank** ***r

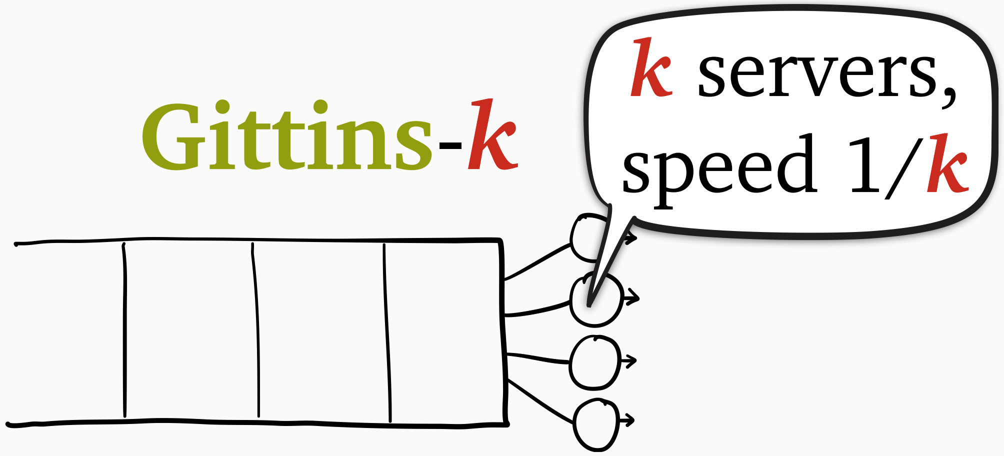
Gittins-*k*



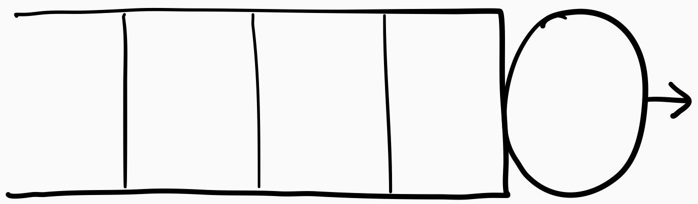
Gittins-1

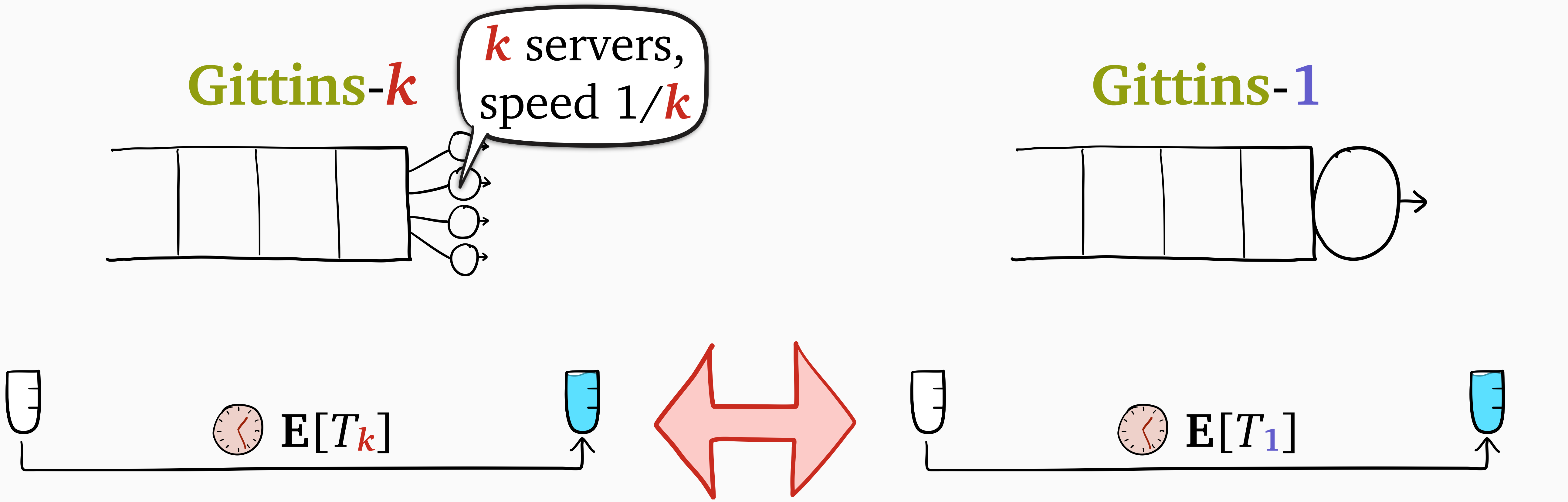


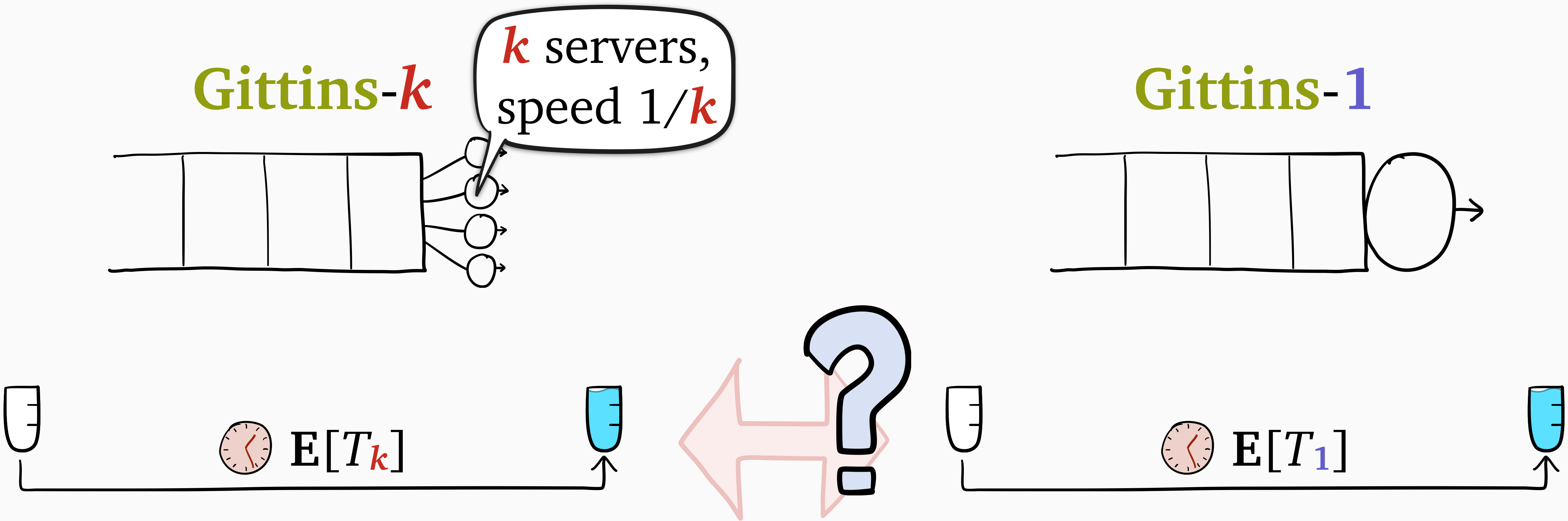
Gittins-*k*

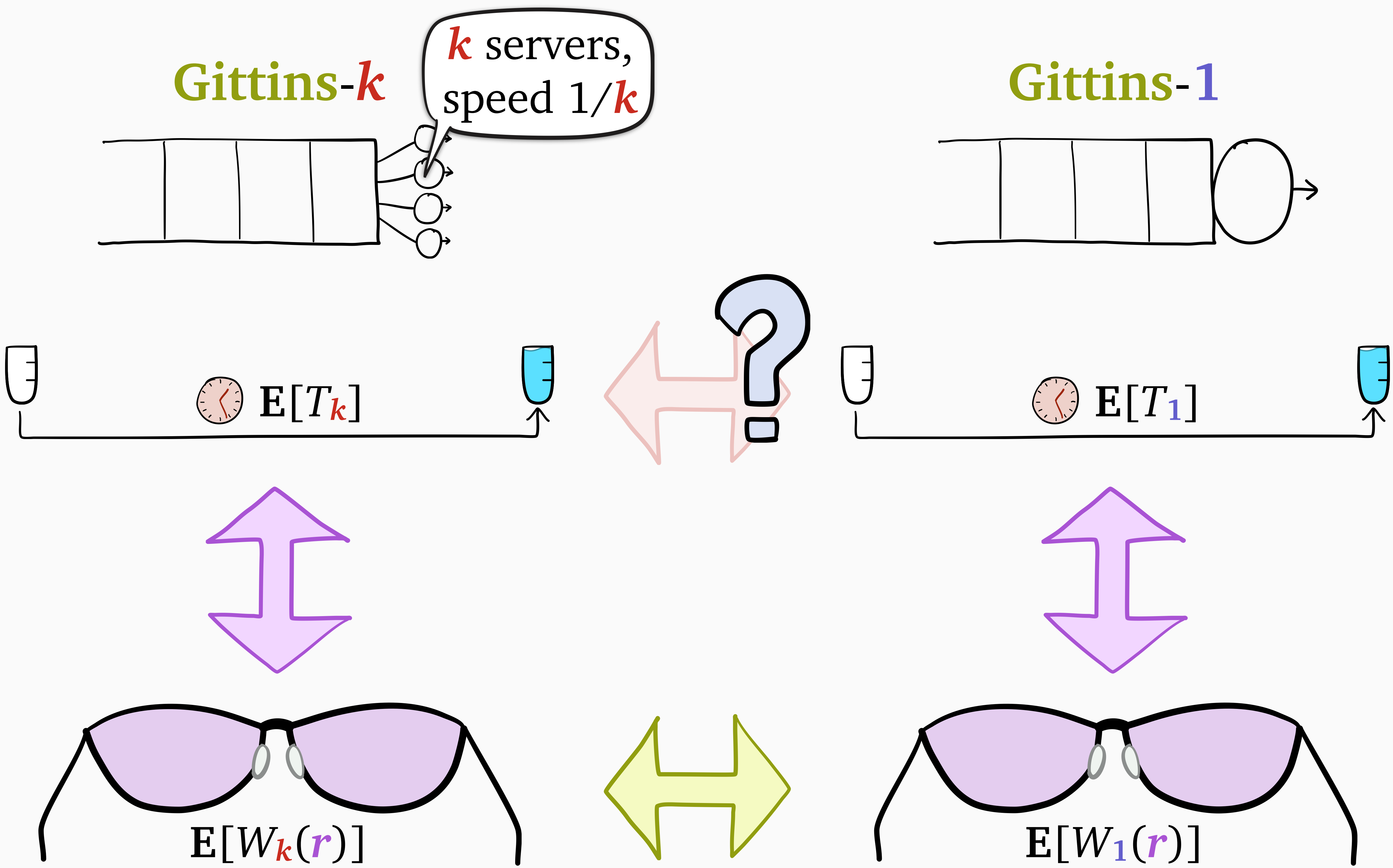


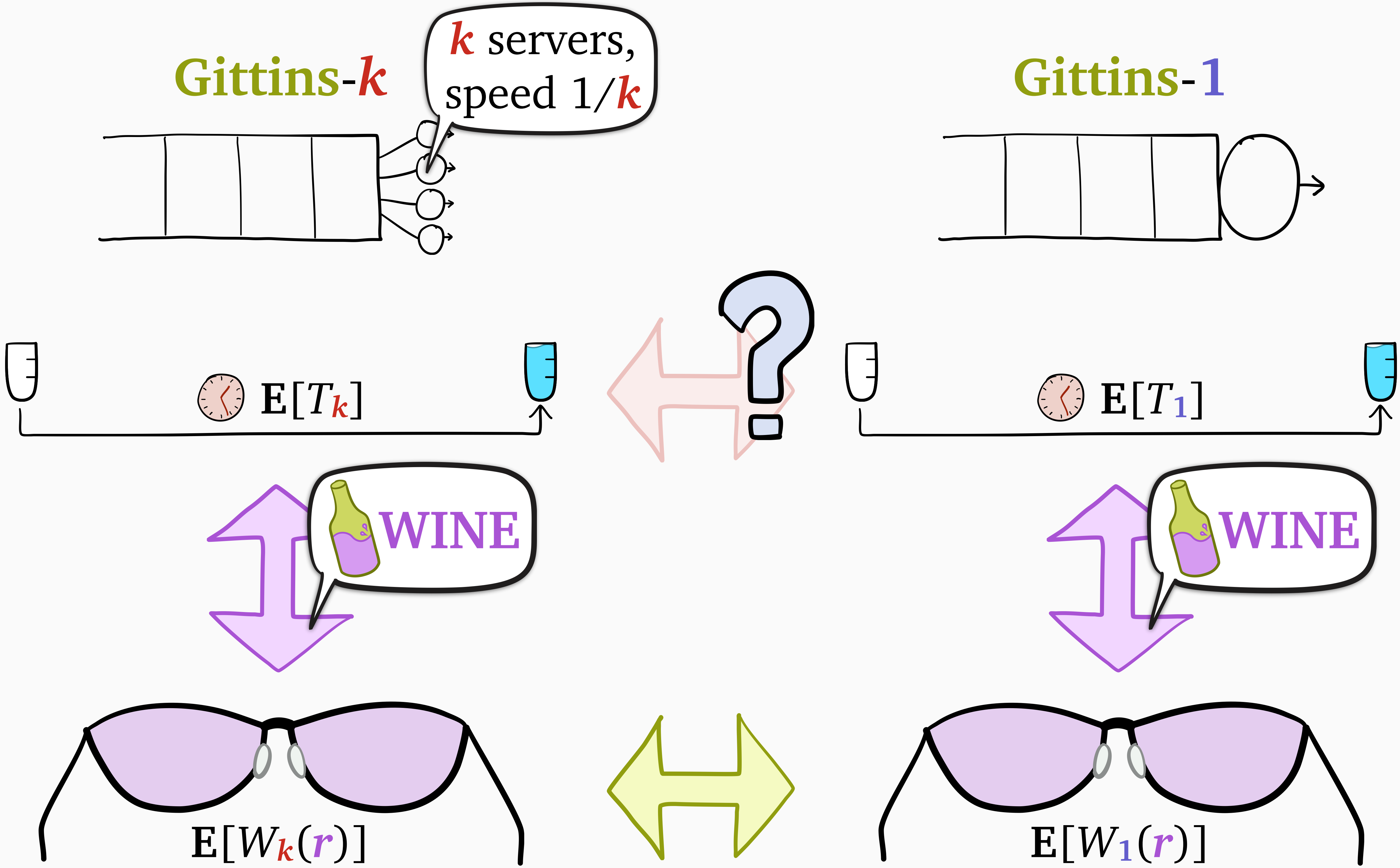
Gittins-1

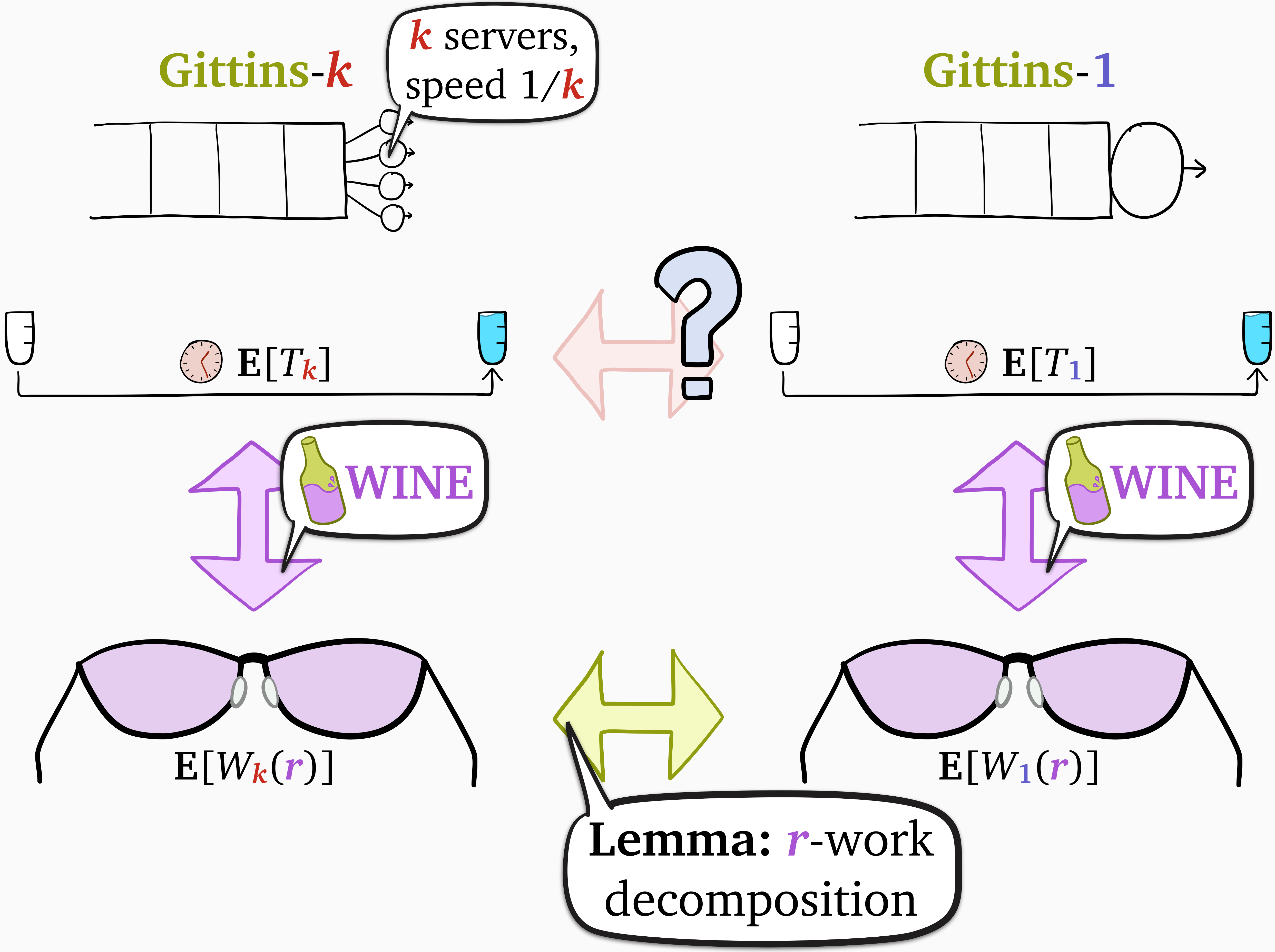



Gittins-k
k servers, speed $1/k$
Gittins-1
$\mathbf{E}[T_k]$
$\mathbf{E}[T_1]$

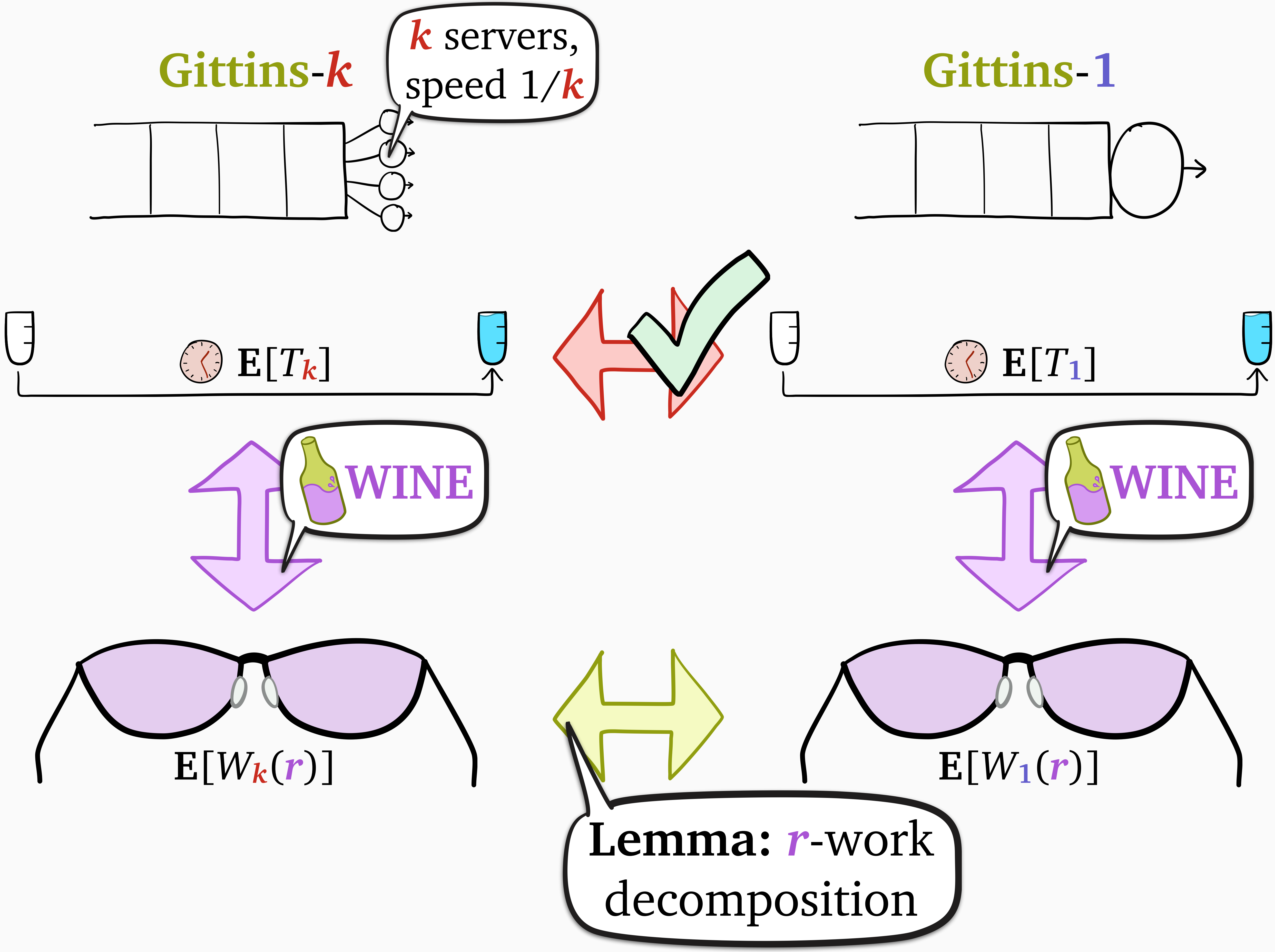

Gittins-k
k servers, speed $1/k$
Gittins-1
$\mathbf{E}[T_k]$
$\mathbf{E}[T_1]$


Gittins-k
k servers, speed $1/k$
Gittins-1
$\mathbf{E}[T_k]$
$\mathbf{E}[T_1]$
$\mathbf{E}[W_k(r)]$
$\mathbf{E}[W_1(r)]$

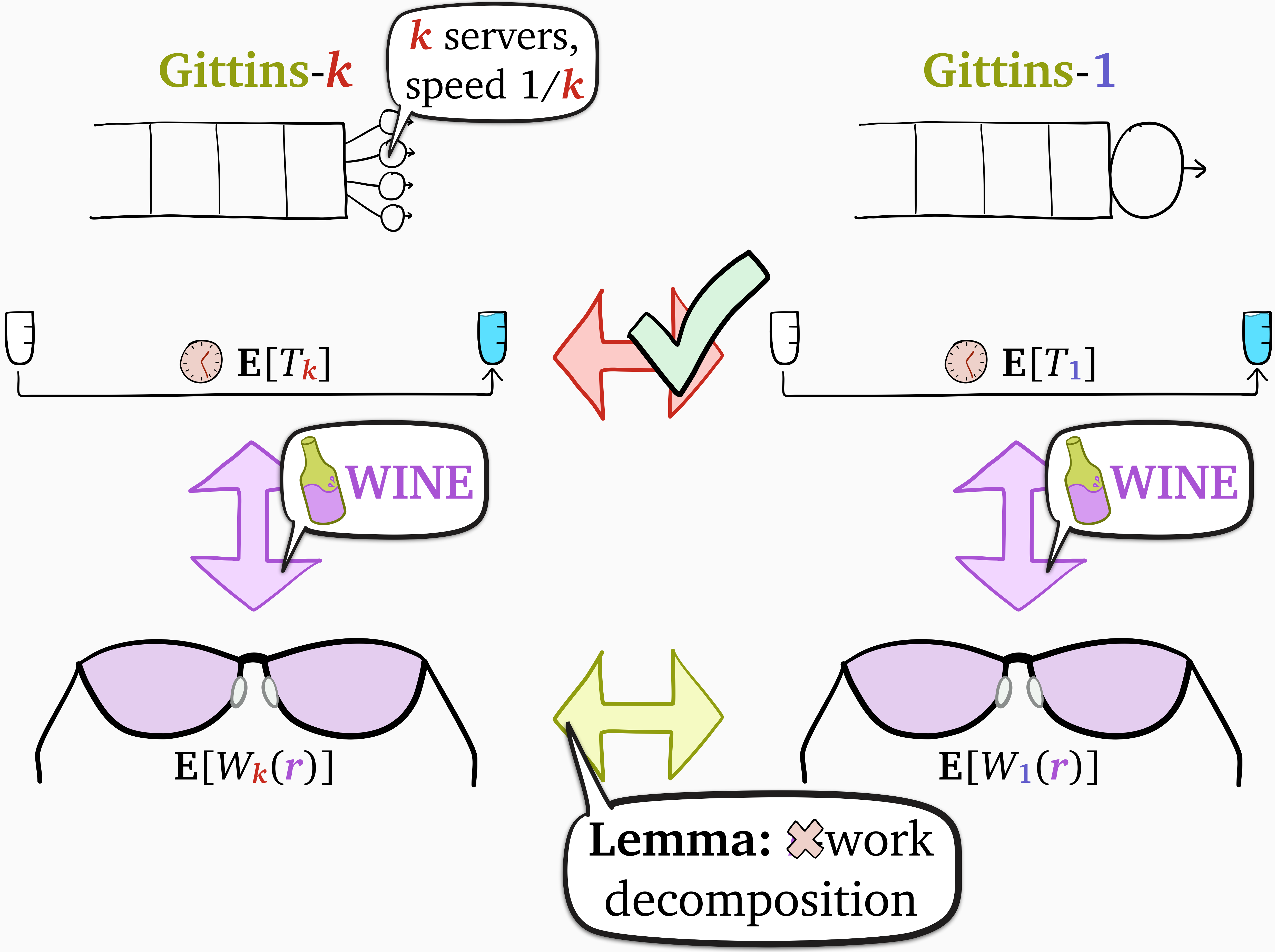

Gittins-k
k servers, speed $1/k$
Gittins-1
$\mathbf{E}[T_k]$
$\mathbf{E}[T_1]$
WINE
WINE
$\mathbf{E}[W_k(r)]$
$\mathbf{E}[W_1(r)]$



Gittins-k
k servers, speed 1/k
Gittins-1
E[T_k]
E[T_1]
WINE
WINE
E[W_k(r)]
E[W_1(r)]
Lemma: r-work decomposition



Gittins-k
k servers, speed $1/k$
Gittins-1
$\mathbf{E}[T_k]$
$\mathbf{E}[T_1]$
WINE
WINE
$\mathbf{E}[W_k(r)]$
$\mathbf{E}[W_1(r)]$
Lemma: r-work decomposition



Gittins-k
k servers, speed $1/k$
Gittins-1
$\mathbf{E}[T_k]$
$\mathbf{E}[T_1]$
WINE
WINE
$\mathbf{E}[W_k(r)]$
$\mathbf{E}[W_1(r)]$
Lemma: work decomposition

Work decomposition law

M/G arrivals:

$$\mathbf{E}[W] = \mathbf{E}[W_{\mathrm{M/G/1}}] + \frac{\mathbf{E}[IW]}{1-\rho}$$

Work decomposition law

W seen by idle server

M/G arrivals:

$$\mathbf{E}[W] = \mathbf{E}[W_{\mathrm{M/G/1}}] + \frac{\mathbf{E}[IW]}{1-\rho}$$

Work decomposition law

W seen by idle server → ℓ_{multi}

M/G arrivals:

$$\mathbf{E}[W] = \mathbf{E}[W_{\text{M/G/1}}] + \frac{\mathbf{E}[IW]}{1-\rho}$$

Work decomposition law

W seen by idle server ⇒ ℓ_{multi}

M/G arrivals:

$$\mathbf{E}[W] = \mathbf{E}[W_{\text{M/G/1}}] + \frac{\mathbf{E}[IW]}{1-\rho}$$

G/G arrivals: NEW!

$$\mathbf{E}[W] = \text{``}\mathbf{E}[W_{\text{G/G/1}}]\text{''} + \frac{\mathbf{E}[IW]}{1-\rho} - \frac{\rho\,\mathbf{E}[IR]}{1-\rho}$$

Work decomposition law

M/G arrivals:

W seen by idle server → ℓ_{multi}

$$\mathbf{E}[W] = \mathbf{E}[W_{\mathrm{M/G/1}}] + \frac{\mathbf{E}[IW]}{1-\rho}$$

G/G arrivals: NEW!

W seen by idle server

$$\mathbf{E}[W] = \text{“}\mathbf{E}[W_{\mathrm{G/G/1}}]\text{”} + \frac{\mathbf{E}[IW]}{1-\rho} - \frac{\rho\mathbf{E}[IR]}{1-\rho}$$

Work decomposition law

M/G arrivals:

W seen by idle server $\Rightarrow \ell_{\text{multi}}$

$$\mathbf{E}[W] = \mathbf{E}[W_{\text{M/G/1}}] + \frac{\mathbf{E}[IW]}{1-\rho}$$

G/G arrivals: NEW!

W seen by idle server $\Rightarrow \ell_{\text{multi}}$

$$\mathbf{E}[W] = \text{“}\mathbf{E}[W_{\text{G/G/1}}]\text{”} + \frac{\mathbf{E}[IW]}{1-\rho} - \frac{\rho\mathbf{E}[IR]}{1-\rho}$$

Work decomposition law

M/G arrivals:

W seen by idle server $\Rightarrow \ell_{\text{multi}}$

$$\mathbf{E}[W] = \mathbf{E}[W_{\mathrm{M/G/1}}] + \frac{\mathbf{E}[IW]}{1-\rho}$$

G/G arrivals: NEW!

W seen by idle server $\Rightarrow \ell_{\text{multi}}$

$$\mathbf{E}[W] = \text{“}\mathbf{E}[W_{\mathrm{G/G/1}}]\text{”} + \frac{\mathbf{E}[IW]}{1-\rho} - \frac{\rho\,\mathbf{E}[IR]}{1-\rho}$$

ρR seen by idle server

Work decomposition law

M/G arrivals:

$$\mathbf{E}[W] = \mathbf{E}[W_{\mathrm{M/G/1}}] + \frac{\mathbf{E}[IW]}{1-\rho}$$

W seen by idle server → ℓ_{multi}

G/G arrivals: NEW!

$$\mathbf{E}[W] = \text{“}\mathbf{E}[W_{\mathrm{G/G/1}}]\text{”} + \frac{\mathbf{E}[IW]}{1-\rho} - \frac{\rho\mathbf{E}[IR]}{1-\rho}$$

W seen by idle server → ℓ_{multi}

ρR seen by idle server → $\ell_{\mathrm{G/G}}$

Work decomposition law

M/G arrivals:

W seen by idle server → ℓ_{multi} and ℓ_{setup}

$$\mathbf{E}[W] = \mathbf{E}[W_{\mathrm{M/G/1}}] + \frac{\mathbf{E}[IW]}{1-\rho}$$

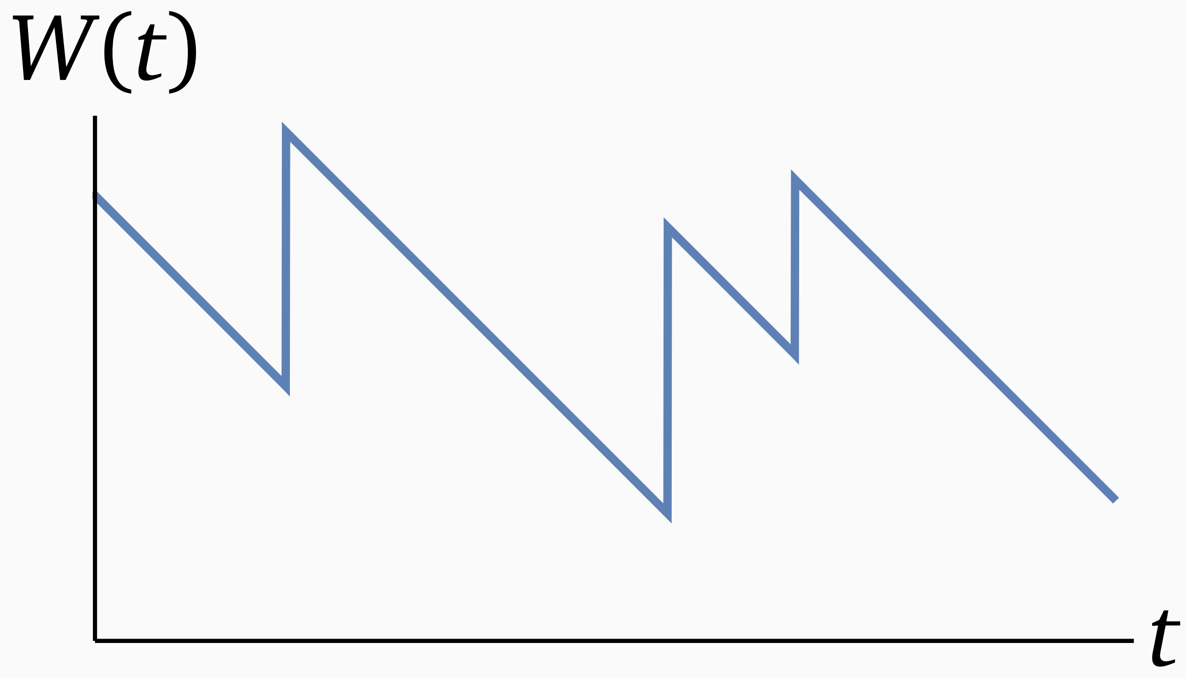
G/G arrivals: NEW!

W seen by idle server → ℓ_{multi} and ℓ_{setup}

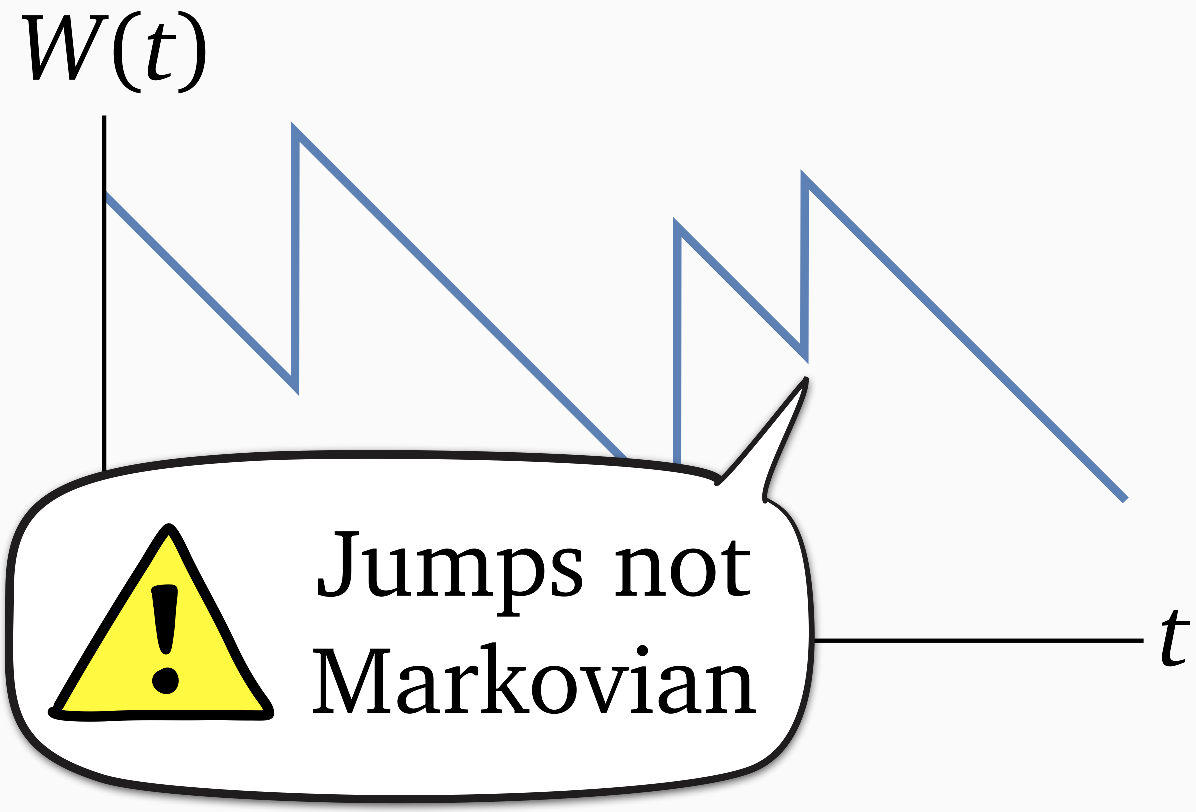
$$\mathbf{E}[W] = \text{“}\mathbf{E}[W_{\mathrm{G/G/1}}]\text{”} + \frac{\mathbf{E}[IW]}{1-\rho} - \frac{\rho\mathbf{E}[IR]}{1-\rho}$$

ρR seen by idle server → $\ell_{\mathrm{G/G}}$

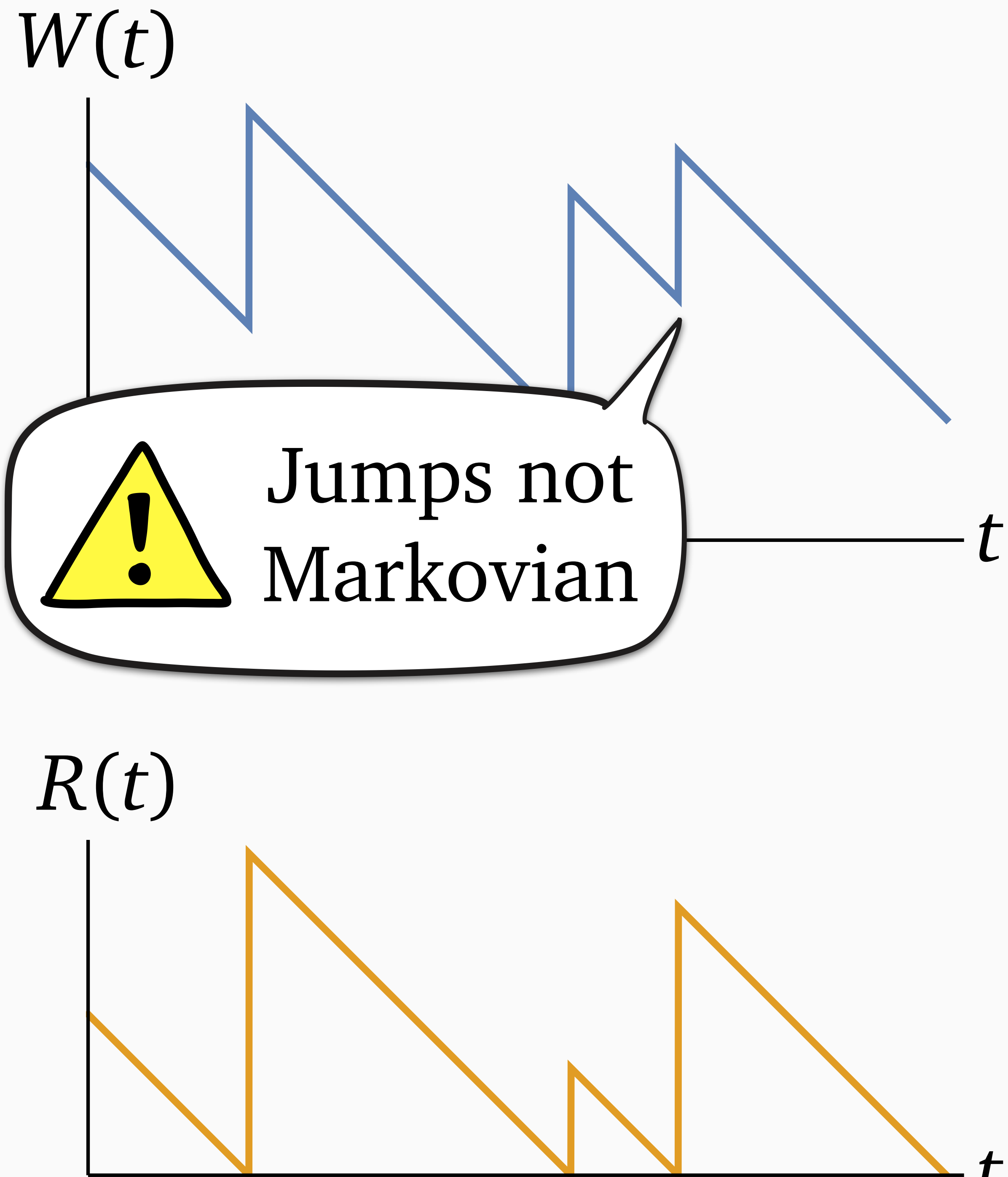
Key idea for G/G arrival proof



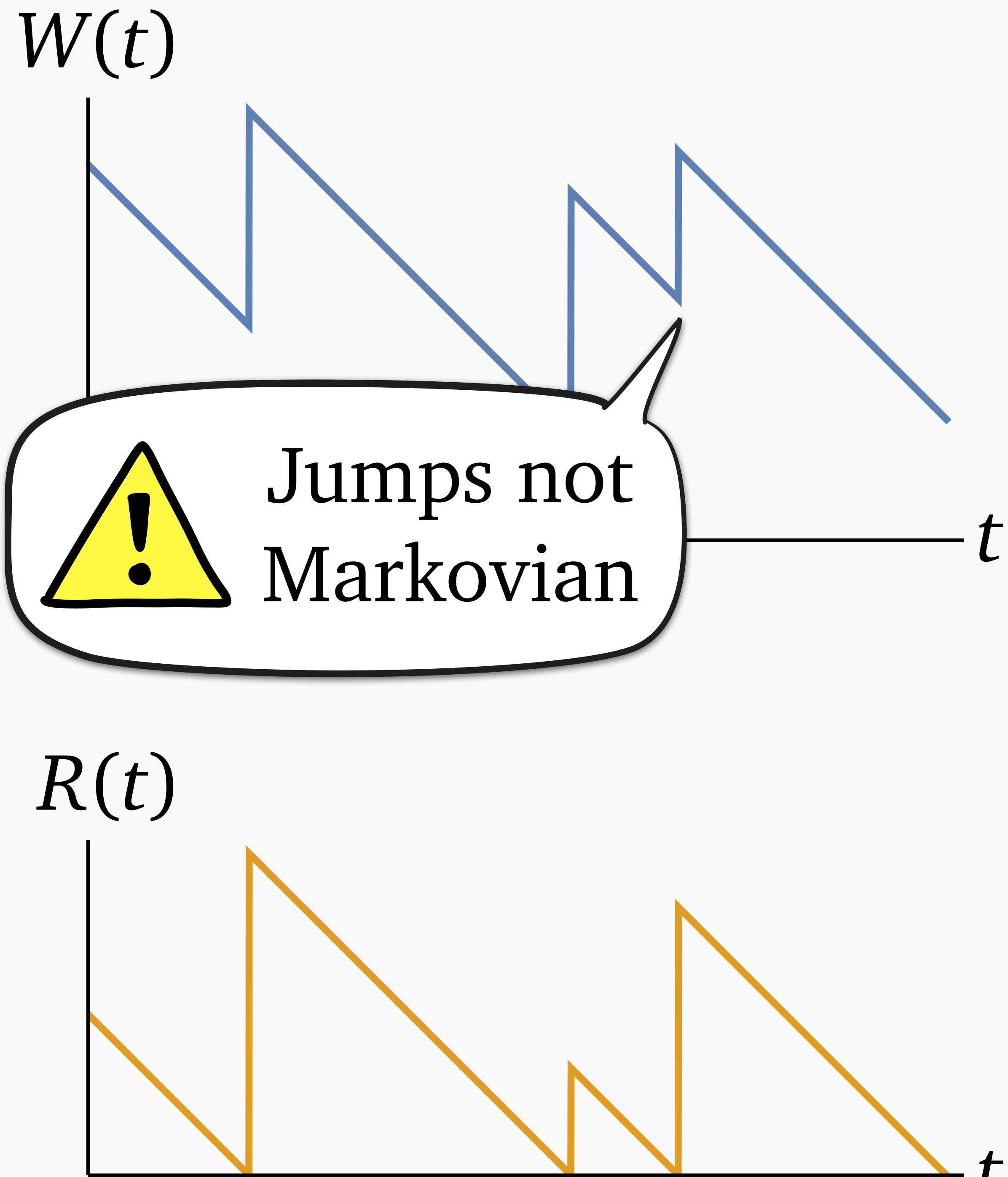
Key idea for G/G arrival proof

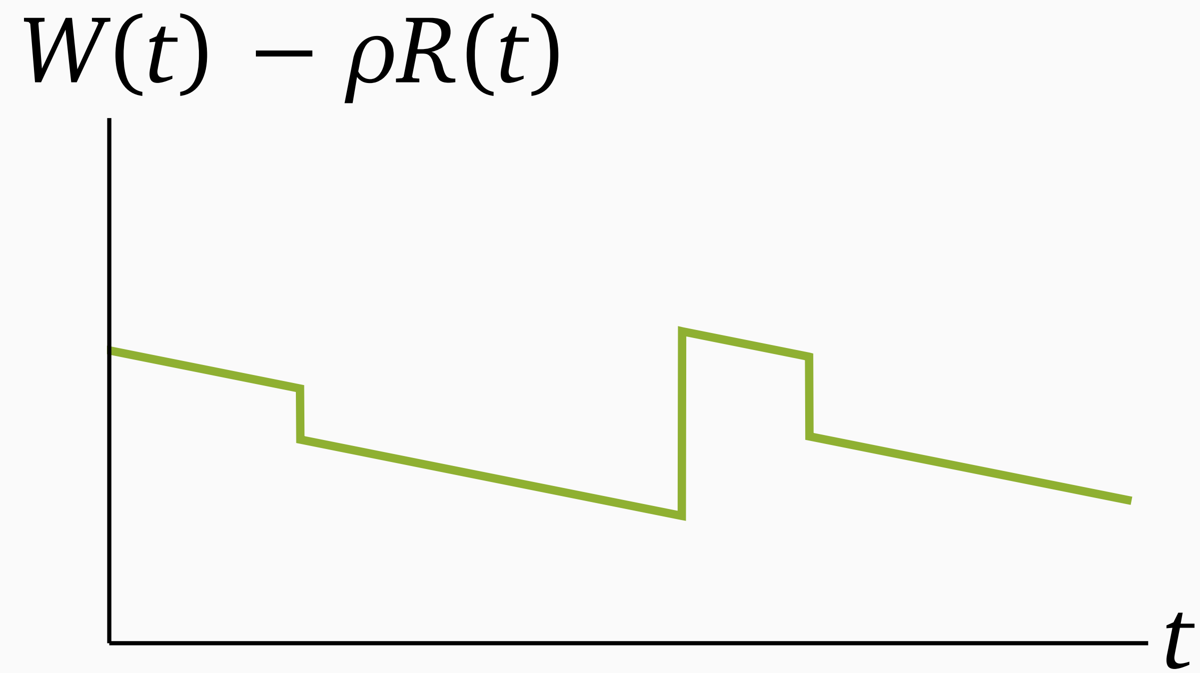


Key idea for G/G arrival proof

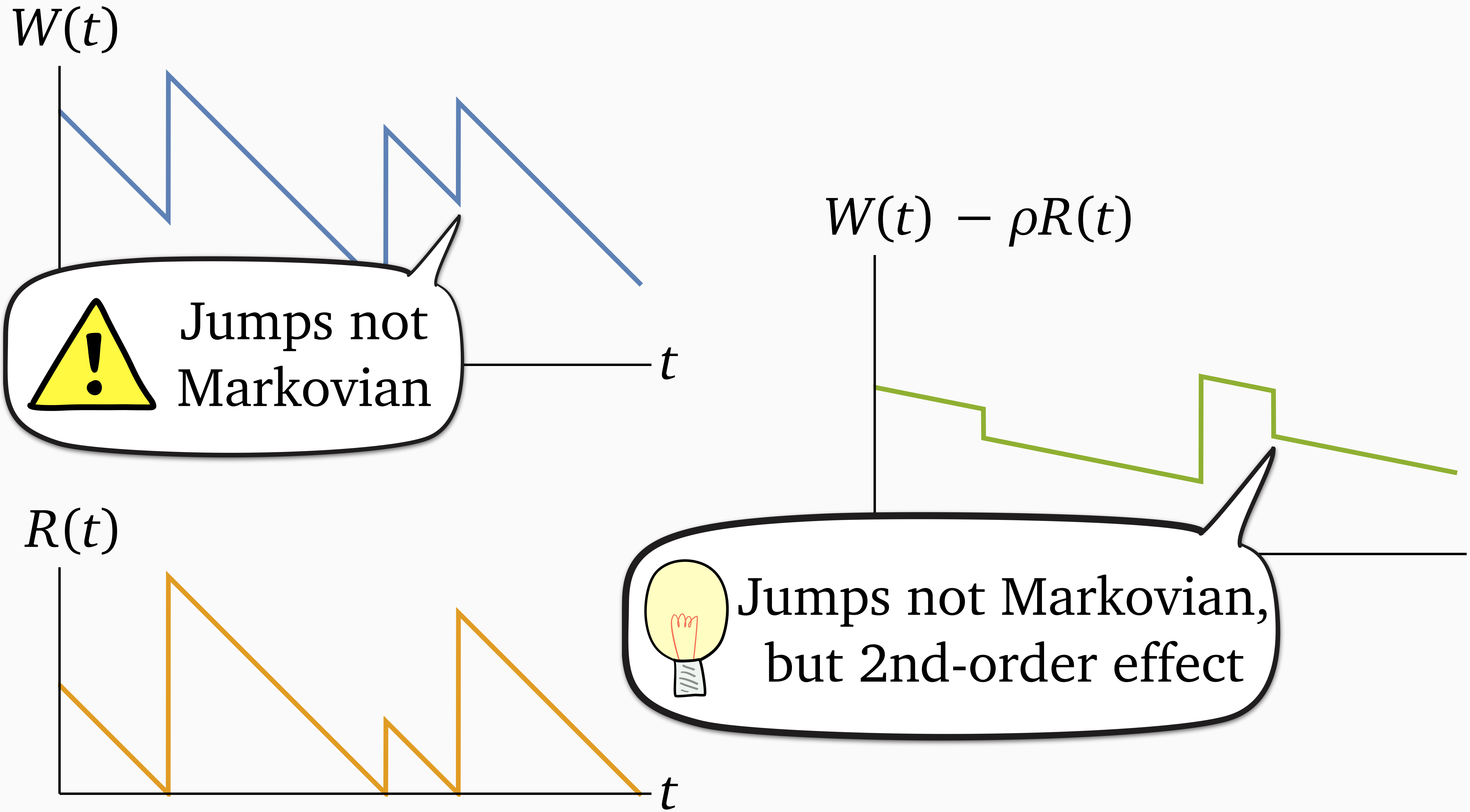


Key idea for G/G arrival proof





Key idea for G/G arrival proof



Strategy 0
just optimality

1966: compute $\mathbf{E}[T]$ (Schrage & Miller)
1968: optimal (Schrage)

Strategy 1
work in system

1970s: optimal (Sevcik; von Olivier; Gittins)
2005: $\mathbf{E}[T]$ in M/M/1 (Whittle)
2018: $\mathbf{E}[T]$ in M/G/1 (Scully et al.)

Strategy 2
states of all jobs
(**Gittins**-specific)

2018: $\mathbf{E}[T]$ bound, "near-optimal" (Grosof et al.)

Strategy 3
work in system
(**Gittins**-specific)

2001: M/M/k (Glazebrook & Niño-Mora)
2003: nonpreemptive M/G/k (Glazebrook)
2020: preemptive M/G/k (Scully et al.)

Strategy 0
just optimality

1966: compute $\mathbf{E}[T]$ (Schrage & Miller)
1968: optimal (Schrage)

Strategy 1
work in system

1970s: optimal (Sevcik; von Olivier; Gittins)
2005: $\mathbf{E}[T]$ in M/M/1 (Whittle)
2018: $\mathbf{E}[T]$ in M/G/1 (Scully et al.)

Strategy 2
states of all jobs
(**Gittins**-specific)

2018: $\mathbf{E}[T]$ bound, "near-optimal" (Grosof et al.)

Strategy 3
work in system
(**Gittins**-specific)

2001: M/M/k (Glazebrook & Niño-Mora)
2003: nonpreemptive M/G/k (Glazebrook)
2020: preemptive M/G/k (Scully et al.)
***2023:* G/G/k/setup** (Hong & Scully)