Performance of Gittins policy in G/G/1 and G/G/k, with and without setup times

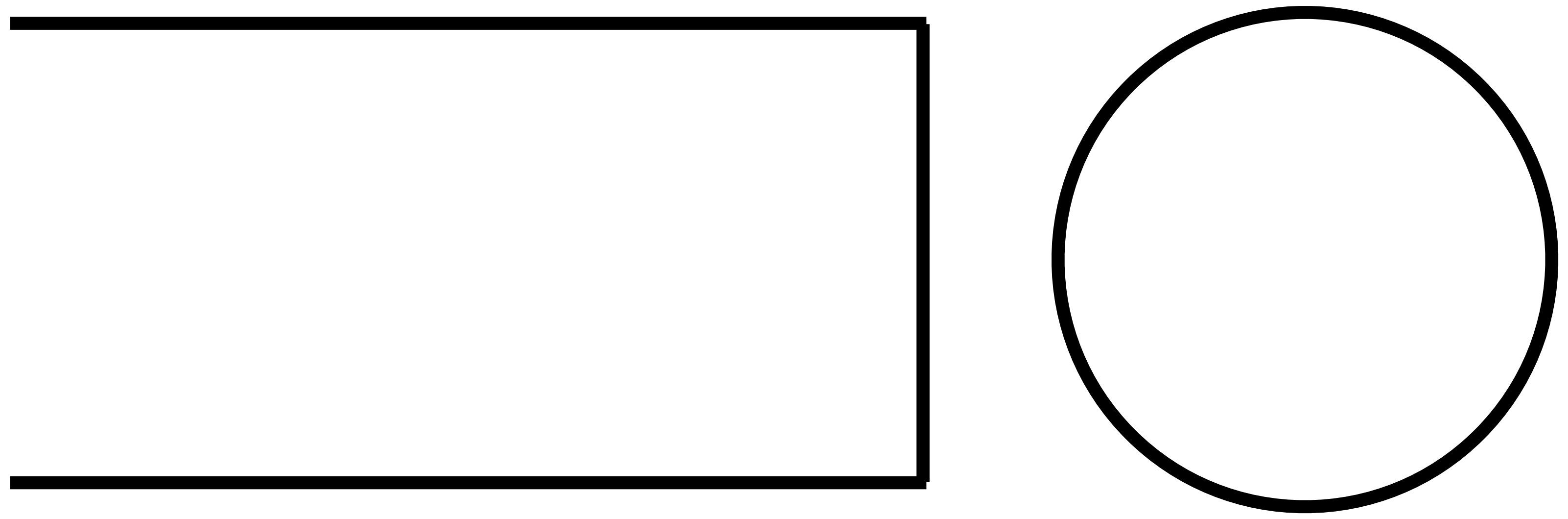
IFIP PERFORMANCE 2023, Nov 14th

Presenter: Yige Hong (Carnegie Mellon University)
Joint work with Ziv Scully (Cornell University)

Motivating question

How should we schedule jobs with unknown sizes in complicated systems?

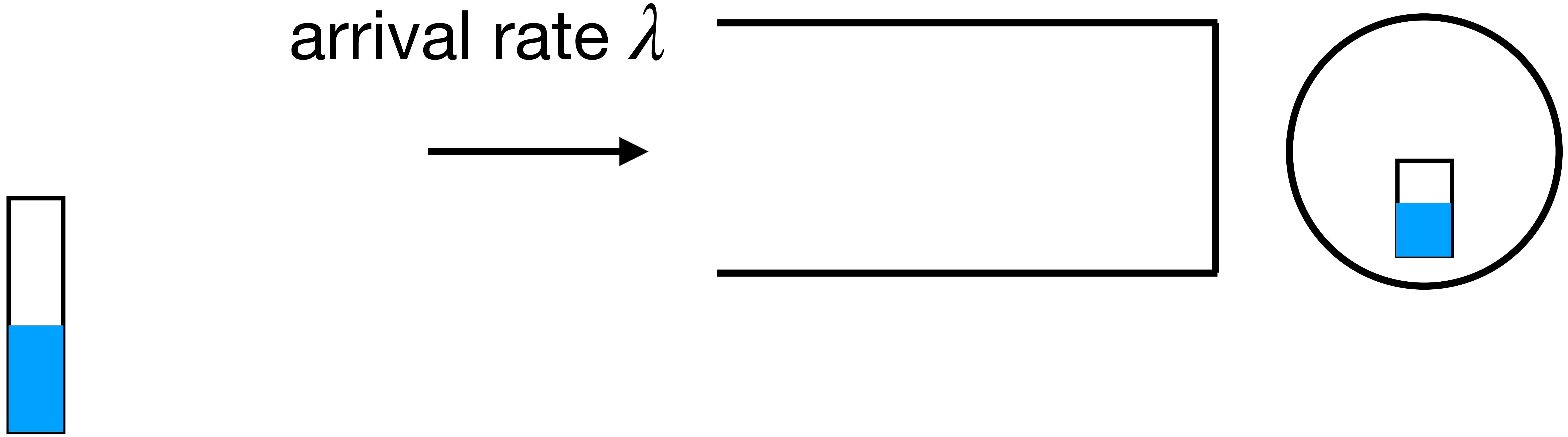
Scheduling problem



Scheduling problem

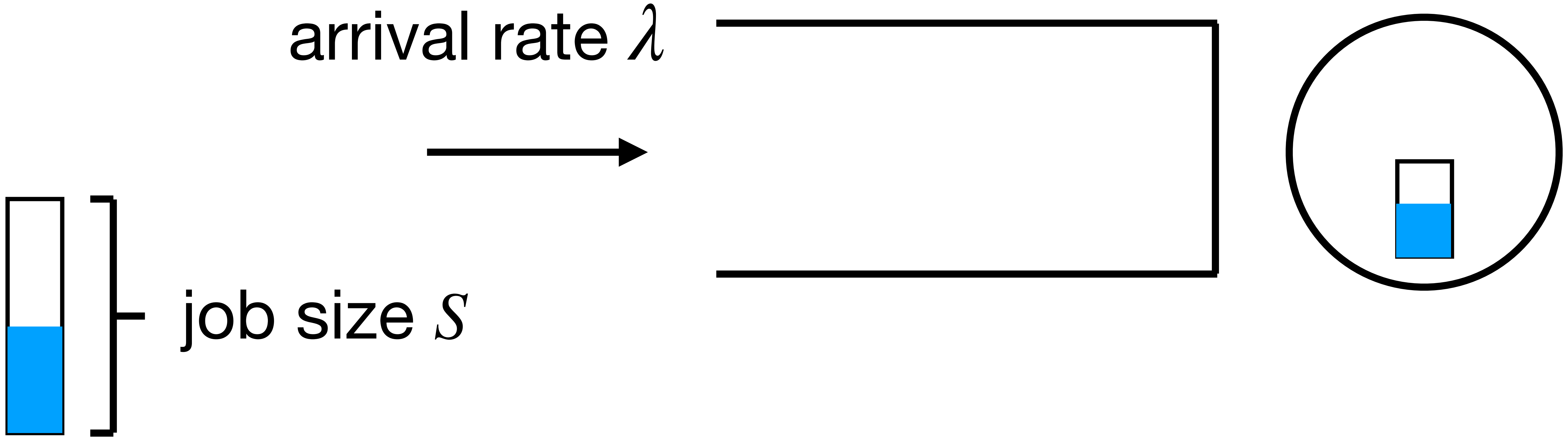
Scheduling problem

Scheduling problem

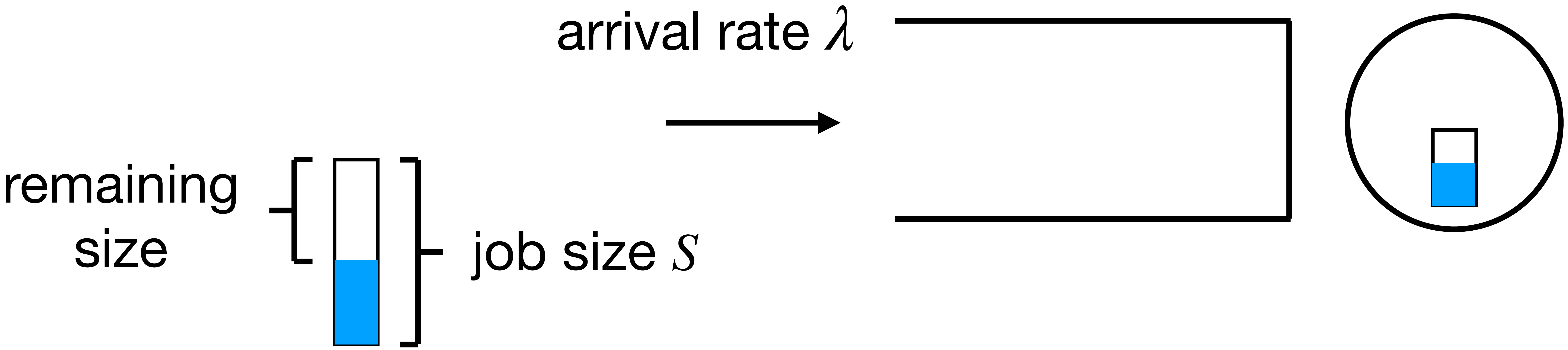


arrival rate λ

Scheduling problem



Scheduling problem

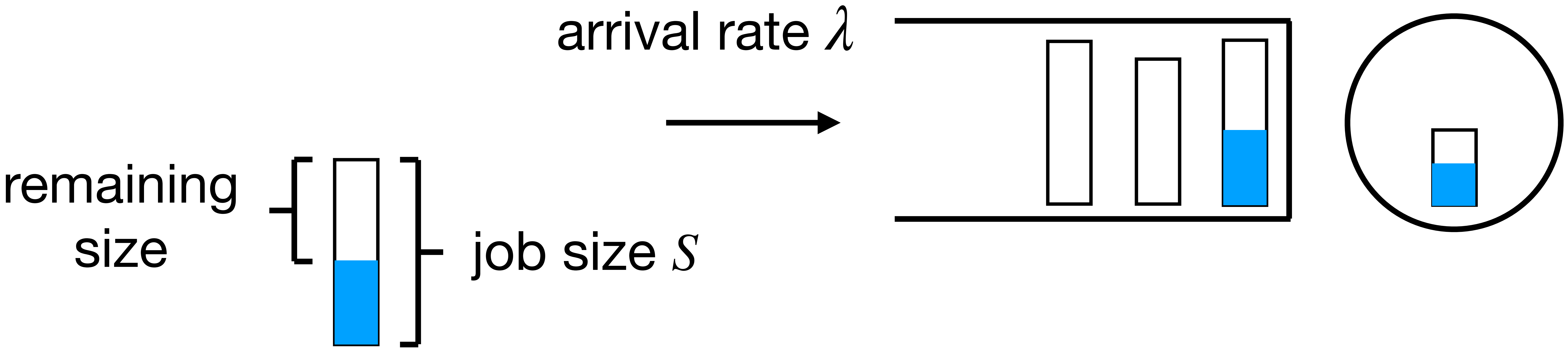


arrival rate λ

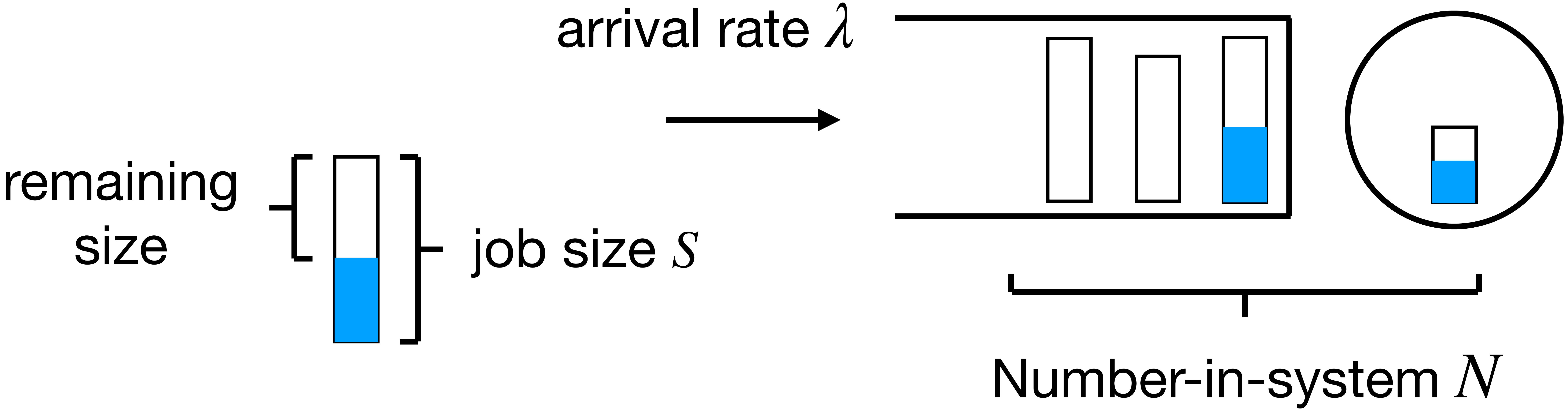
remaining size

job size S

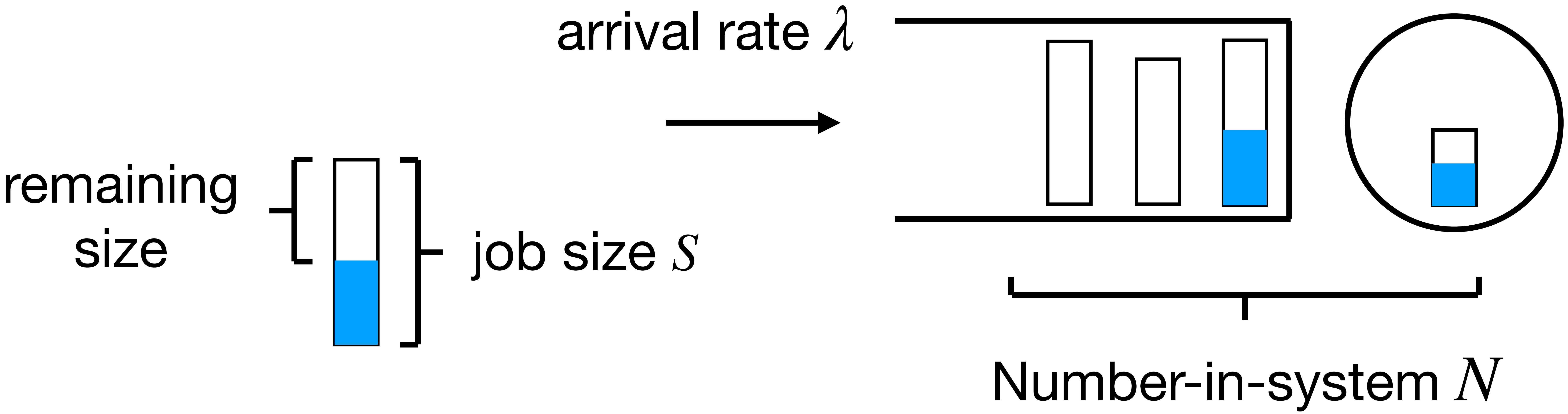
Scheduling problem



Scheduling problem

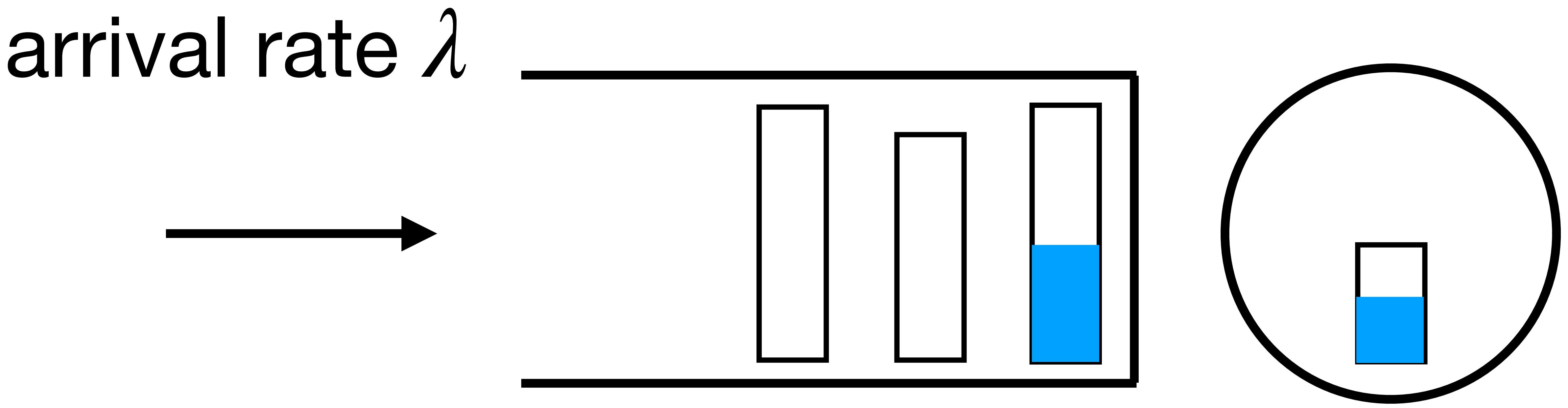


Scheduling problem



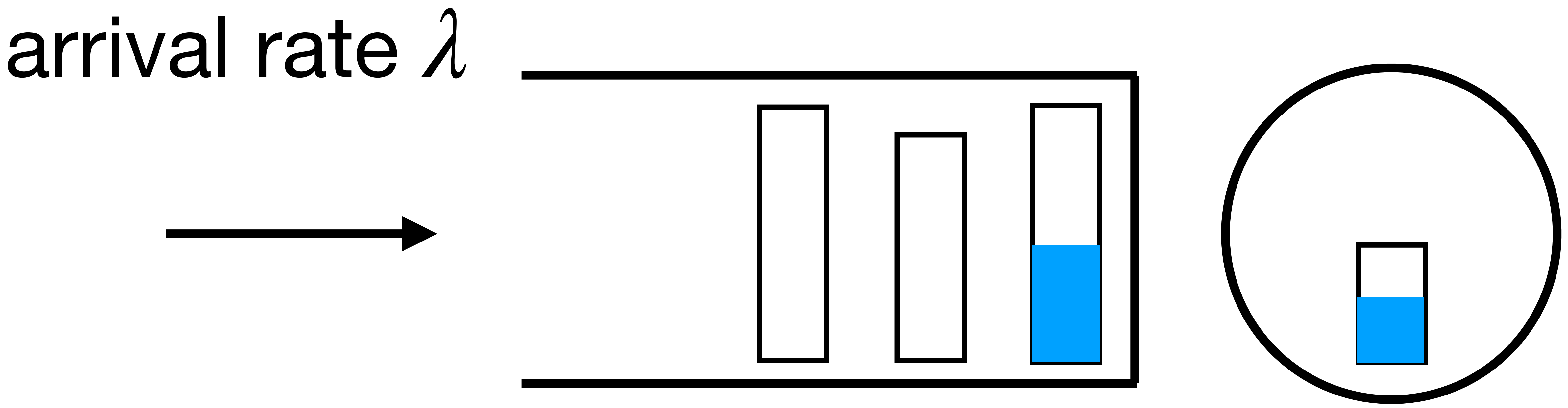
Goal:
Find the scheduling policy (order of serving jobs)
that minimizes steady-state $\mathbb{E}[N]$

Simpler case: known job sizes



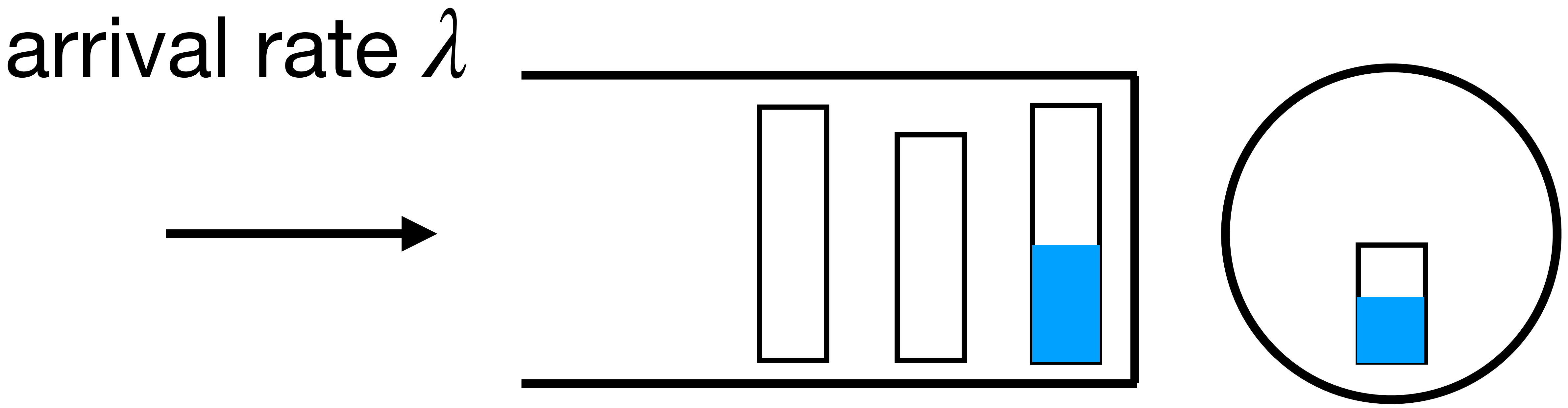
- **Shortest Remaining Processing Time (SRPT)**

Simpler case: known job sizes



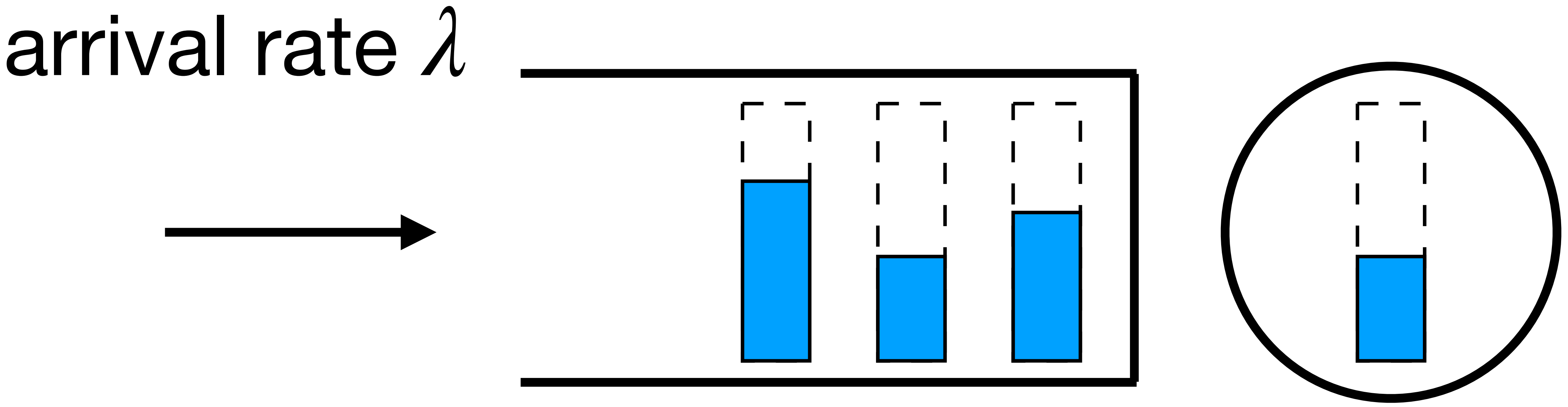
- **Shortest Remaining Processing Time (SRPT)**
  - Minimizes $\mathbb{E}[N]$

Simpler case: known job sizes

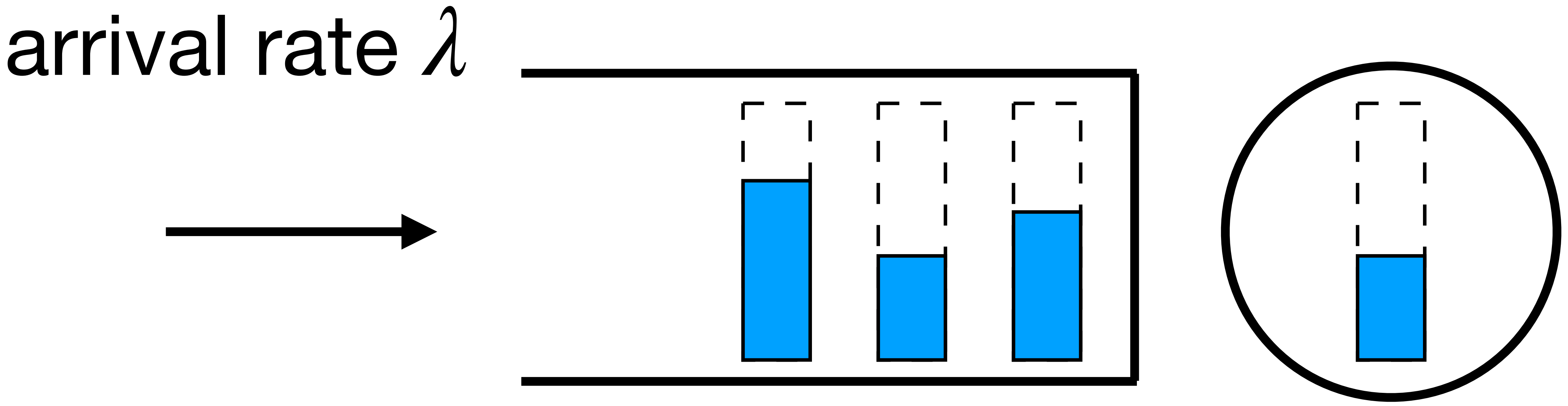


- **Shortest Remaining Processing Time (SRPT)**
  - Minimizes $\mathbb{E}[N]$
  - Why: decreases the number as fast as possible

Unknown job sizes

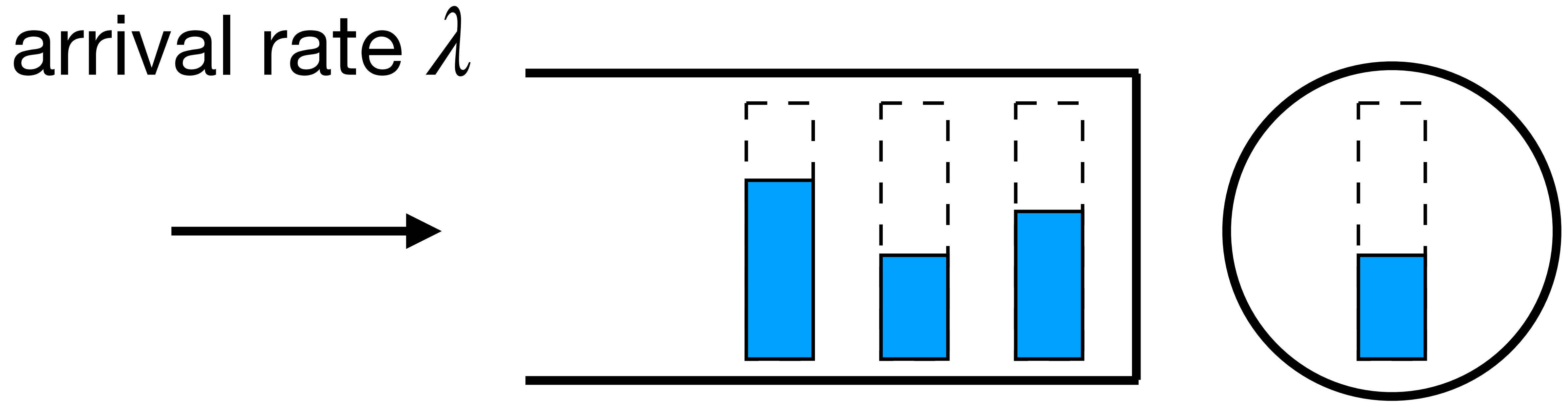


Unknown job sizes



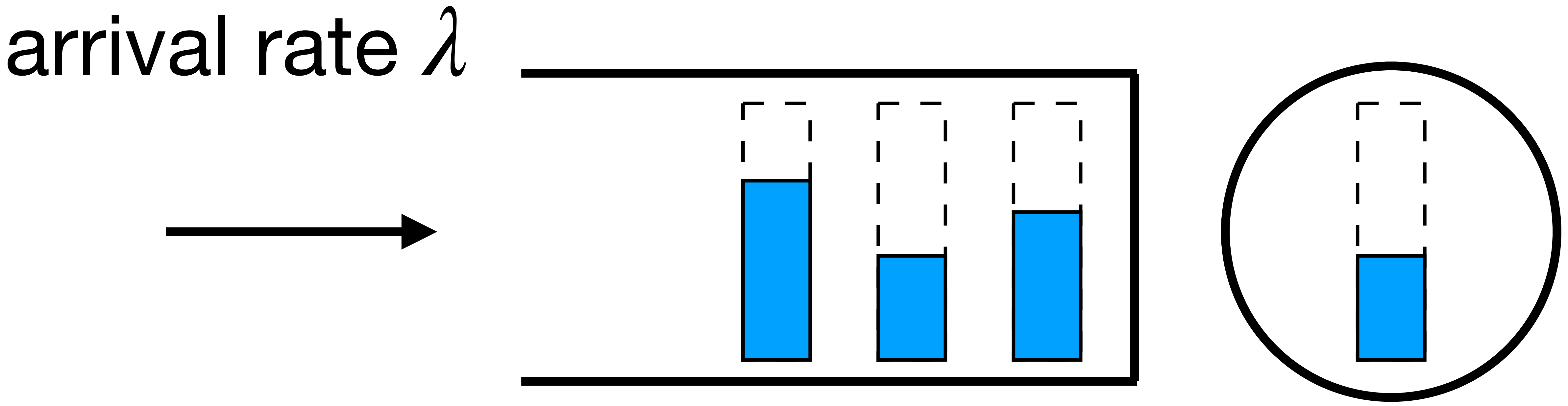
- Sampled from a known distribution

Unknown job sizes



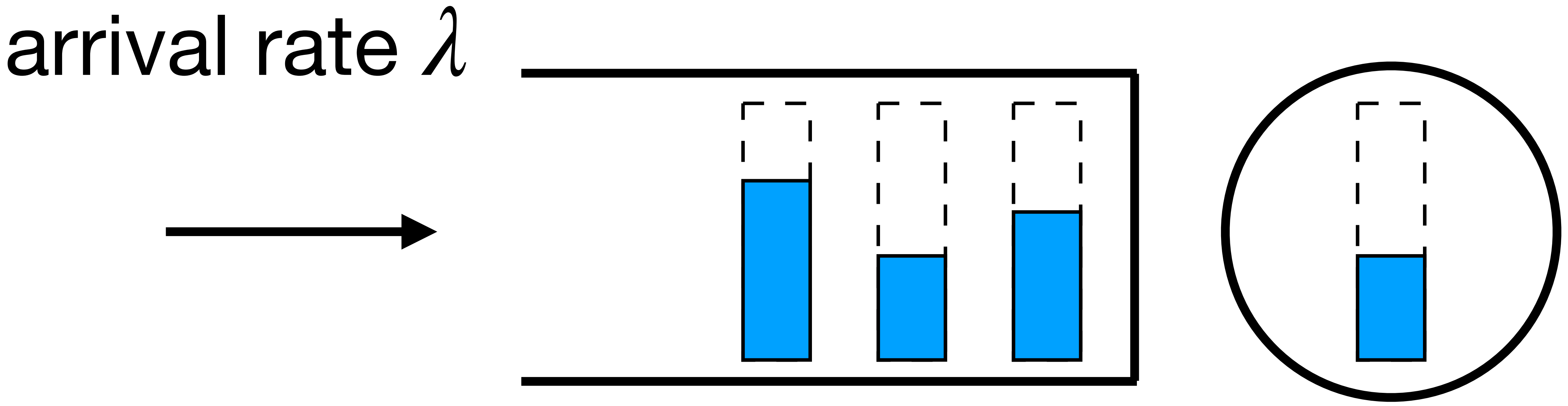
- Sampled from a known distribution
- Prioritize jobs that are *likely* to have small remaining size

Unknown job sizes

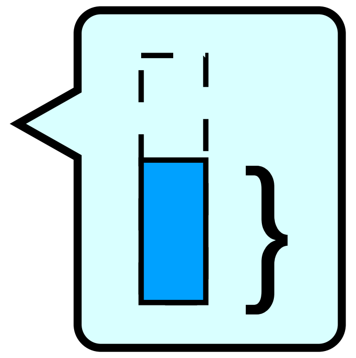


- Sampled from a known distribution
- Prioritize jobs that are *likely* to have small remaining size
- Infer remaining size from *age*

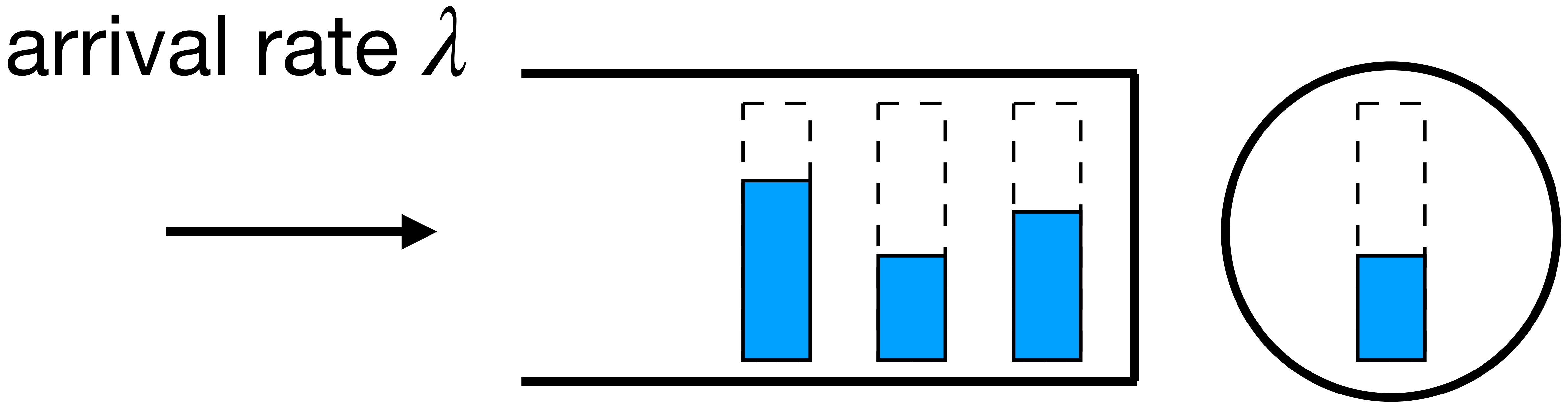
Unknown job sizes

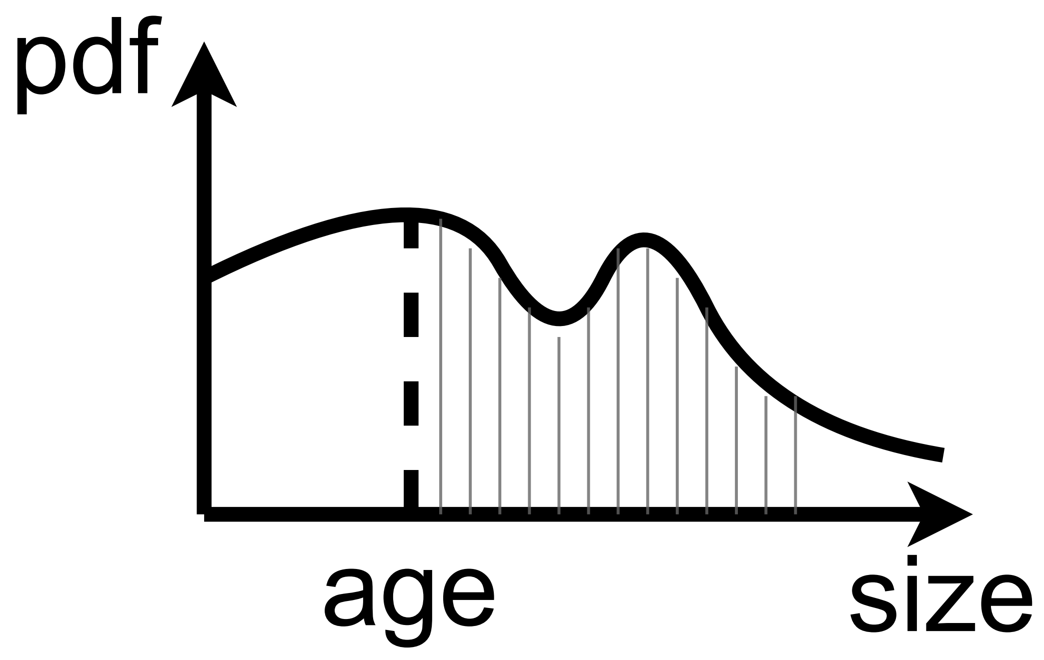


- Sampled from a known distribution
- Prioritize jobs that are *likely* to have small remaining size
- Infer remaining size from *age*

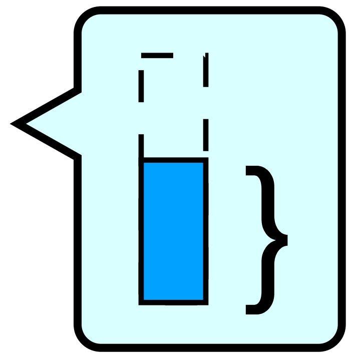


Unknown job sizes

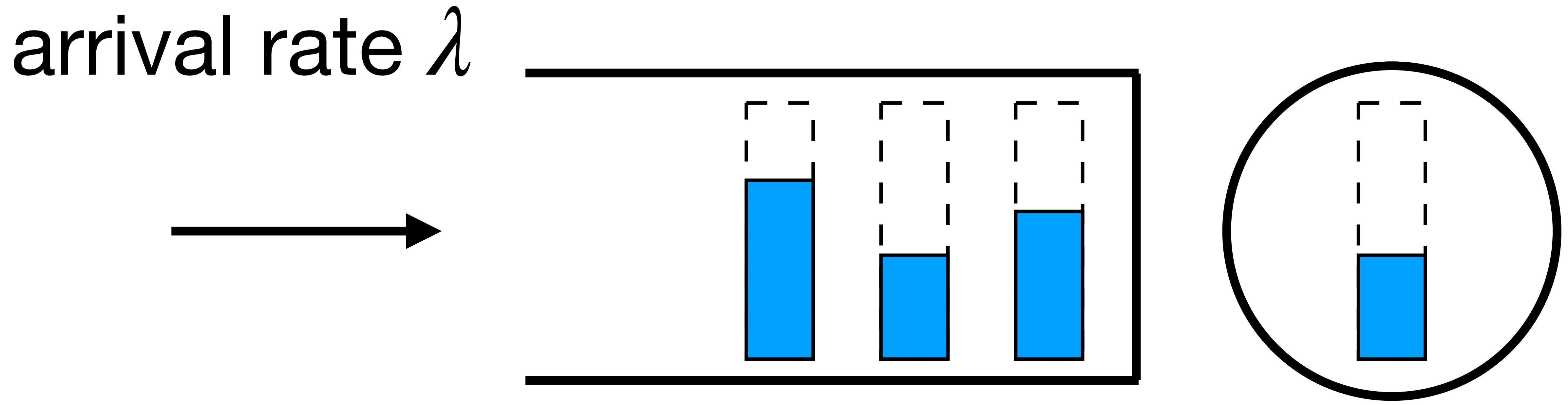




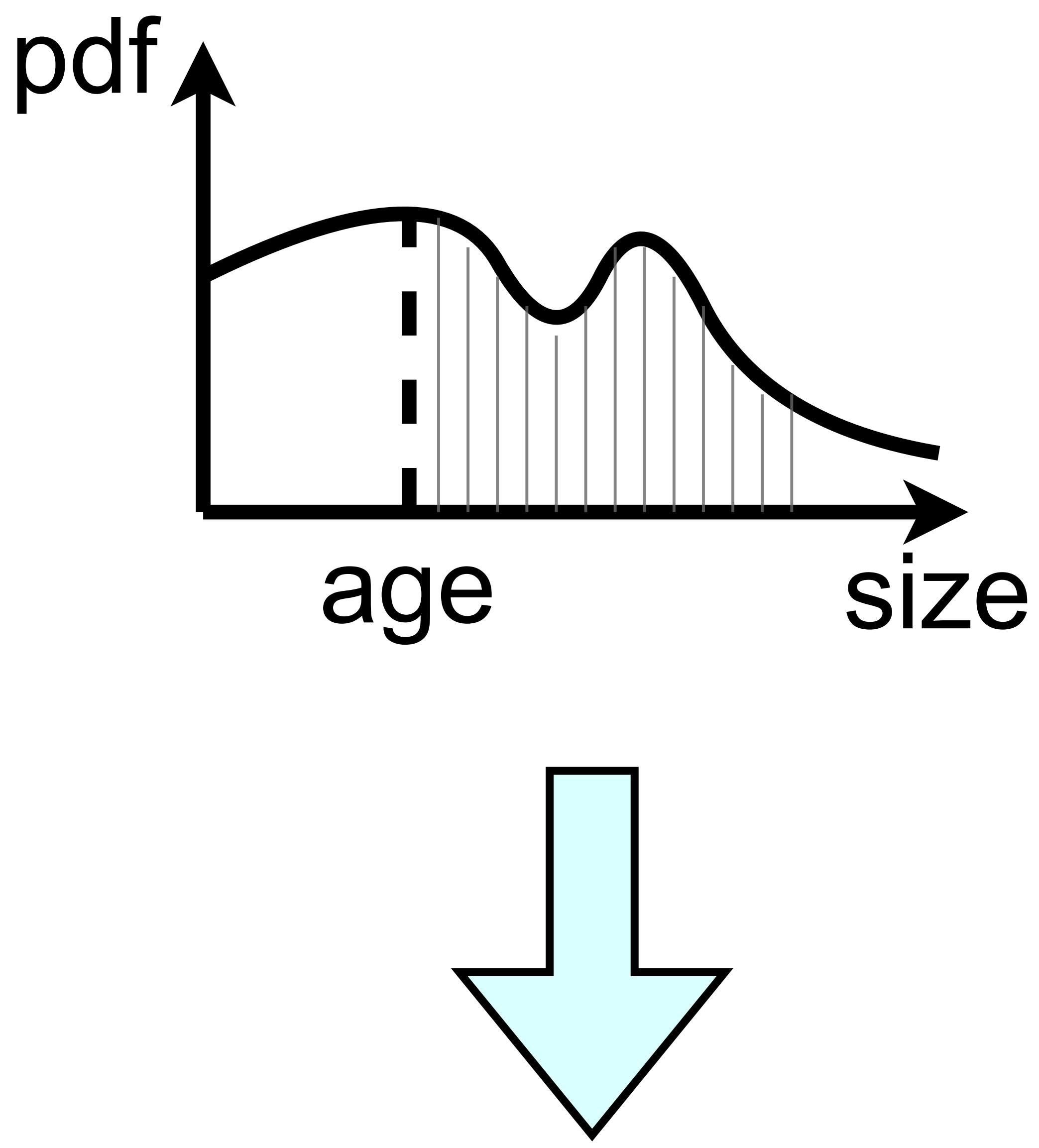
- Sampled from a known distribution
- Prioritize jobs that are *likely* to have small remaining size
- Infer remaining size from *age*



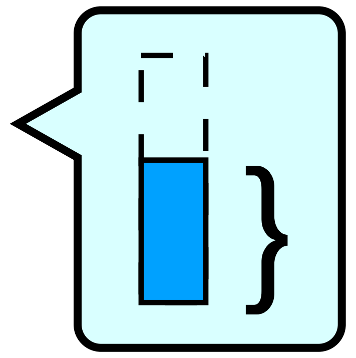
Unknown job sizes



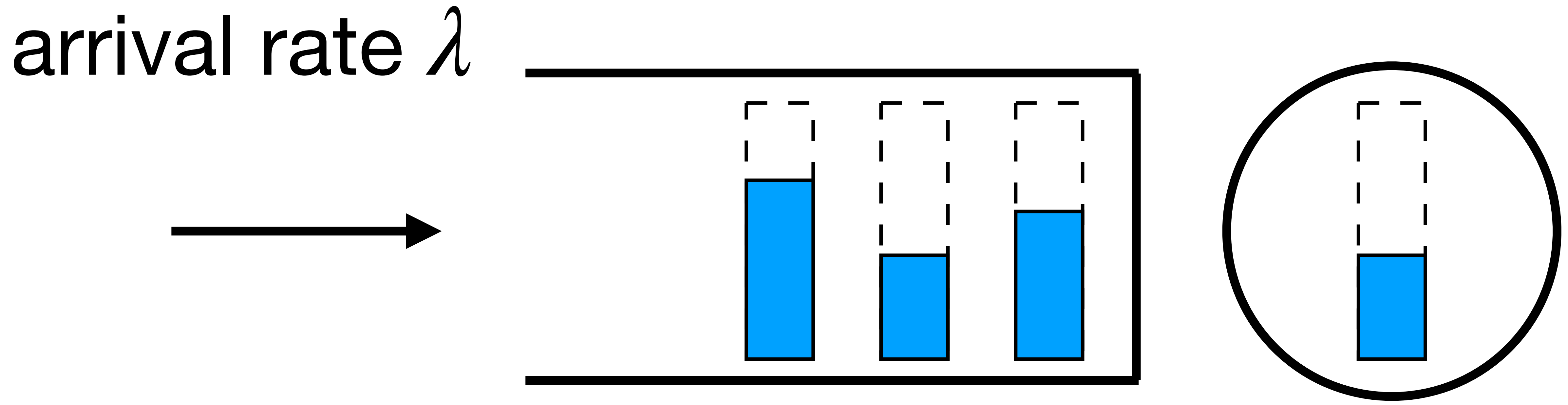
arrival rate λ



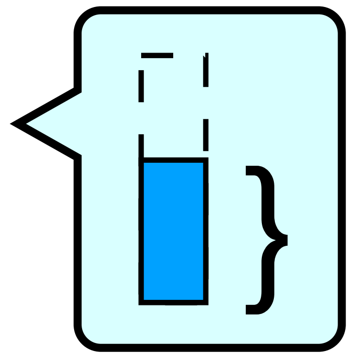
- Sampled from a known distribution
- Prioritize jobs that are *likely* to have small remaining size
- Infer remaining size from *age*

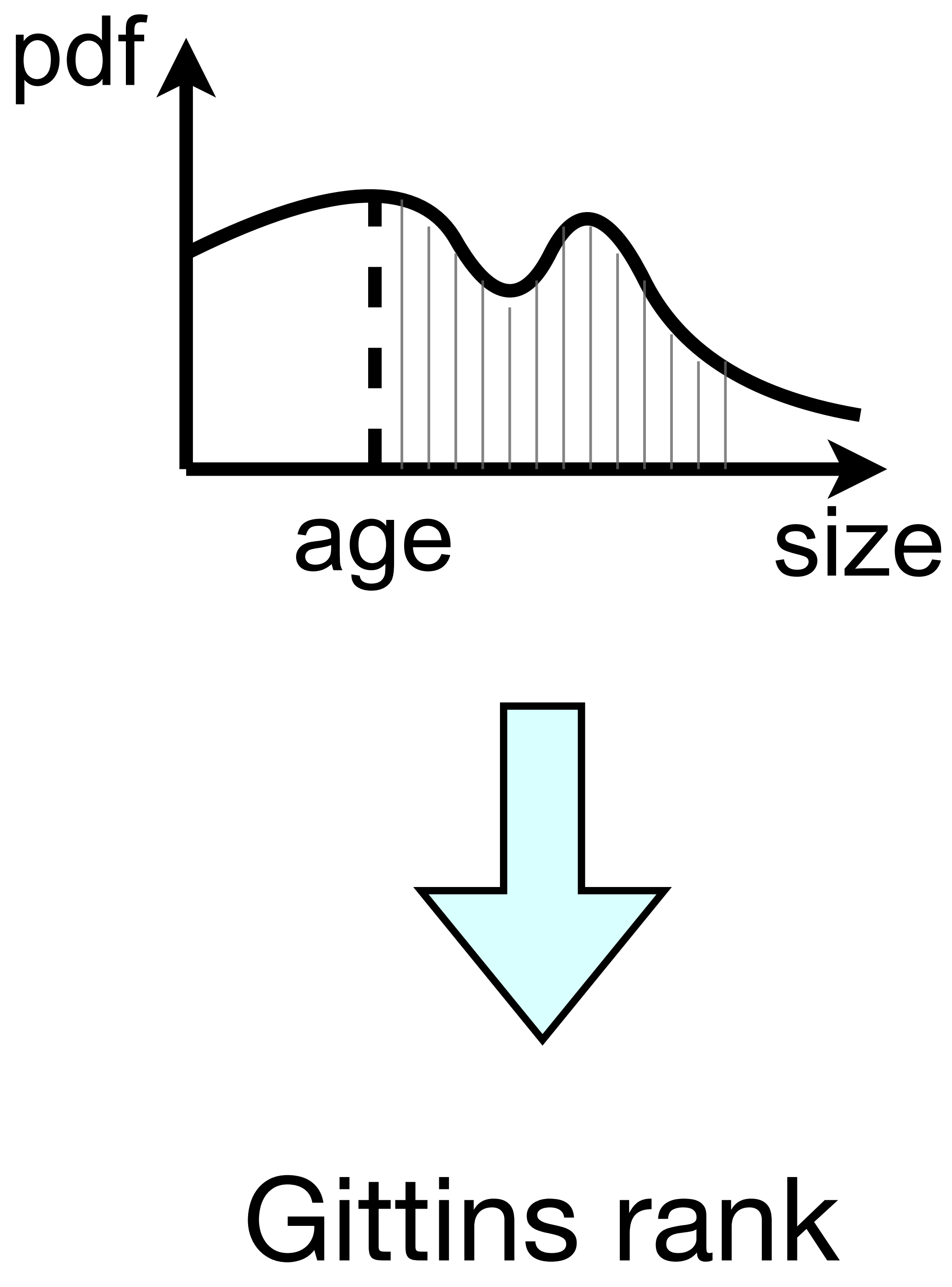


Unknown job sizes

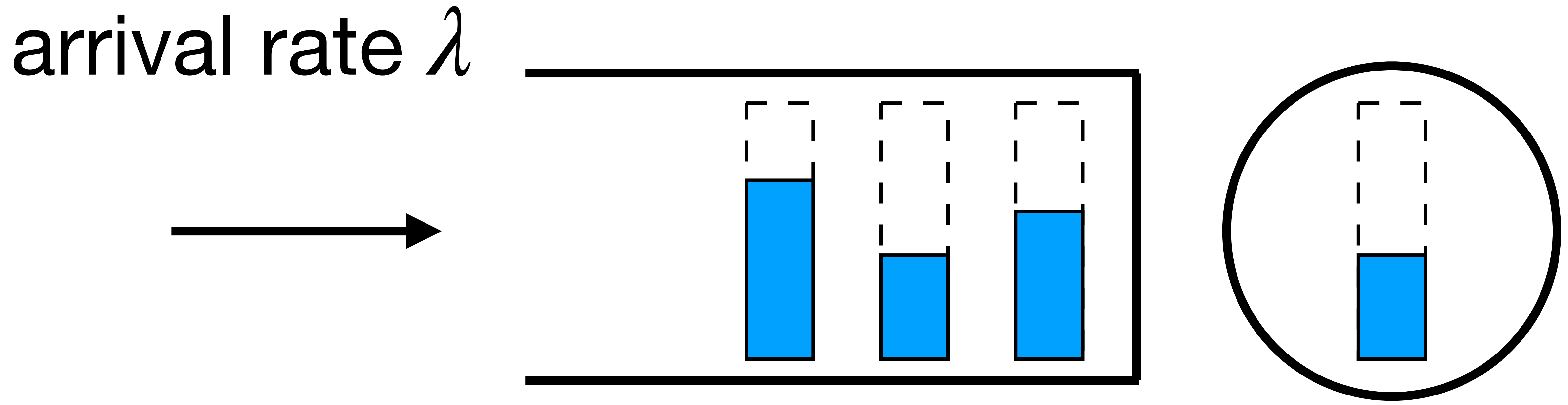


- Sampled from a known distribution
- Prioritize jobs that are *likely* to have small remaining size
- Infer remaining size from *age*

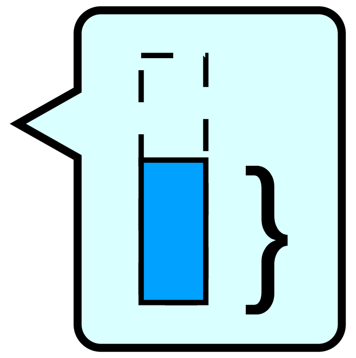




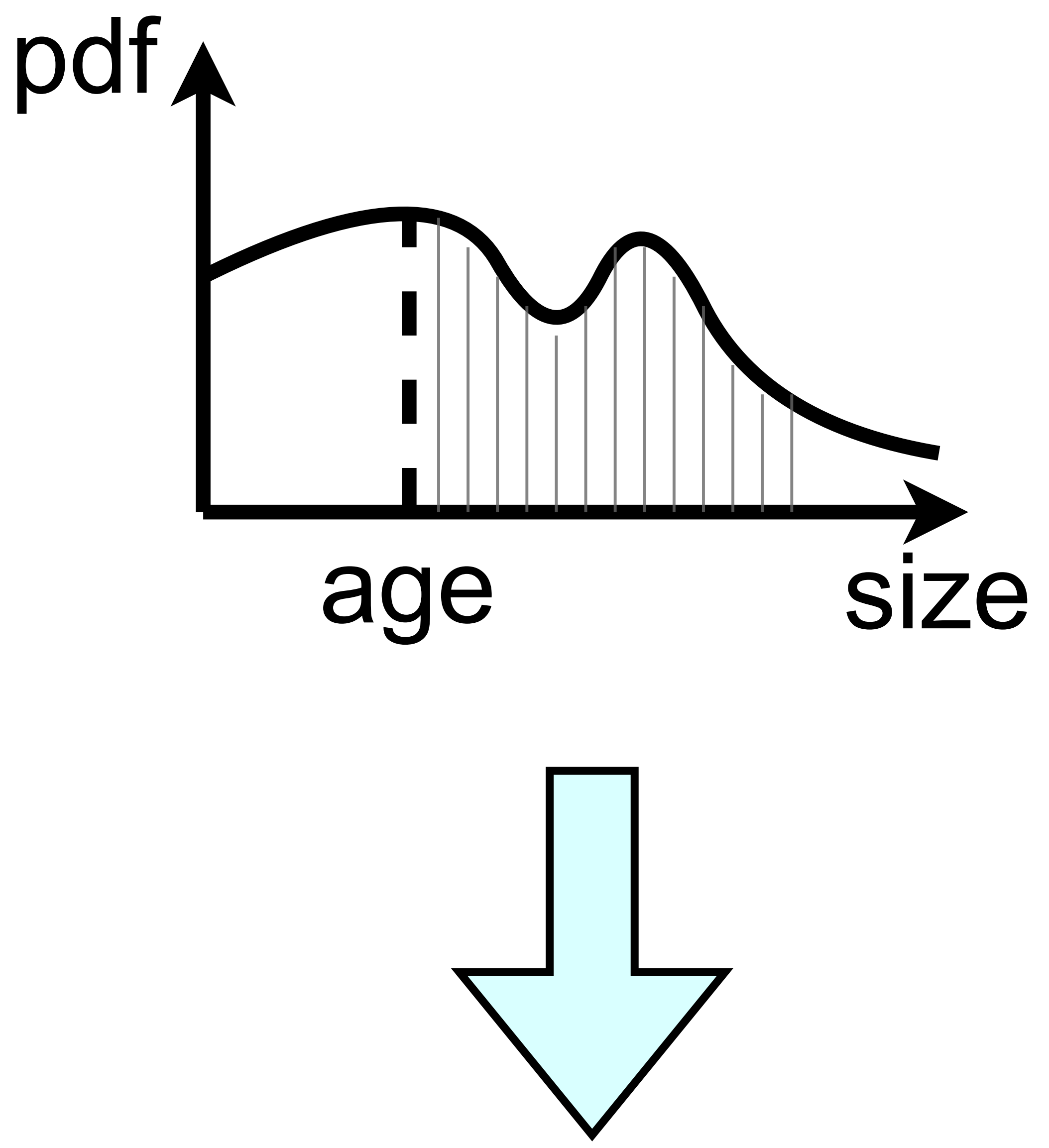
Unknown job sizes



- Sampled from a known distribution
- Prioritize jobs that are *likely* to have small remaining size
- Infer remaining size from *age*

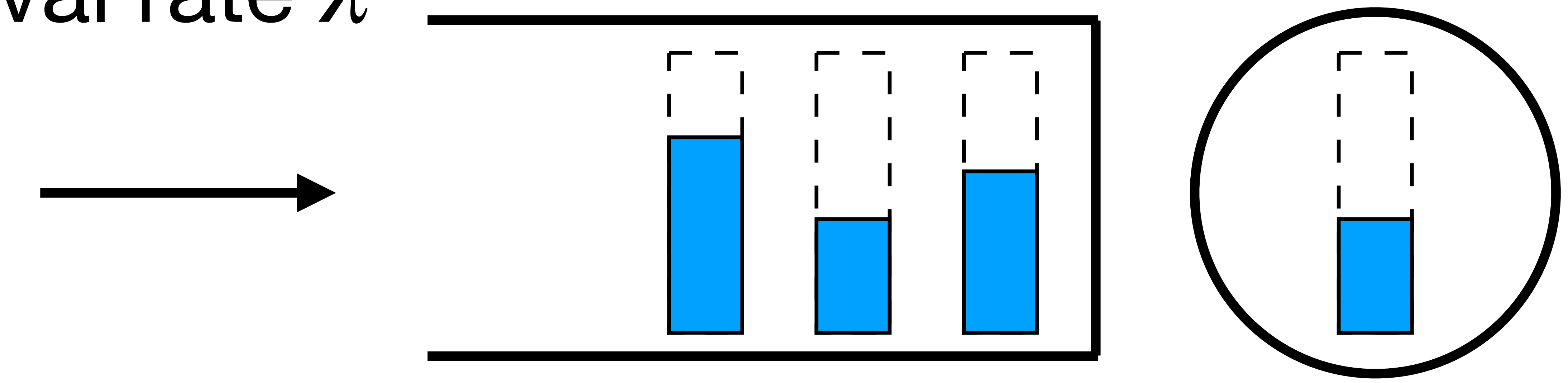


- **Gittins policy**

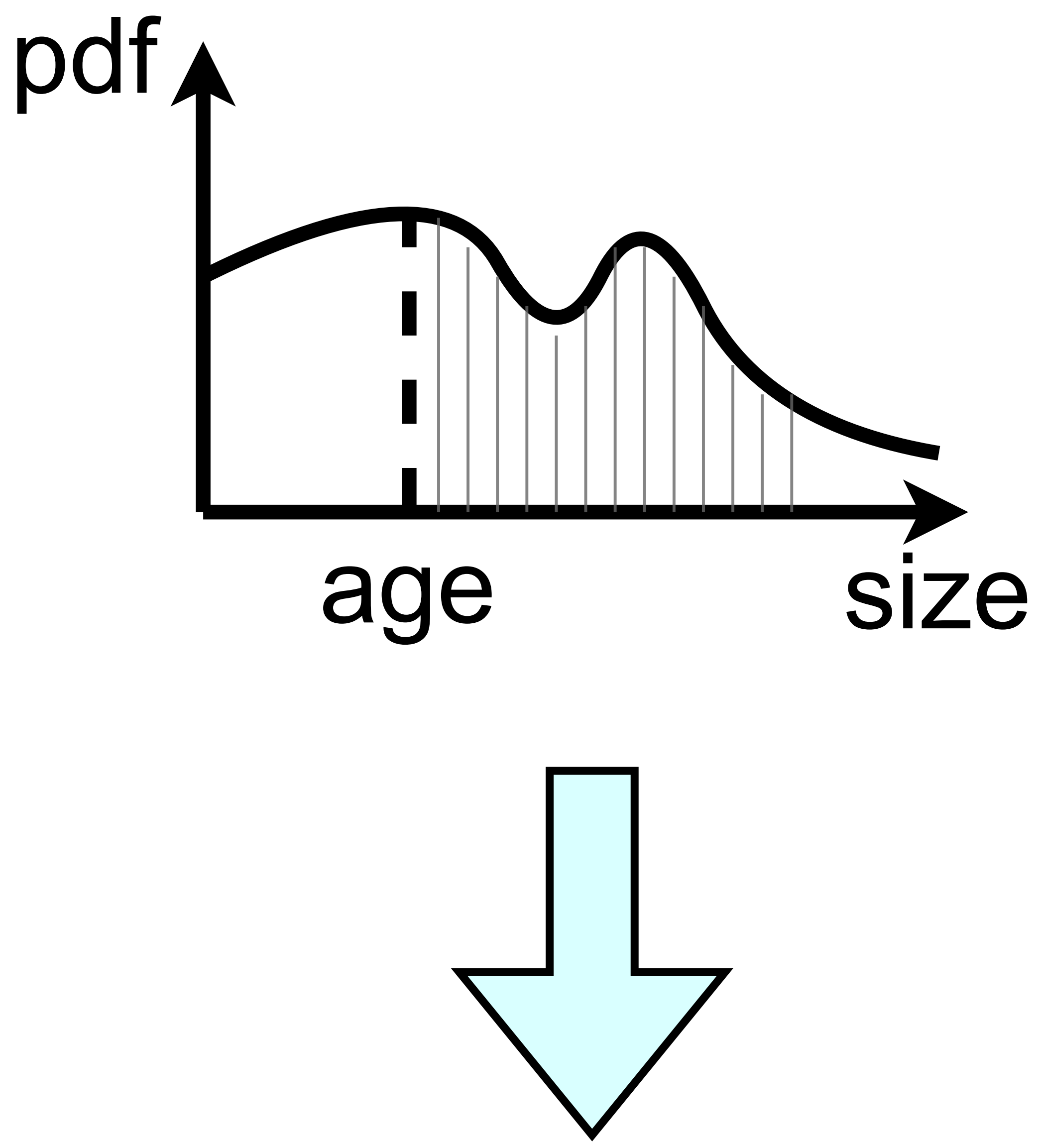


Gittins rank

Unknown job sizes

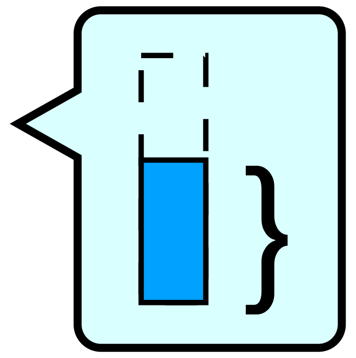


arrival rate λ



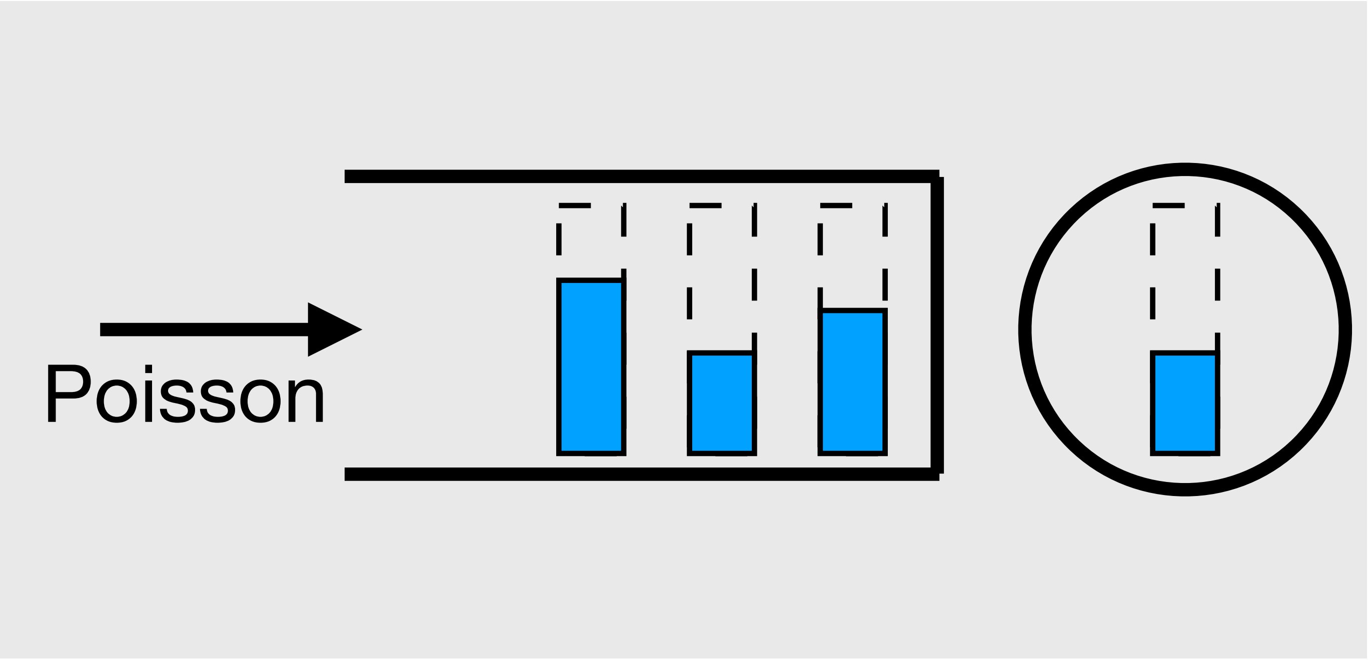
Gittins rank

- Sampled from a known distribution
- Prioritize jobs that are *likely* to have small remaining size
- Infer remaining size from *age*

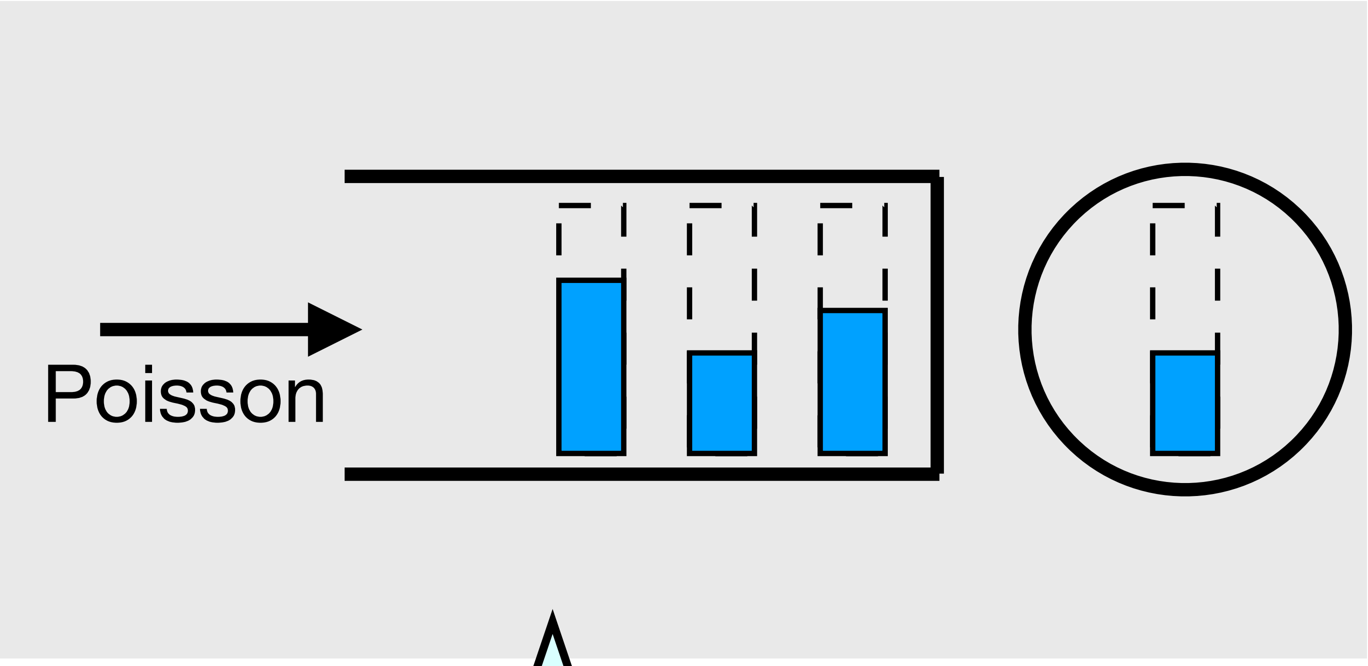


- **Gittins policy**
- Optimal if arrivals are Poisson (M/G/1 system)

M/G/1



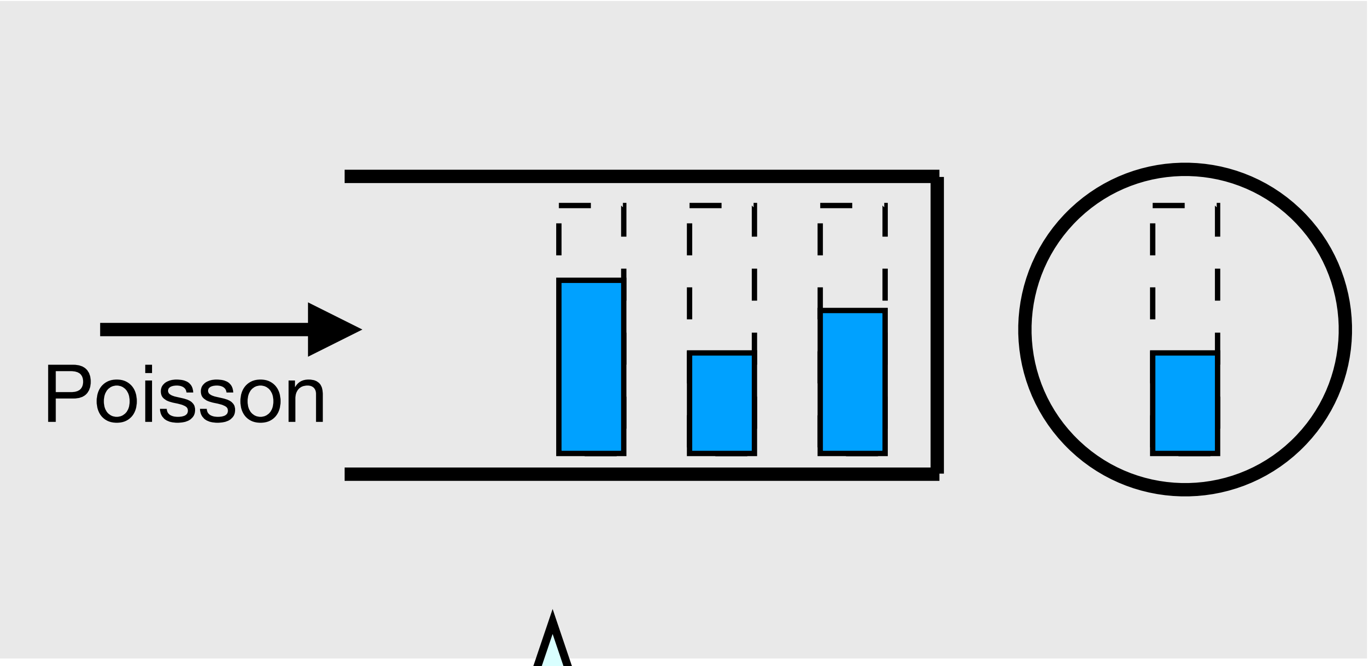
M/G/1



[Gittins79]:
Gittins policy optimal

Complicated systems

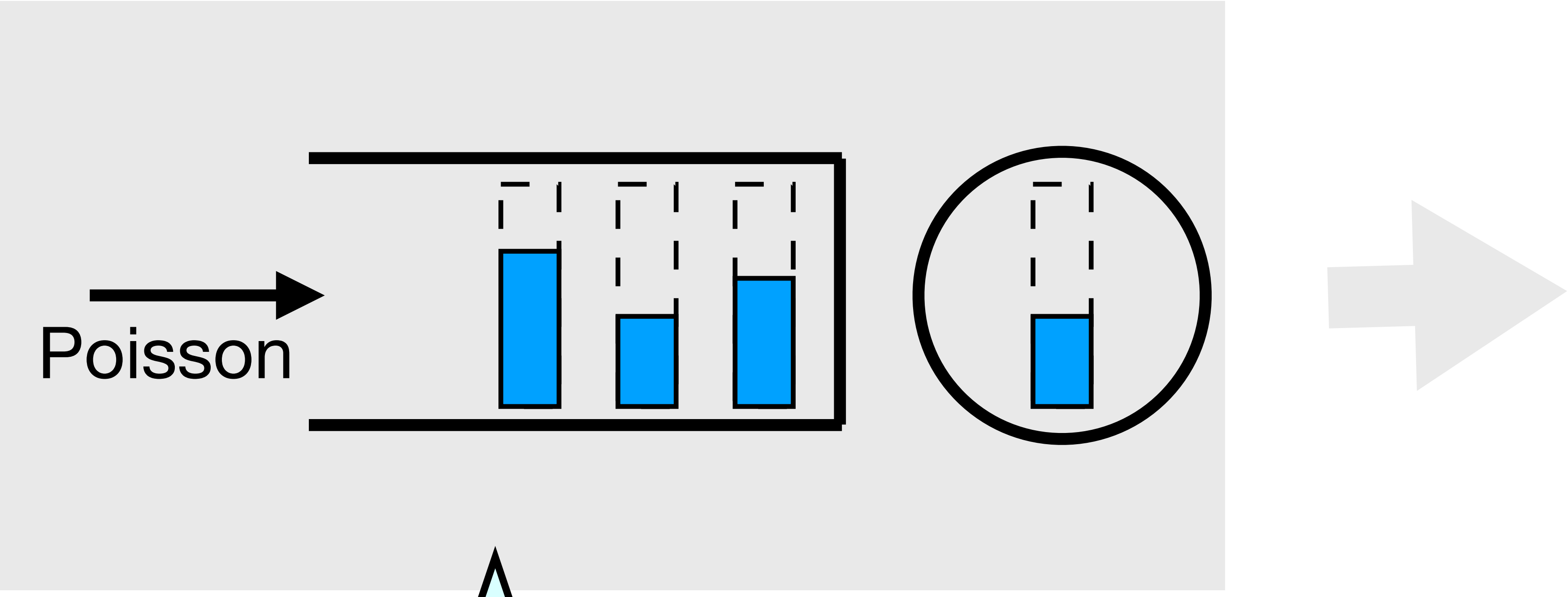
M/G/1



[Gittins79]:
Gittins policy optimal

Complicated systems

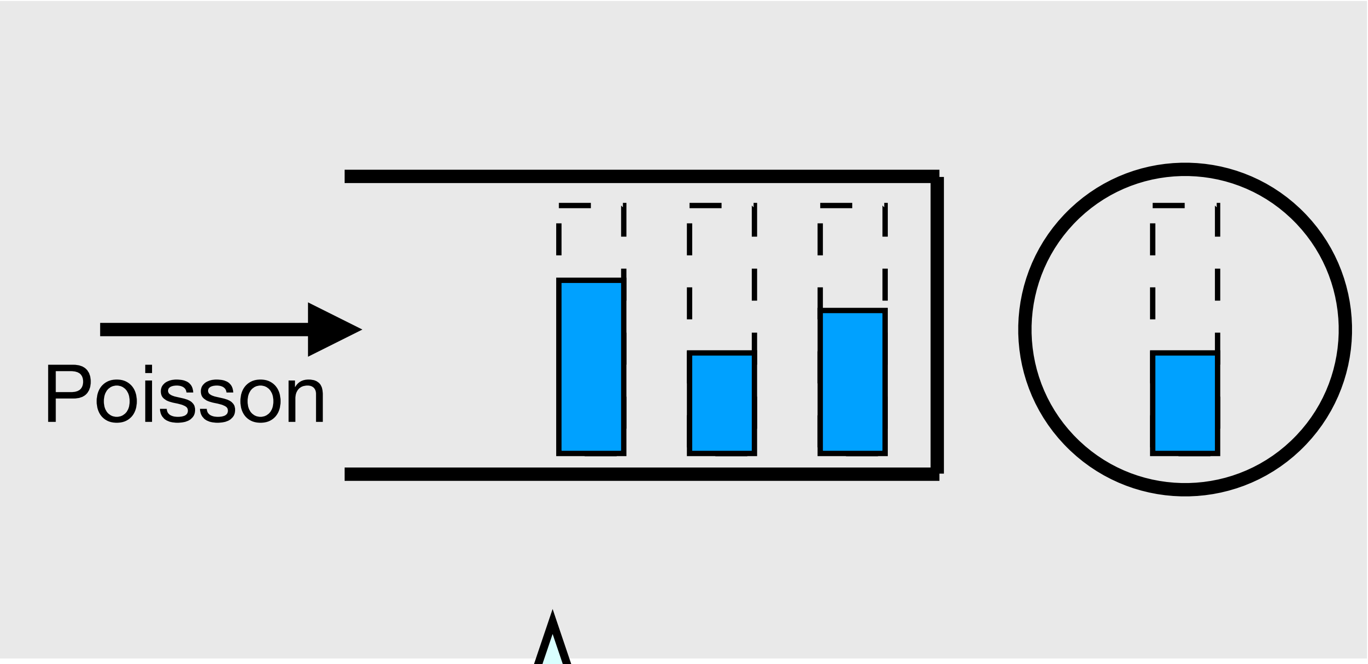
M/G/1



[Gittins79]:
Gittins policy optimal

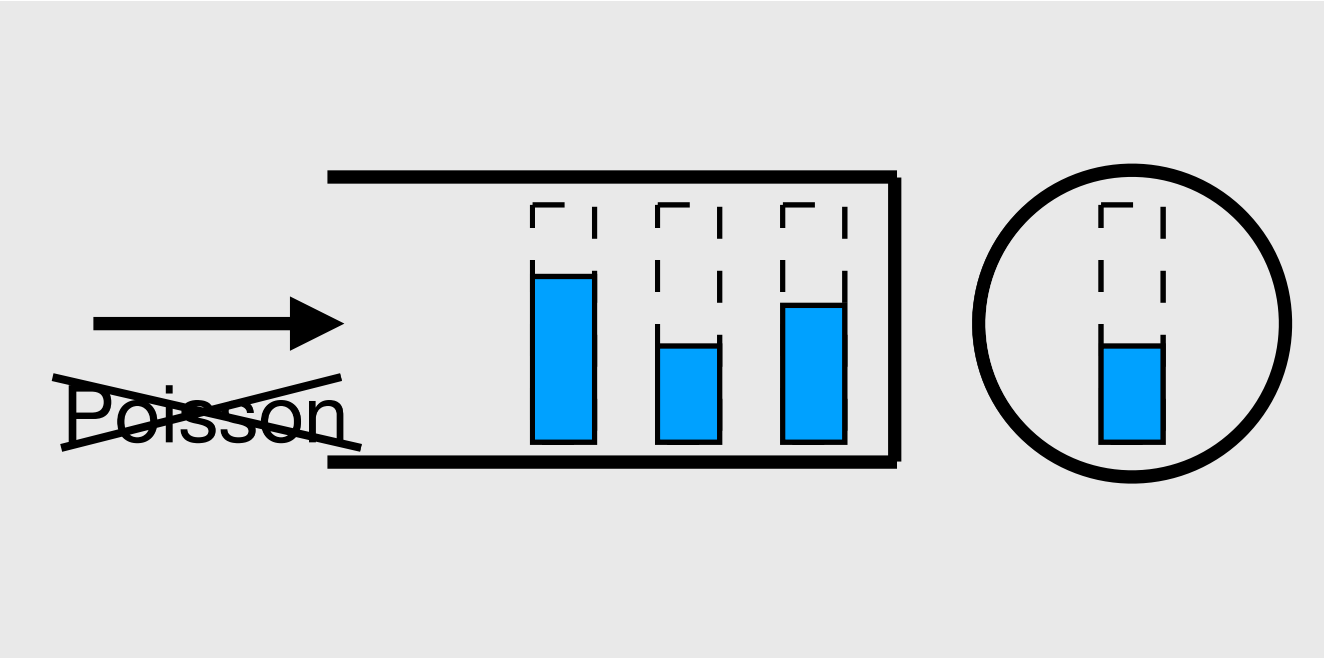
Complicated systems

M/G/1

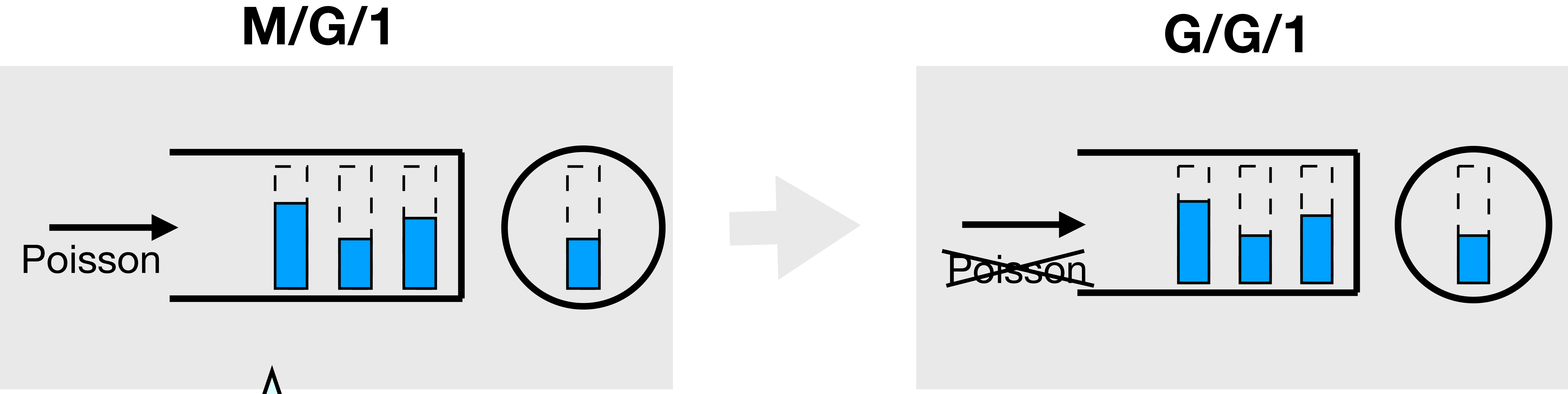


[Gittins79]:
Gittins policy optimal

G/G/1



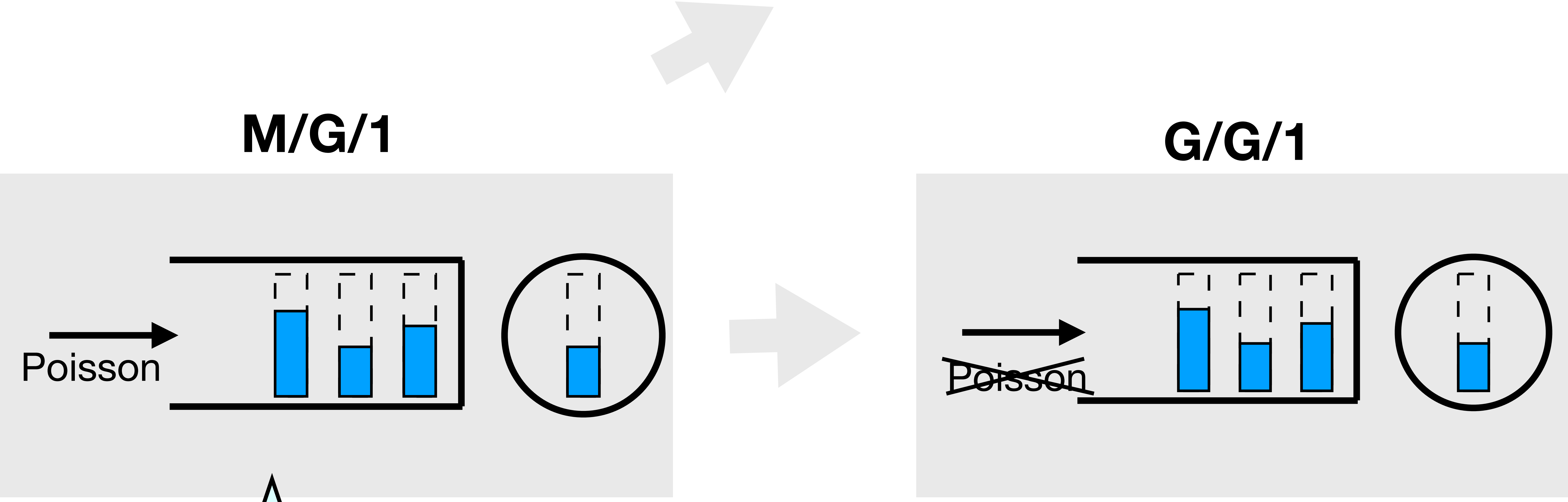
Complicated systems



[Gittins79]:
Gittins policy optimal

Optimal unknown

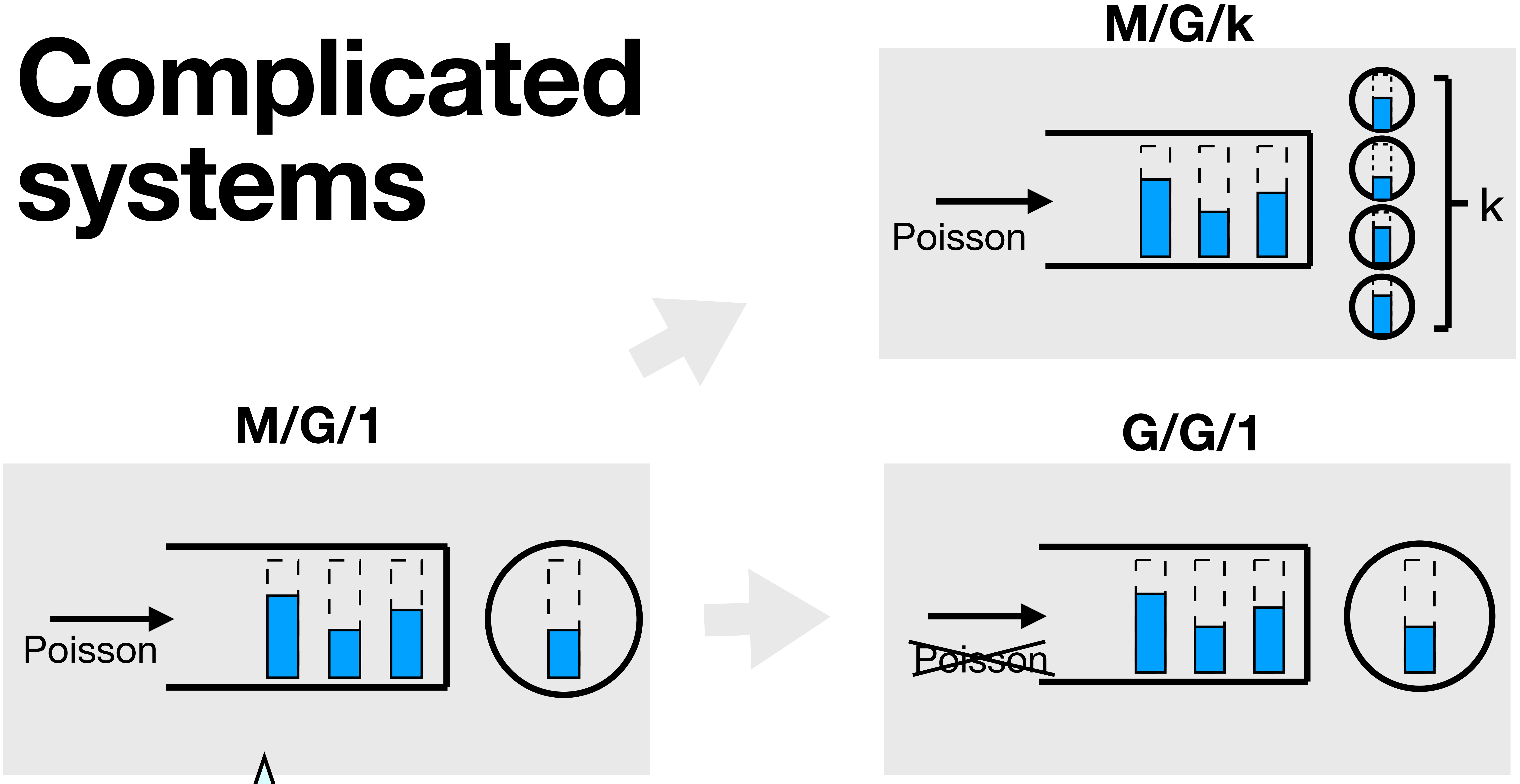
Complicated systems



[Gittins79]:
Gittins policy optimal

Optimal unknown

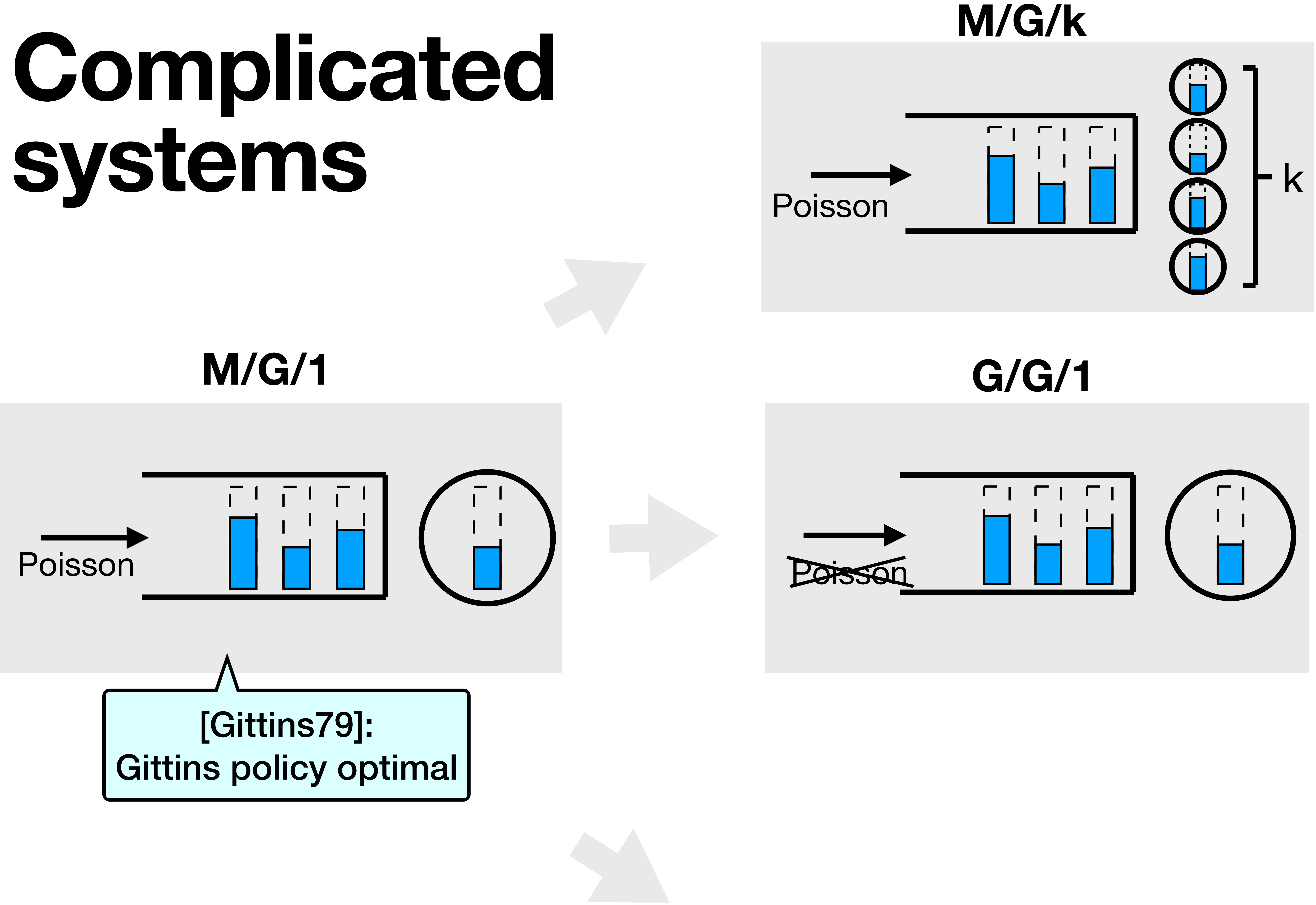
Complicated systems



[Gittins79]:
Gittins policy optimal

Optimal
unknown

Complicated systems

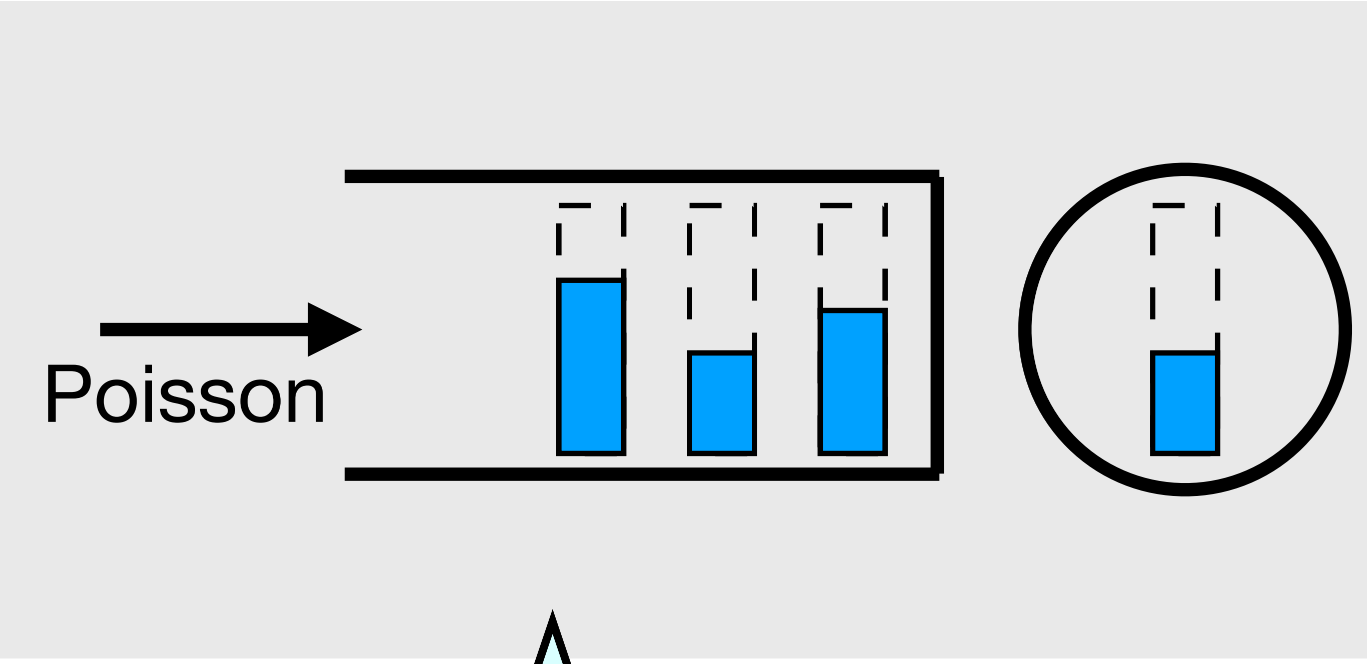


[Gittins79]:
Gittins policy optimal

Optimal unknown

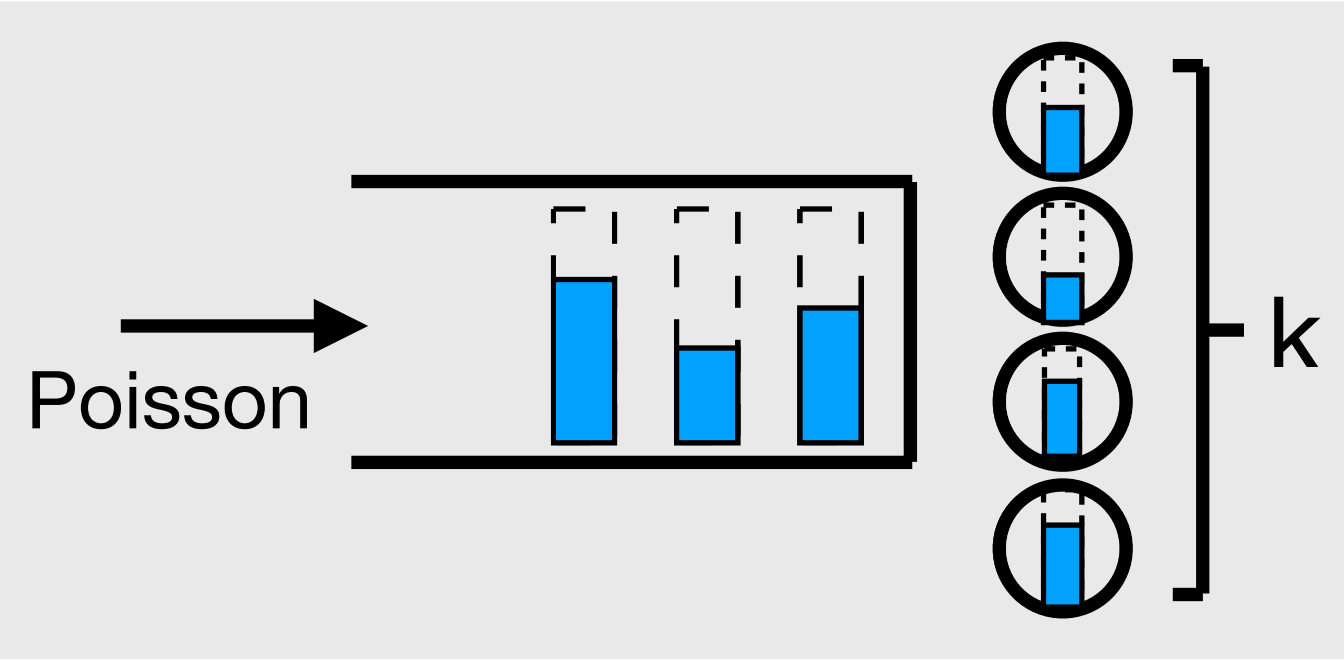
Complicated systems

M/G/1

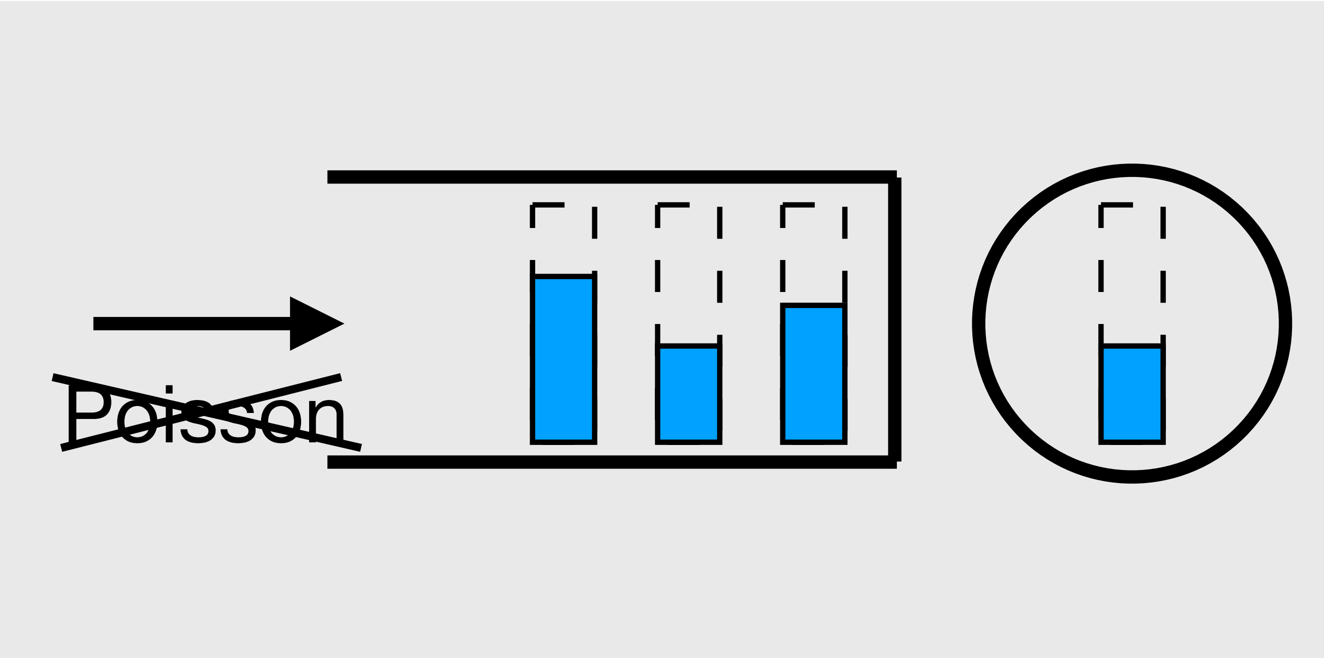


[Gittins79]:
Gittins policy optimal

M/G/k

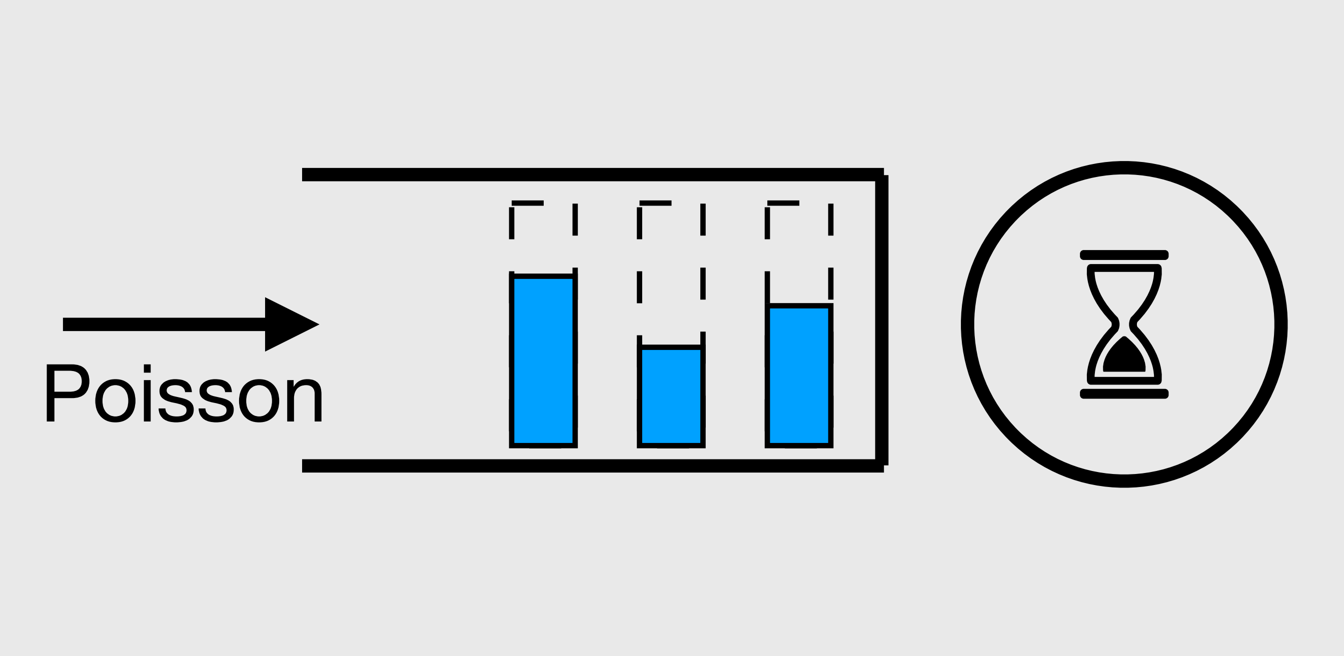


G/G/1

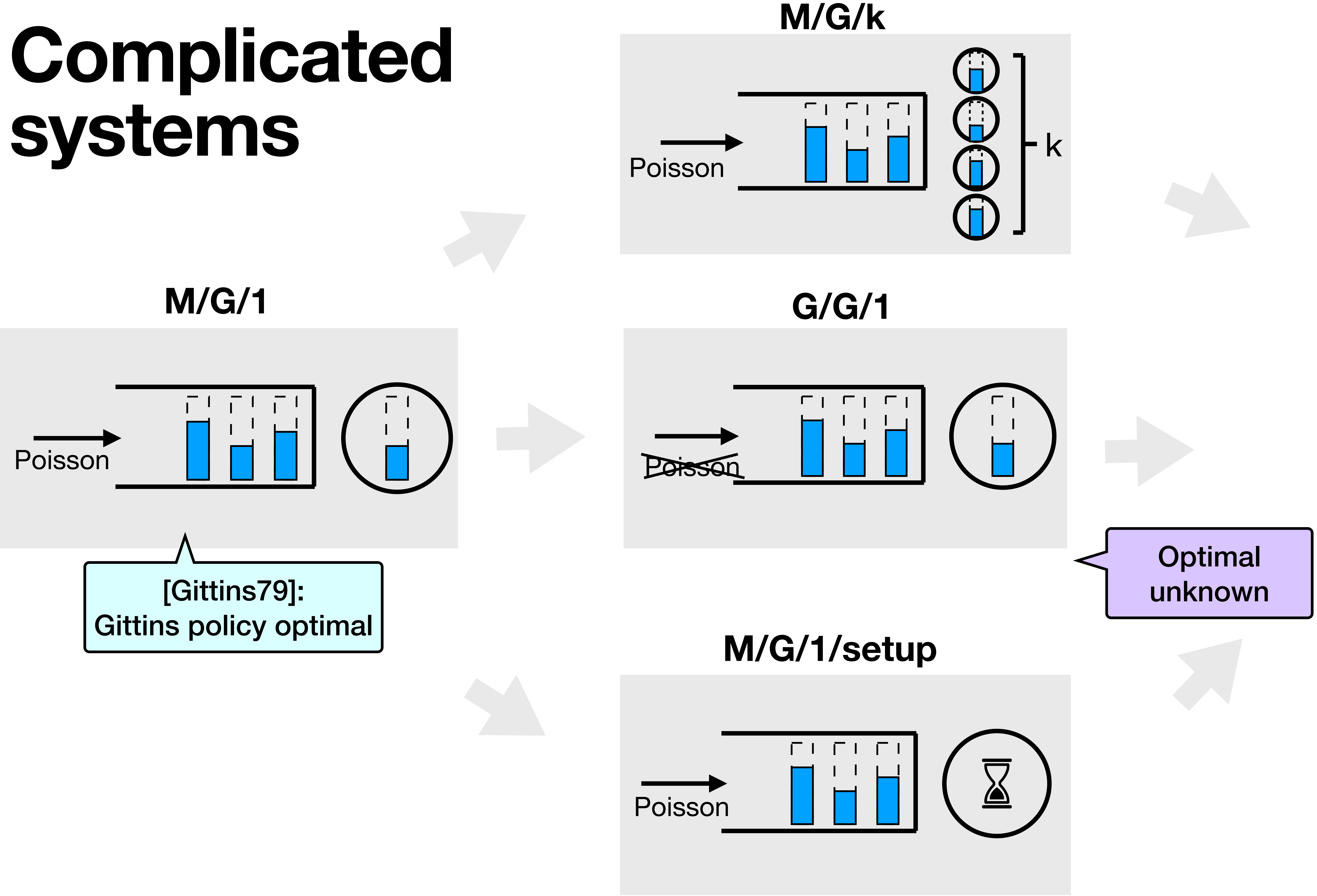


Optimal
unknown

M/G/1/setup



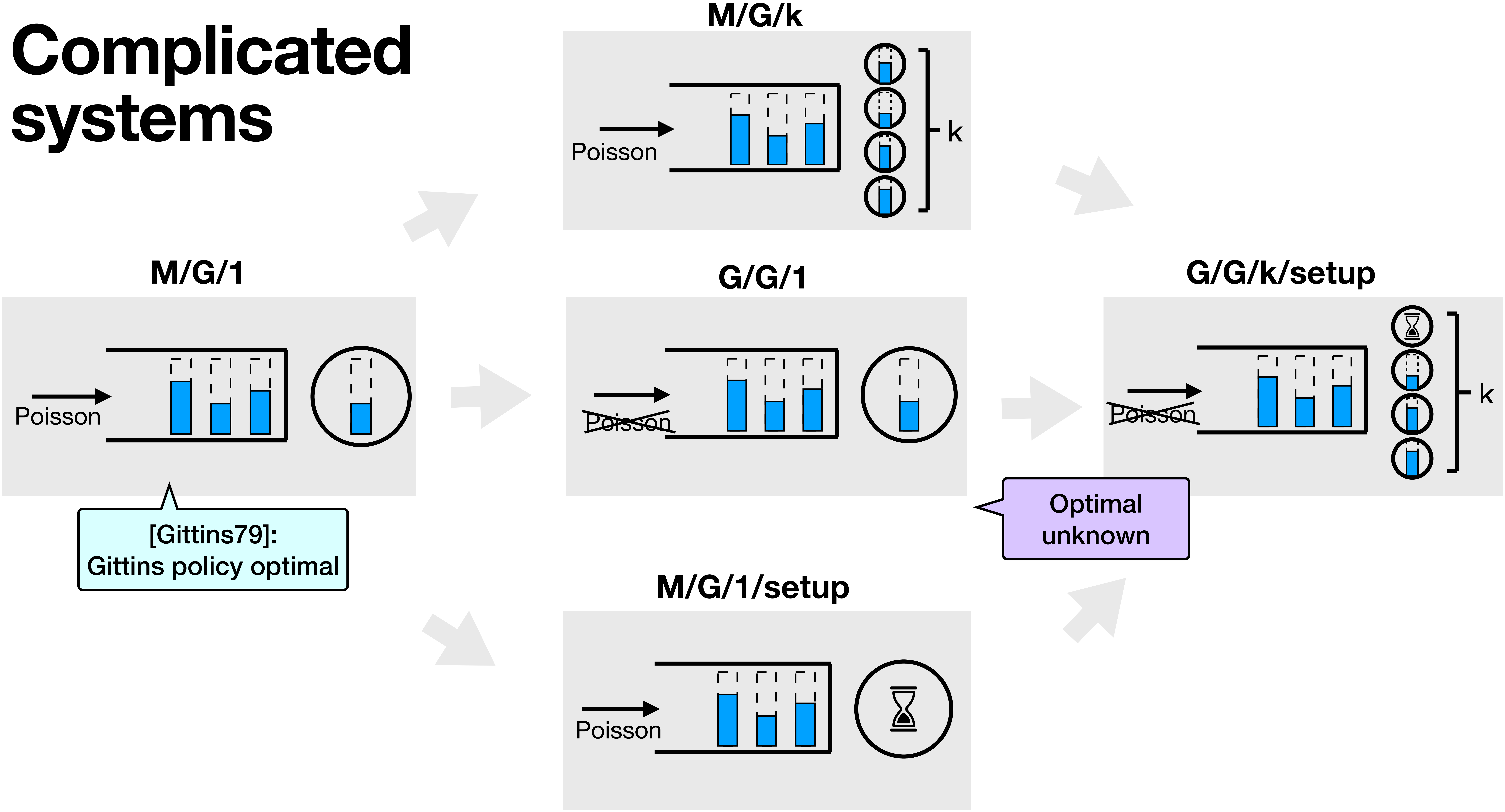
Complicated systems



[Gittins79]:
Gittins policy optimal

Optimal
unknown

Complicated systems



M/G/1

Poisson

[Gittins79]:
Gittins policy optimal

M/G/k

Poisson

k

G/G/1

~~Poisson~~

M/G/1/setup

Poisson

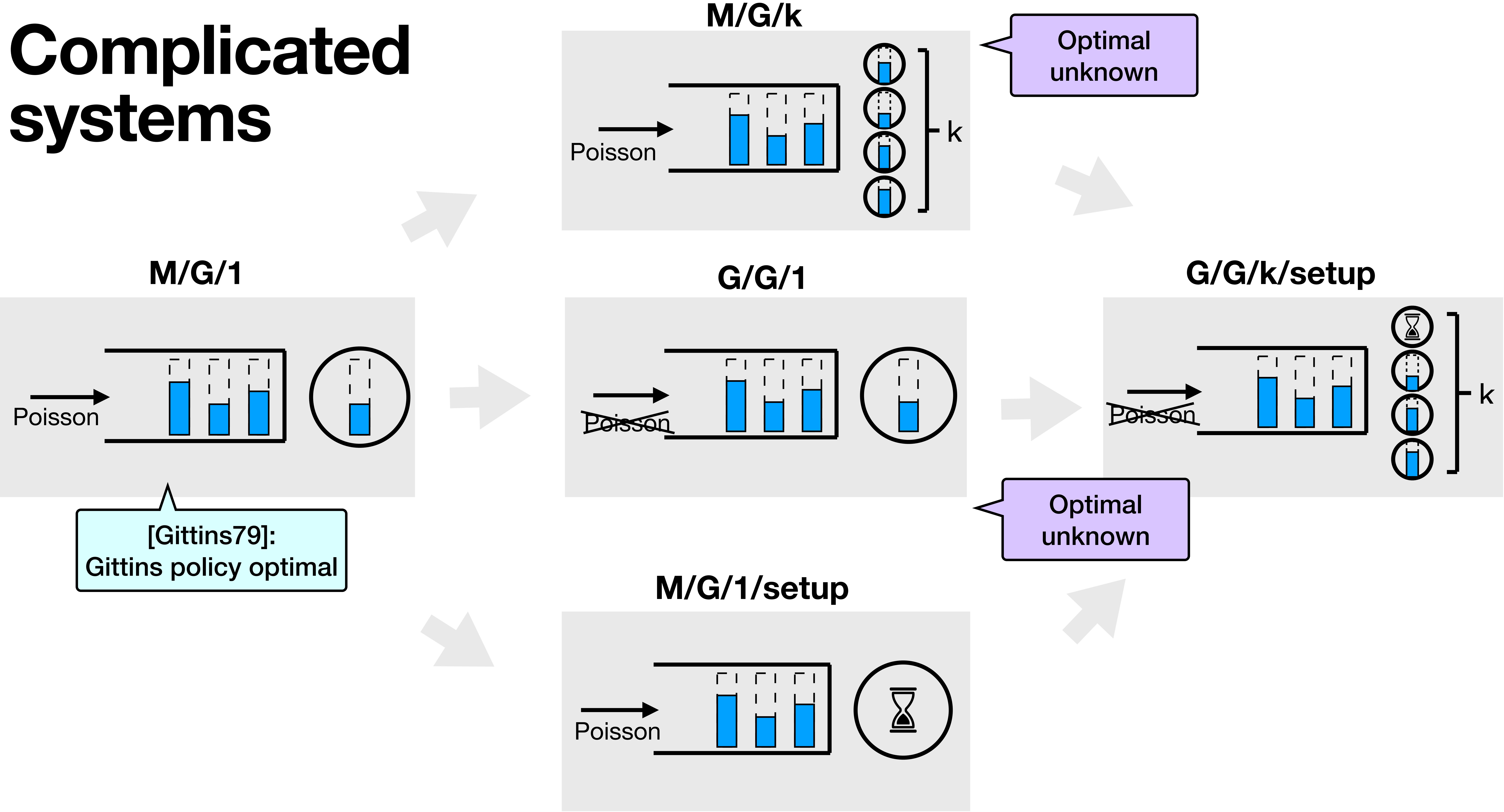
G/G/k/setup

~~Poisson~~

k

Optimal
unknown

Complicated systems

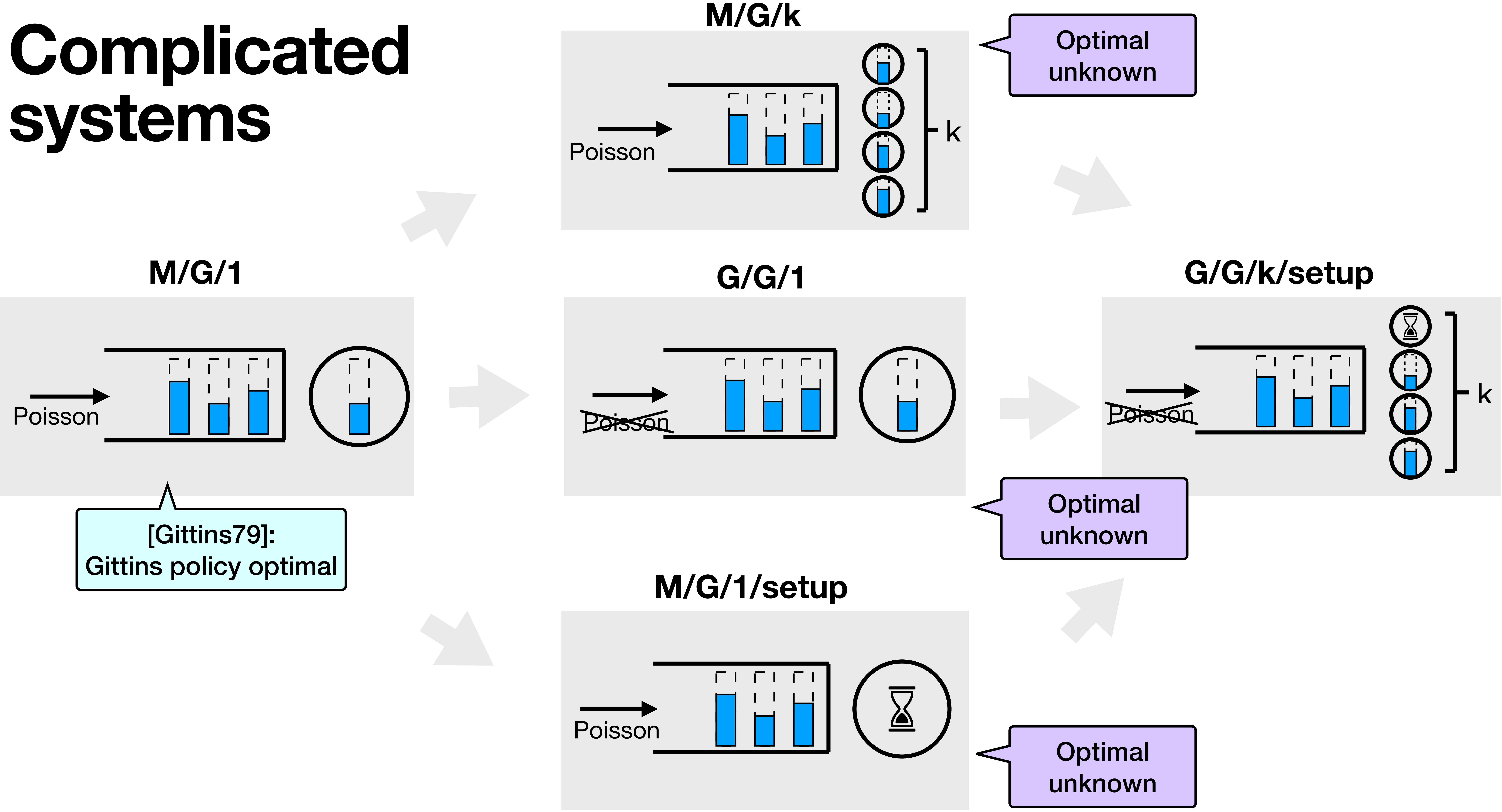


[Gittins79]:
Gittins policy optimal

Optimal unknown

Optimal unknown

Complicated systems



M/G/1

[Gittins79]:
Gittins policy optimal

M/G/k

Optimal unknown

G/G/1

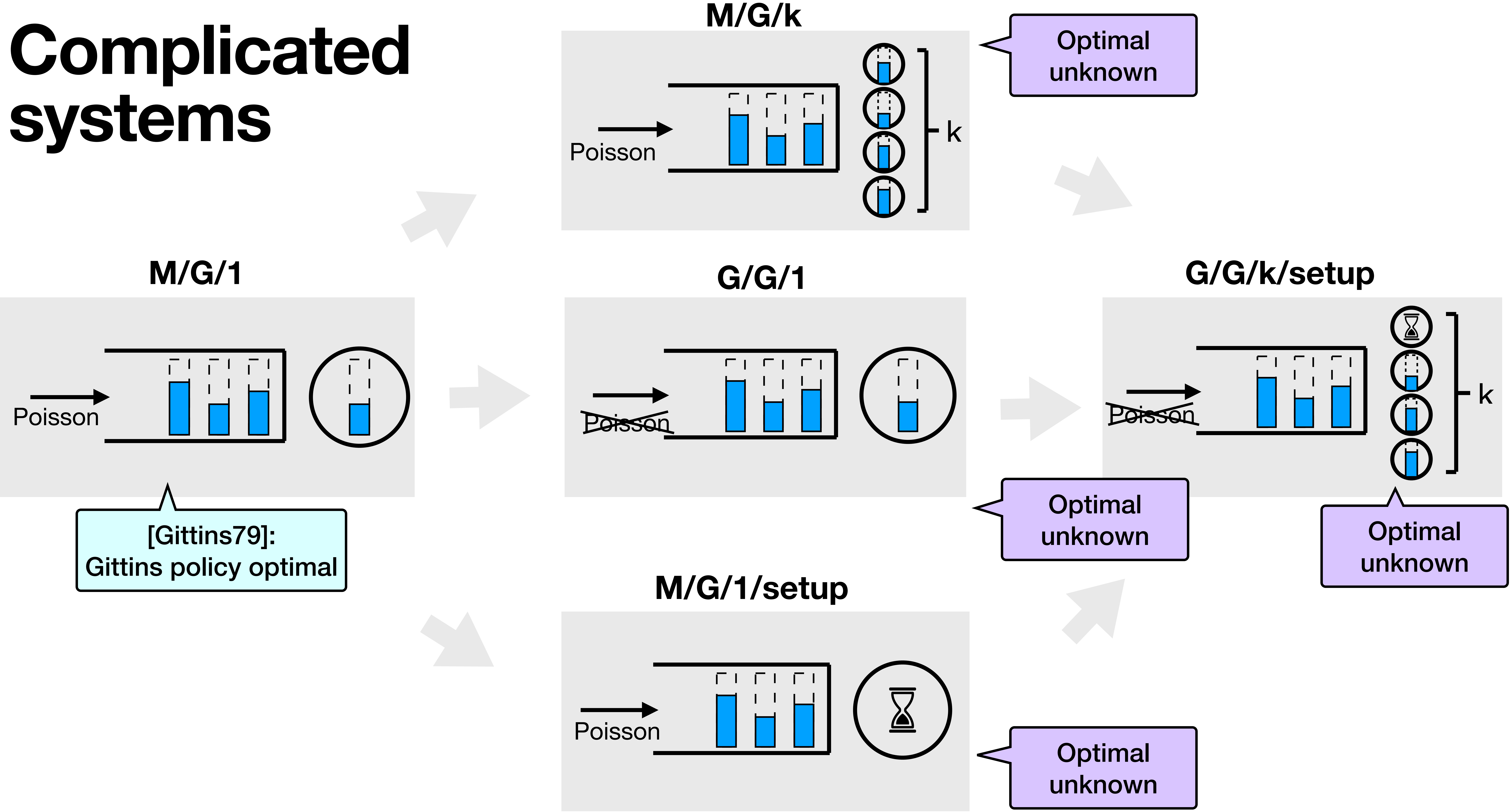
Optimal unknown

M/G/1/setup

Optimal unknown

G/G/k/setup

Complicated systems



M/G/1

[Gittins79]:
Gittins policy optimal

M/G/k

Optimal unknown

G/G/1

Optimal unknown

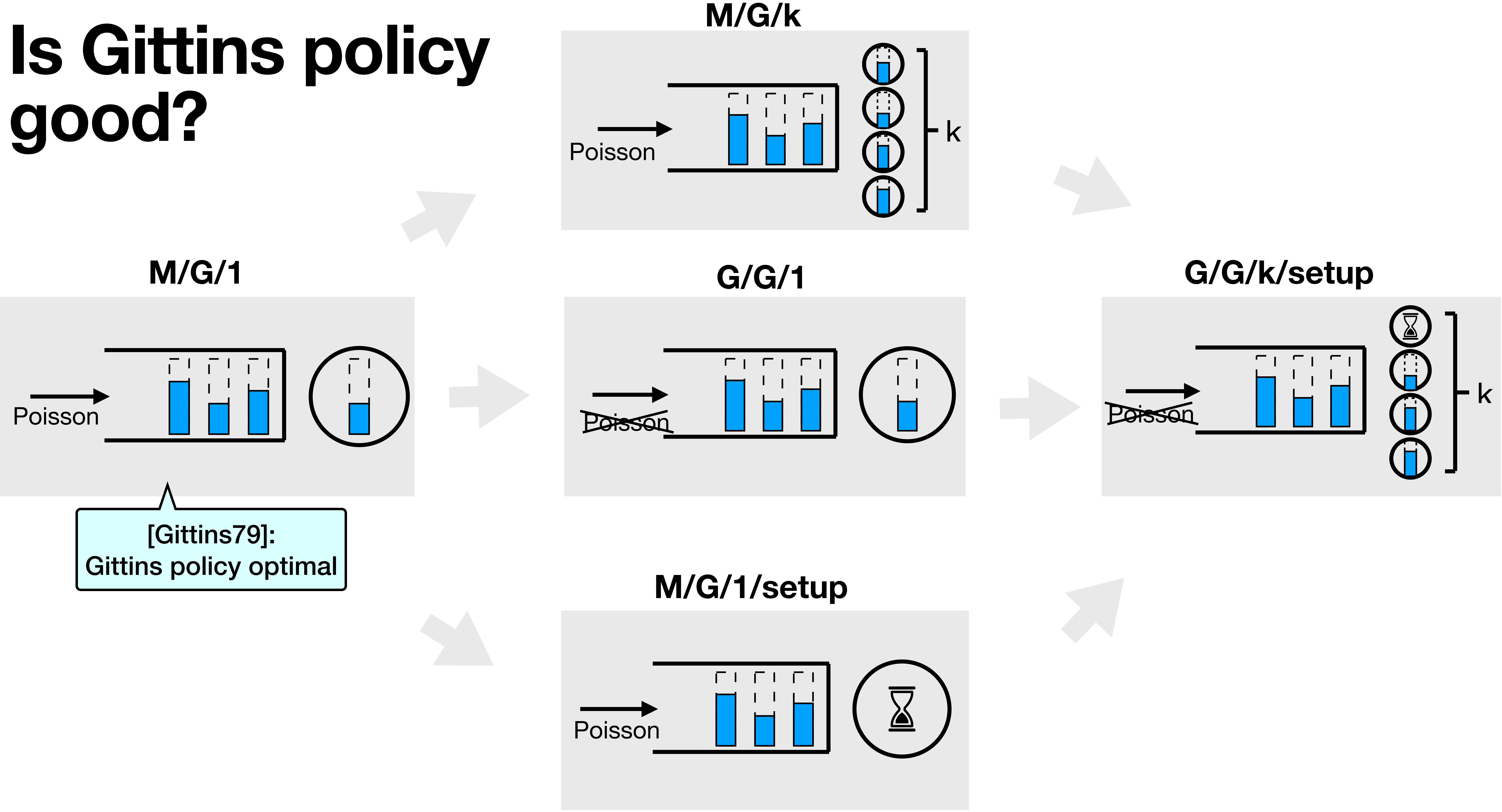
M/G/1/setup

Optimal unknown

G/G/k/setup

Optimal unknown

Is Gittins policy good?



M/G/1

Poisson

[Gittins79]:
Gittins policy optimal

M/G/k

Poisson

k

G/G/1

~~Poisson~~

M/G/1/setup

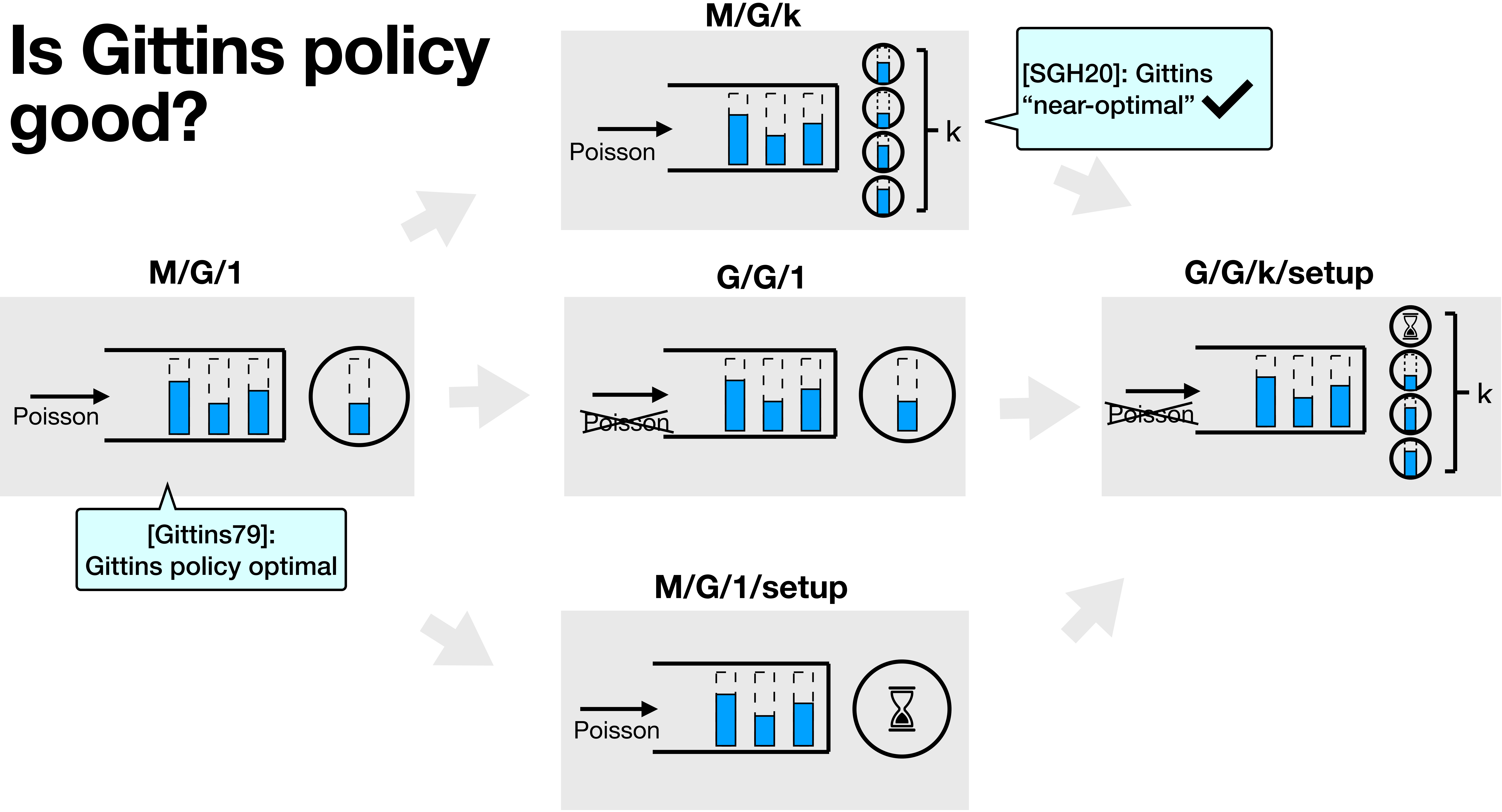
Poisson

G/G/k/setup

~~Poisson~~

k

Is Gittins policy good?


M/G/1

[Gittins79]:
Gittins policy optimal

M/G/k

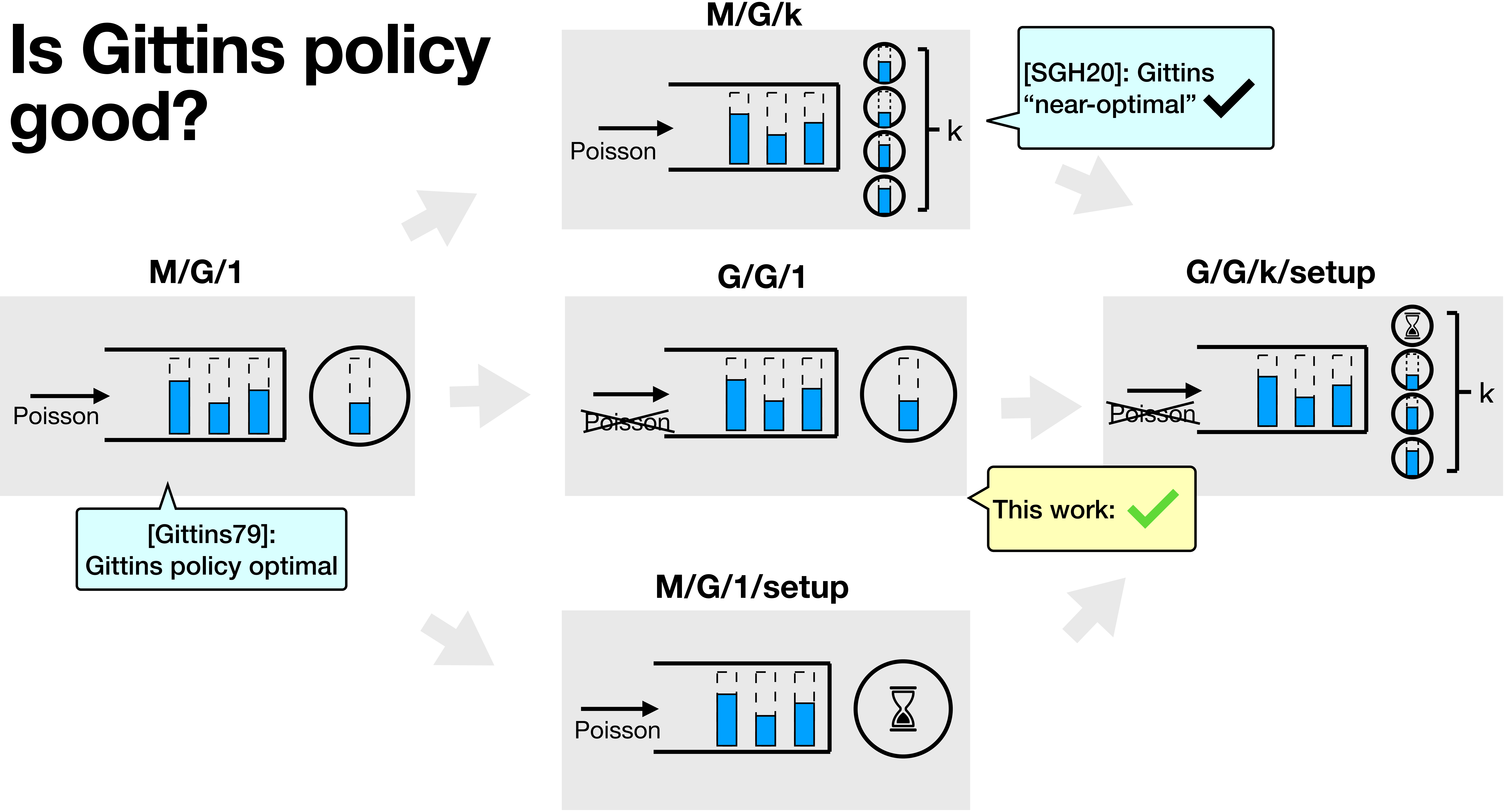
[SGH20]: Gittins
“near-optimal” ✓

G/G/1

M/G/1/setup

G/G/k/setup

Is Gittins policy good?

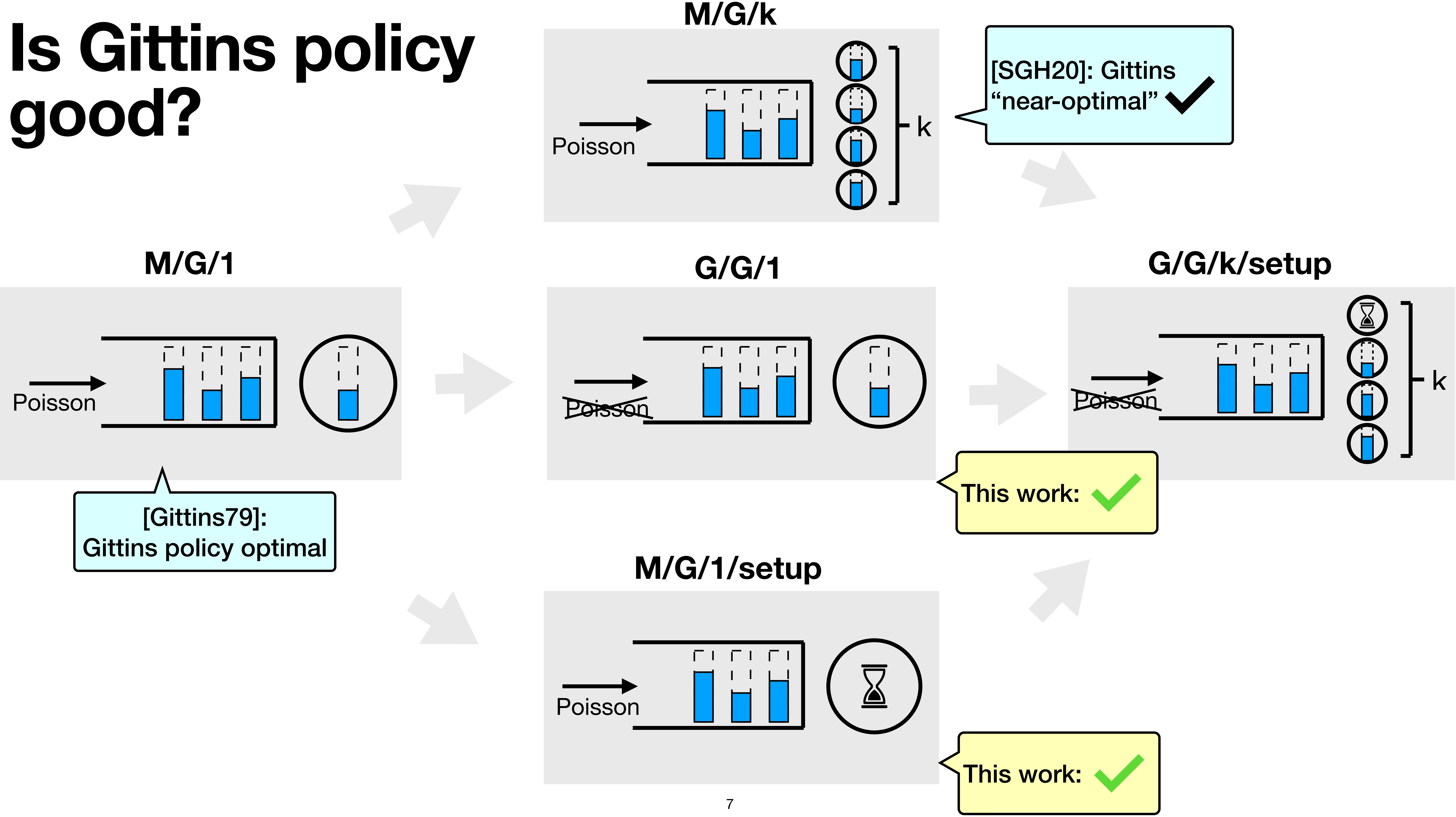


[Gittins79]:
Gittins policy optimal

[SGH20]: Gittins
“near-optimal” ✓

This work: ✓

Is Gittins policy good?



[SGH20]: Gittins “near-optimal” ✓

[Gittins79]: Gittins policy optimal

This work: ✓

This work: ✓

Is Gittins policy good?

M/G/1

[Gittins79]: Gittins policy optimal

M/G/k

[SGH20]: Gittins "near-optimal" ✓

G/G/1

This work: ✓

M/G/1/setup

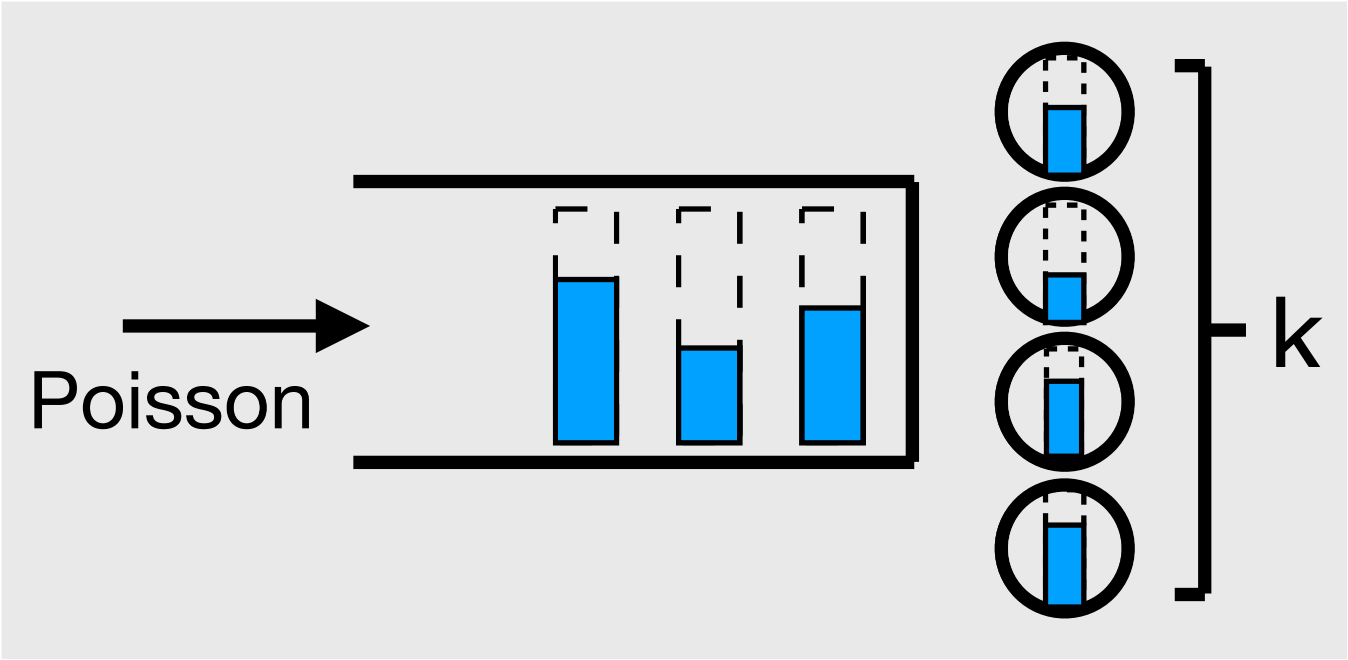
This work: ✓

G/G/k/setup

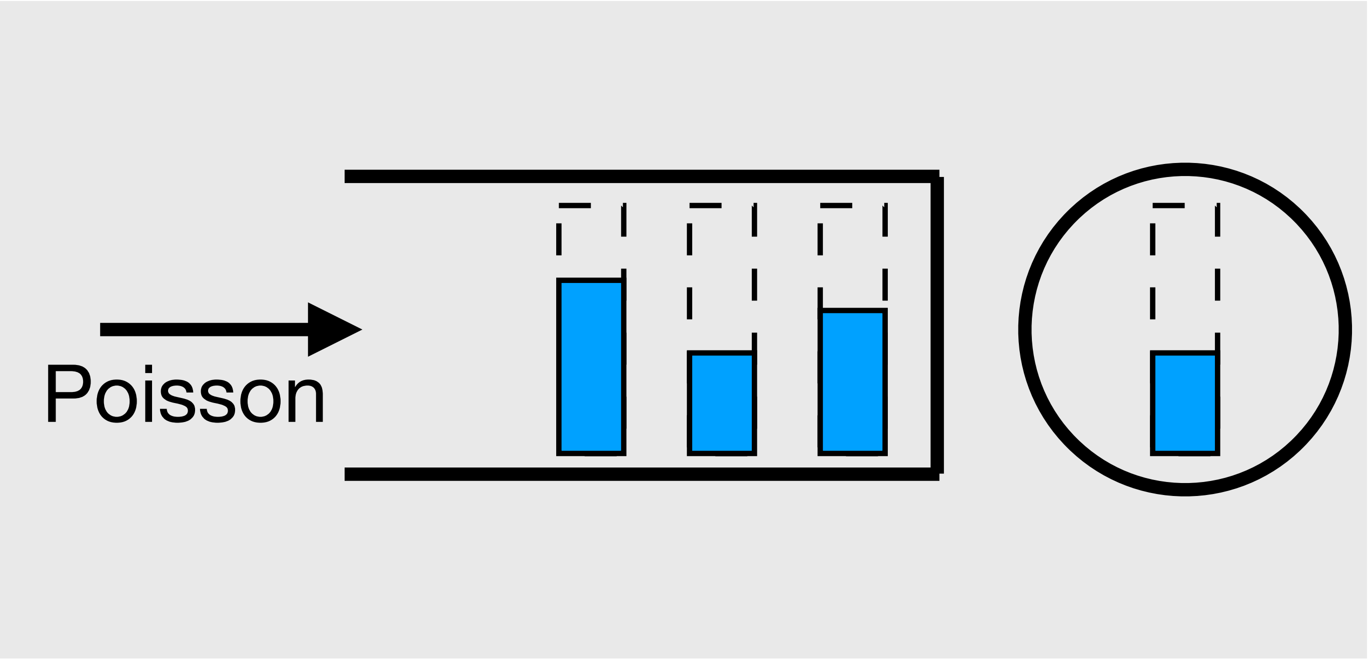
This work: ✓

Suboptimality gaps

M/G/k

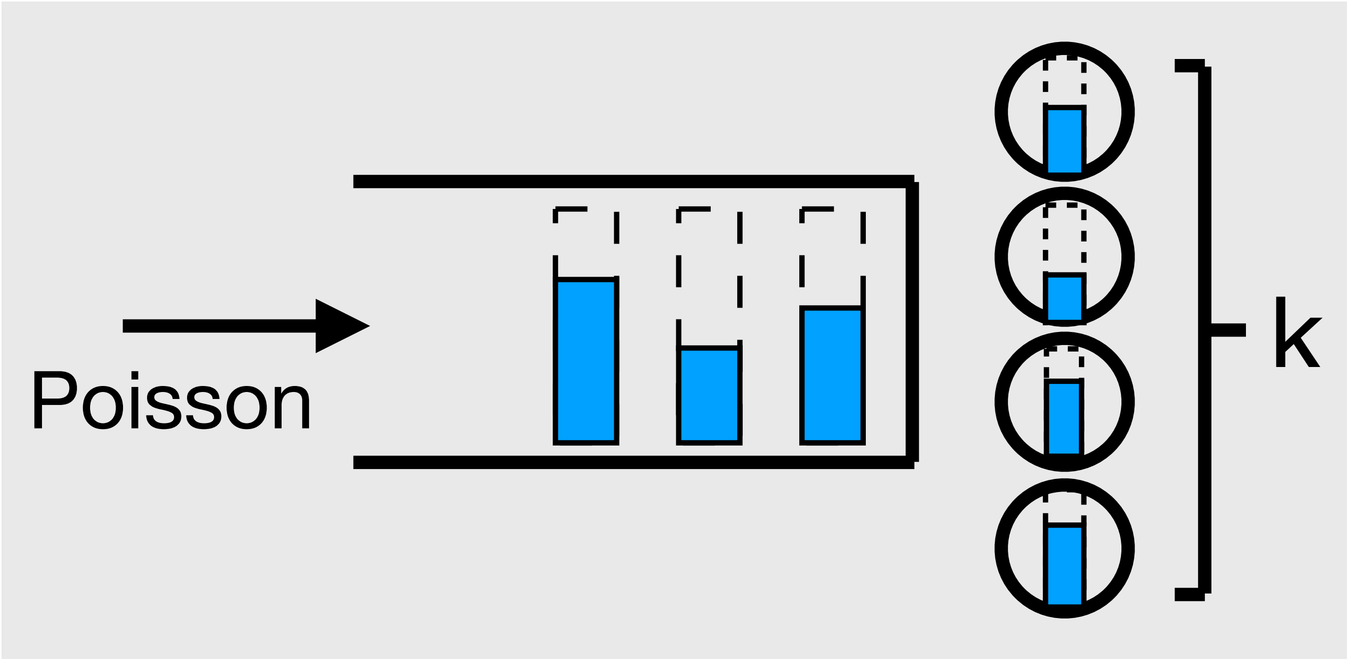


M/G/1

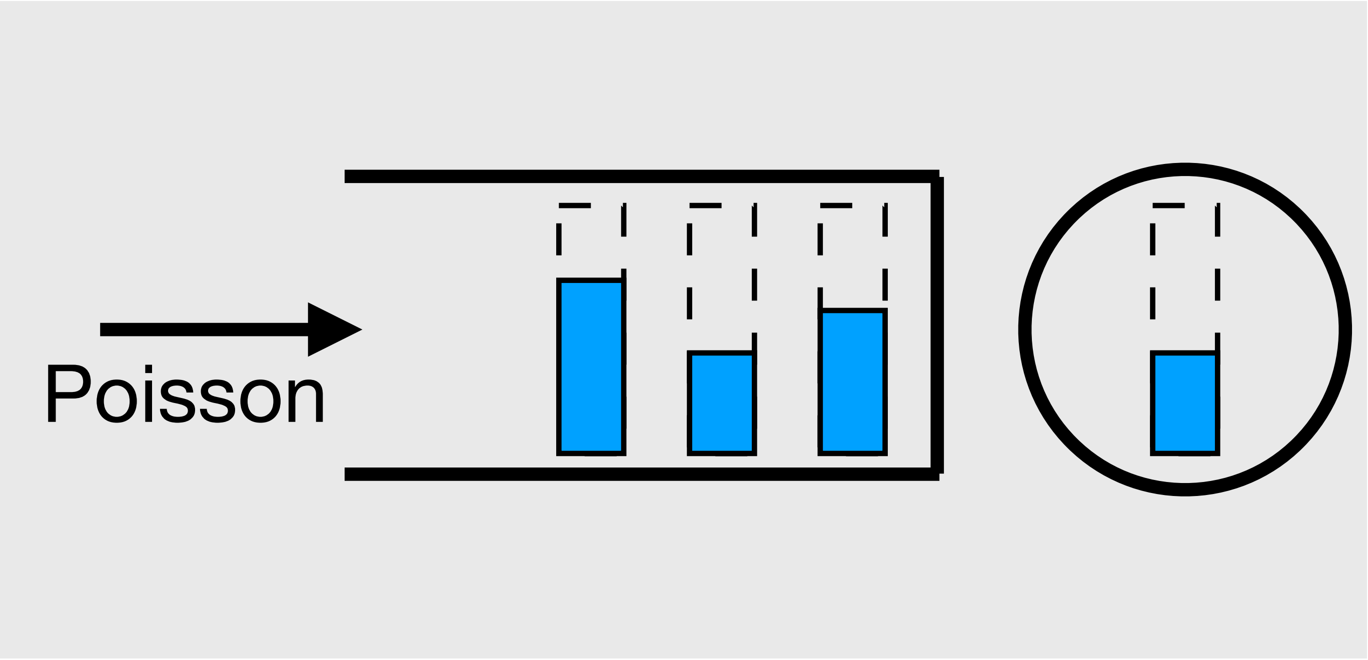


Suboptimality gaps

M/G/k



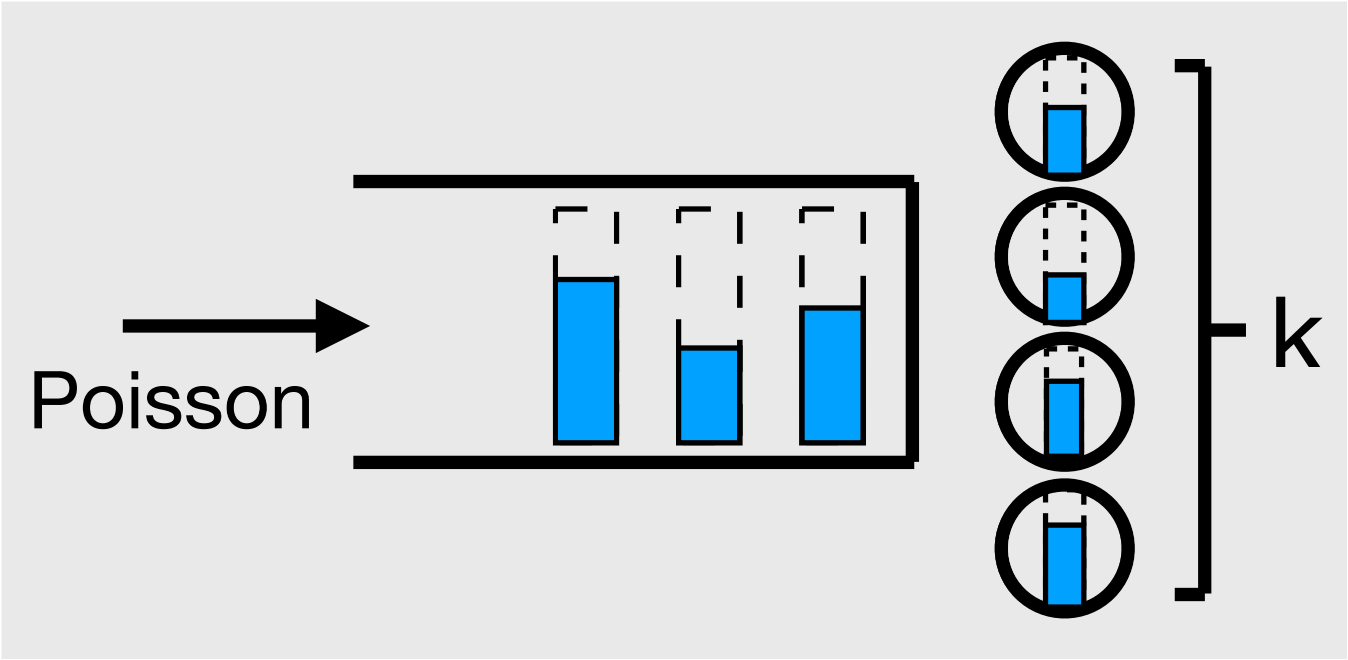
M/G/1



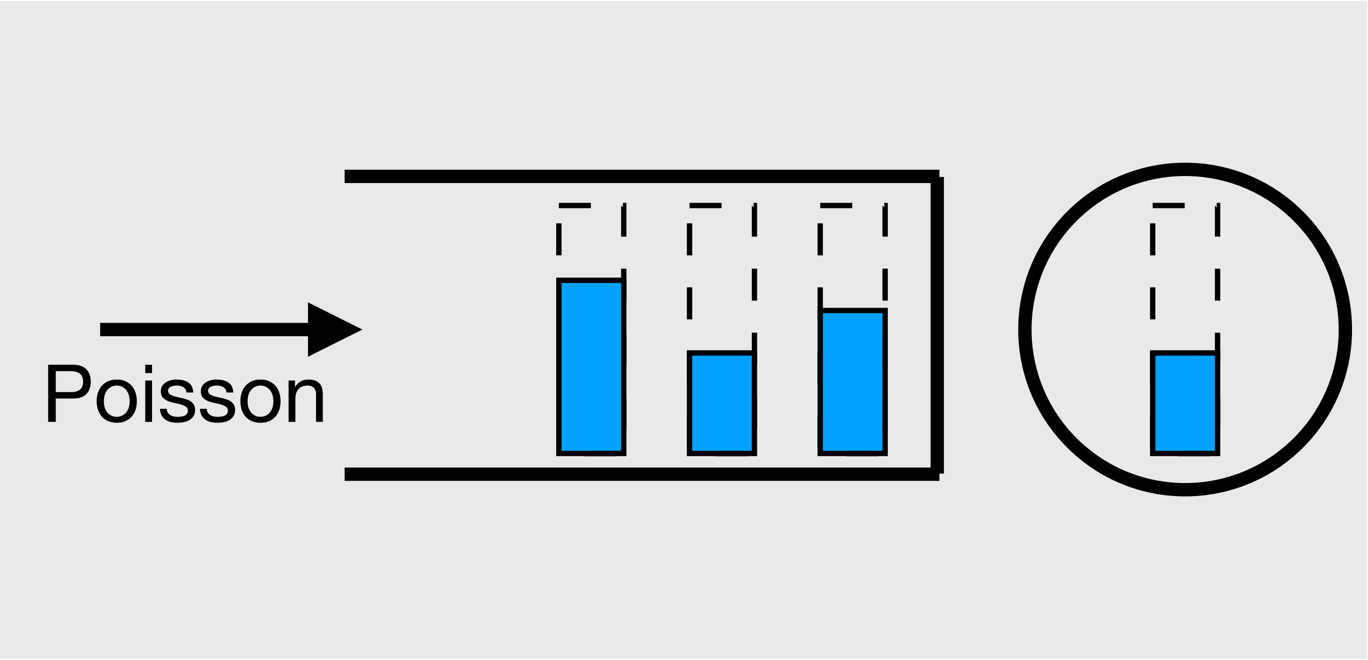
[Scully, Grosof, Harchol-Balter 20; Scully 22]

Suboptimality gaps

M/G/k



M/G/1

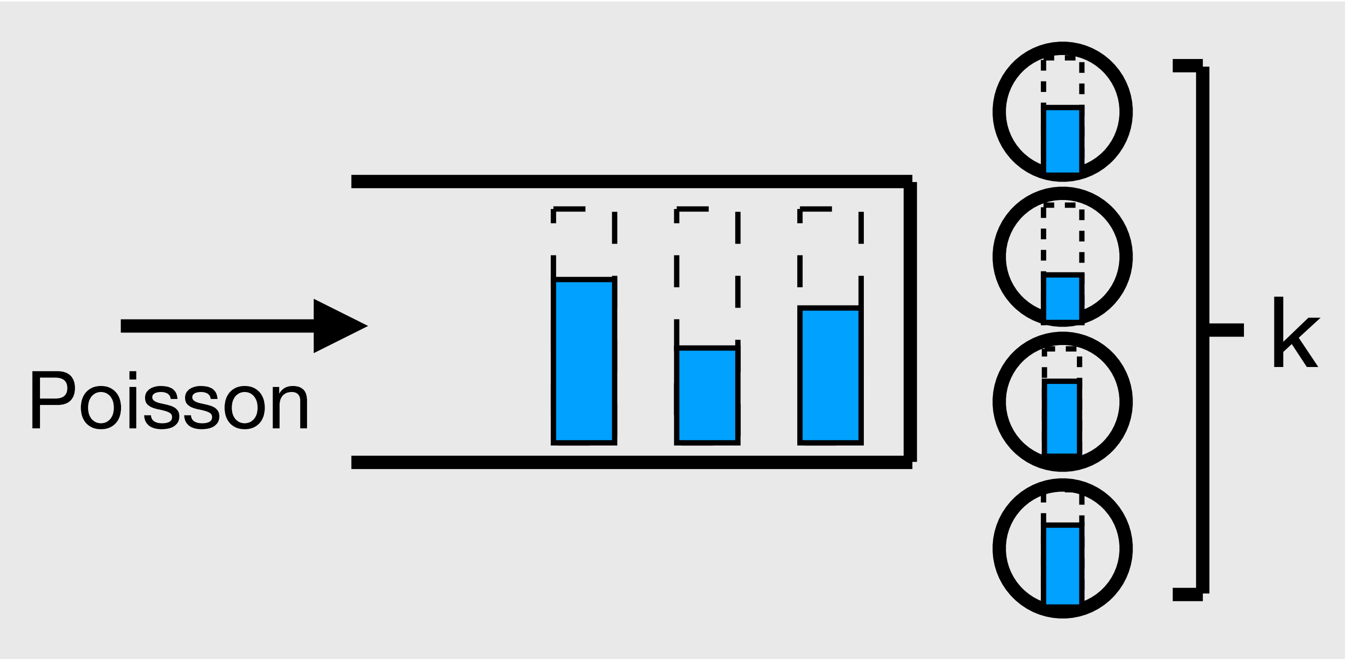


[Scully, Grosof, Harchol-Balter 20; Scully 22]

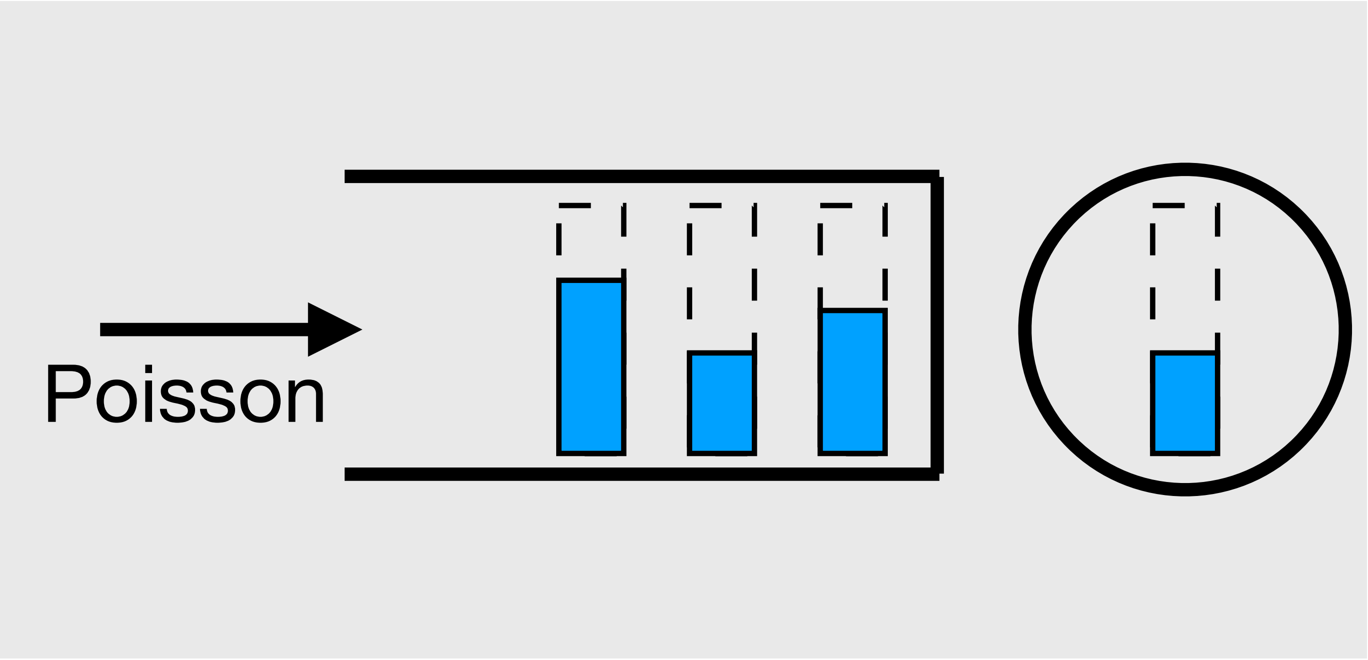
1. $\mathbb{E}[N]^{\text{Gittins}} \leq \mathbb{E}[N]^{\text{OPT}} + \ell_{(a)}, \ \ell_{(a)} = 3.8(k-1)\log\frac{1}{1-\rho}$

Suboptimality gaps

M/G/k



M/G/1



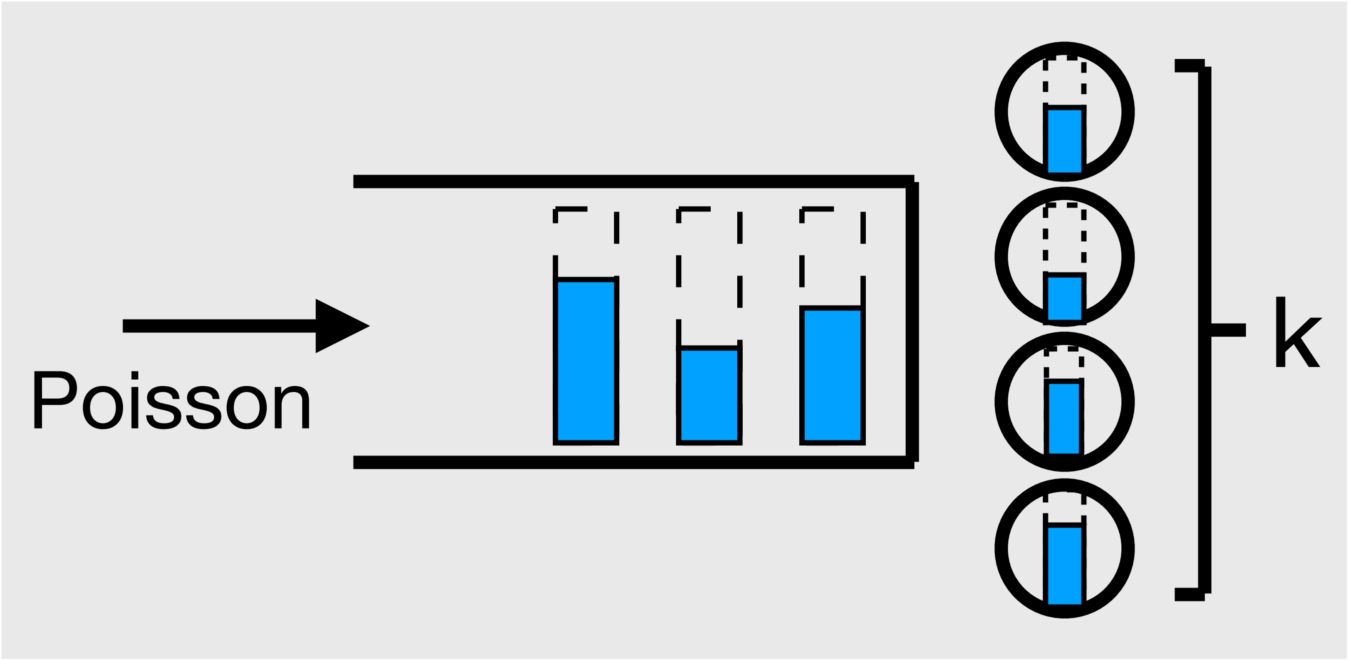
[Scully, Grosof, Harchol-Balter 20; Scully 22]

1. $\mathbb{E}[N]^{\text{Gittins}} \leq \mathbb{E}[N]^{\text{OPT}} + \ell_{(a)},\ \ \ell_{(a)} = 3.8(k-1)\log\frac{1}{1-\rho}$

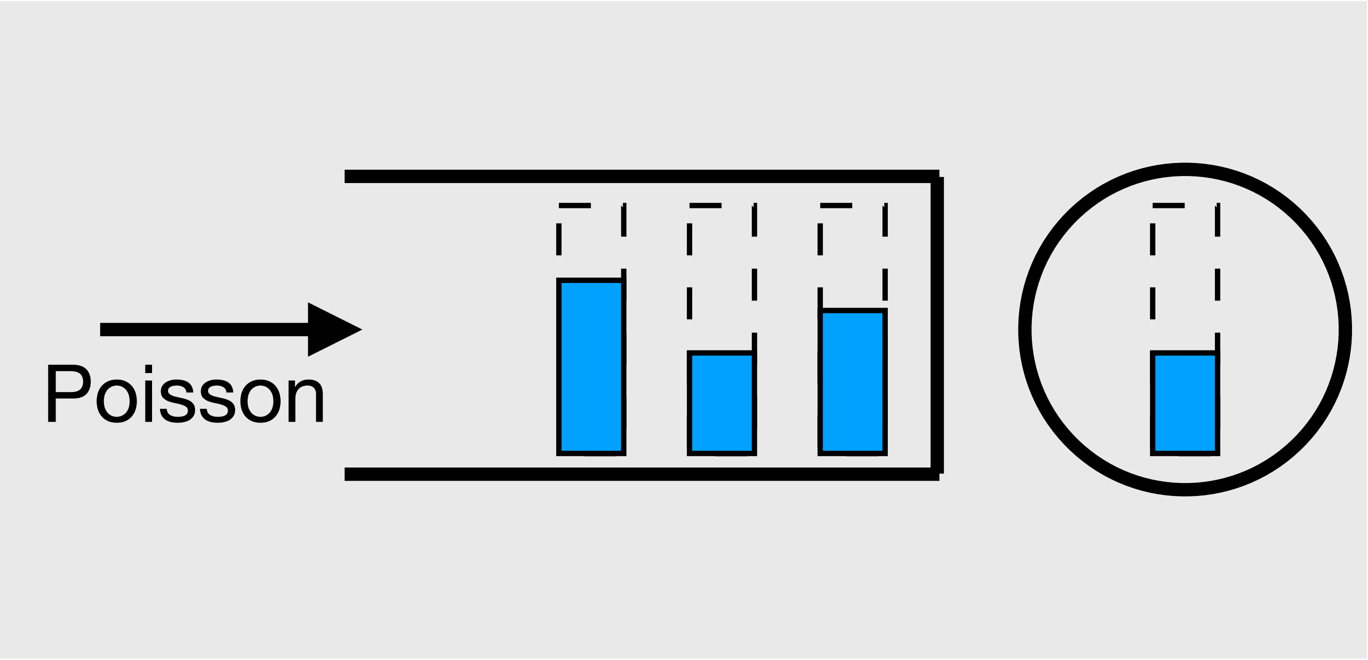
(load $\rho \triangleq \lambda\mathbb{E}[S]$ = expected fraction of busy servers)

Suboptimality gaps

M/G/k



M/G/1



[Scully, Grosof, Harchol-Balter 20; Scully 22]

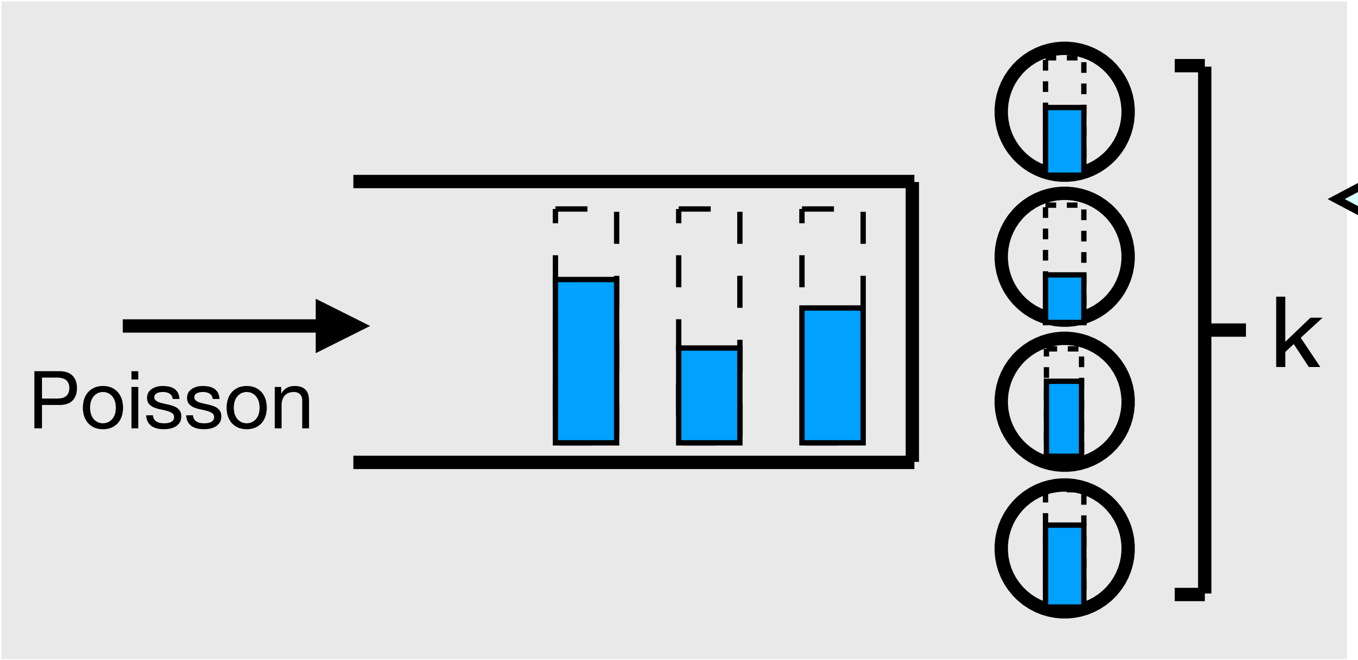
1. $\mathbb{E}[N]^{\text{Gittins}} \leq \mathbb{E}[N]^{\text{OPT}} + \ell_{(a)},\ \ell_{(a)} = 3.8(k-1)\log\frac{1}{1-\rho}$

   (load $\rho \triangleq \lambda\mathbb{E}[S]$ = expected fraction of busy servers)

2. Heavy traffic opt: $\frac{\mathbb{E}[N]^{\text{Gittins}}}{\mathbb{E}[N]^{\text{OPT}}} \to 1$ when $\rho \to 1$

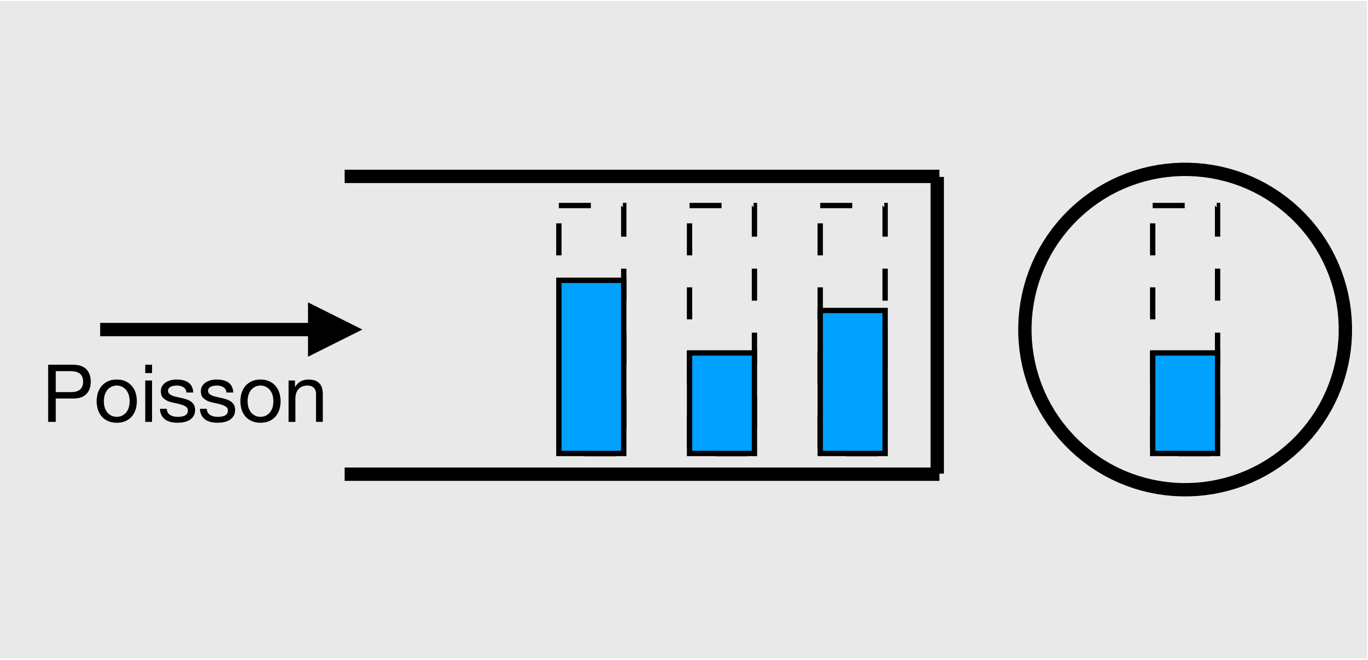
Suboptimality gaps

M/G/k



[SGH20]:
gap $\leq \ell_{(a)}$
heavy-traffic opt

M/G/1



[Scully, Grosof, Harchol-Balter 20; Scully 22]

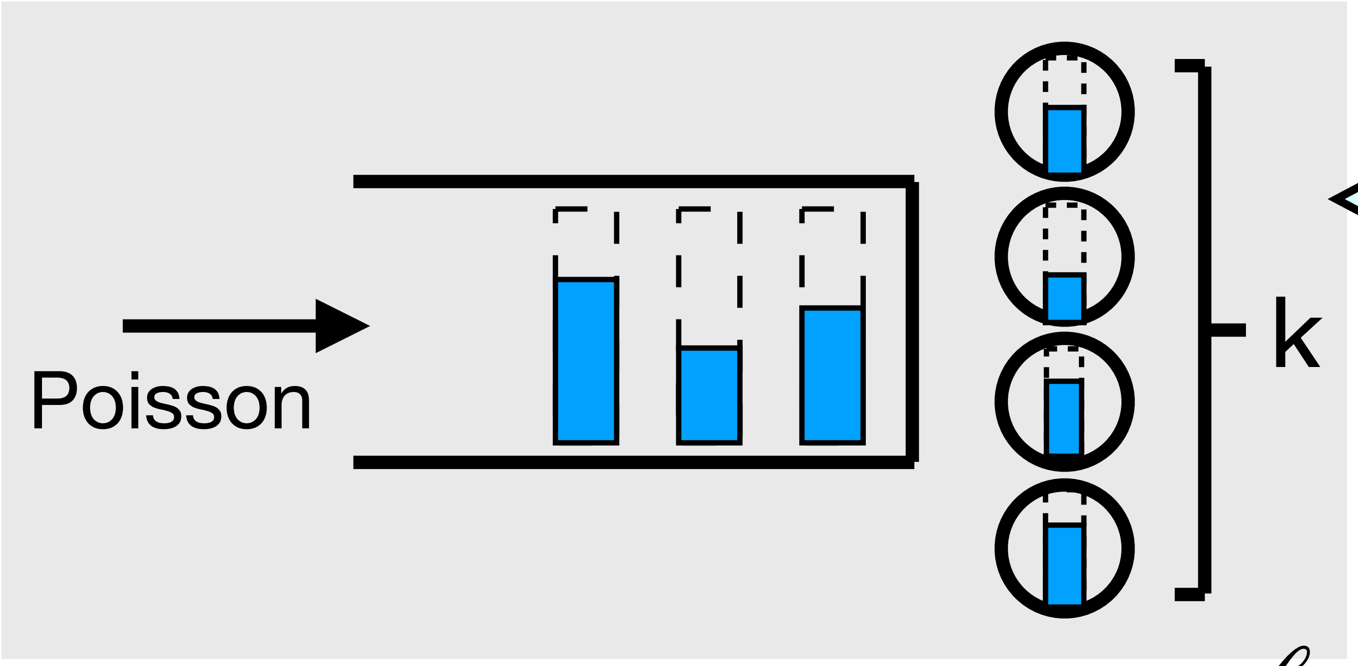
1. $\mathbb{E}[N]^{\text{Gittins}} \leq \mathbb{E}[N]^{\text{OPT}} + \ell_{(a)}, \ \ell_{(a)} = 3.8(k-1)\log\frac{1}{1-\rho}$

   (load $\rho \triangleq \lambda\mathbb{E}[S]$ = expected fraction of busy servers)

2. Heavy traffic opt: $\frac{\mathbb{E}[N]^{\text{Gittins}}}{\mathbb{E}[N]^{\text{OPT}}} \to 1$ when $\rho \to 1$

Suboptimality gaps

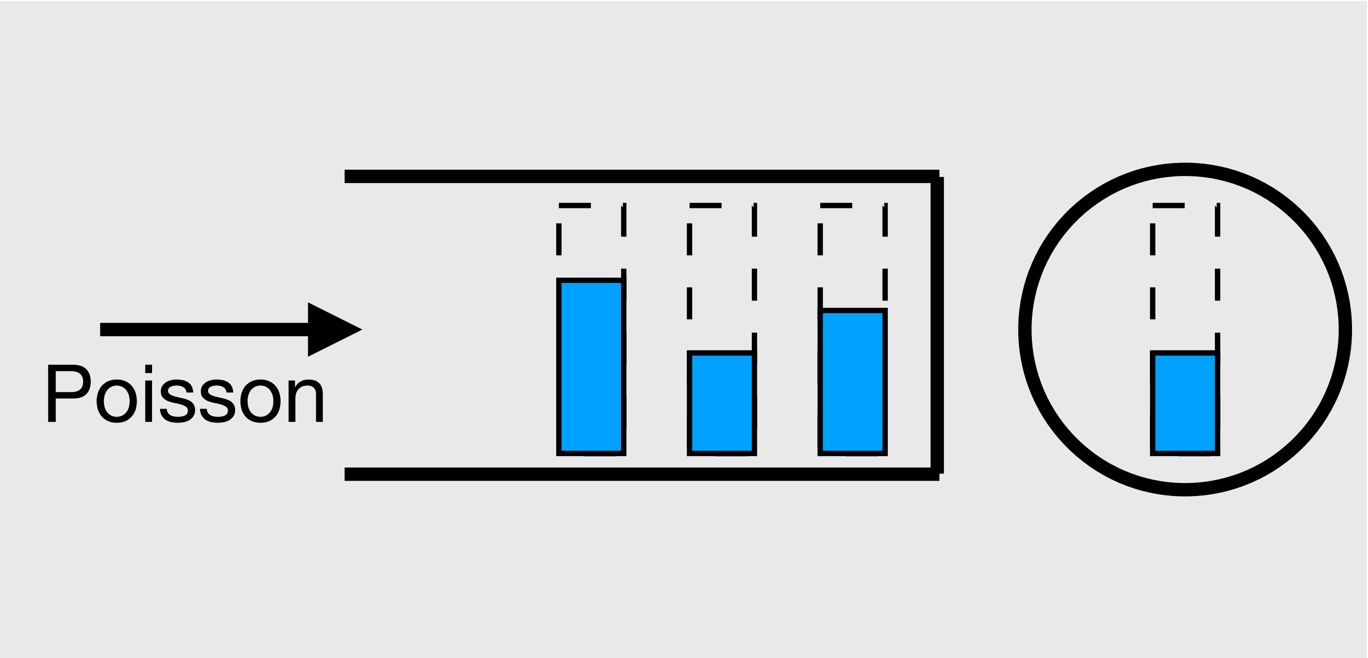
M/G/k



[SGH20]:
gap $\leq \ell_{(a)}$
heavy-traffic opt

$$\ell_{(a)} = O\left(\log \frac{1}{1-\rho}\right)$$

M/G/1



[Scully, Grosof, Harchol-Balter 20; Scully 22]

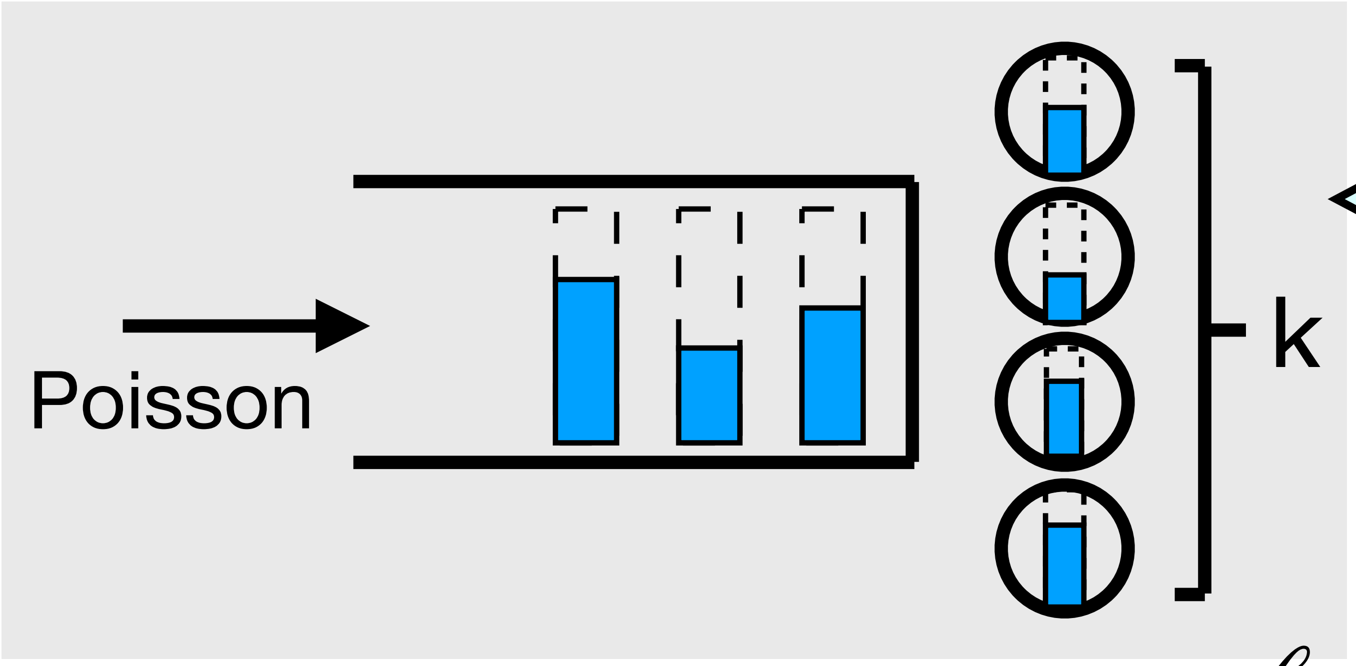
1. $\mathbb{E}[N]^{\text{Gittins}} \leq \mathbb{E}[N]^{\text{OPT}} + \ell_{(a)},\ \ \ell_{(a)} = 3.8(k-1)\log\frac{1}{1-\rho}$

   (load $\rho \triangleq \lambda\mathbb{E}[S]$ = expected fraction of busy servers)

2. Heavy traffic opt: $\dfrac{\mathbb{E}[N]^{\text{Gittins}}}{\mathbb{E}[N]^{\text{OPT}}} \to 1$ when $\rho \to 1$

Suboptimality gaps

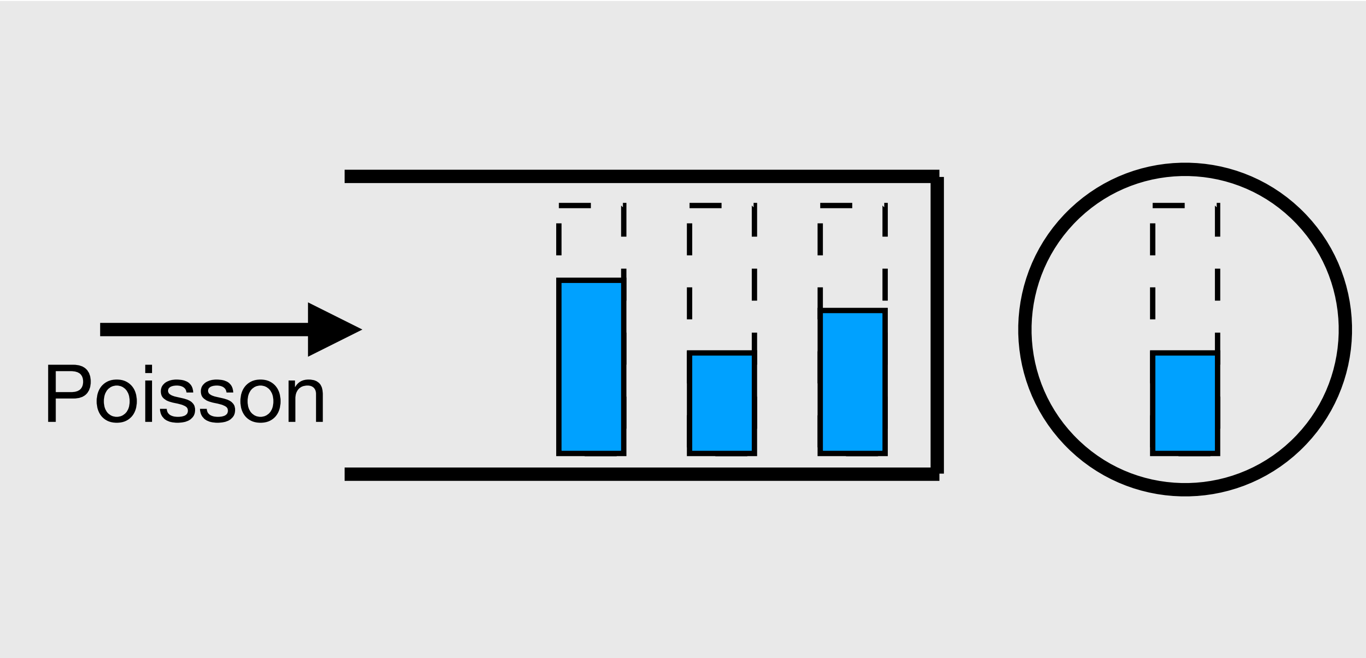
M/G/k



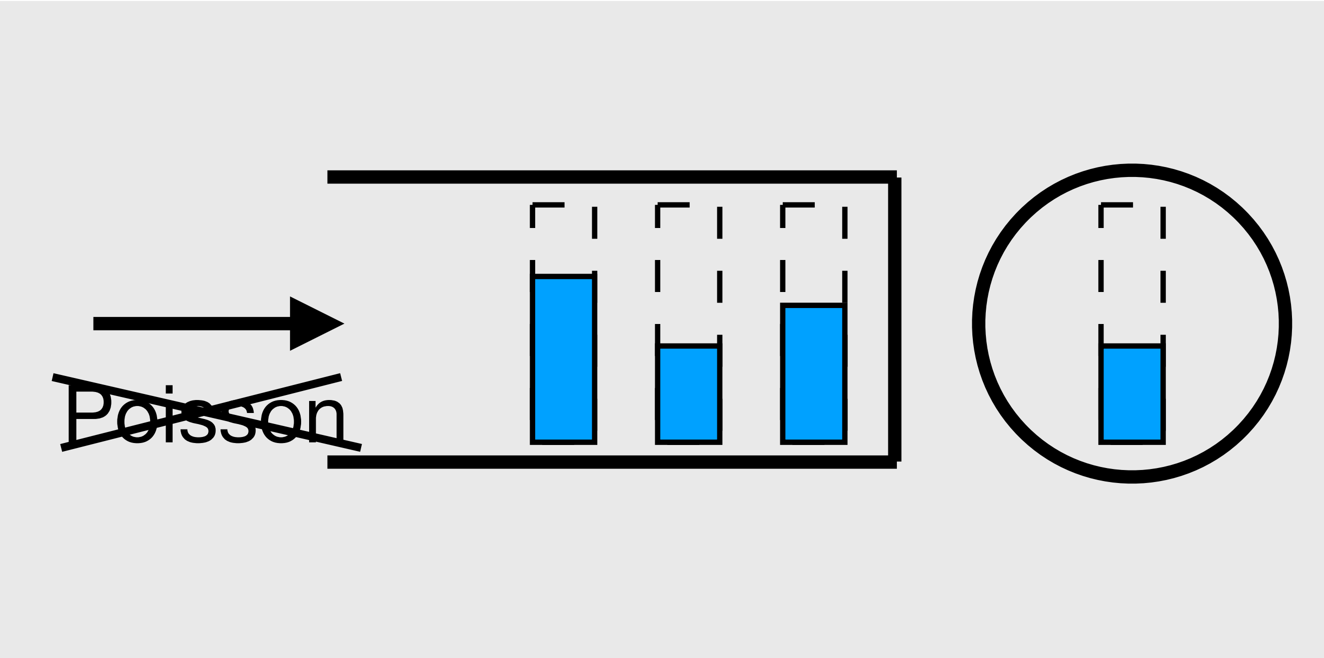
[SGH20]:
gap $\leq \ell_{(a)}$
heavy-traffic opt

$$\ell_{(a)} = O\left(\log \frac{1}{1-\rho}\right)$$

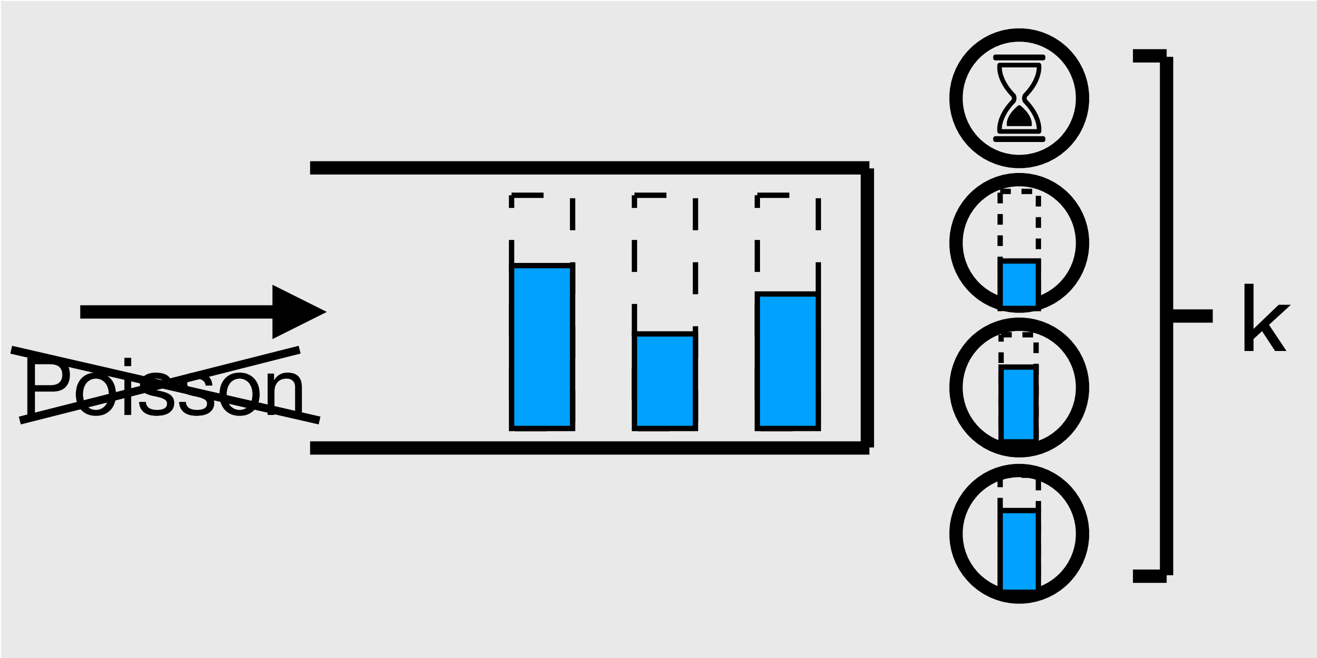
M/G/1



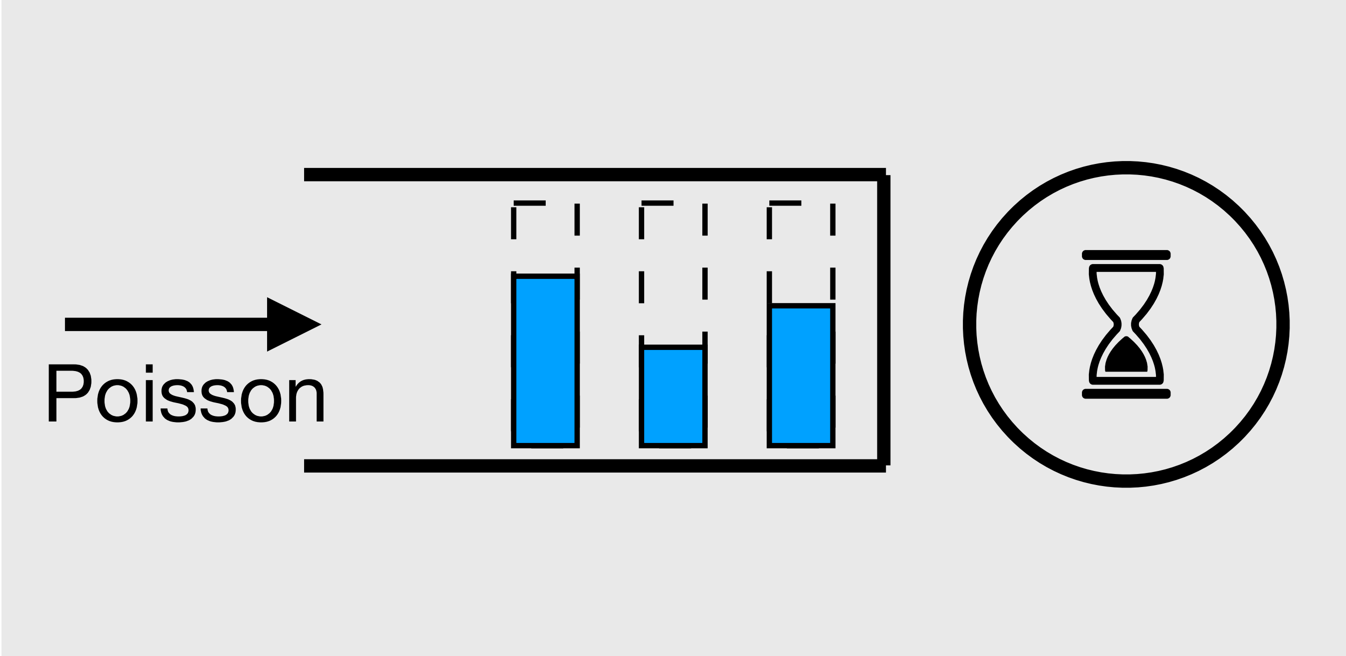
G/G/1



G/G/k/setup

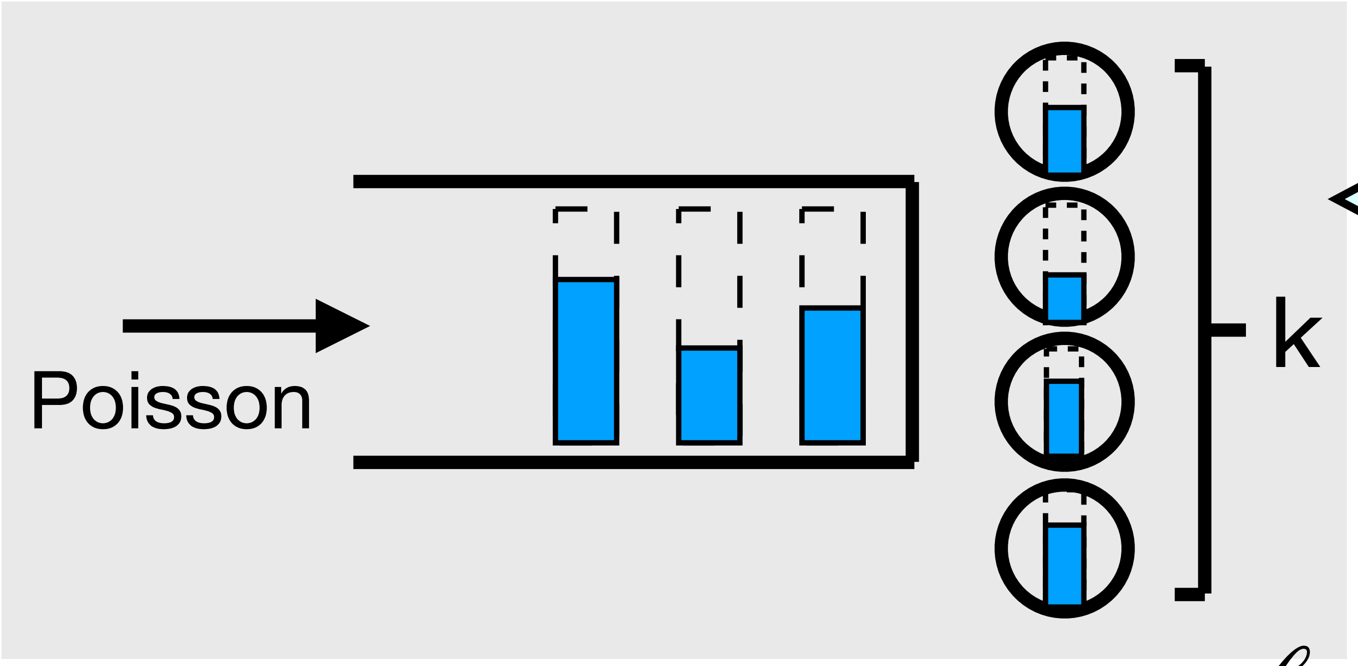


M/G/1/setup



Suboptimality gaps

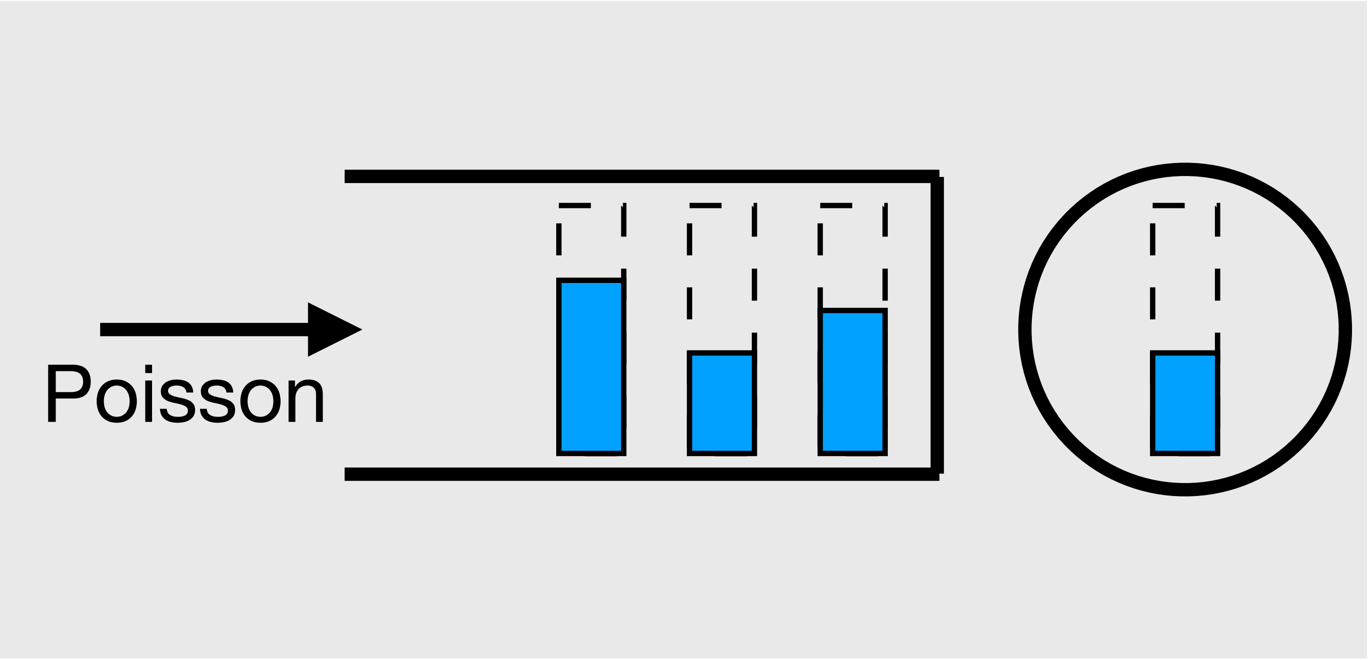
M/G/k



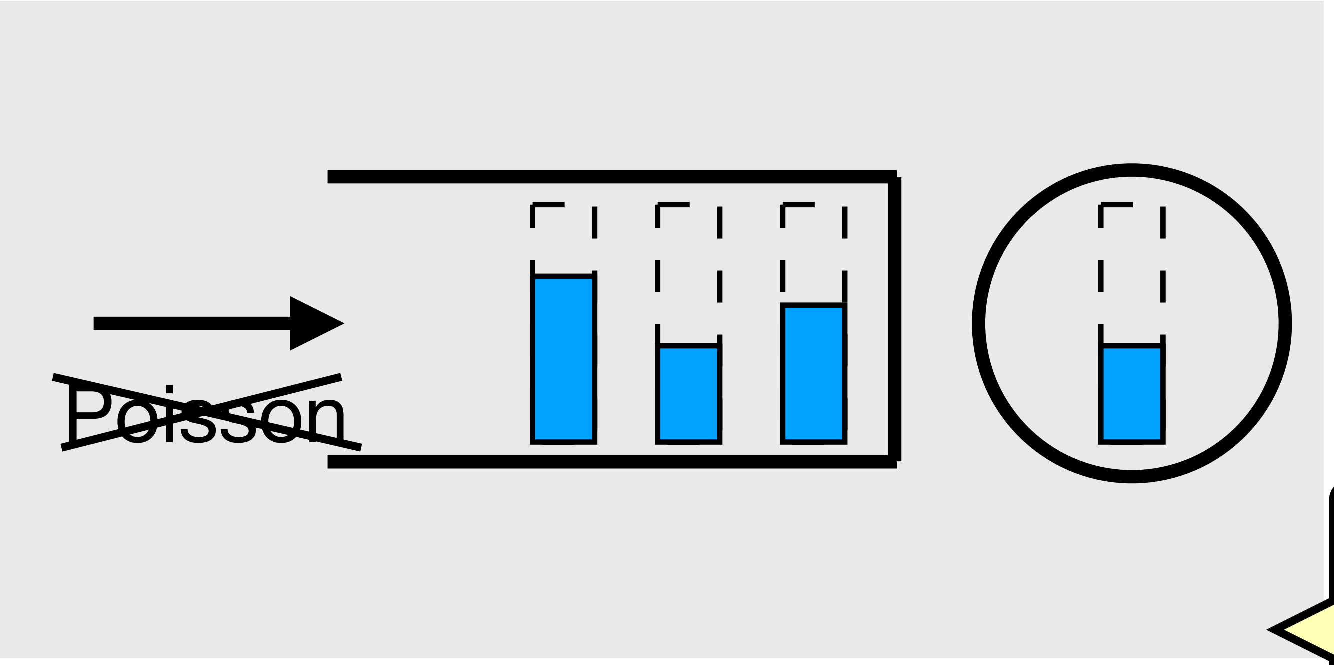
[SGH20]:
gap $\le \ell_{(a)}$
heavy-traffic opt

$$\ell_{(a)} = O\left(\log \frac{1}{1-\rho}\right)$$

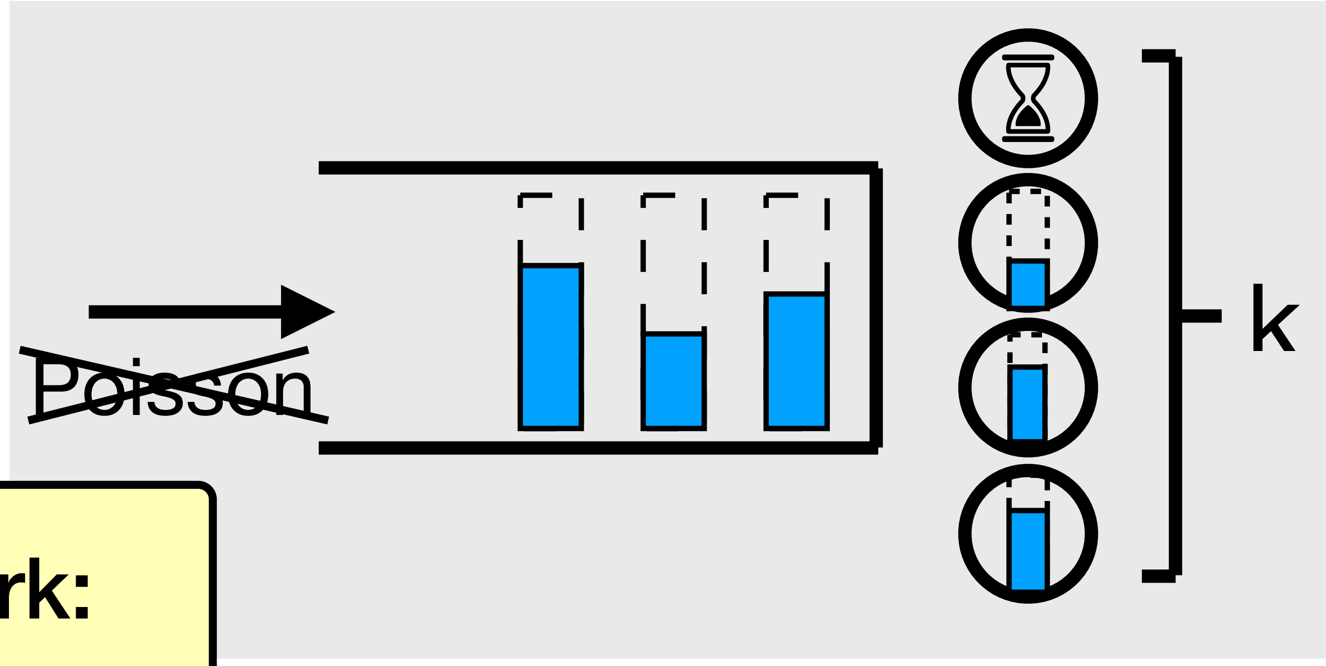
M/G/1



G/G/1

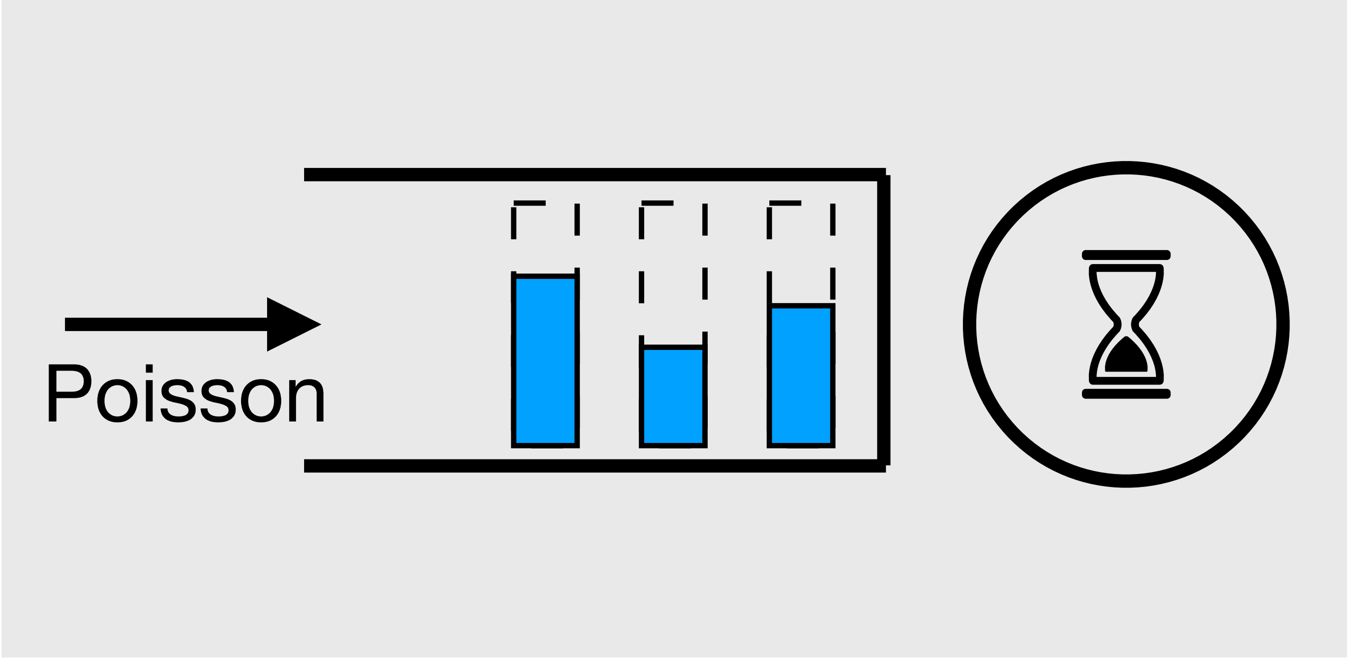


G/G/k/setup



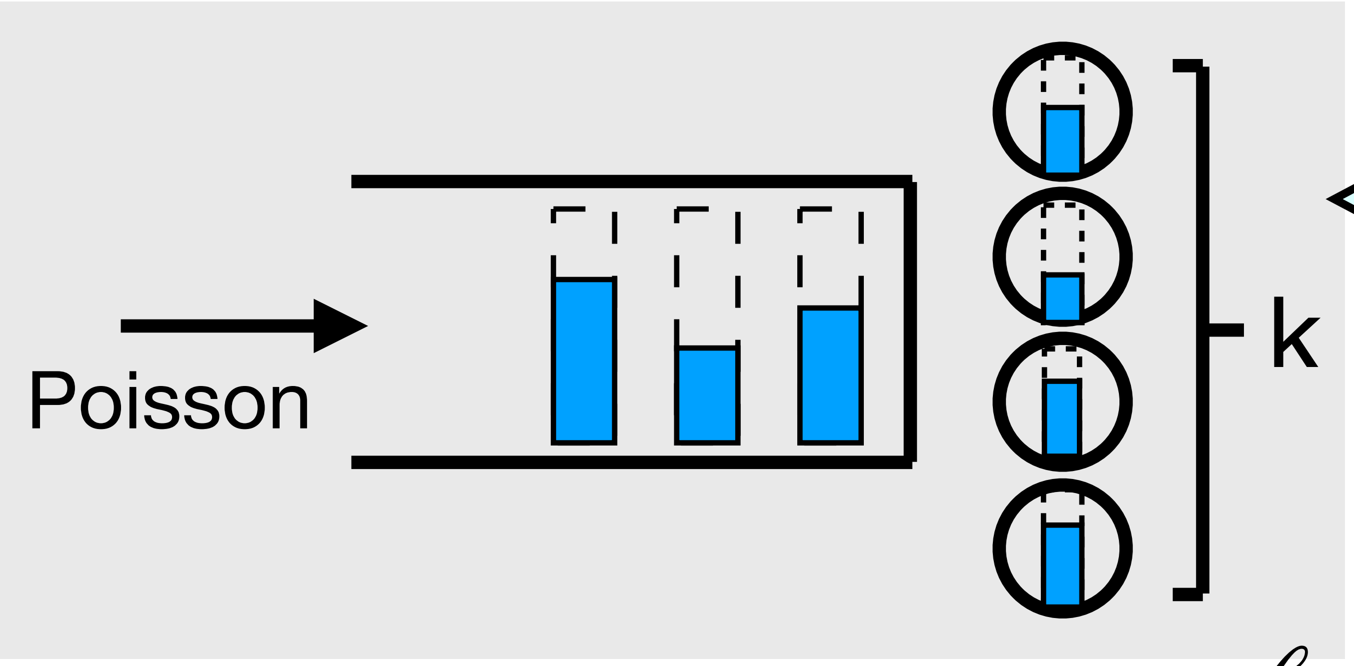
This work:
gap $\le \ell_{(b)}$

M/G/1/setup



Suboptimality gaps

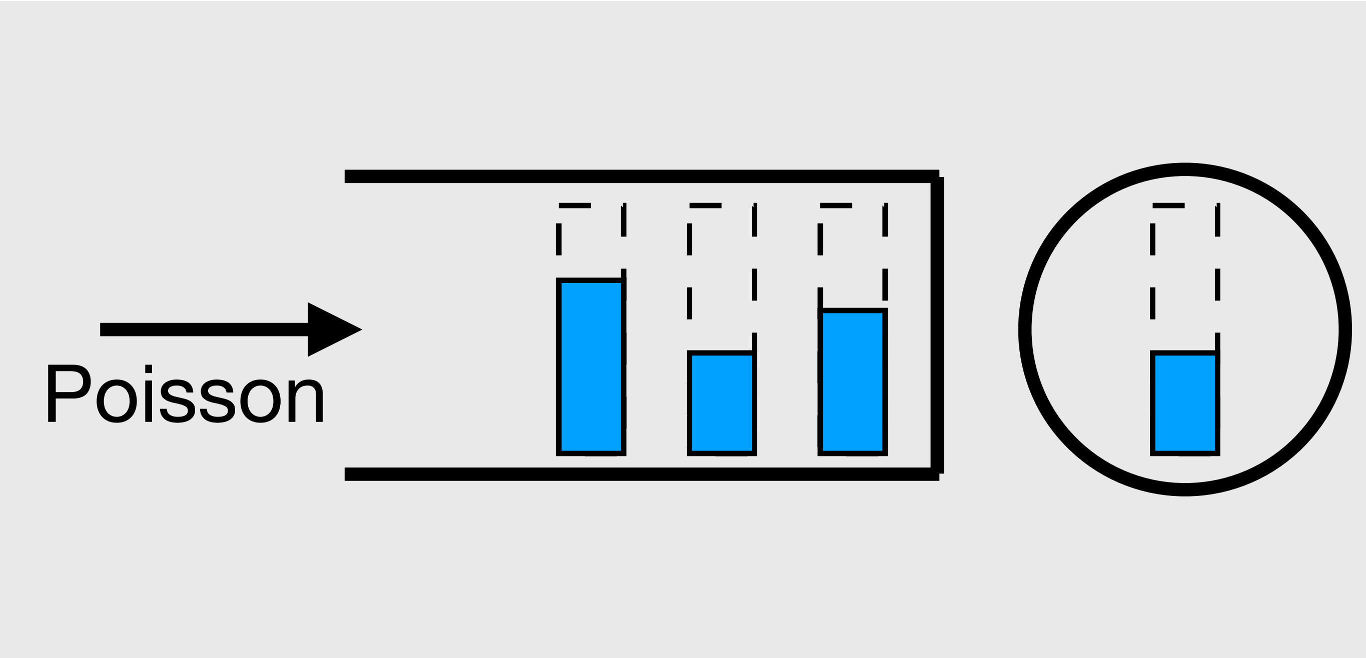
M/G/k



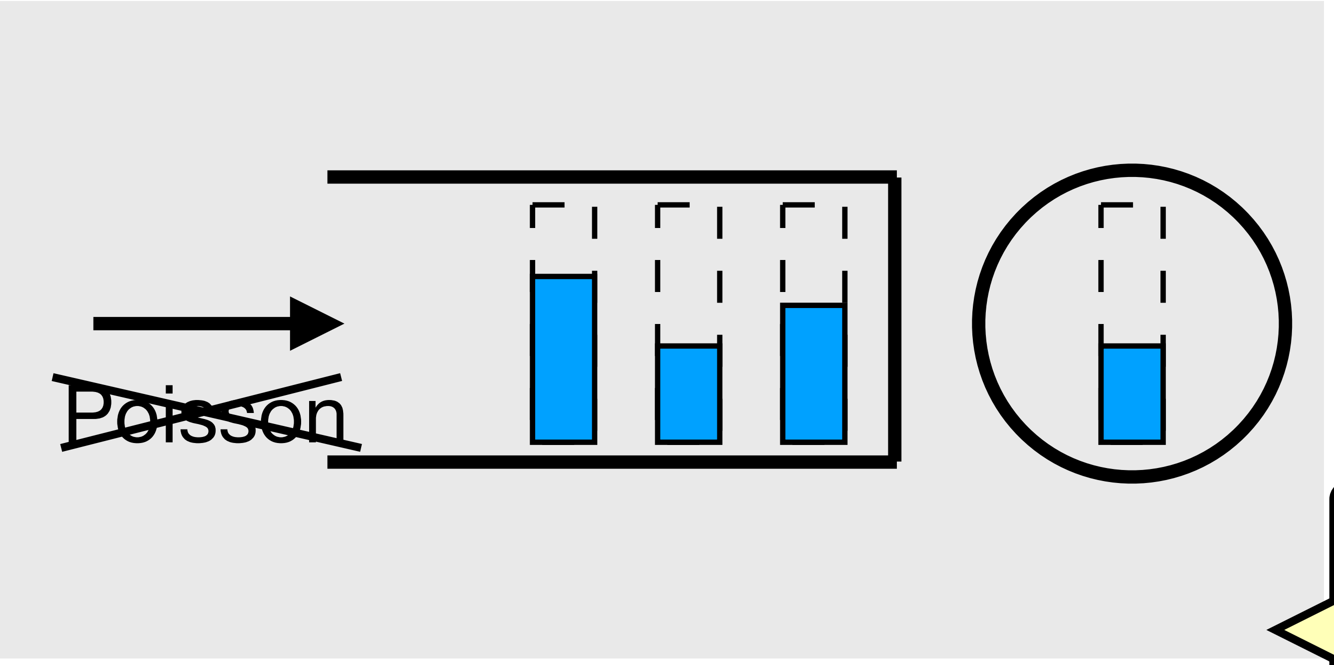
[SGH20]:
gap $\leq \ell_{(a)}$
heavy-traffic opt

$$\ell_{(a)} = O\left(\log \frac{1}{1-\rho}\right)$$

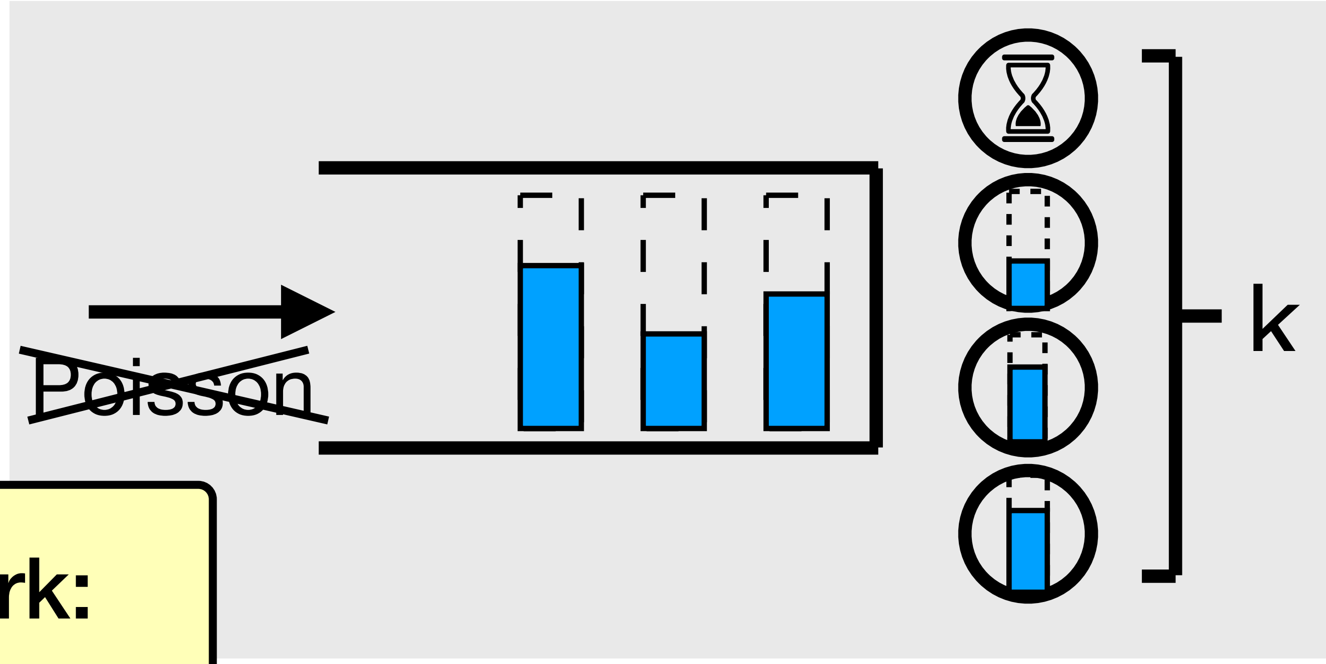
M/G/1



G/G/1



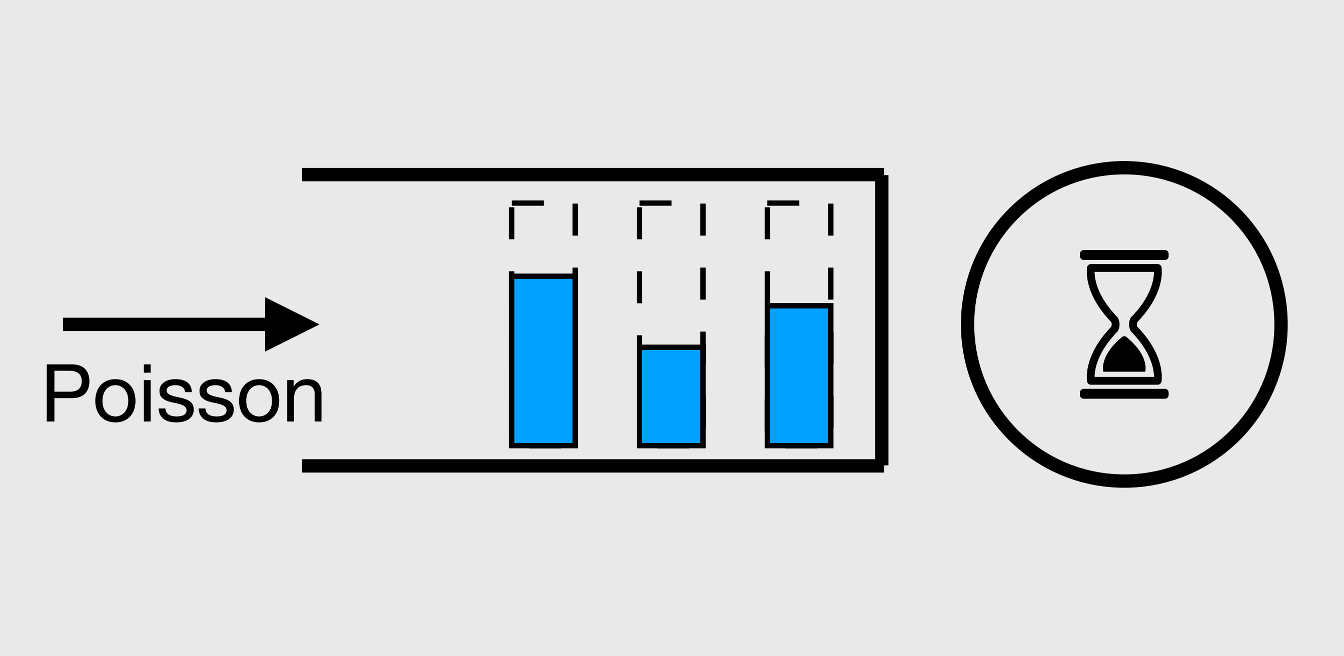
G/G/k/setup



This work:
gap $\leq \ell_{(b)}$

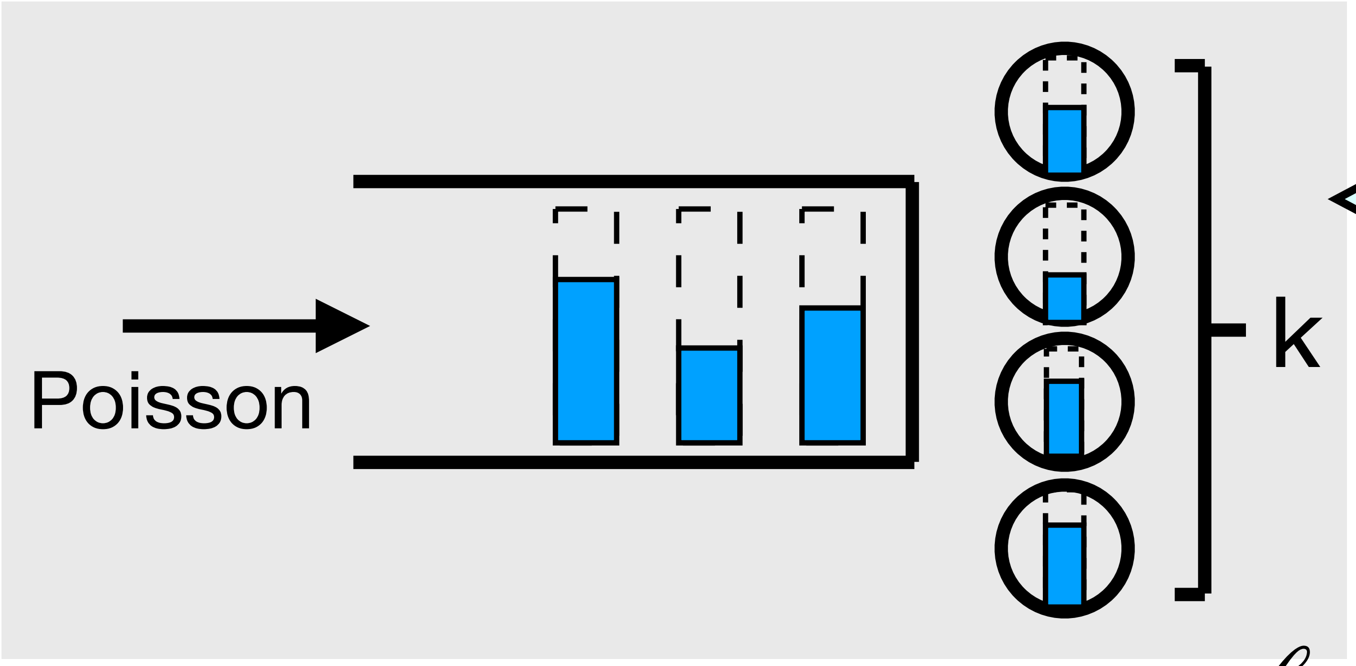
$$\ell_{(b)} = O(1)$$

M/G/1/setup



Suboptimality gaps

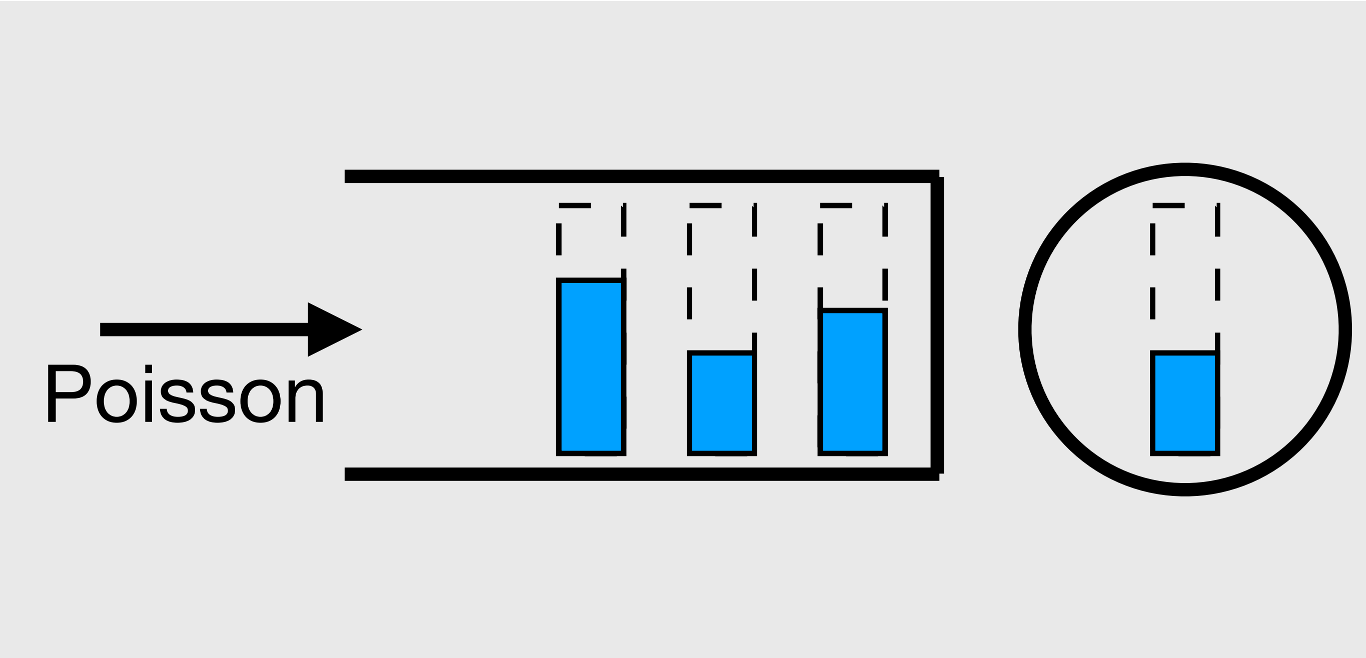
M/G/k



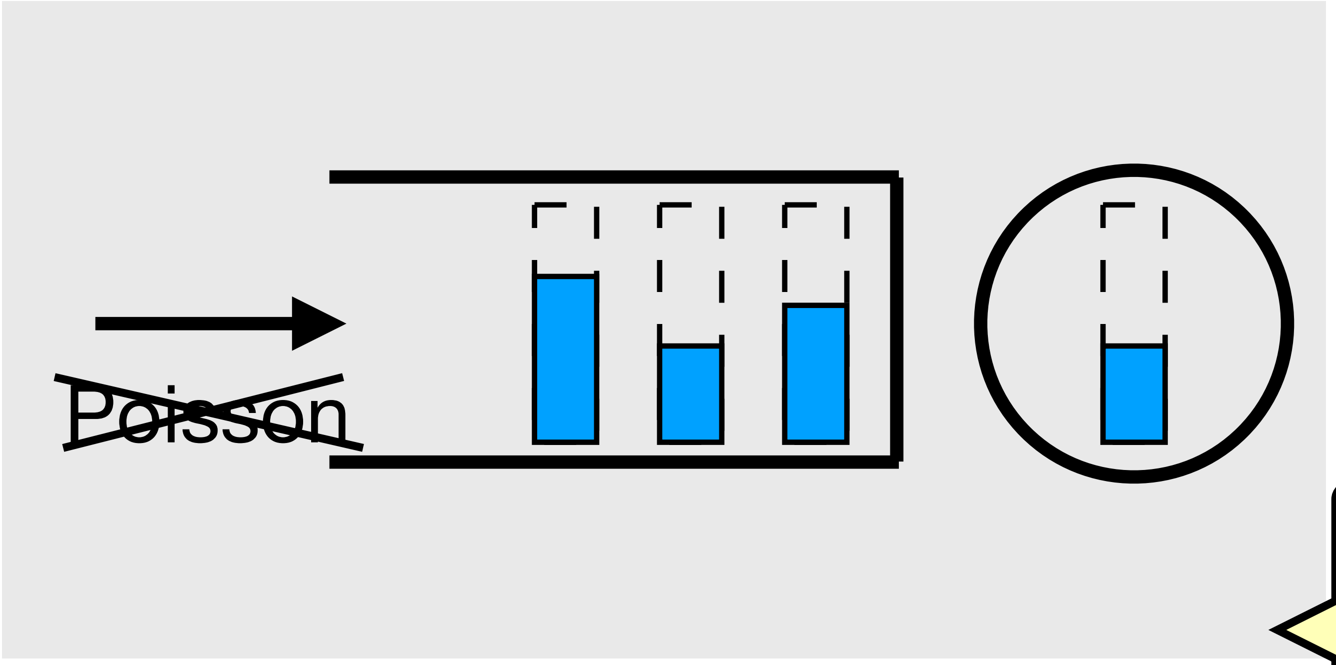
[SGH20]:
gap $\leq \ell_{(a)}$
heavy-traffic opt

$$\ell_{(a)} = O\left(\log \frac{1}{1-\rho}\right)$$

M/G/1

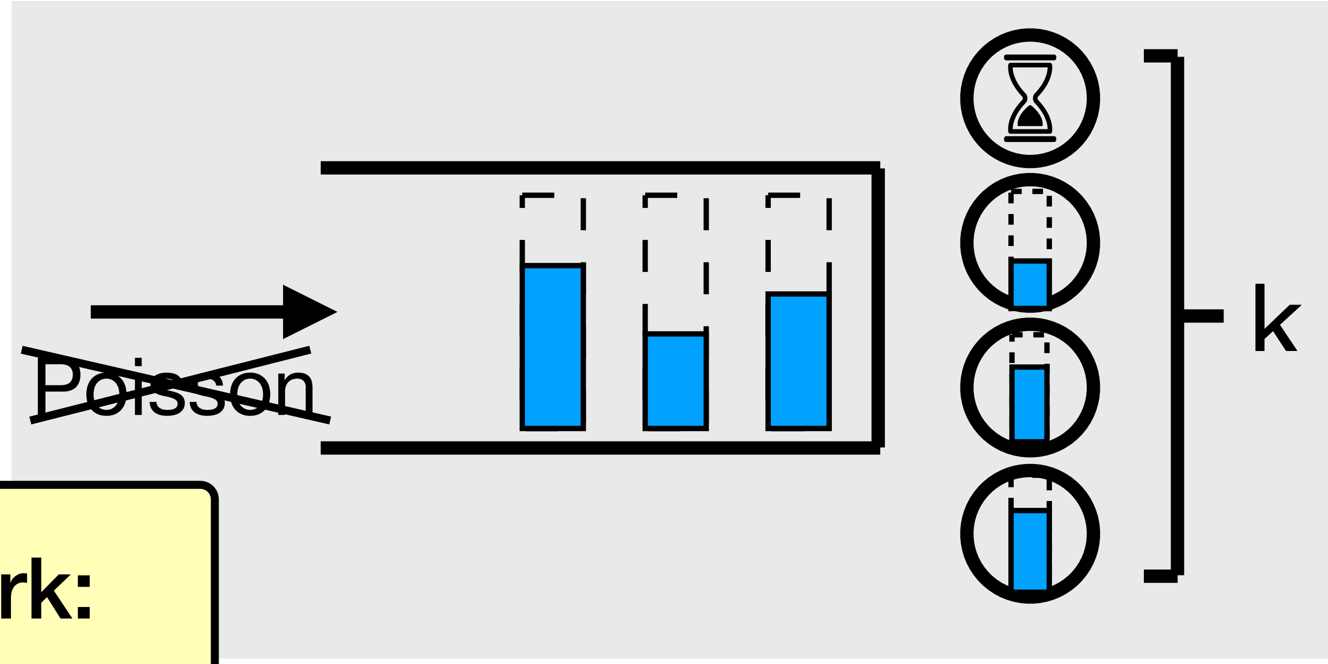


G/G/1



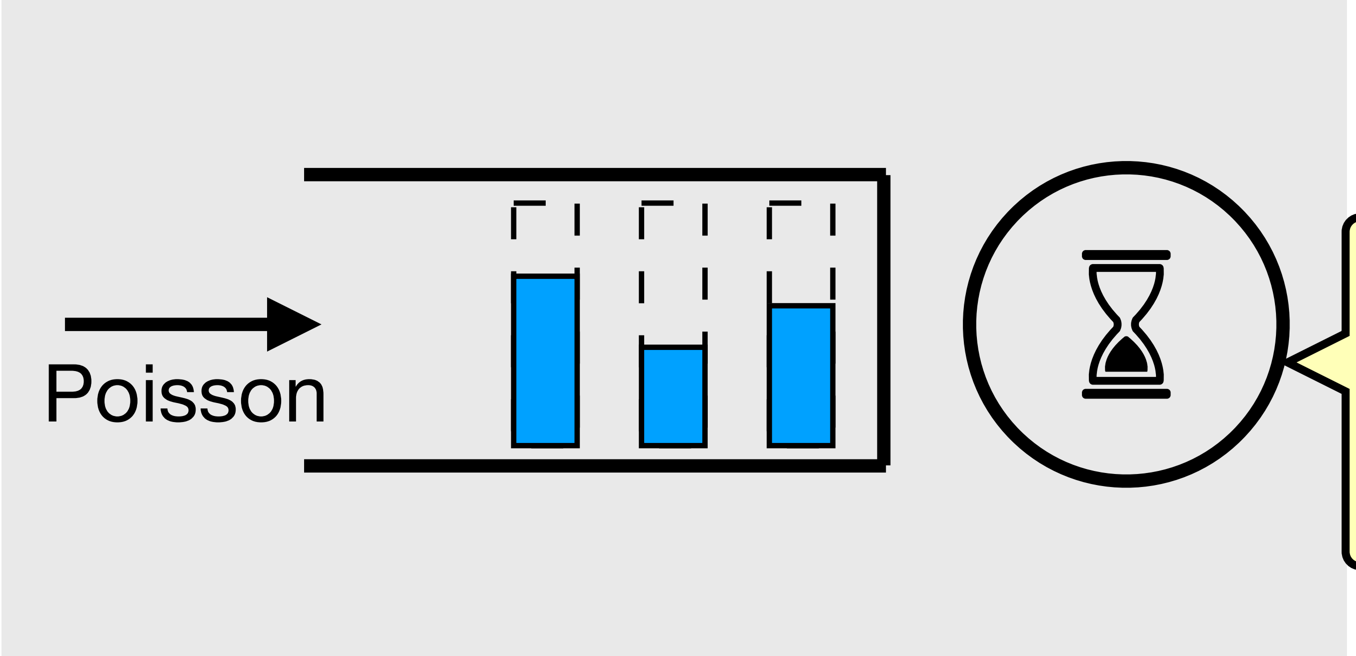
This work:
gap $\leq \ell_{(b)}$

G/G/k/setup



$$\ell_{(b)} = O(1)$$

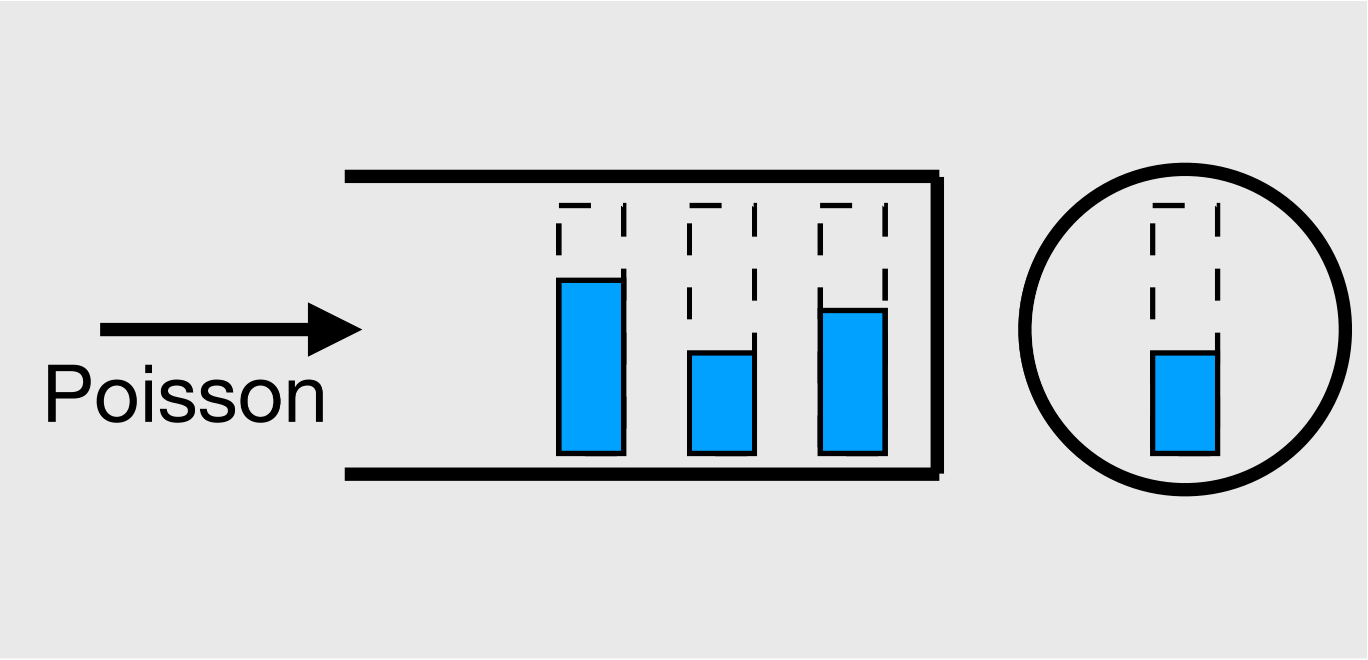
M/G/1/setup



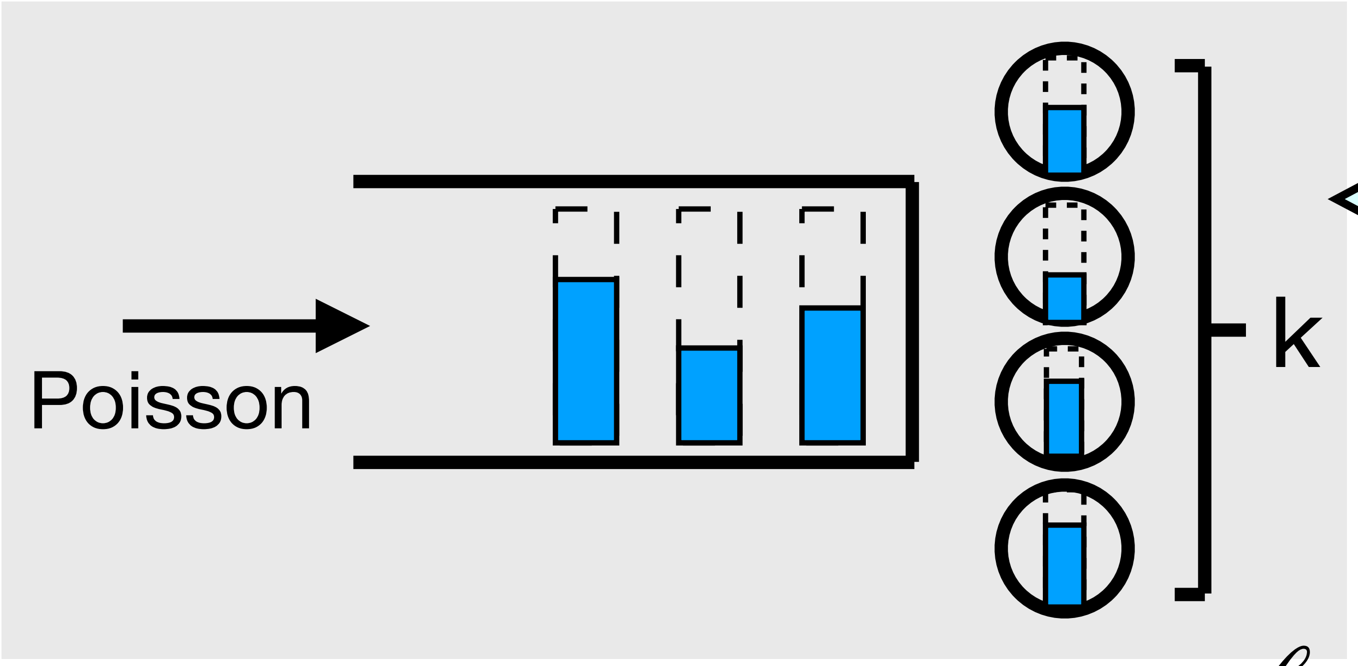
This work:
gap $\leq \ell_{(c)}$

Suboptimality gaps

M/G/1



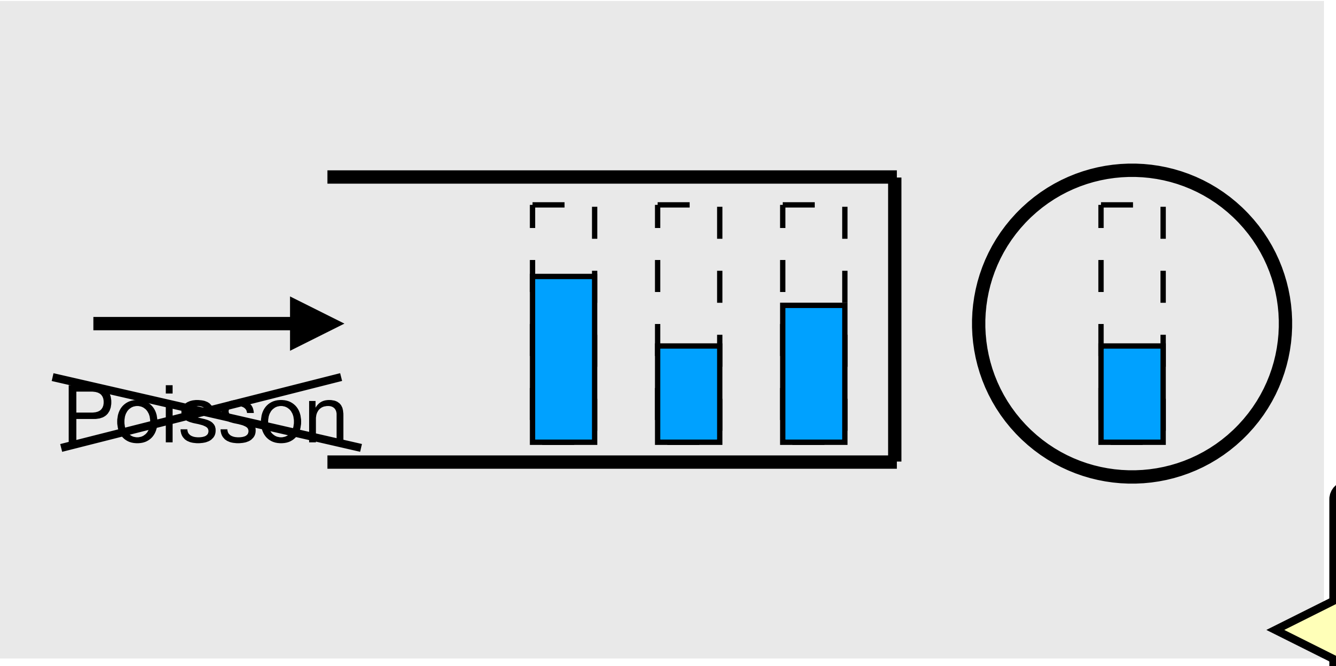
M/G/k



[SGH20]:
gap $\leq \ell_{(a)}$
heavy-traffic opt

$$\ell_{(a)} = O\left(\log \frac{1}{1-\rho}\right)$$

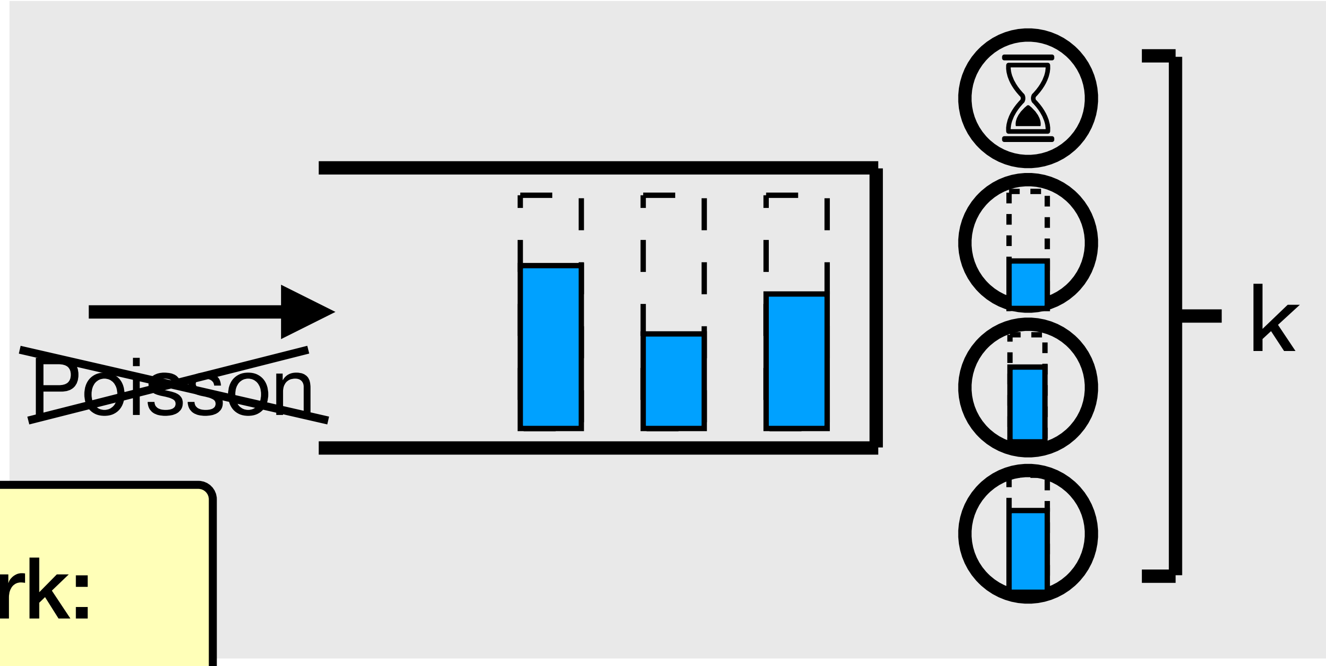
G/G/1



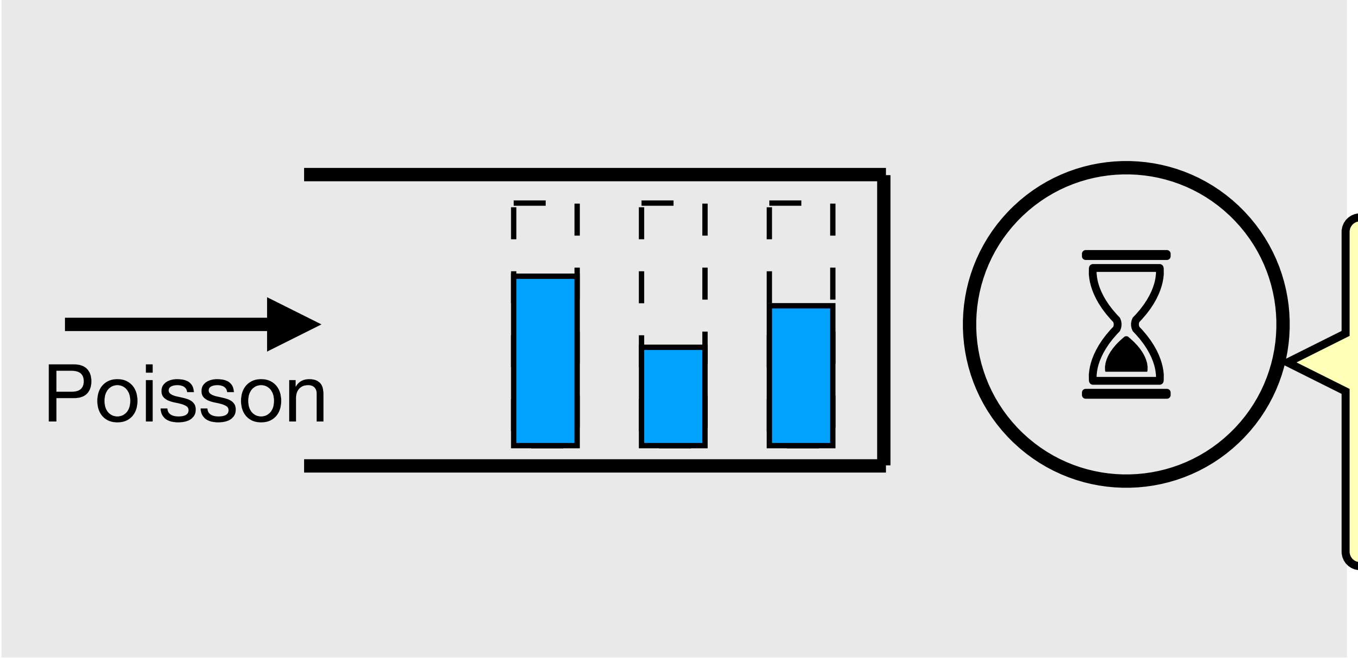
This work:
gap $\leq \ell_{(b)}$

$$\ell_{(b)} = O(1)$$

G/G/k/setup



M/G/1/setup

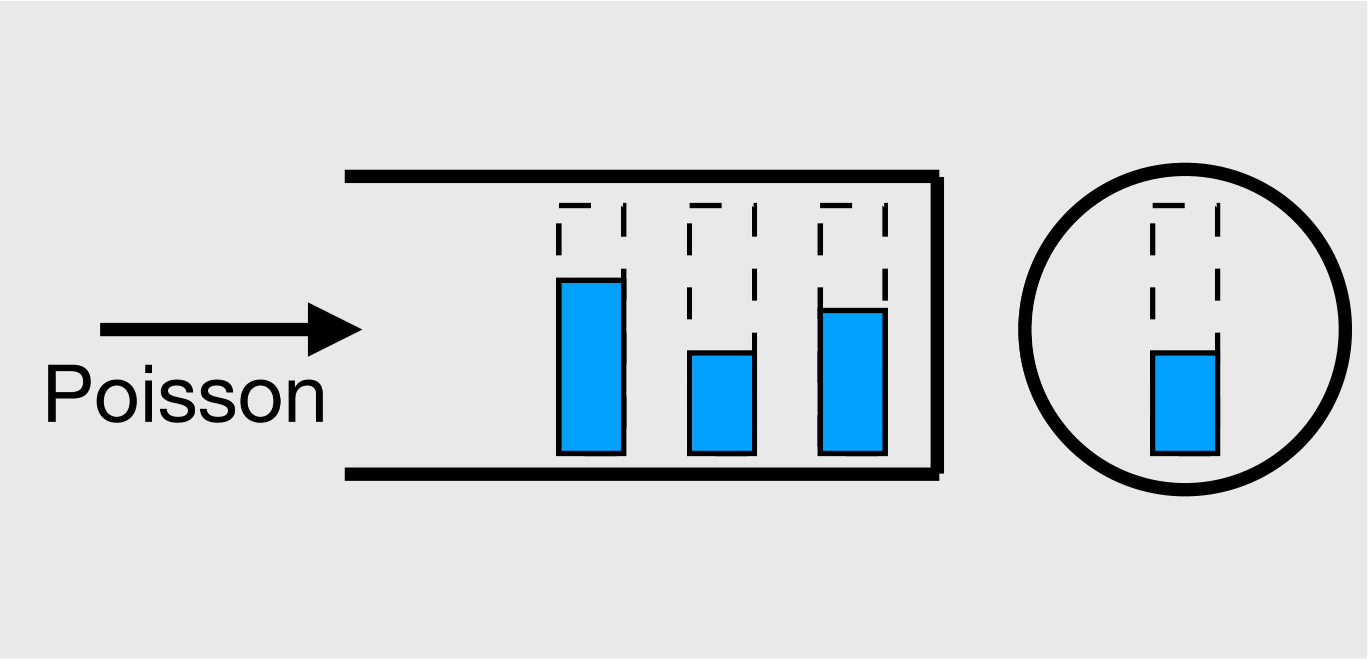


This work:
gap $\leq \ell_{(c)}$

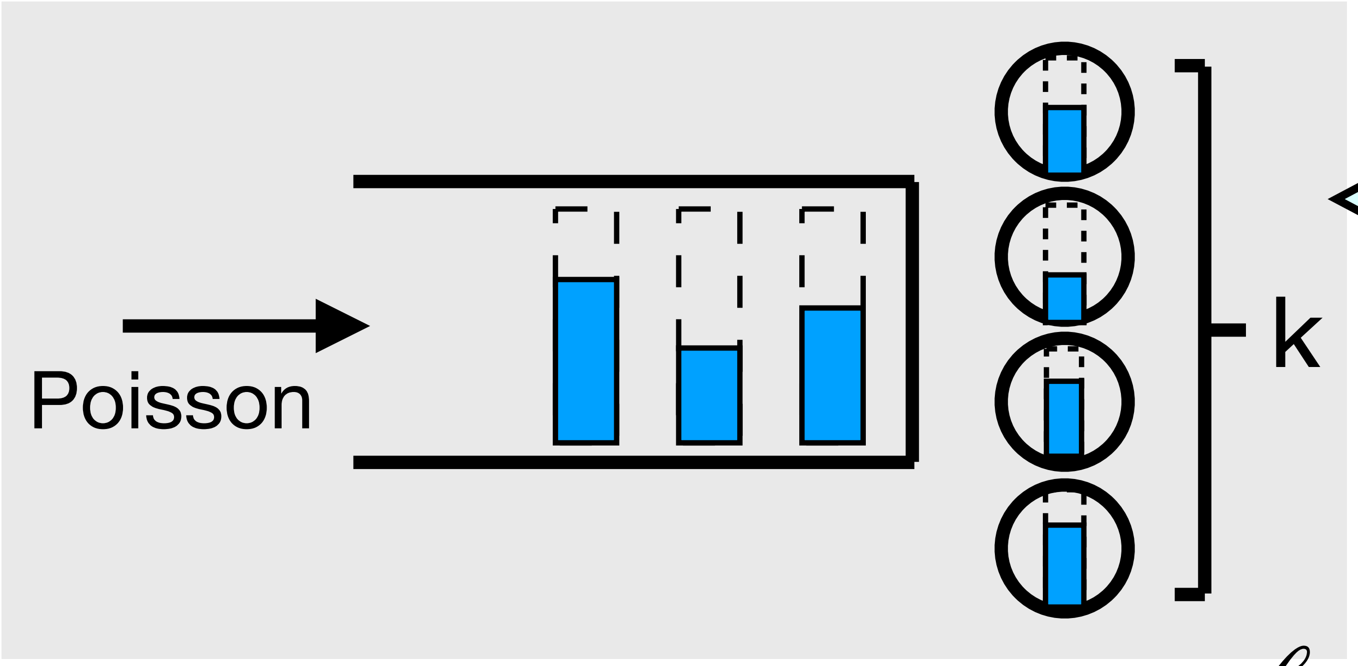
$$\ell_{(c)} = O(1)$$

Suboptimality gaps

M/G/1



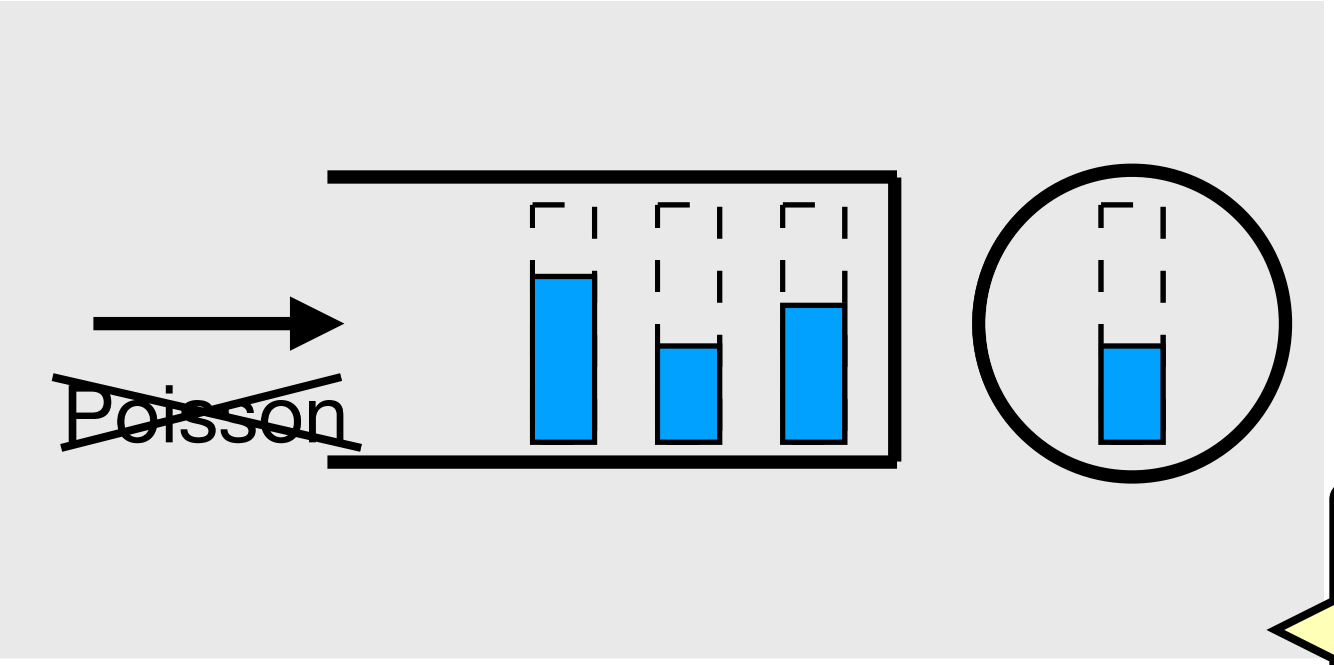
M/G/k



[SGH20]:
gap $\leq \ell_{(a)}$
heavy-traffic opt

$$\ell_{(a)} = O\left(\log \frac{1}{1-\rho}\right)$$

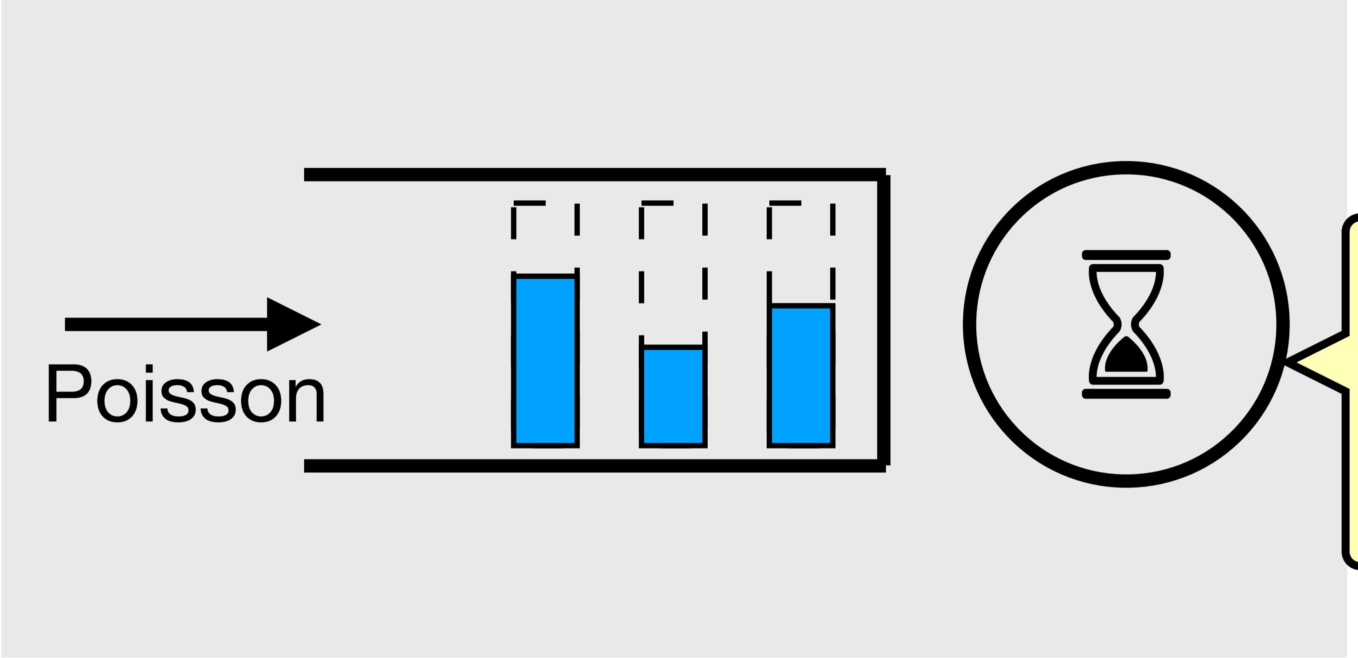
G/G/1



This work:
gap $\leq \ell_{(b)}$

$$\ell_{(b)} = O(1)$$

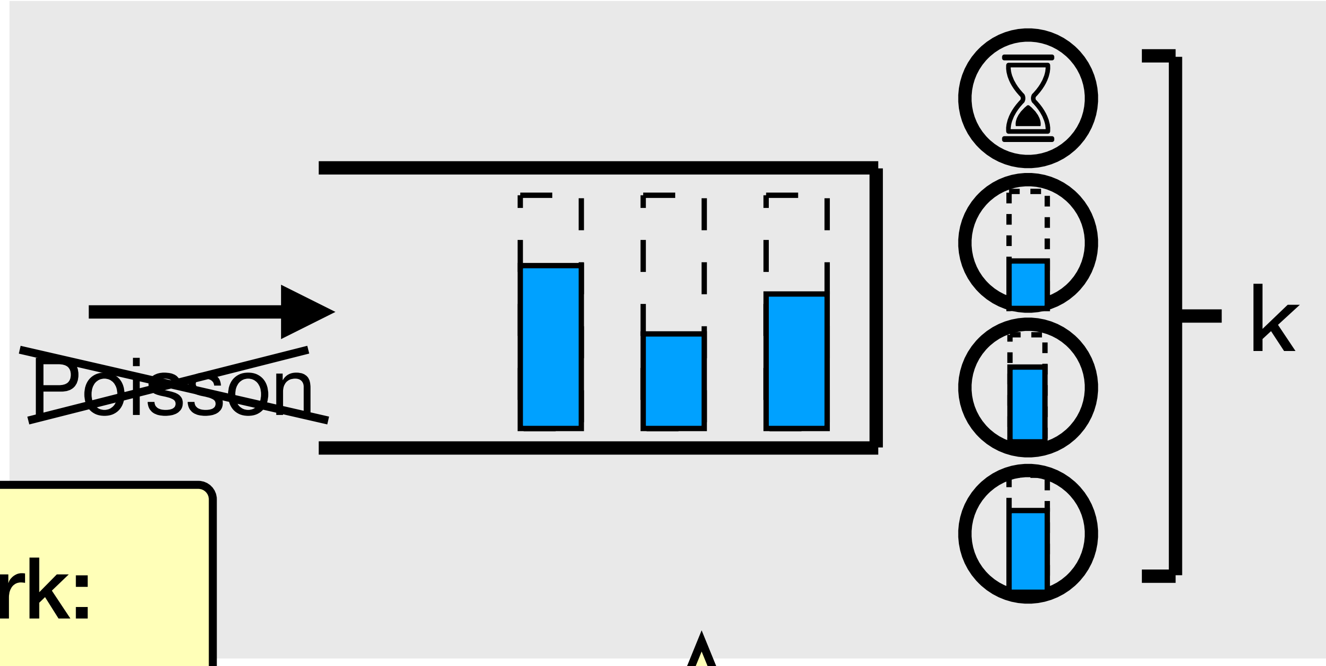
M/G/1/setup



This work:
gap $\leq \ell_{(c)}$

$$\ell_{(c)} = O(1)$$

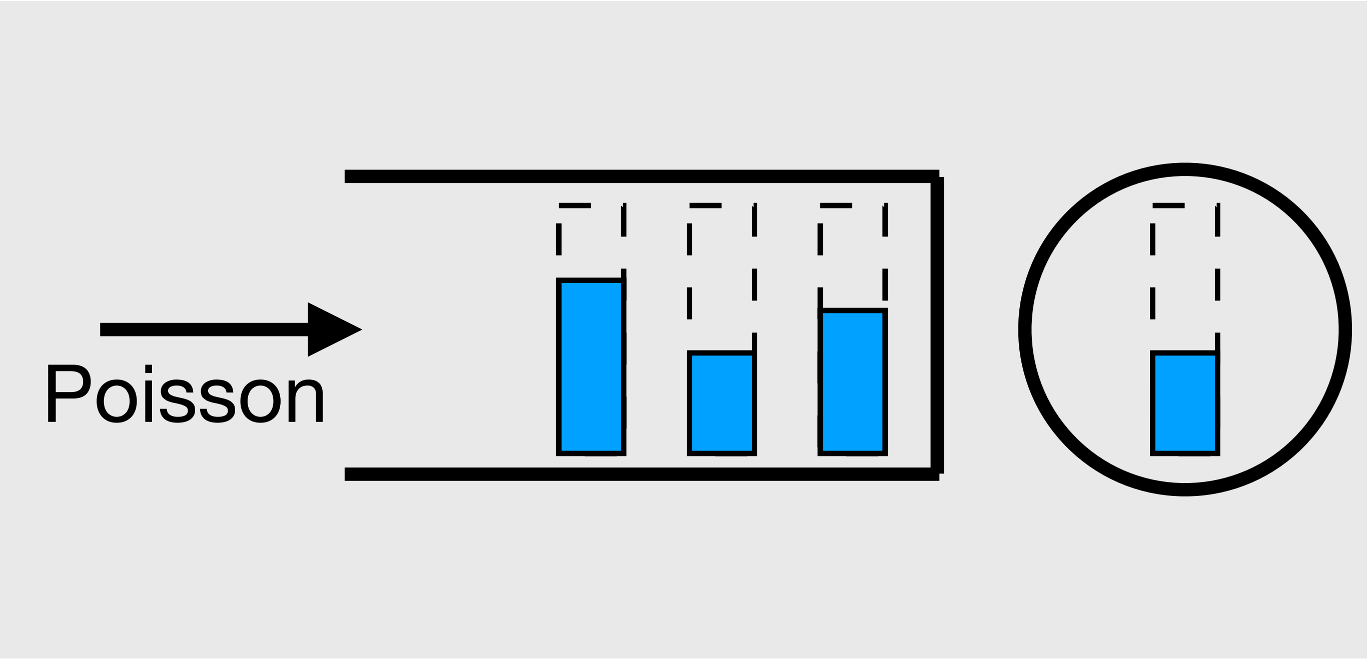
G/G/k/setup



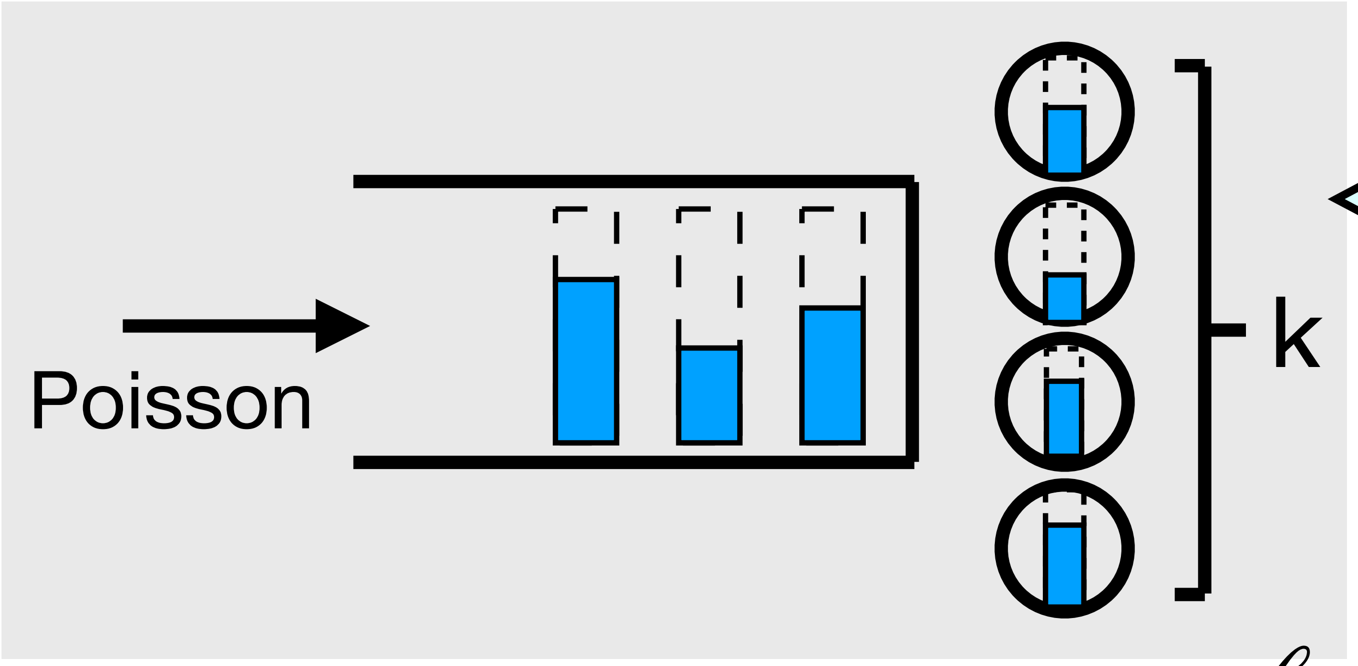
This work:
gap $\leq \ell_{(a)} + \ell_{(b)} + \ell_{(c)}$

Suboptimality gaps

M/G/1



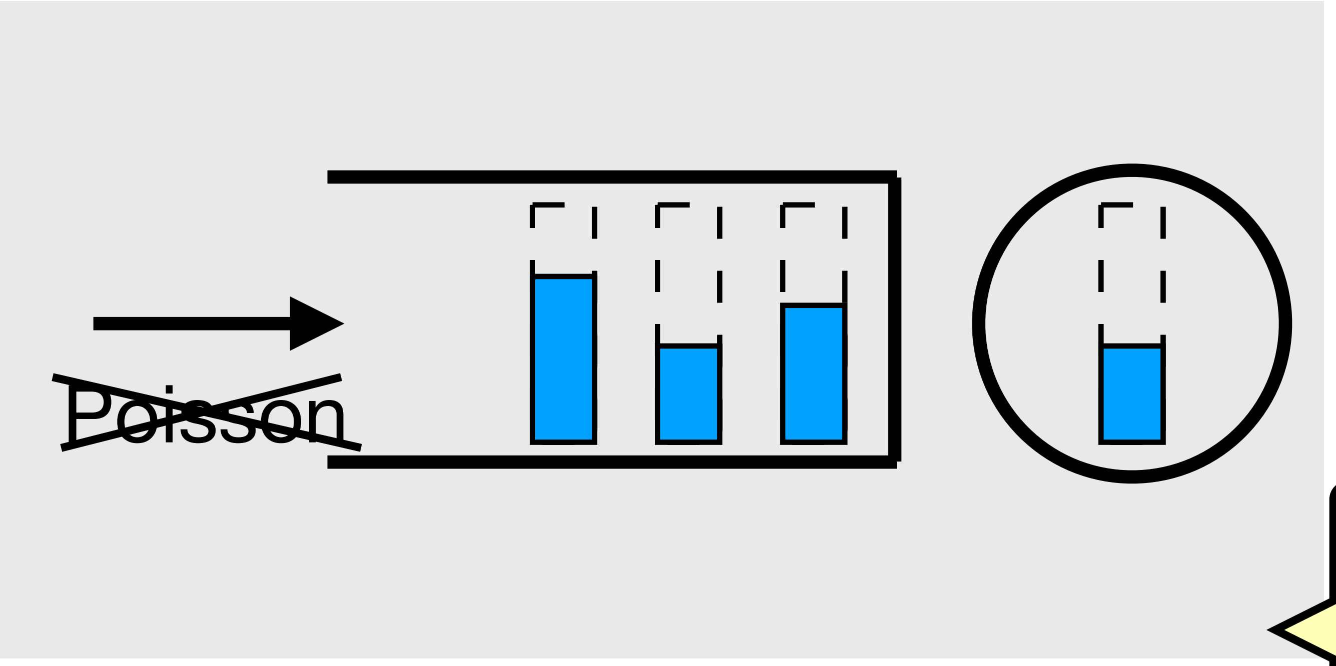
M/G/k



[SGH20]:
gap $\leq \ell_{(a)}$
heavy-traffic opt

$$\ell_{(a)} = O\left(\log \frac{1}{1-\rho}\right)$$

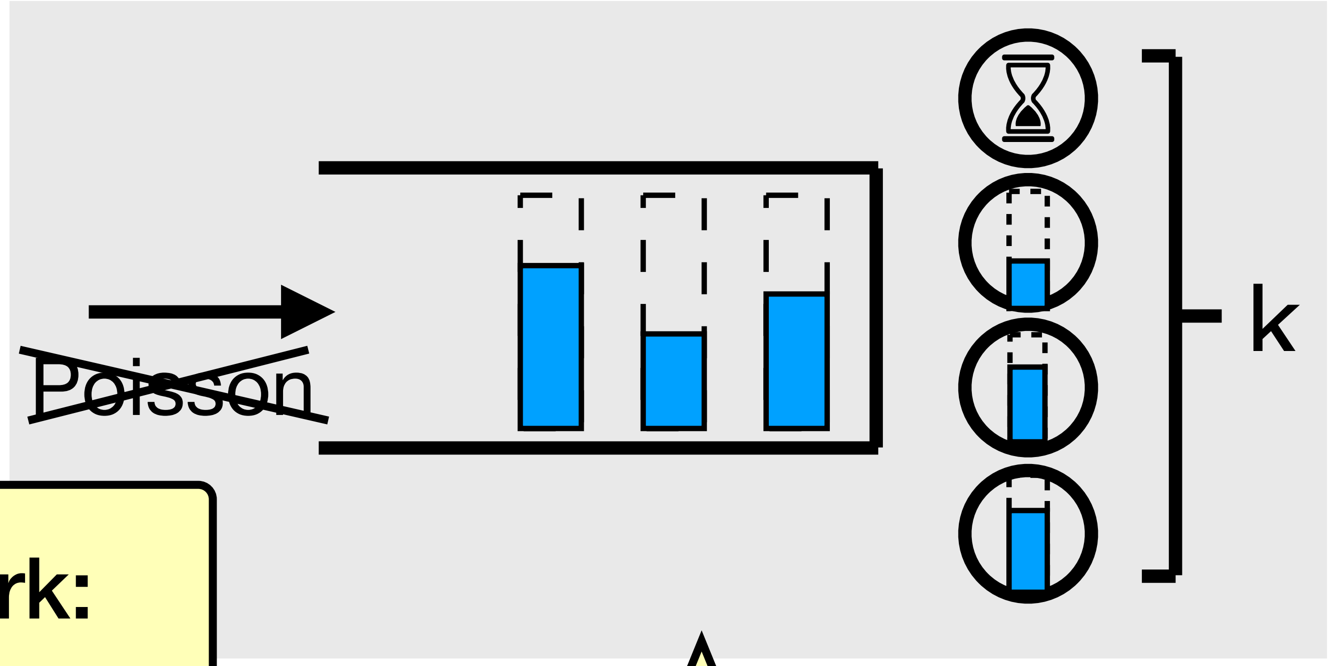
G/G/1



This work:
gap $\leq \ell_{(b)}$

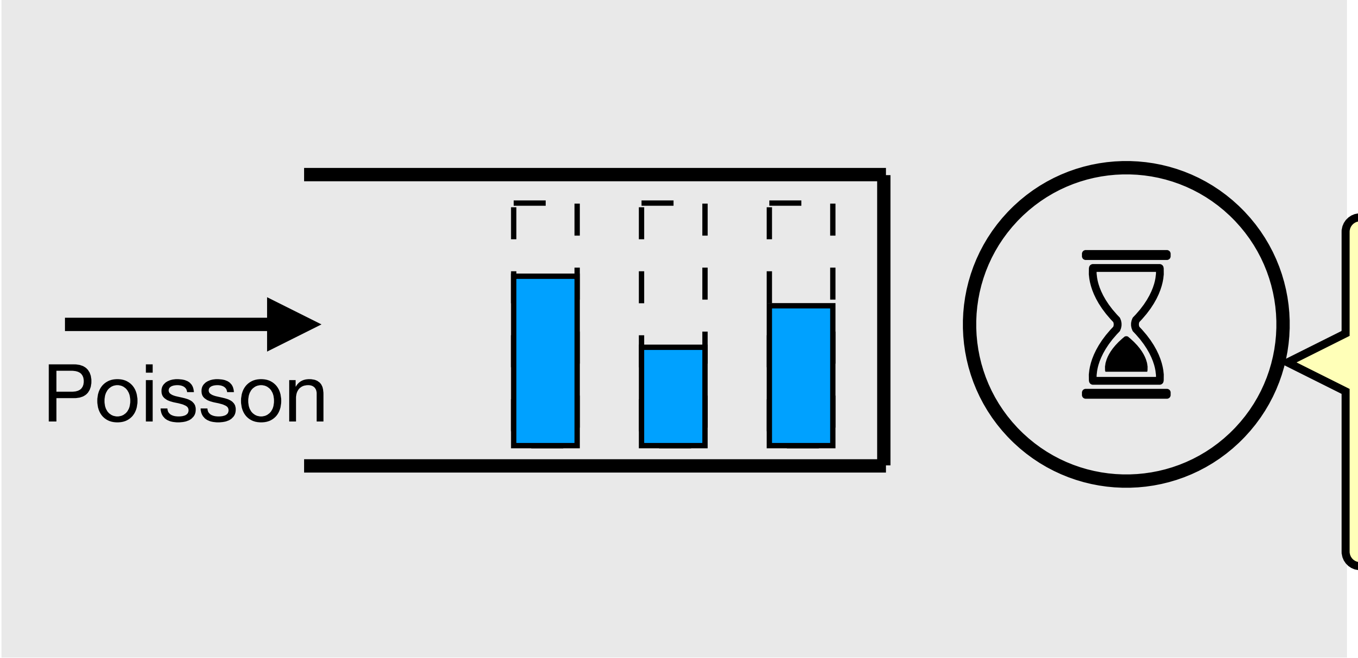
$$\ell_{(b)} = O(1)$$

G/G/k/setup



This work:
gap $\leq \ell_{(a)} + \ell_{(b)} + \ell_{(c)}$

M/G/1/setup



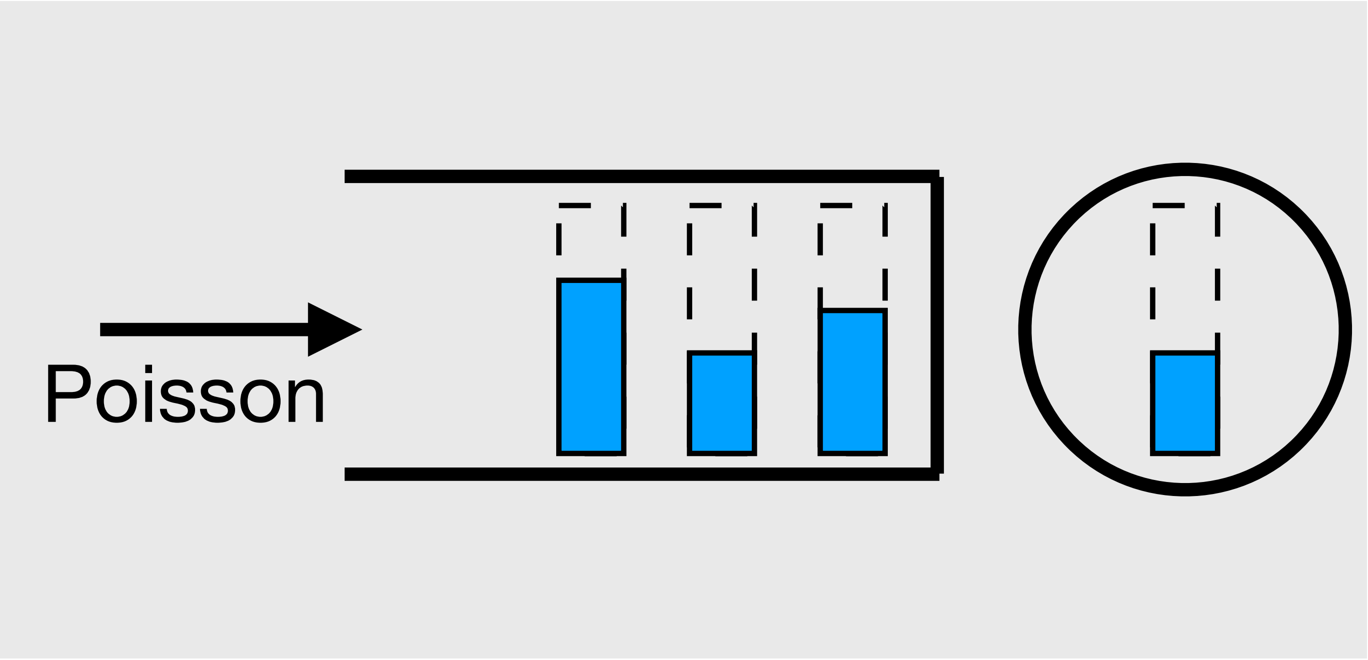
This work:
gap $\leq \ell_{(c)}$

$$\ell_{(c)} = O(1)$$

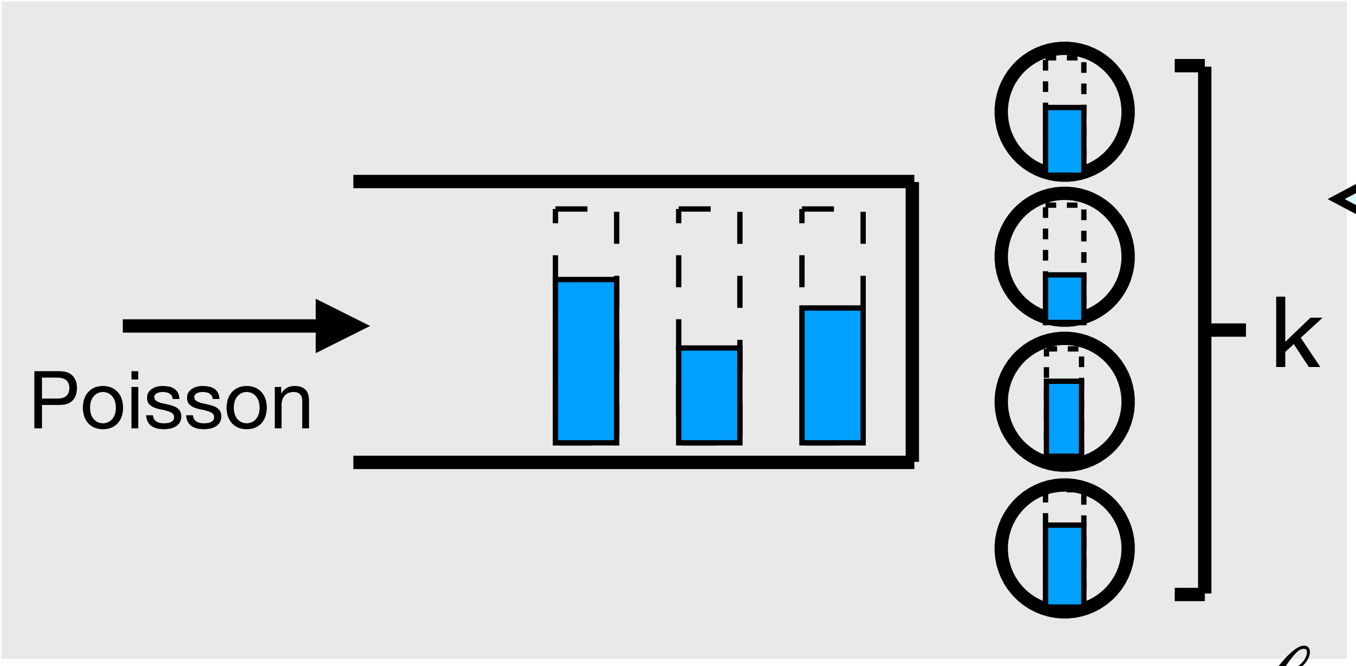
All cases:
heavy-traffic
opt

Suboptimality gaps

M/G/1



M/G/k

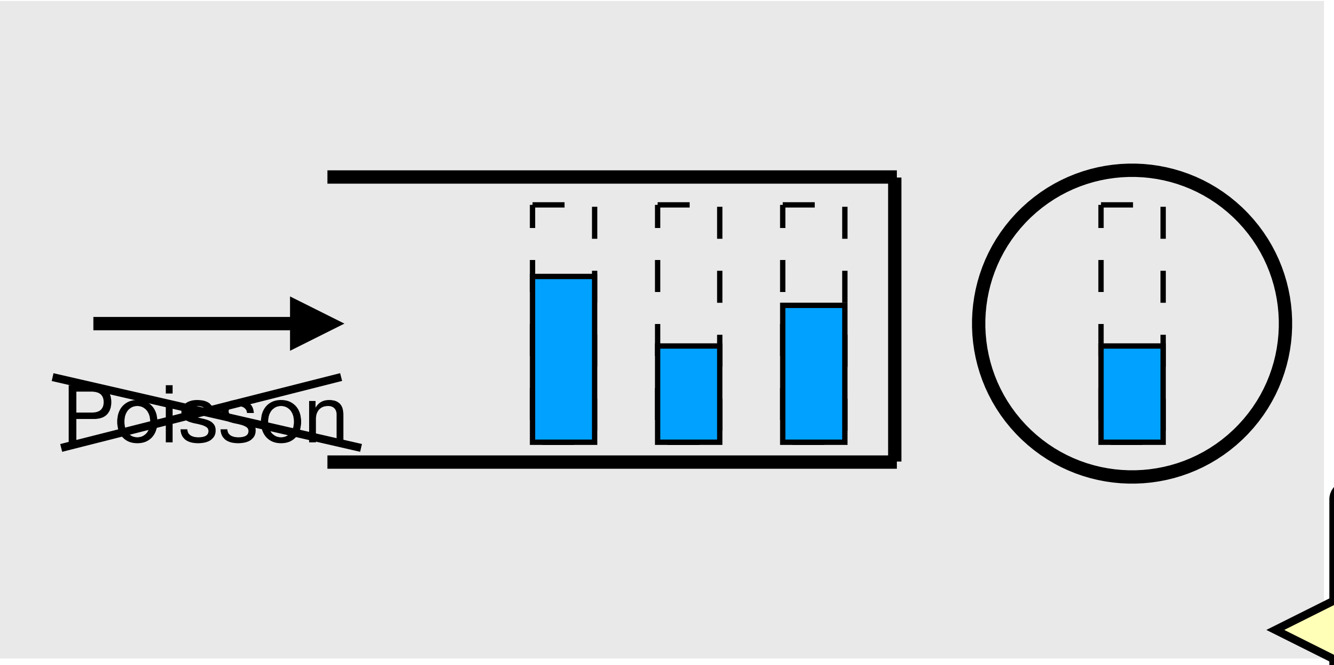


[SGH20]:
gap $\leq \ell_{(a)}$
heavy-traffic opt

Assumption:
interarrival times
light-tailed

$$\ell_{(a)} = O\left(\log \frac{1}{1-\rho}\right)$$

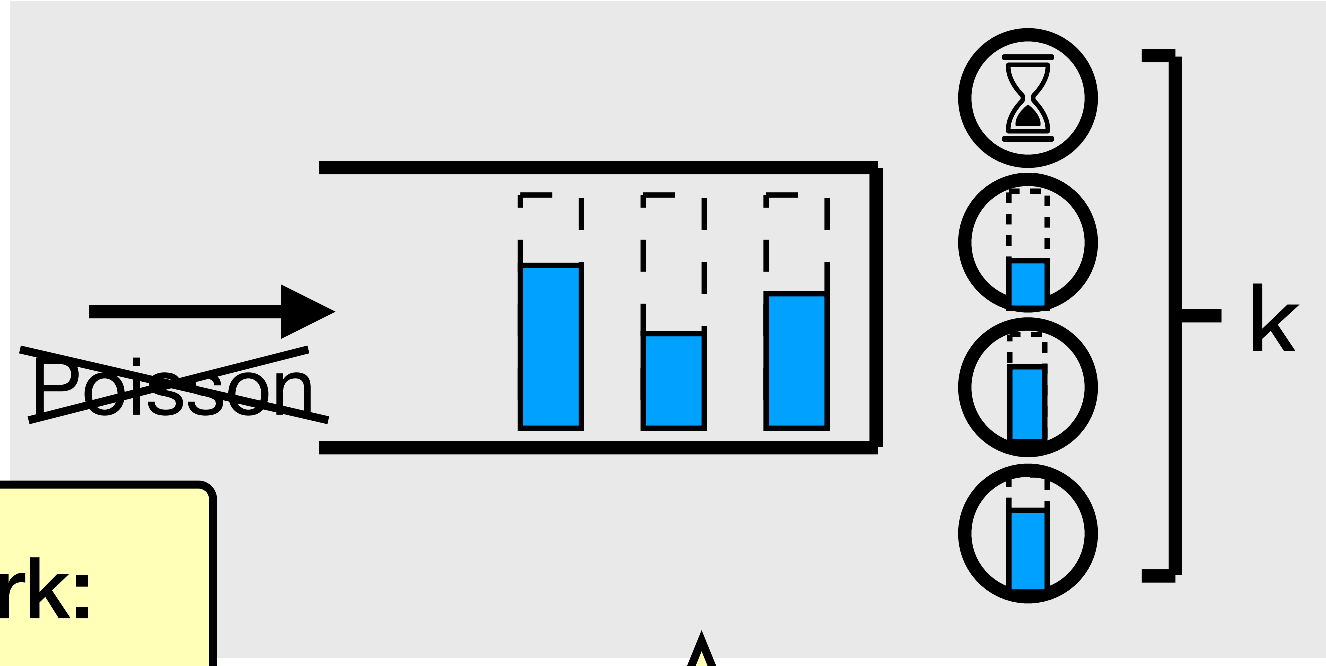
G/G/1



This work:
gap $\leq \ell_{(b)}$

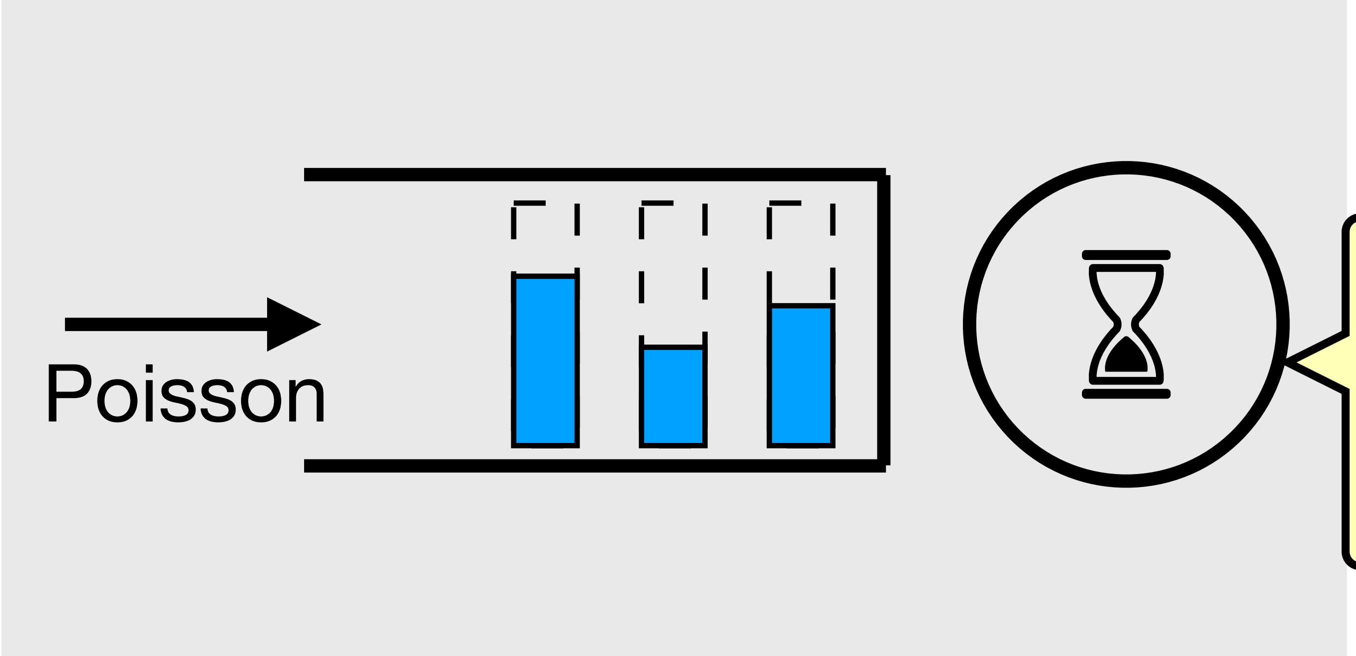
$$\ell_{(b)} = O(1)$$

G/G/k/setup



This work:
gap $\leq \ell_{(a)} + \ell_{(b)} + \ell_{(c)}$

M/G/1/setup



This work:
gap $\leq \ell_{(c)}$

$$\ell_{(c)} = O(1)$$

All cases:
heavy-traffic
opt

Outline for the rest of the talk

- What is our problem and result? ✓

Outline for the rest of the talk

G/G/k/setup-Gittins

- What is our problem and result? ✓

Outline for the rest of the talk

G/G/k/setup-Gittins

suboptimality gaps
+ heavy-traffic opt

- What is our problem and result? ✓

Outline for the rest of the talk

G/G/k/setup-Gittins

suboptimality gaps
+ heavy-traffic opt

- What is our problem and result? ✓
- How does our G/G/k/setup analysis work?

Outline for the rest of the talk

G/G/k/setup-Gittins

suboptimality gaps
+ heavy-traffic opt

- What is our problem and result? ✓
- How does our G/G/k/setup analysis work?
- What is the main obstacle and how do we solve it?

Analysis roadmap of G/G/k/setup

Analysis roadmap of G/G/k/setup

based on [SGH20]

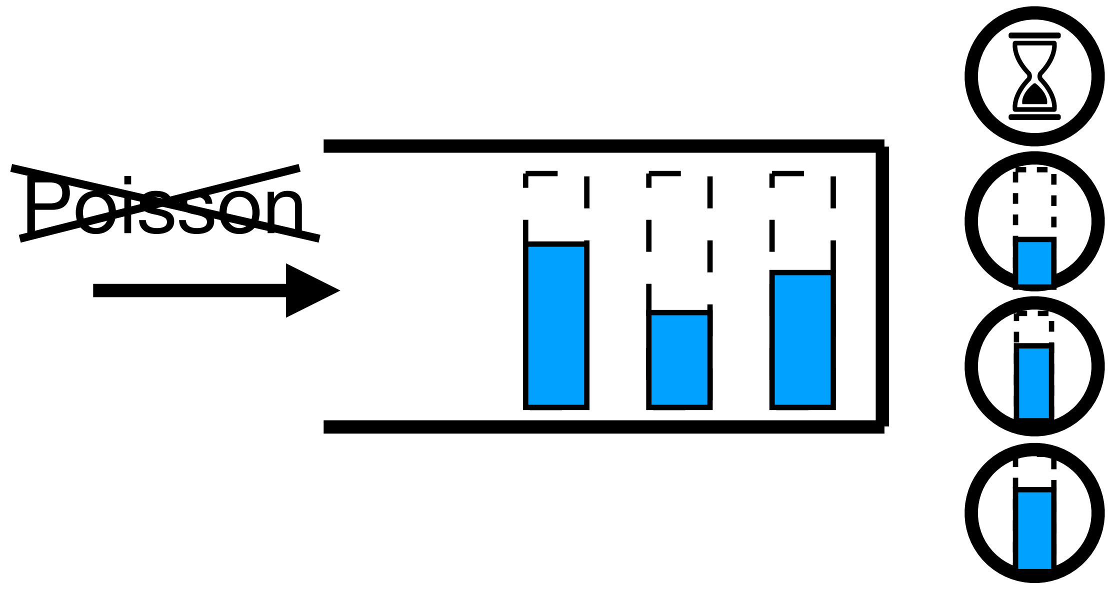
Analysis roadmap of G/G/k/setup

based on [SGH20]

number in

Analysis roadmap of G/G/k/setup

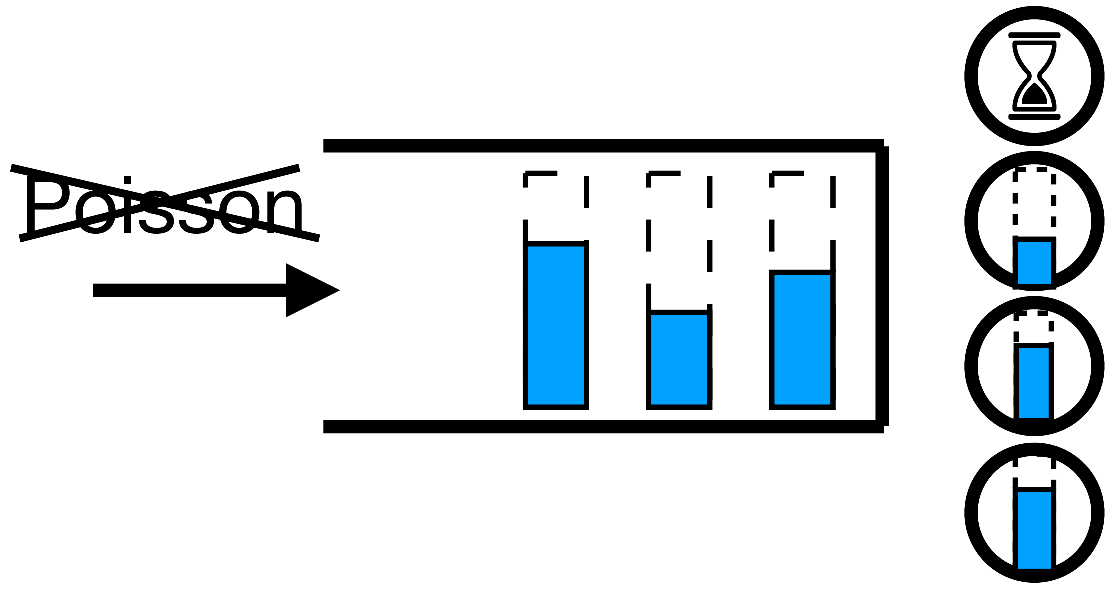
based on [SGH20]



number in **G/G/k/setup-Gittins**

Analysis roadmap of G/G/k/setup

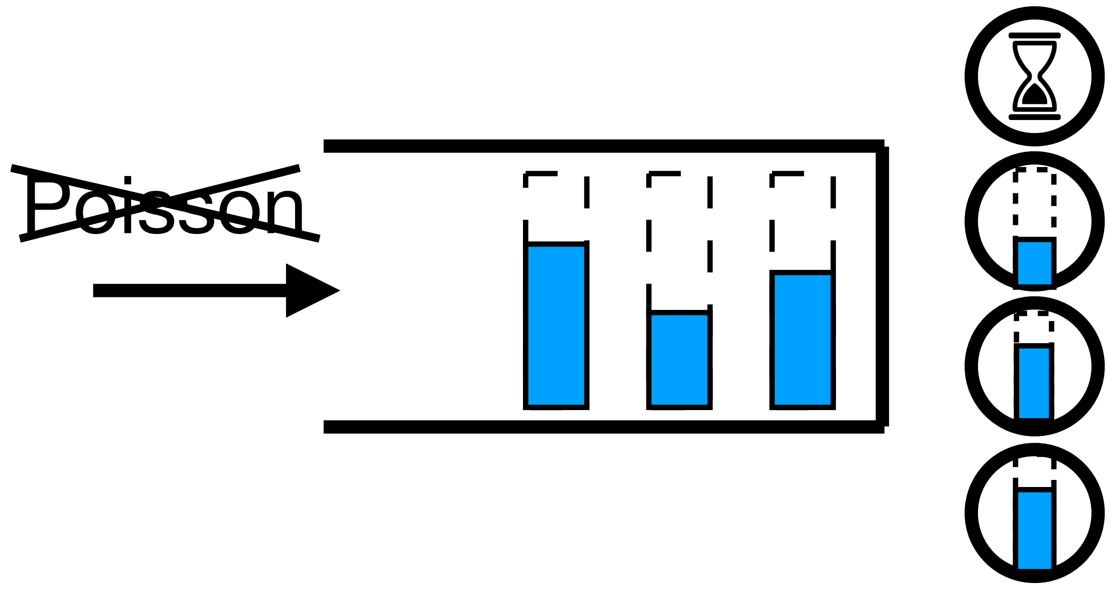
based on [SGH20]

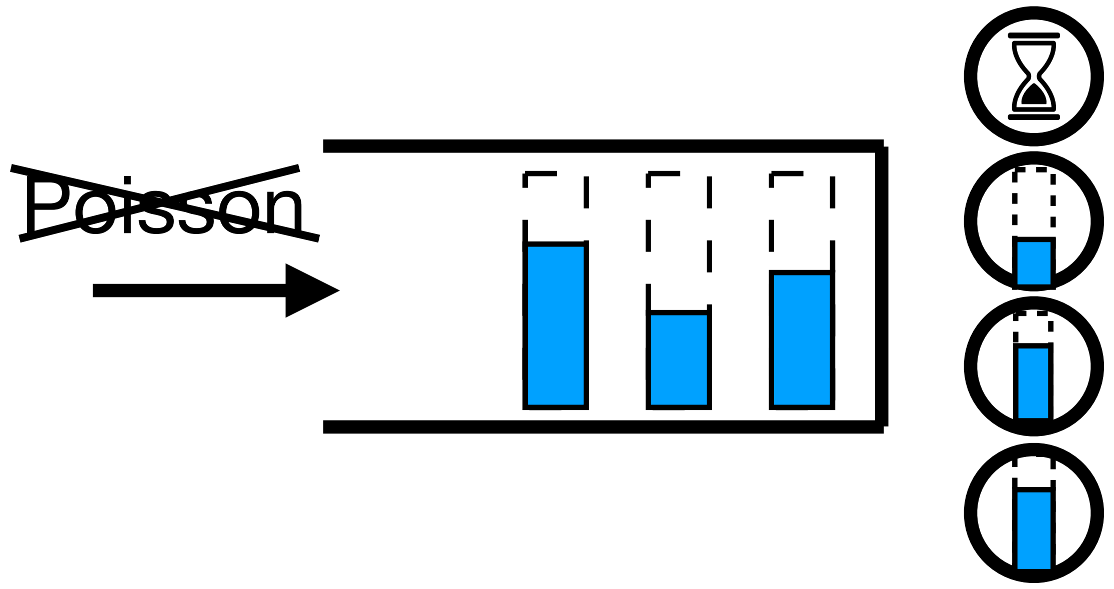


number in **G/G/k/setup-Gittins** $\geq$

Analysis roadmap of G/G/k/setup

based on [SGH20]

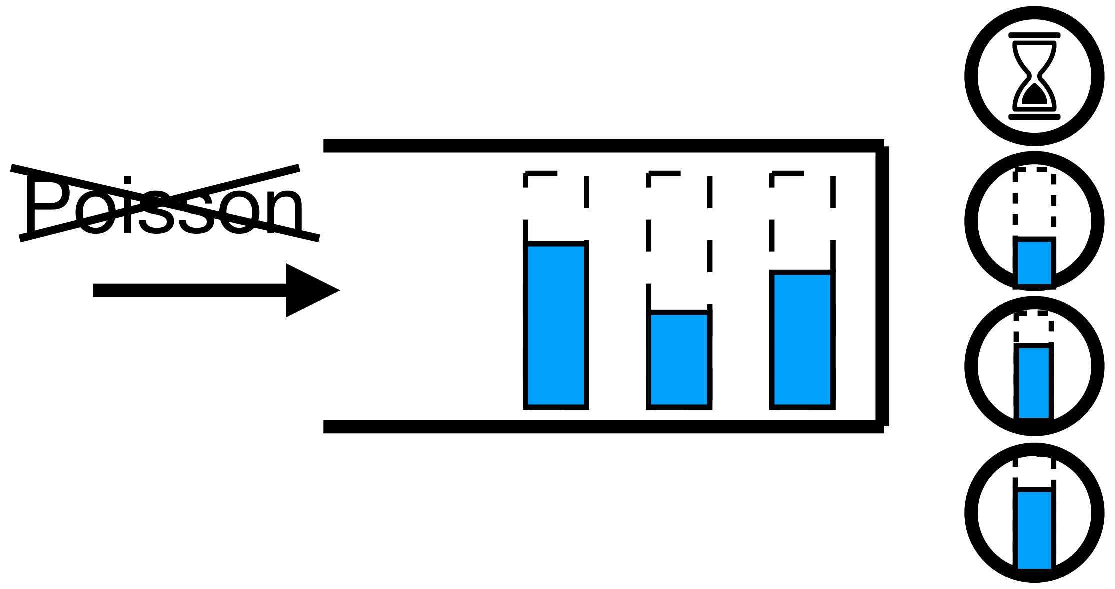


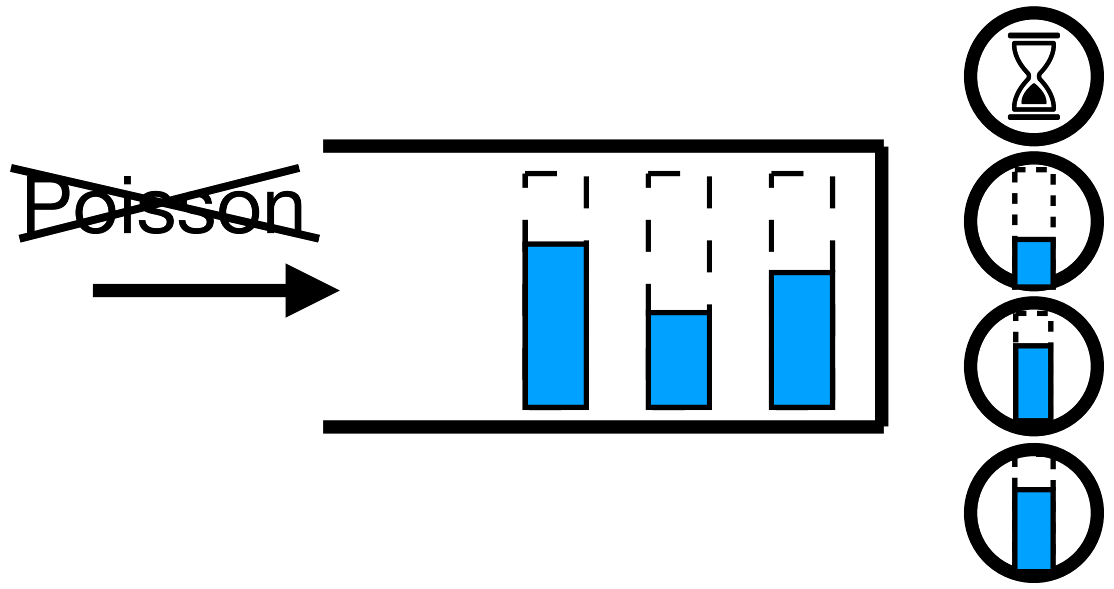


number in **G/G/k/setup-Gittins** $\geq$ **G/G/k/setup-OPT**

Analysis roadmap of G/G/k/setup

based on [SGH20]



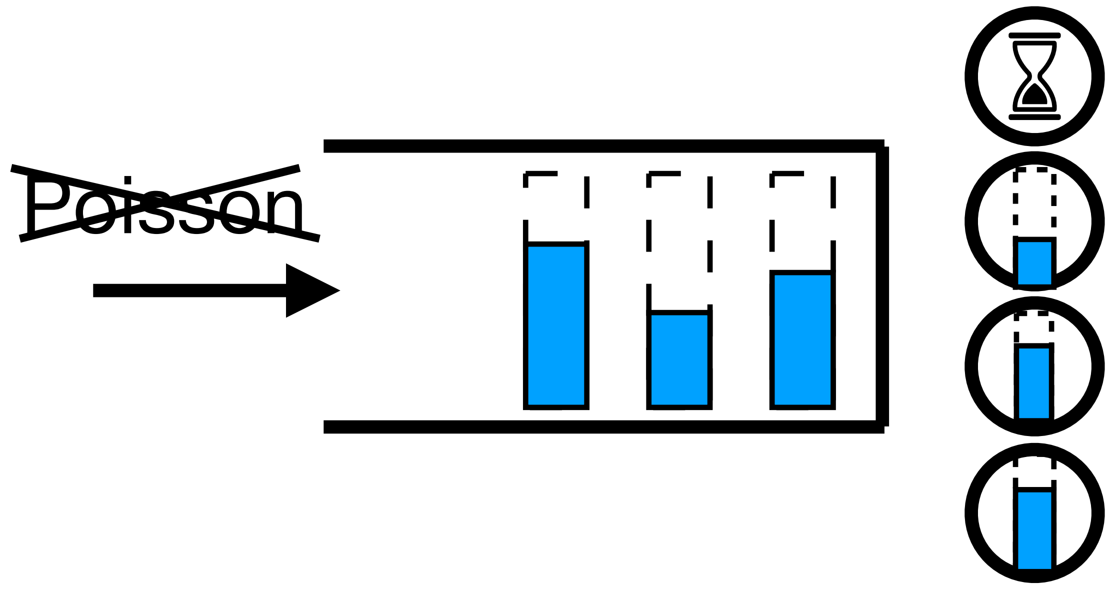


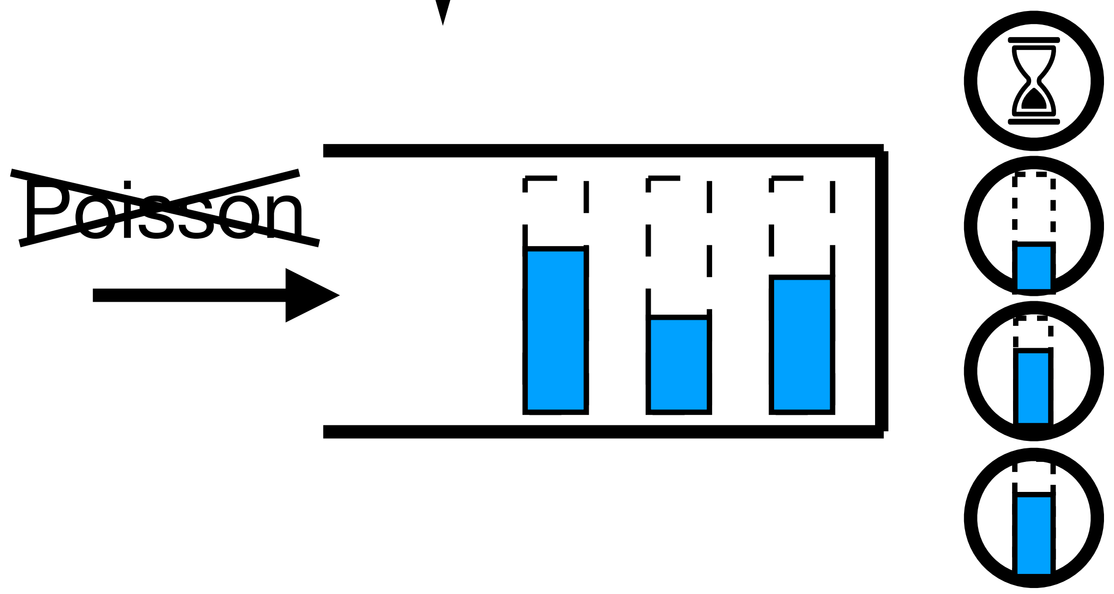
number in $\underbrace{\textbf{G/G/k/setup-Gittins} \geq \textbf{G/G/k/setup-OPT}}_{\text{Gap}}$

Analysis roadmap of G/G/k/setup

based on [SGH20]

Unknown and complicated



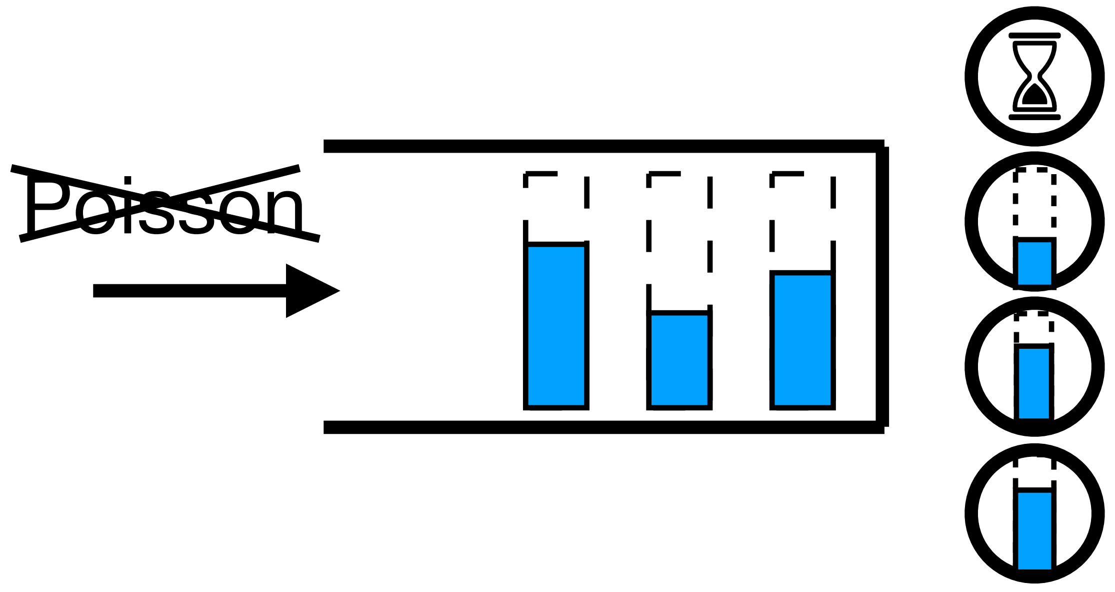


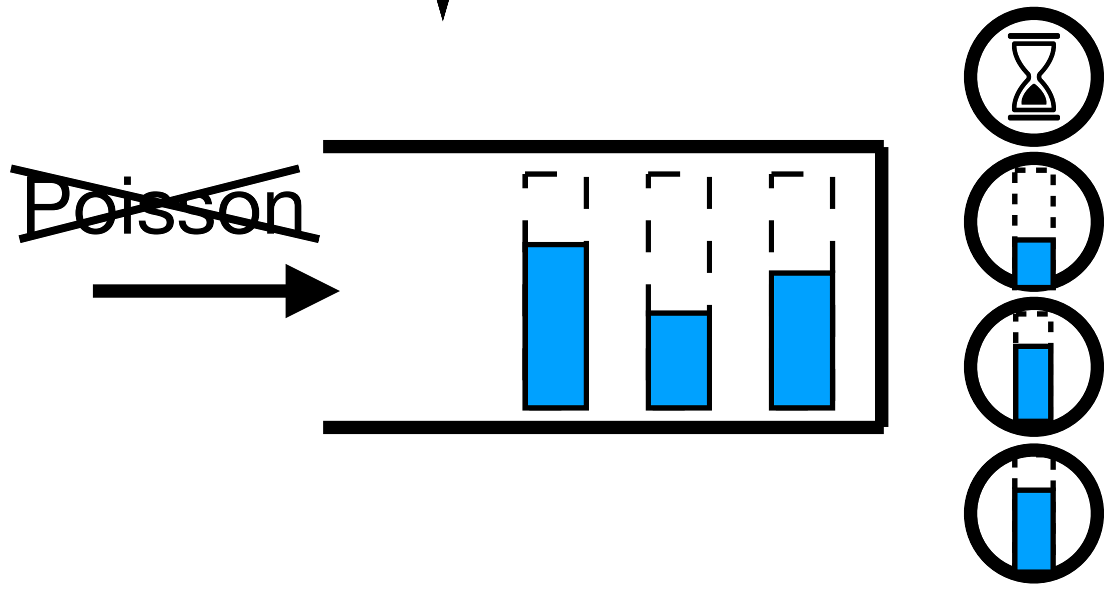
number in $\underbrace{\textbf{G/G/k/setup-Gittins} \geq \textbf{G/G/k/setup-OPT}}_{\text{Gap}}$

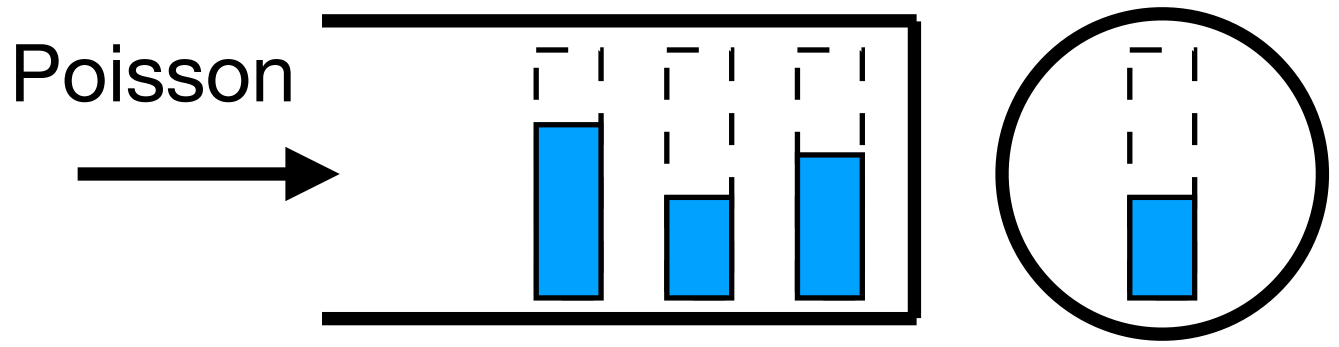
Analysis roadmap of G/G/k/setup

based on [SGH20]

Unknown and complicated







number in **G/G/k/setup-Gittins** ≥ **G/G/k/setup-OPT** ≥ **G/G/1-OPT**

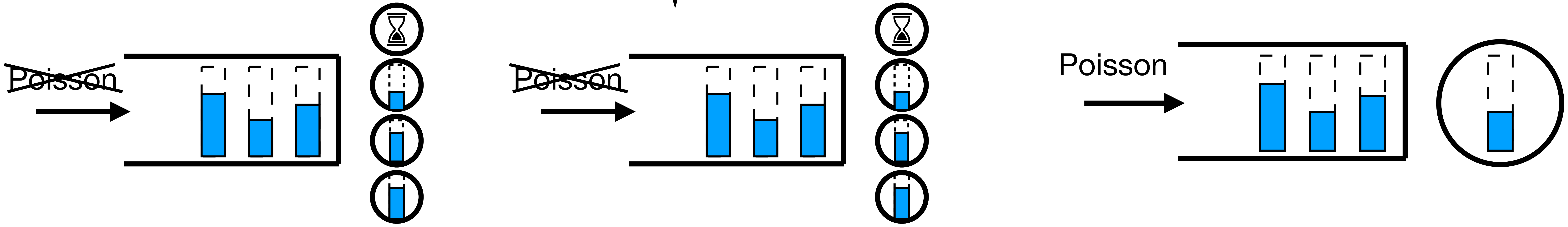
Gap

Analysis roadmap of G/G/k/setup

based on [SGH20]

Unknown and complicated

Simpler

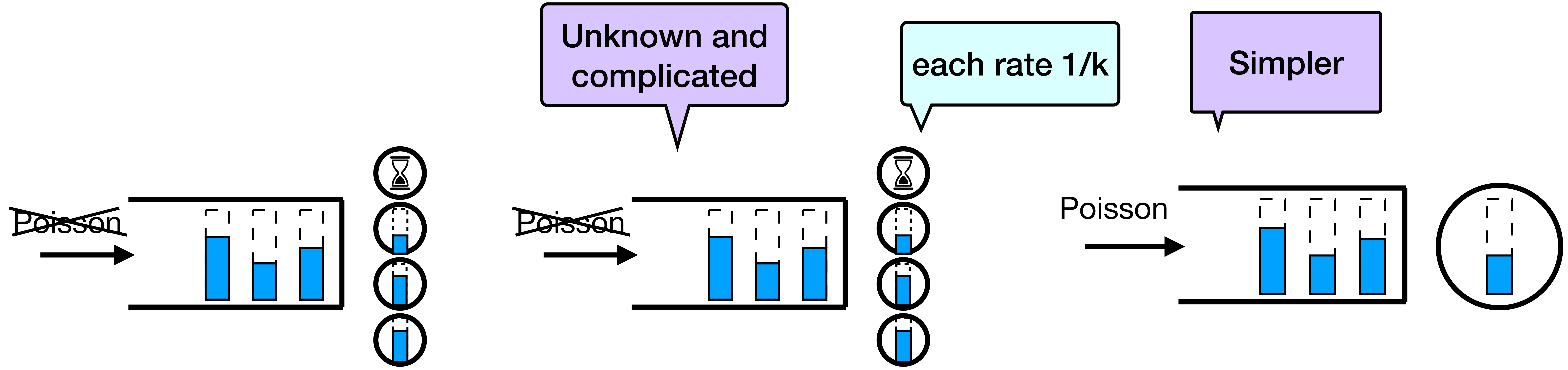


number in **G/G/k/setup-Gittins** $\geq$ **G/G/k/setup-OPT** $\geq$ **G/G/1-OPT**

$$\underbrace{\text{G/G/k/setup-Gittins} \geq \text{G/G/k/setup-OPT}}_{\text{Gap}}$$

Analysis roadmap of G/G/k/setup

based on [SGH20]

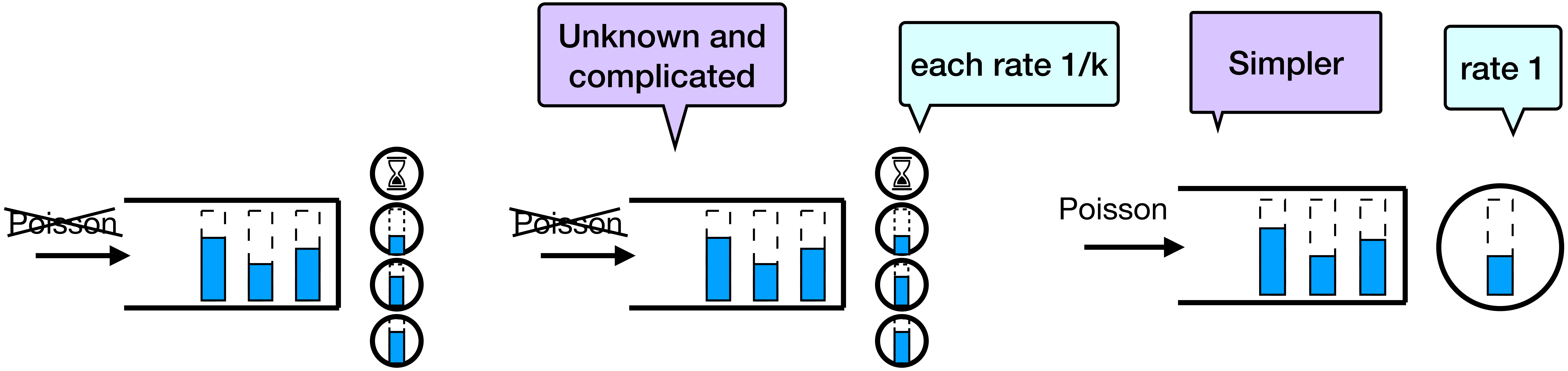


number in **G/G/k/setup-Gittins** ≥ **G/G/k/setup-OPT** ≥ **G/G/1-OPT**

Gap

Analysis roadmap of G/G/k/setup

based on [SGH20]

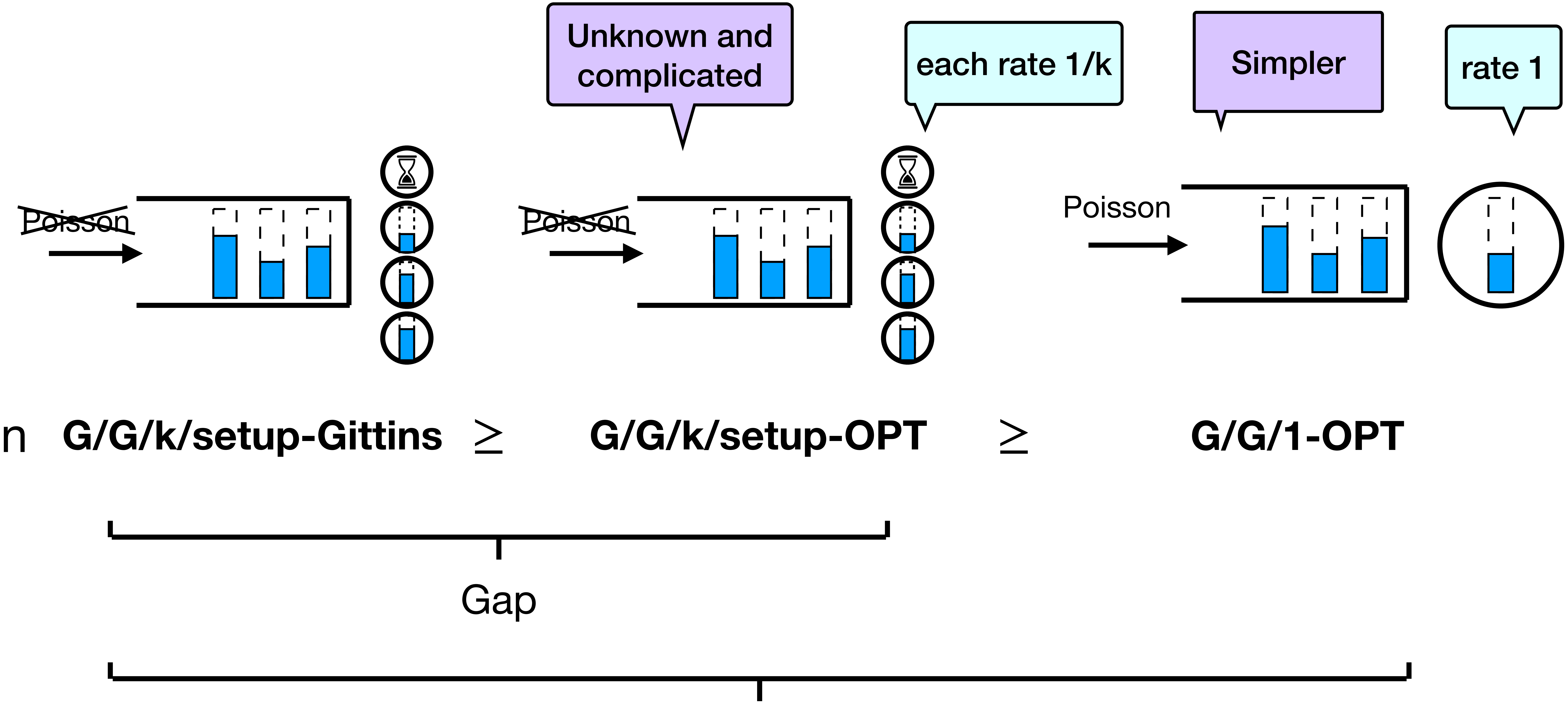


number in **G/G/k/setup-Gittins** ≥ **G/G/k/setup-OPT** ≥ **G/G/1-OPT**

Gap

Analysis roadmap of G/G/k/setup

based on [SGH20]



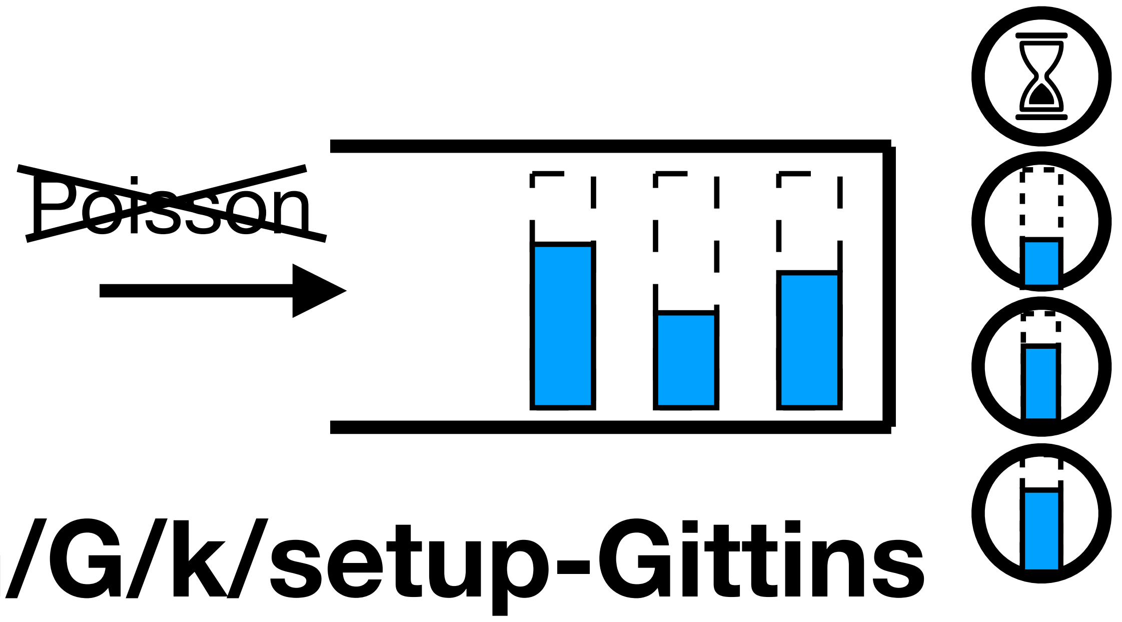
number in **G/G/k/setup-Gittins** $\geq$ **G/G/k/setup-OPT** $\geq$ **G/G/1-OPT**

Gap

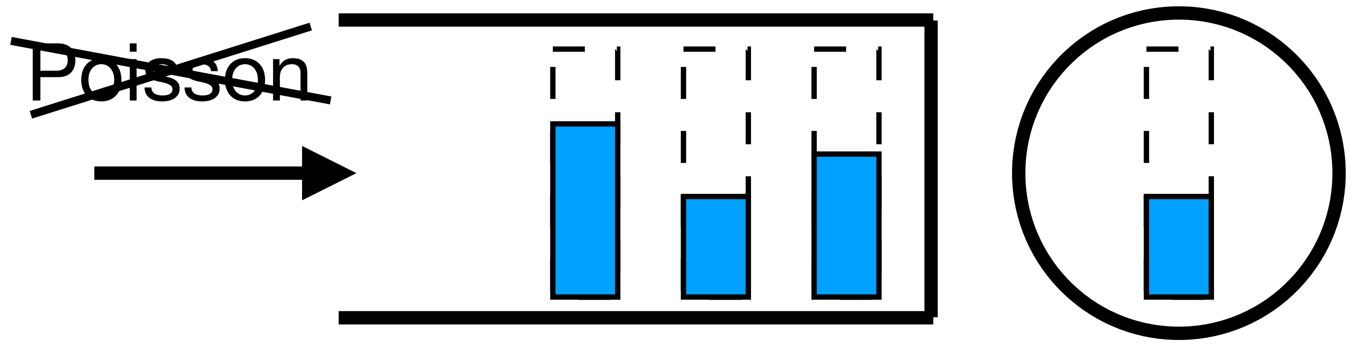
Enough to bound the difference

Analysis roadmap of G/G/k/setup

based on [SGH20]



G/G/k/setup-Gittins

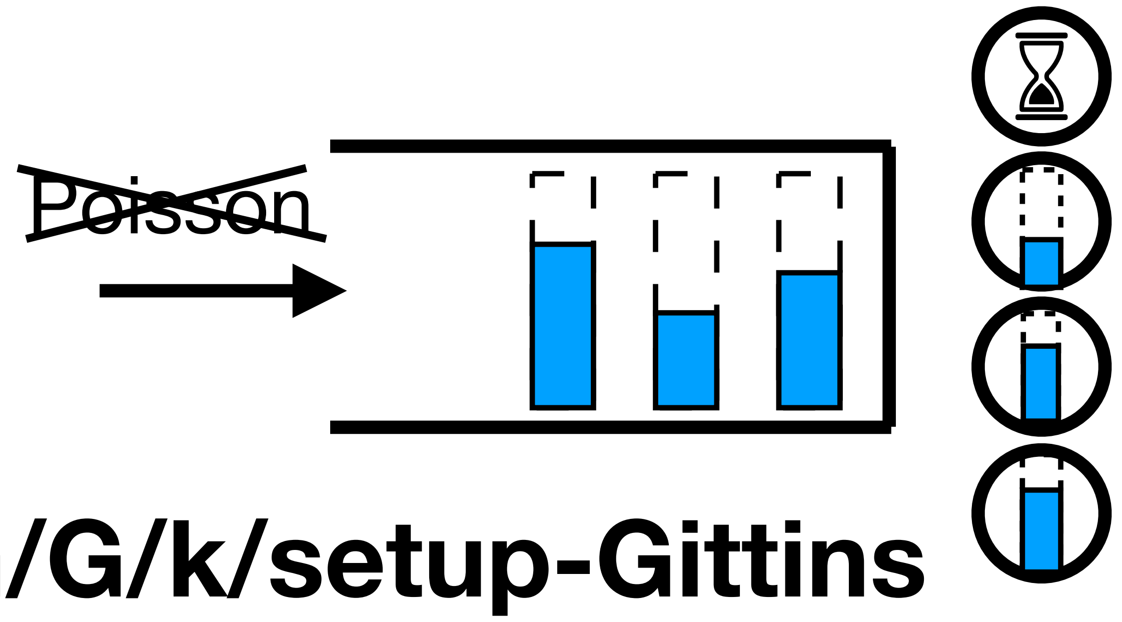


G/G/1-OPT

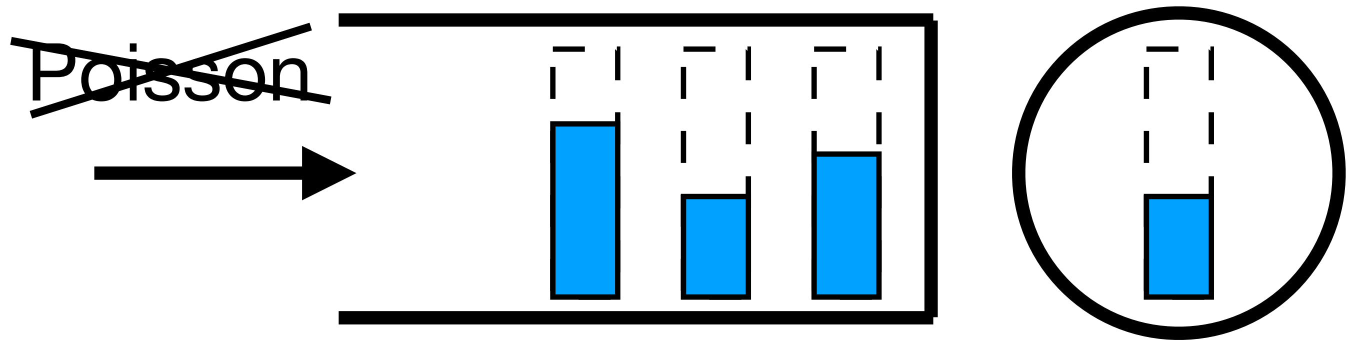
number in G/G/k-setup ←- - - gap - - -→ number in G/G/1

Analysis roadmap of G/G/k/setup

based on [SGH20]



G/G/k/setup-Gittins



G/G/1-OPT

number in G/G/k-setup ↔ gap ↔ number in G/G/1

↕

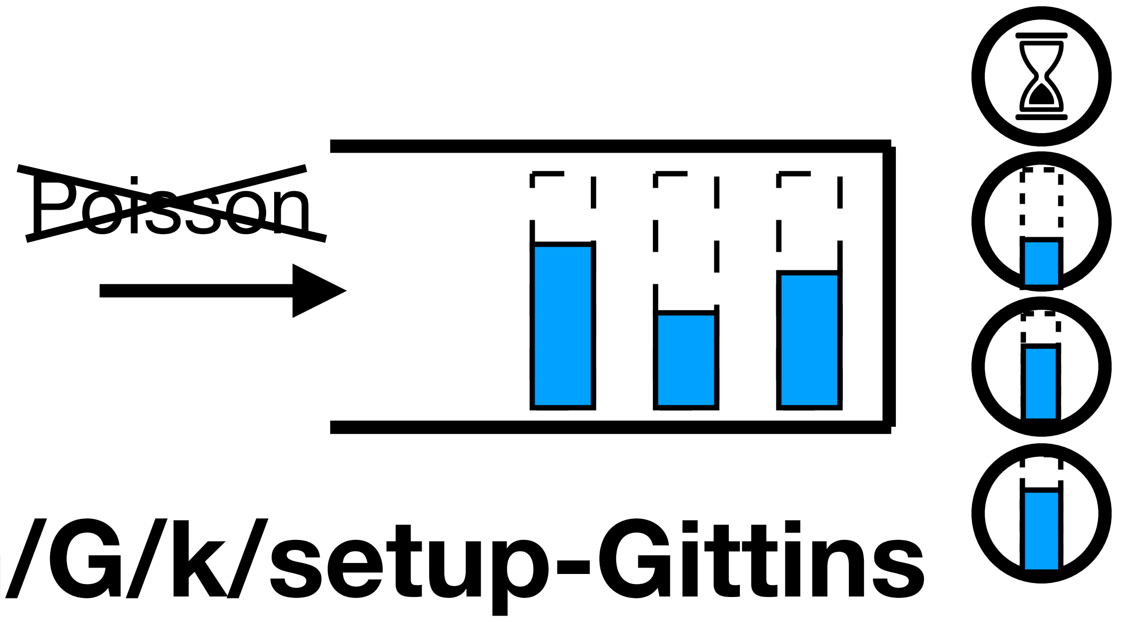
work in G/G/k-setup

↕

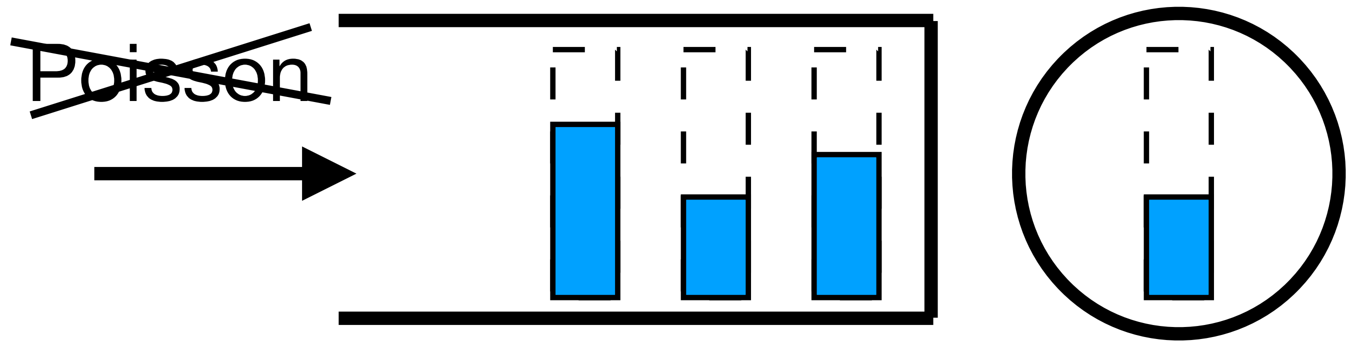
work in G/G/1

Analysis roadmap of G/G/k/setup

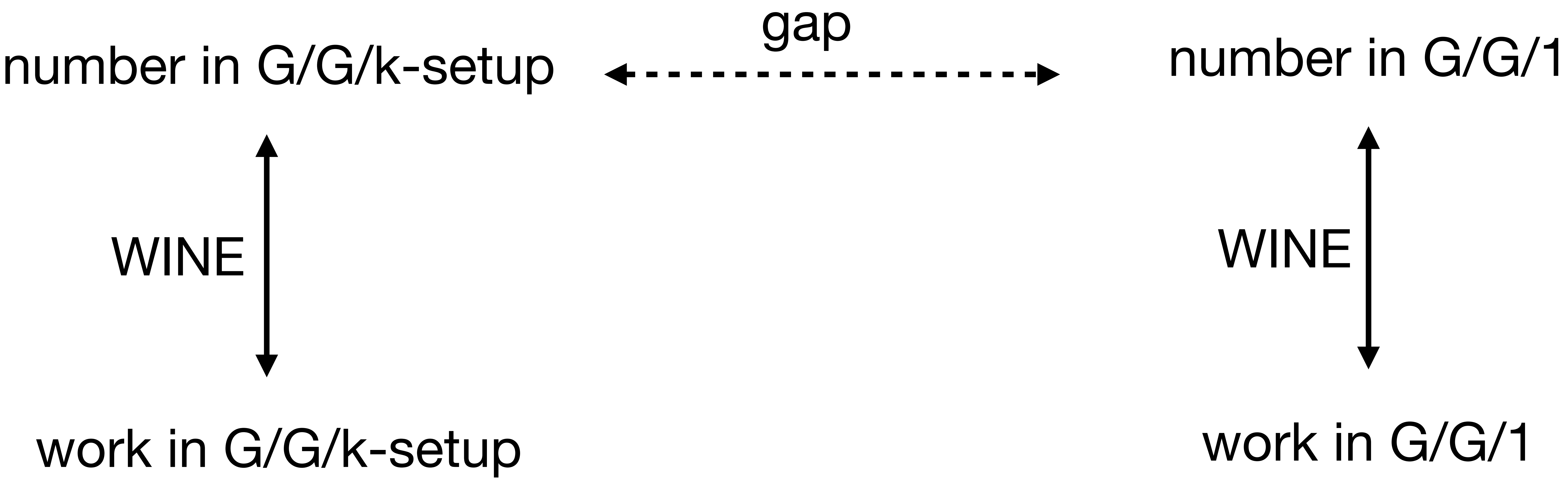
based on [SGH20]



G/G/k/setup-Gittins



G/G/1-OPT



number in G/G/k-setup

gap

number in G/G/1

WINE

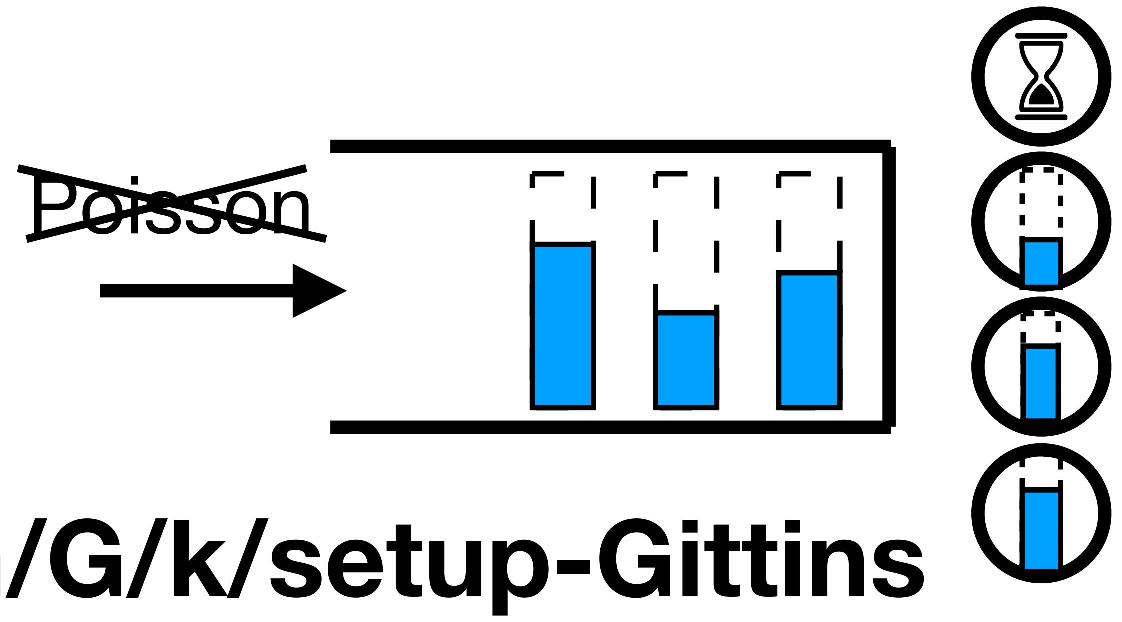
WINE

work in G/G/k-setup

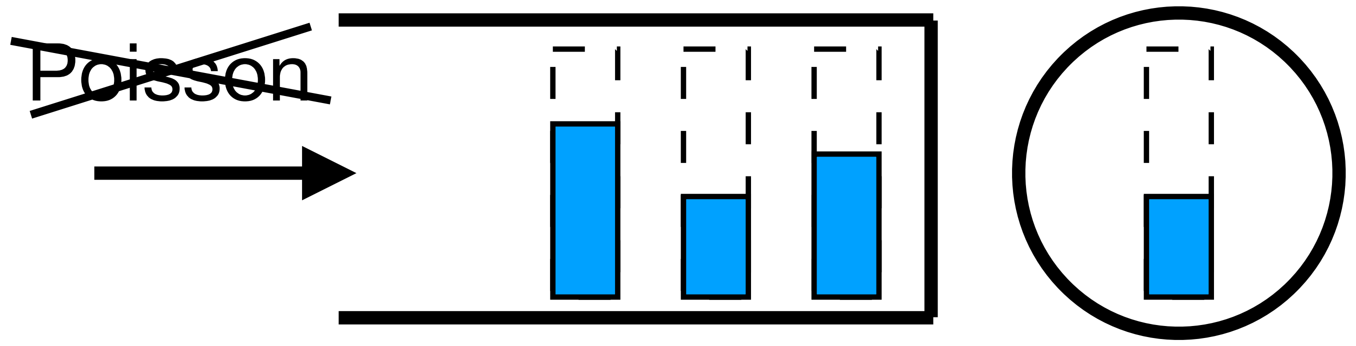
work in G/G/1

Analysis roadmap of G/G/k/setup

based on [SGH20]



G/G/k/setup-Gittins



G/G/1-OPT

number in G/G/k-setup ← gap → number in G/G/1

SGH20, general

WINE ↕

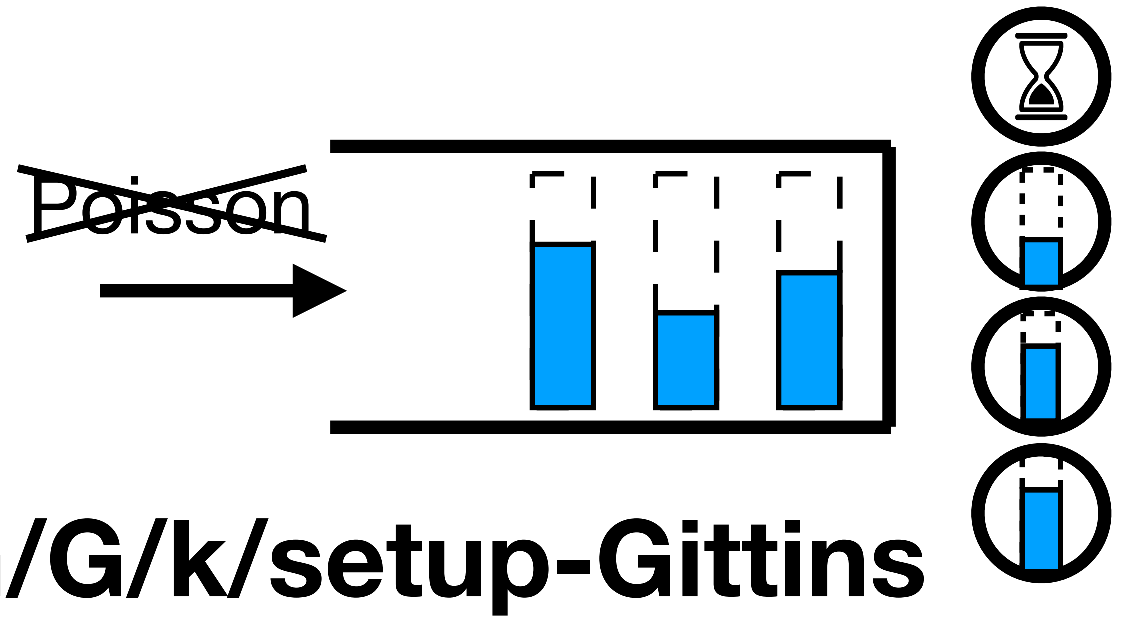
WINE ↕

work in G/G/k-setup

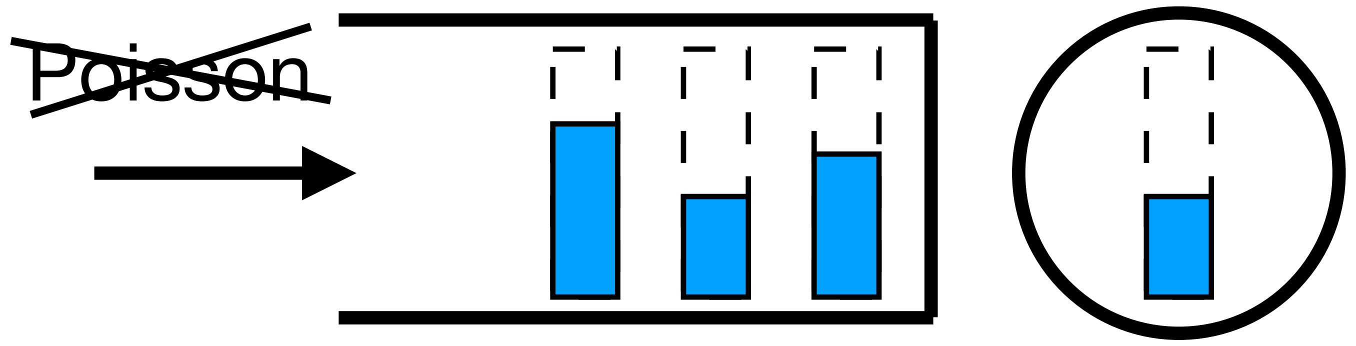
work in G/G/1

Analysis roadmap of G/G/k/setup

based on [SGH20]



G/G/k/setup-Gittins



G/G/1-OPT

number in G/G/k-setup ←gap→ number in G/G/1

SGH20, general

WINE ↕

work in G/G/k-setup

r-work

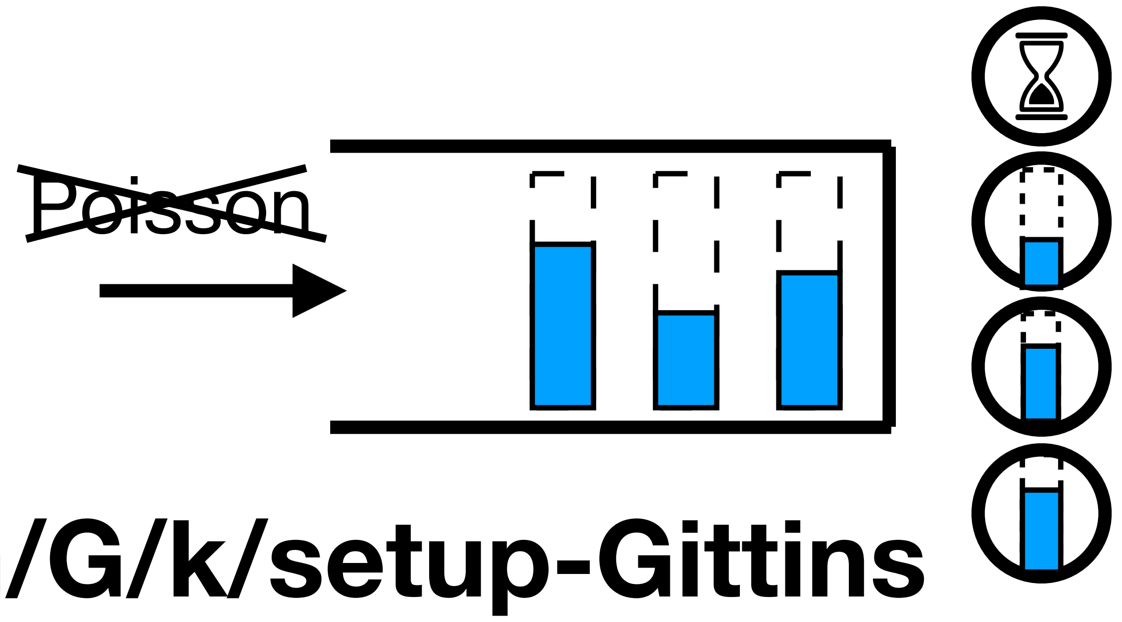
WINE ↕

work in G/G/1

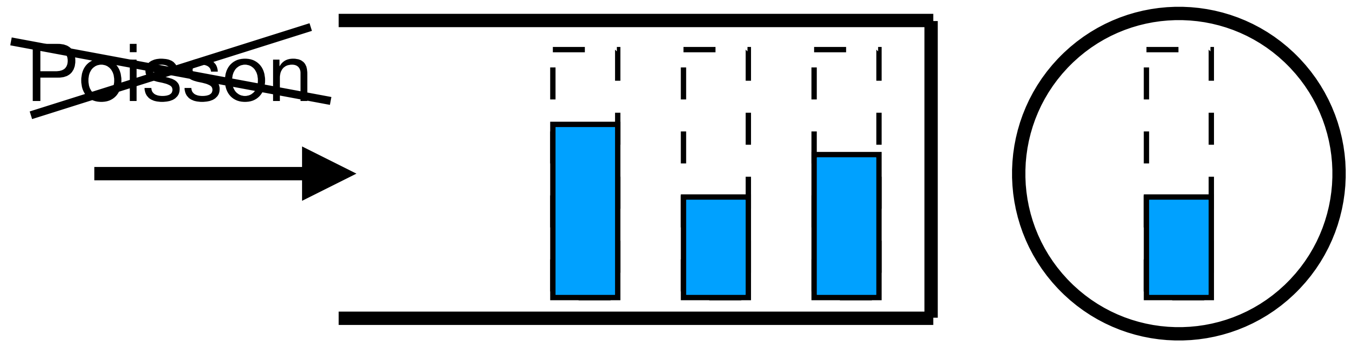
r-work

Analysis roadmap of G/G/k/setup

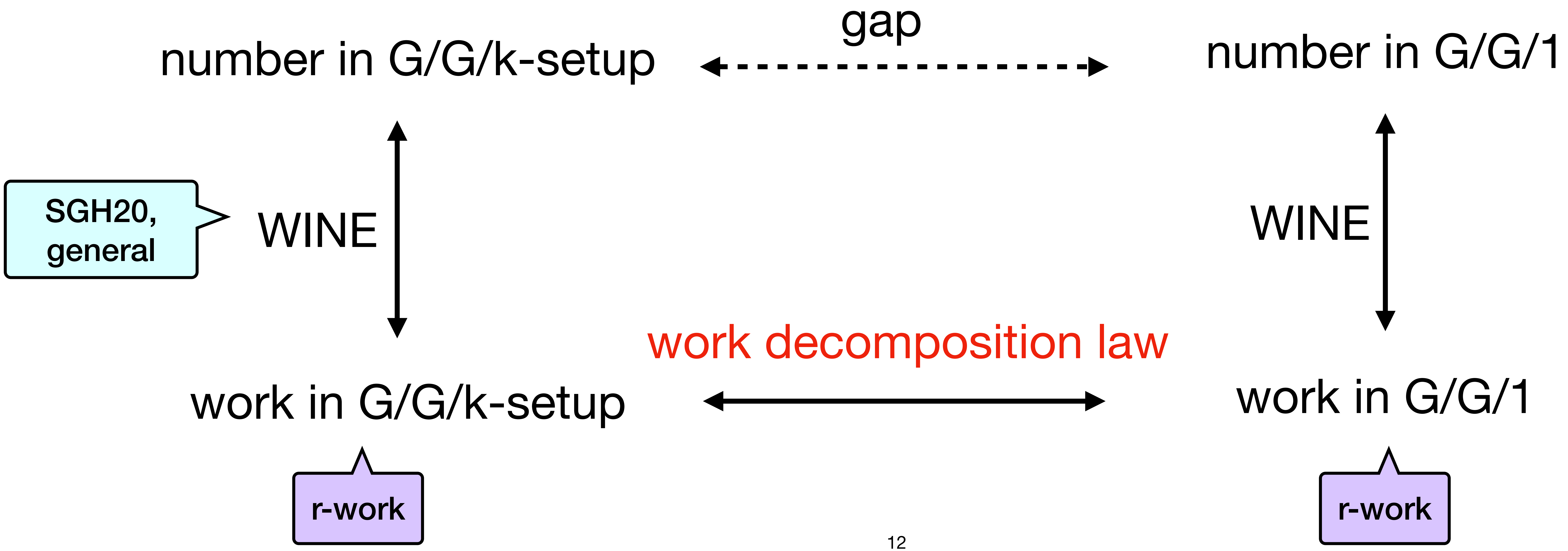
based on [SGH20]



G/G/k/setup-Gittins



G/G/1-OPT



number in G/G/k-setup

gap

number in G/G/1

SGH20, general

WINE

WINE

work decomposition law

work in G/G/k-setup

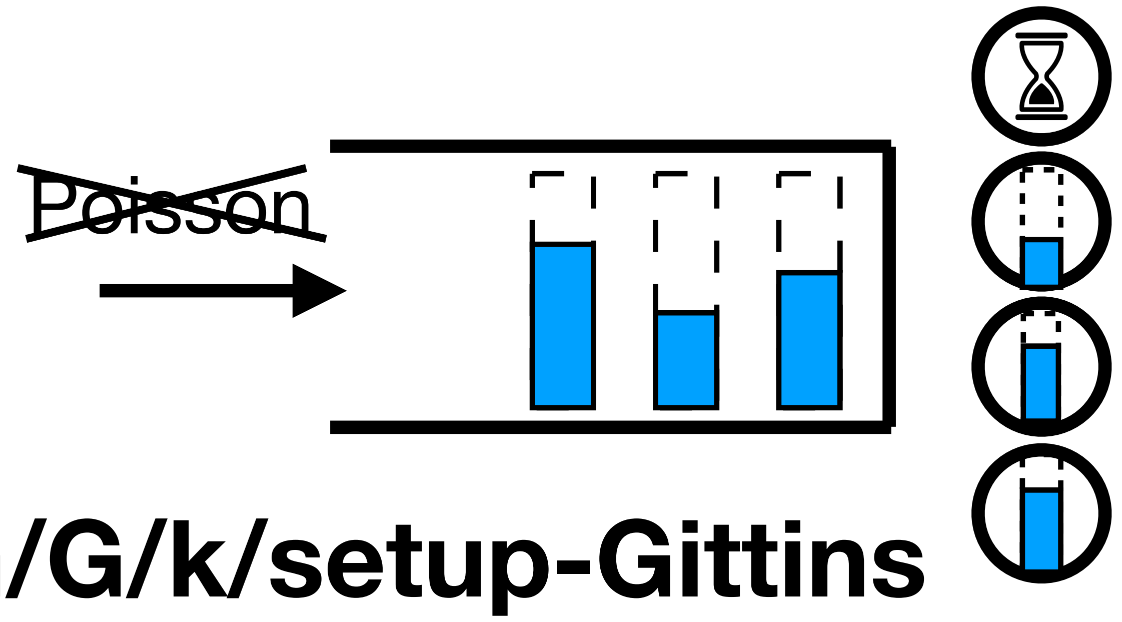
work in G/G/1

r-work

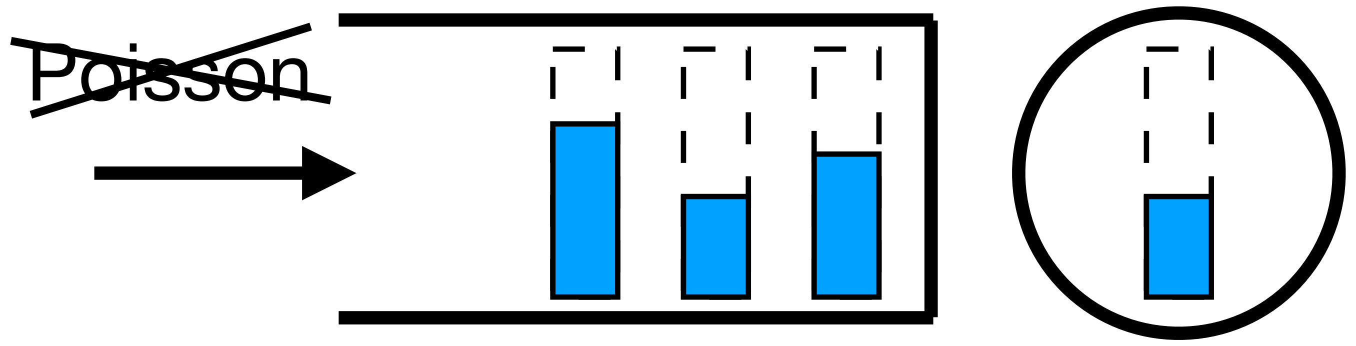
r-work

Analysis roadmap of G/G/k/setup

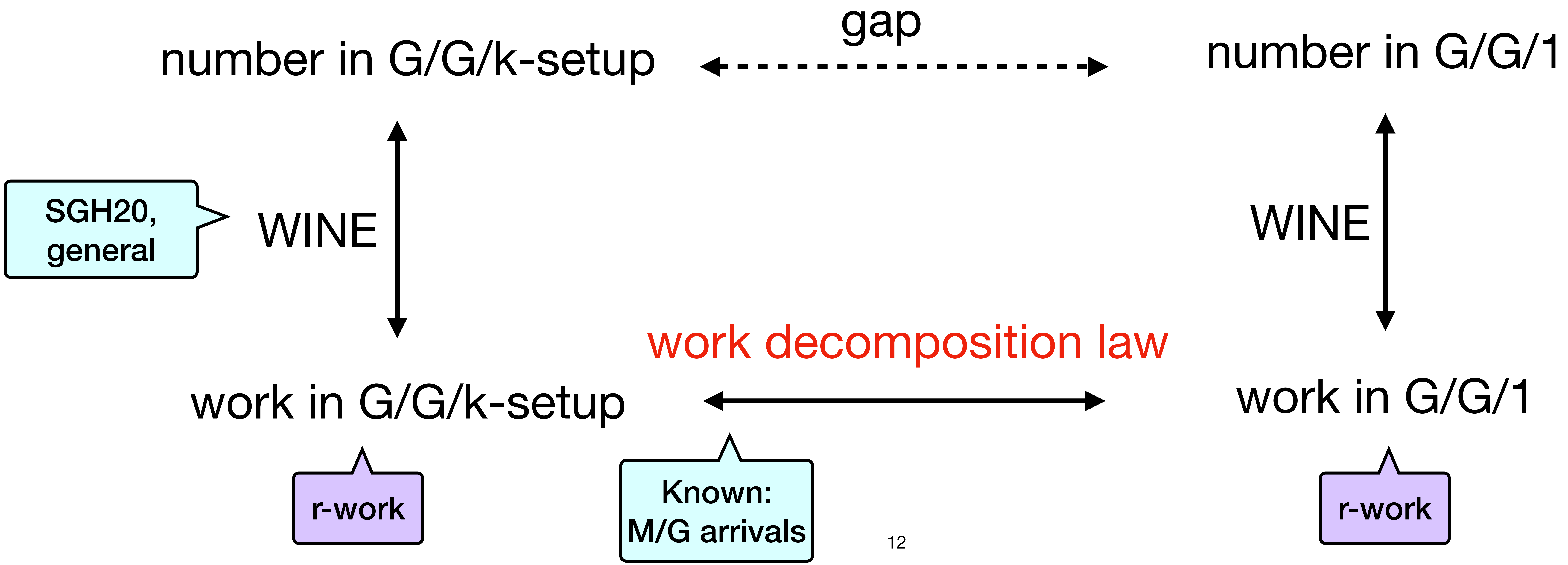
based on [SGH20]



G/G/k/setup-Gittins



G/G/1-OPT



number in G/G/k-setup

gap

number in G/G/1

SGH20, general

WINE

WINE

work decomposition law

work in G/G/k-setup

work in G/G/1

r-work

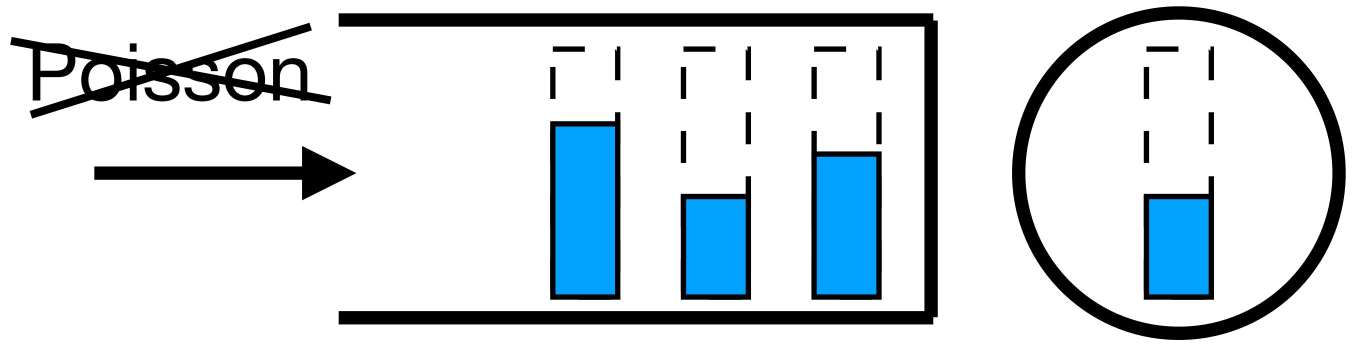
Known: M/G arrivals

r-work

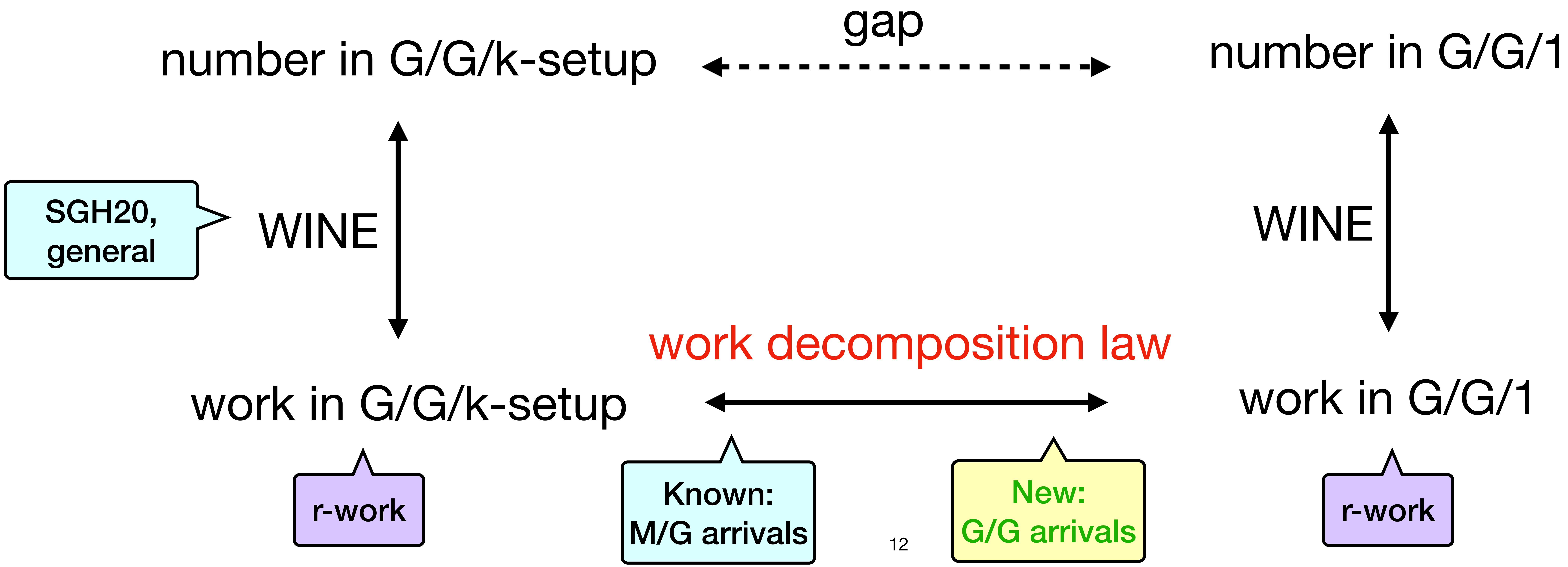
Analysis roadmap of G/G/k/setup

based on [SGH20]

G/G/k/setup-Gittins



G/G/1-OPT



gap

number in G/G/k-setup

number in G/G/1

SGH20, general

WINE

WINE

work decomposition law

work in G/G/k-setup

work in G/G/1

r-work

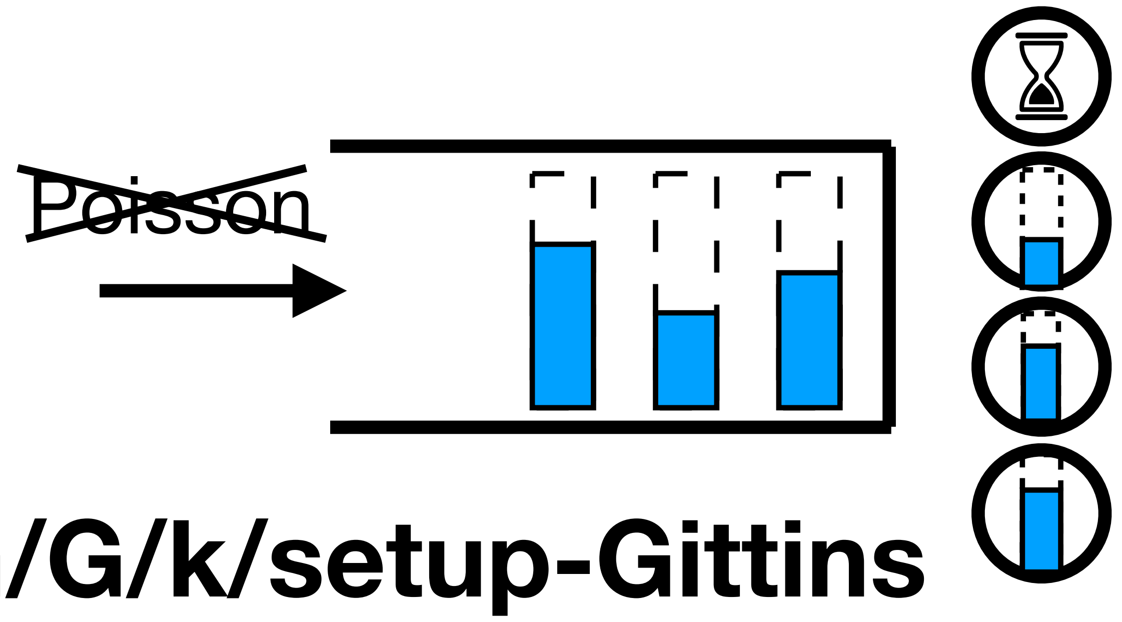
Known:
M/G arrivals

New:
G/G arrivals

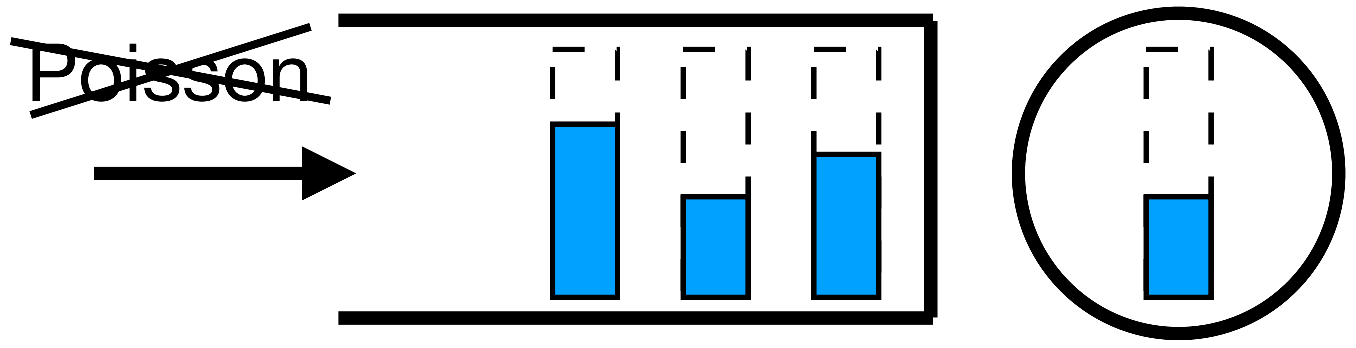
r-work

Analysis roadmap of G/G/k/setup

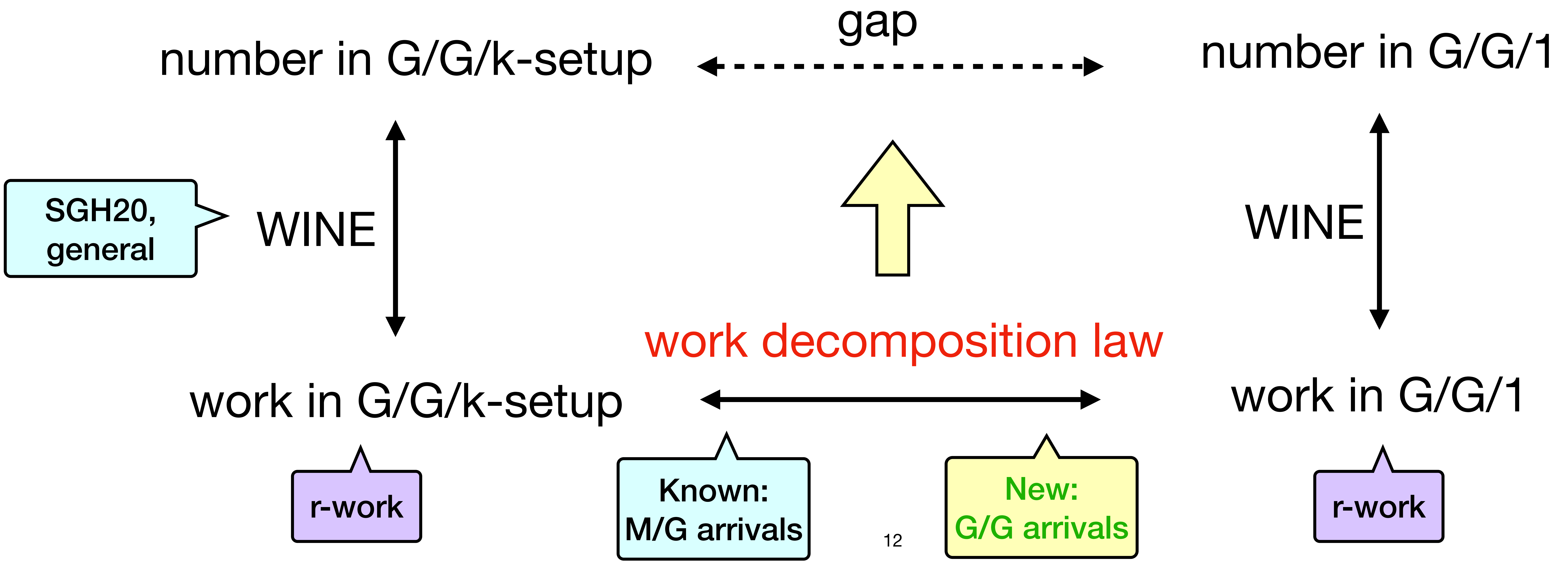
based on [SGH20]



G/G/k/setup-Gittins



G/G/1-OPT



number in G/G/k-setup

gap

number in G/G/1

SGH20, general

WINE

WINE

work decomposition law

work in G/G/k-setup

work in G/G/1

r-work

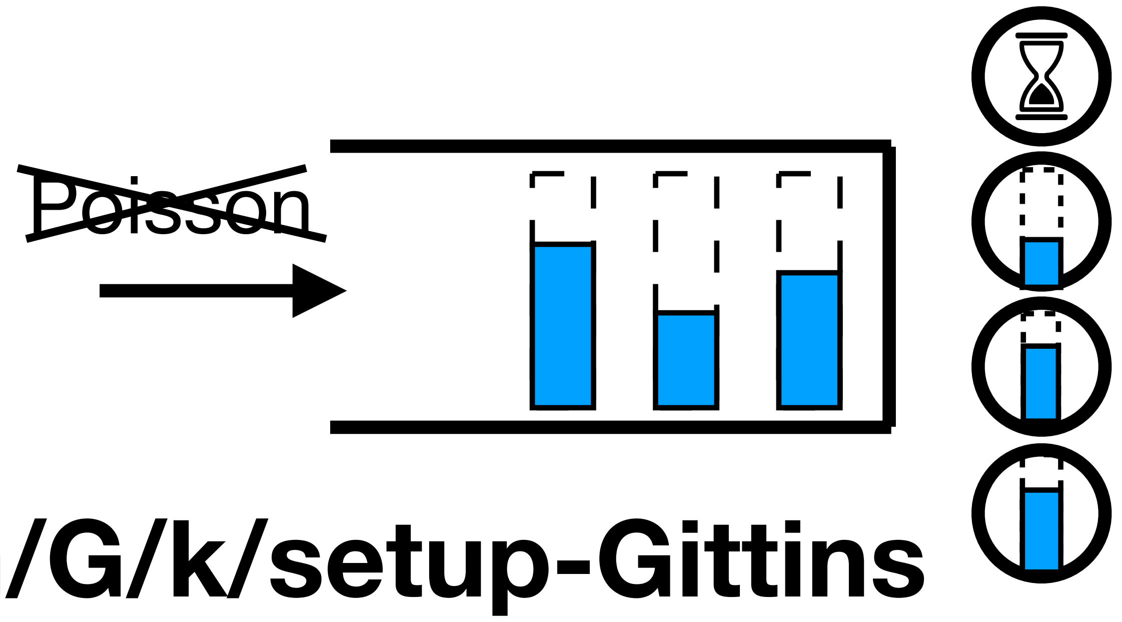
Known: M/G arrivals

New: G/G arrivals

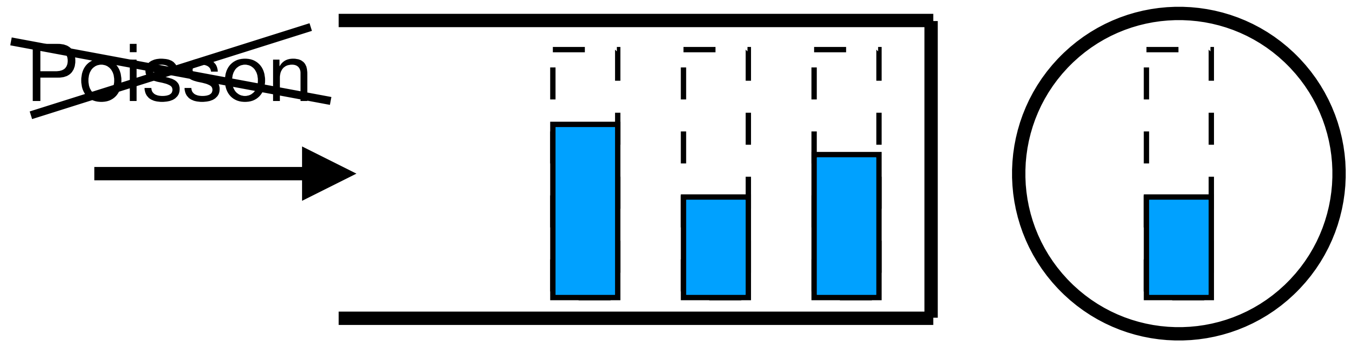
r-work

Analysis roadmap of G/G/k/setup

based on [SGH20]



G/G/k/setup-Gittins



G/G/1-OPT

number in G/G/k-setup ↔ gap ↔ number in G/G/1

$\leq \ell_{(a)} + \ell_{(b)} + \ell_{(c)}$

SGH20, general

WINE

WINE

work decomposition law

work in G/G/k-setup ↔ work in G/G/1

r-work

Known: M/G arrivals

New: G/G arrivals

r-work

G/G work-decomposition law

$$
\begin{aligned}
\mathbb{E}[W] = {} & \frac{\left(c_A^2 + \rho c_S^2\right)\mathbb{E}[S]}{2(1-\rho)} + \frac{\mathbb{E}[S]}{2} \\
& + \frac{\mathbb{E}[I_{off}\,W]}{1-\rho} \\
& + \frac{\mathbb{E}[I_{setup}\,W]}{1-\rho} \\
& - \frac{\rho}{1-\rho}\mathbb{E}[I\,R]
\end{aligned}
$$

G/G work-decomposition law

$$\mathbb{E}[W] = \frac{\left(c_A^2 + \rho c_S^2\right)\mathbb{E}[S]}{2(1-\rho)} + \frac{\mathbb{E}[S]}{2}$$

$$+\frac{\mathbb{E}[I_{off}\,W]}{1-\rho}$$

$$+\frac{\mathbb{E}[I_{setup}\,W]}{1-\rho}$$

$$-\frac{\rho}{1-\rho}\mathbb{E}[I\,R]$$

same for all G/G systems & policies

G/G work-decomposition law

$$\mathbb{E}[W] = \frac{\left(c_A^2 + \rho c_S^2\right)\mathbb{E}[S]}{2(1-\rho)} + \frac{\mathbb{E}[S]}{2}$$

same for all G/G systems & policies

$$+\frac{\mathbb{E}[I_{off}\,W]}{1-\rho}$$

$$+\frac{\mathbb{E}[I_{setup}\,W]}{1-\rho}$$

$$-\frac{\rho}{1-\rho}\mathbb{E}[I\,R]$$

depend on system & policy

G/G work-decomposition law

$$\mathbb{E}[W] = \frac{\left(c_A^2 + \rho c_S^2\right)\mathbb{E}[S]}{2(1-\rho)} + \frac{\mathbb{E}[S]}{2}$$

same for all G/G systems & policies

$$+\frac{\mathbb{E}[I_{off}W]}{1-\rho} \longrightarrow$$

$$+\frac{\mathbb{E}[I_{setup}W]}{1-\rho}$$

$$-\frac{\rho}{1-\rho}\mathbb{E}[IR]$$

depend on system & policy

G/G work-decomposition law

$$\mathbb{E}[W] = \frac{\left(c_A^2 + \rho c_S^2\right)\mathbb{E}[S]}{2(1-\rho)} + \frac{\mathbb{E}[S]}{2}$$

same for all G/G systems & policies

$$+\frac{\mathbb{E}[I_{off}\,W]}{1-\rho}$$

→ Captures the effect of multiple servers

$$+\frac{\mathbb{E}[I_{setup}\,W]}{1-\rho}$$

$$-\frac{\rho}{1-\rho}\mathbb{E}[I\,R]$$

depend on system & policy

G/G work-decomposition law

$$\mathbb{E}[W] = \frac{\left(c_A^2 + \rho c_S^2\right)\mathbb{E}[S]}{2(1-\rho)} + \frac{\mathbb{E}[S]}{2}$$

same for all G/G systems & policies

$$+\frac{\mathbb{E}[I_{off}\,W]}{1-\rho}$$

Captures the effect of multiple servers

- I_{off}: fraction of servers that are off

$$+\frac{\mathbb{E}[I_{setup}\,W]}{1-\rho}$$

$$-\frac{\rho}{1-\rho}\mathbb{E}[I\,R]$$

depend on system & policy

G/G work-decomposition law

$$\mathbb{E}[W] = \frac{\left(c_A^2 + \rho c_S^2\right)\mathbb{E}[S]}{2(1-\rho)} + \frac{\mathbb{E}[S]}{2}$$

same for all G/G systems & policies

$$+\frac{\mathbb{E}[I_{off}W]}{1-\rho}$$

$$+\frac{\mathbb{E}[I_{setup}W]}{1-\rho}$$

$$-\frac{\rho}{1-\rho}\mathbb{E}[IR]$$

depend on system & policy

Captures the effect of multiple servers

- I_{off}: fraction of servers that are off
- $k = 1$:

G/G work-decomposition law

$$\mathbb{E}[W] = \frac{\left(c_A^2 + \rho c_S^2\right)\mathbb{E}[S]}{2(1-\rho)} + \frac{\mathbb{E}[S]}{2}$$

same for all G/G systems & policies

$$+\frac{\mathbb{E}[I_{off}\,W]}{1-\rho}$$

Captures the effect of multiple servers

- I_{off}: fraction of servers that are off
- $k = 1$:
  - whenever $W > 0$, $I_{off} = 0$

depend on system & policy

$$+\frac{\mathbb{E}[I_{setup}\,W]}{1-\rho}$$

$$-\frac{\rho}{1-\rho}\mathbb{E}[I\,R]$$

G/G work-decomposition law

$$\mathbb{E}[W] = \frac{\left(c_A^2 + \rho c_S^2\right)\mathbb{E}[S]}{2(1-\rho)} + \frac{\mathbb{E}[S]}{2}$$

same for all G/G systems & policies

$$+\frac{\mathbb{E}[I_{off}W]}{1-\rho}$$

$$+\frac{\mathbb{E}[I_{setup}W]}{1-\rho}$$

$$-\frac{\rho}{1-\rho}\mathbb{E}[IR]$$

depend on system & policy

→ Captures the effect of multiple servers

- I_{off}: fraction of servers that are off
- $k = 1$:
  - whenever $W > 0,\ I_{off} = 0$
- k > 1:

G/G work-decomposition law

$$\mathbb{E}[W] = \frac{\left(c_A^2 + \rho c_S^2\right)\mathbb{E}[S]}{2(1-\rho)} + \frac{\mathbb{E}[S]}{2}$$

same for all G/G systems & policies

$$+\frac{\mathbb{E}[I_{off}\,W]}{1-\rho}$$

$$+\frac{\mathbb{E}[I_{setup}\,W]}{1-\rho}$$

$$-\frac{\rho}{1-\rho}\mathbb{E}[I\,R]$$

depend on system & policy

Captures the effect of multiple servers

- I_{off} : fraction of servers that are off
- $k = 1$:
  - whenever $W > 0$, $I_{off} = 0$
- k > 1:
  - $I_{off}W$ can be positive

G/G work-decomposition law

$$\mathbb{E}[W] = \frac{\left(c_A^2 + \rho c_S^2\right)\mathbb{E}[S]}{2(1-\rho)} + \frac{\mathbb{E}[S]}{2}$$

same for all G/G systems & policies

$$+\frac{\mathbb{E}[I_{off}W]}{1-\rho}$$

$$+\frac{\mathbb{E}[I_{setup}W]}{1-\rho}$$

$$-\frac{\rho}{1-\rho}\mathbb{E}[IR]$$

depend on system & policy

Captures the effect of multiple servers

- I_{off}: fraction of servers that are off
- $k = 1$:
  - whenever $W > 0$, $I_{off} = 0$
- k > 1:
  - $I_{off}W$ can be positive
  - whenever $\geq k$ jobs, $I_{off} = 0$, so $\mathbb{E}[I_{off}W]$ small

G/G work-decomposition law

$$\mathbb{E}[W] = \frac{\left(c_A^2 + \rho c_S^2\right)\mathbb{E}[S]}{2(1-\rho)} + \frac{\mathbb{E}[S]}{2}$$

same for all G/G systems & policies

$$+\frac{\mathbb{E}[I_{off}\,W]}{1-\rho}$$

depend on system & policy

$$+\frac{\mathbb{E}[I_{setup}\,W]}{1-\rho}$$

$$-\frac{\rho}{1-\rho}\mathbb{E}[I\,R]$$

G/G work-decomposition law

$$\mathbb{E}[W] = \frac{\left(c_A^2 + \rho c_S^2\right)\mathbb{E}[S]}{2(1-\rho)} + \frac{\mathbb{E}[S]}{2}$$

$$+\frac{\mathbb{E}[I_{off}\,W]}{1-\rho}$$

$$+\frac{\mathbb{E}[I_{setup}\,W]}{1-\rho}$$

$$-\frac{\rho}{1-\rho}\mathbb{E}[I\,R]$$

same for all G/G systems & policies

depend on system & policy

G/G work-decomposition law

$$\mathbb{E}[W] = \frac{\left(c_A^2 + \rho c_S^2\right)\mathbb{E}[S]}{2(1-\rho)} + \frac{\mathbb{E}[S]}{2}$$

same for all G/G systems & policies

$$+\frac{\mathbb{E}[I_{off} W]}{1-\rho}$$

$$+\frac{\mathbb{E}[I_{setup} W]}{1-\rho}$$

Captures the effect of setup times

$$-\frac{\rho}{1-\rho}\mathbb{E}[I R]$$

depend on system & policy

G/G work-decomposition law

$$\mathbb{E}[W] = \frac{\left(c_A^2 + \rho c_S^2\right)\mathbb{E}[S]}{2(1-\rho)} + \frac{\mathbb{E}[S]}{2}$$

same for all G/G systems & policies

$$+\frac{\mathbb{E}[I_{off}\,W]}{1-\rho}$$

$$+\frac{\mathbb{E}[I_{setup}\,W]}{1-\rho}$$

$$-\frac{\rho}{1-\rho}\mathbb{E}[I\,R]$$

depend on system & policy

Captures the effect of setup times

- I_{setup} : fraction of servers setting up

G/G work-decomposition law

$$\mathbb{E}[W] = \frac{\left(c_A^2 + \rho c_S^2\right)\mathbb{E}[S]}{2(1-\rho)} + \frac{\mathbb{E}[S]}{2}$$

same for all G/G systems & policies

$$+\frac{\mathbb{E}[I_{off}W]}{1-\rho}$$

$$+\frac{\mathbb{E}[I_{setup}W]}{1-\rho}$$

$$-\frac{\rho}{1-\rho}\mathbb{E}[IR]$$

depend on system & policy

Captures the effect of setup times

- I_{setup} : fraction of servers setting up
- Without setup times,

G/G work-decomposition law

$$\mathbb{E}[W] = \frac{\left(c_A^2 + \rho c_S^2\right)\mathbb{E}[S]}{2(1-\rho)} + \frac{\mathbb{E}[S]}{2}$$

$$+\frac{\mathbb{E}[I_{off}W]}{1-\rho}$$

$$+\frac{\mathbb{E}[I_{setup}W]}{1-\rho}$$

$$-\frac{\rho}{1-\rho}\mathbb{E}[IR]$$

same for all G/G systems & policies

depend on system & policy

Captures the effect of setup times

- I_{setup} : fraction of servers setting up
- Without setup times,
  - $I_{setup} = 0$

G/G work-decomposition law

$$\mathbb{E}[W] = \frac{\left(c_A^2 + \rho c_S^2\right)\mathbb{E}[S]}{2(1-\rho)} + \frac{\mathbb{E}[S]}{2}$$

same for all G/G systems & policies

$$+\frac{\mathbb{E}[I_{off}W]}{1-\rho}$$

$$+\frac{\mathbb{E}[I_{setup}W]}{1-\rho}$$

$$-\frac{\rho}{1-\rho}\mathbb{E}[IR]$$

depend on system & policy

Captures the effect of setup times

- I_{setup} : fraction of servers setting up
- Without setup times,
  - $I_{setup} = 0$
- With setup times,

G/G work-decomposition law

$$\mathbb{E}[W] = \frac{\left(c_A^2 + \rho c_S^2\right)\mathbb{E}[S]}{2(1-\rho)} + \frac{\mathbb{E}[S]}{2}$$

same for all G/G systems & policies

$$+\frac{\mathbb{E}[I_{off}W]}{1-\rho}$$

$$+\frac{\mathbb{E}[I_{setup}W]}{1-\rho}$$

$$-\frac{\rho}{1-\rho}\mathbb{E}[IR]$$

depend on system & policy

Captures the effect of setup times

- I_{setup} : fraction of servers setting up
- Without setup times,
  - $I_{setup} = 0$
- With setup times,
  - Long busy cycle, servers only set up at the beginning, $\mathbb{E}[I_{setup}W]$ small

G/G work-decomposition law

$$\mathbb{E}[W] = \frac{\left(c_A^2 + \rho c_S^2\right)\mathbb{E}[S]}{2(1-\rho)} + \frac{\mathbb{E}[S]}{2}$$

same for all G/G systems & policies

$$+\frac{\mathbb{E}[I_{off}\,W]}{1-\rho}$$

$$+\frac{\mathbb{E}[I_{setup}\,W]}{1-\rho}$$

$$-\frac{\rho}{1-\rho}\mathbb{E}[I\,R]$$

depend on system & policy

G/G work-decomposition law

$$\mathbb{E}[W] = \frac{\left(c_A^2 + \rho c_S^2\right)\mathbb{E}[S]}{2(1-\rho)} + \frac{\mathbb{E}[S]}{2}$$

same for all G/G systems & policies

$$+\frac{\mathbb{E}[I_{off} W]}{1-\rho}$$

$$+\frac{\mathbb{E}[I_{setup} W]}{1-\rho}$$

$$-\frac{\rho}{1-\rho}\mathbb{E}[I R]$$

depend on system & policy

G/G work-decomposition law

$$\mathbb{E}[W] = \frac{\left(c_A^2 + \rho c_S^2\right)\mathbb{E}[S]}{2(1-\rho)} + \frac{\mathbb{E}[S]}{2}$$

$$+\frac{\mathbb{E}[I_{off}W]}{1-\rho}$$

$$+\frac{\mathbb{E}[I_{setup}W]}{1-\rho}$$

$$-\frac{\rho}{1-\rho}\mathbb{E}[IR]$$

same for all G/G systems & policies

depend on system & policy

Captures the effect of G/G arrivals

G/G work-decomposition law

$$\mathbb{E}[W] = \frac{\left(c_A^2 + \rho c_S^2\right)\mathbb{E}[S]}{2(1-\rho)} + \frac{\mathbb{E}[S]}{2}$$

same for all G/G systems & policies

$$+\frac{\mathbb{E}[I_{off}\,W]}{1-\rho}$$

$$+\frac{\mathbb{E}[I_{setup}\,W]}{1-\rho}$$

$$-\frac{\rho}{1-\rho}\mathbb{E}[I\,R]$$

depend on system & policy

Captures the effect of G/G arrivals

- $I = I_{off} + I_{setup}$

G/G work-decomposition law

$$\mathbb{E}[W] = \frac{\left(c_A^2 + \rho c_S^2\right)\mathbb{E}[S]}{2(1-\rho)} + \frac{\mathbb{E}[S]}{2}$$

$$+\frac{\mathbb{E}[I_{off} W]}{1-\rho}$$

$$+\frac{\mathbb{E}[I_{setup} W]}{1-\rho}$$

$$-\frac{\rho}{1-\rho}\mathbb{E}[I R]$$

same for all G/G systems & policies

depend on system & policy

Captures the effect of G/G arrivals

- $I = I_{off} + I_{setup}$
- R residue arrival time

G/G work-decomposition law

$$\mathbb{E}[W] = \frac{\left(c_A^2 + \rho c_S^2\right)\mathbb{E}[S]}{2(1-\rho)} + \frac{\mathbb{E}[S]}{2}$$

same for all G/G systems & policies

$$+\frac{\mathbb{E}[I_{off}\,W]}{1-\rho}$$

$$+\frac{\mathbb{E}[I_{setup}\,W]}{1-\rho}$$

depend on system & policy

$$-\frac{\rho}{1-\rho}\mathbb{E}[I\,R]$$

Captures the effect of G/G arrivals

- $I = I_{off} + I_{setup}$
- R residue arrival time
- When Poisson arrivals,

G/G work-decomposition law

$$
\begin{aligned}
\mathbb{E}[W] = {} & \frac{\left(c_A^2 + \rho c_S^2\right)\mathbb{E}[S]}{2(1-\rho)} + \frac{\mathbb{E}[S]}{2} \\
& + \frac{\mathbb{E}[I_{off}\, W]}{1-\rho} \\
& + \frac{\mathbb{E}[I_{setup}\, W]}{1-\rho} \\
& - \frac{\rho}{1-\rho}\mathbb{E}[I\, R]
\end{aligned}
$$

same for all G/G systems & policies

depend on system & policy

Captures the effect of G/G arrivals

- $I = I_{off} + I_{setup}$
- R residue arrival time
- When Poisson arrivals,
  - R exponential and indep. of I

G/G work-decomposition law

$$\mathbb{E}[W] = \frac{\left(c_A^2 + \rho c_S^2\right)\mathbb{E}[S]}{2(1-\rho)} + \frac{\mathbb{E}[S]}{2}$$

same for all G/G systems & policies

$$+\frac{\mathbb{E}[I_{off}\,W]}{1-\rho}$$

$$+\frac{\mathbb{E}[I_{setup}\,W]}{1-\rho}$$

depend on system & policy

$$-\frac{\rho}{1-\rho}\mathbb{E}[I\,R]$$

Captures the effect of G/G arrivals

- $I = I_{off} + I_{setup}$
- R residue arrival time
- When Poisson arrivals,
  - R exponential and indep. of I
  - $\mathbb{E}[I\,R] = \mathbb{E}[I]\mathbb{E}[R]$, fixed

G/G work-decomposition law

$$\mathbb{E}[W] = \frac{\left(c_A^2 + \rho c_S^2\right)\mathbb{E}[S]}{2(1-\rho)} + \frac{\mathbb{E}[S]}{2}$$

same for all G/G systems & policies

$$+\frac{\mathbb{E}[I_{off} W]}{1-\rho}$$

$$+\frac{\mathbb{E}[I_{setup} W]}{1-\rho}$$

depend on system & policy

$$-\frac{\rho}{1-\rho}\mathbb{E}[I R]$$

Captures the effect of G/G arrivals

- $I = I_{off} + I_{setup}$
- R residue arrival time
- When Poisson arrivals,
  - R exponential and indep. of I
  - $\mathbb{E}[I R] = \mathbb{E}[I]\mathbb{E}[R]$, fixed
- In general,

G/G work-decomposition law

$$\mathbb{E}[W] = \frac{\left(c_A^2 + \rho c_S^2\right)\mathbb{E}[S]}{2(1-\rho)} + \frac{\mathbb{E}[S]}{2}$$

same for all G/G systems & policies

$$+\frac{\mathbb{E}[I_{off} W]}{1-\rho}$$

$$+\frac{\mathbb{E}[I_{setup} W]}{1-\rho}$$

depend on system & policy

$$-\frac{\rho}{1-\rho}\mathbb{E}[I R]$$

Captures the effect of G/G arrivals

- $I = I_{off} + I_{setup}$
- R residue arrival time
- When Poisson arrivals,
  - R exponential and indep. of I
  - $\mathbb{E}[I R] = \mathbb{E}[I]\mathbb{E}[R]$, fixed
- In general,
  - assume $\exists R_{min}, R_{\max}$ $R_{\min} \leq \mathbb{E}[R \mid A_{age}] \leq R_{\max}$

G/G work-decomposition law + WINE in G/G/k/setup

$$\mathbb{E}[W] = \frac{\left(c_A^2 + \rho c_S^2\right)\mathbb{E}[S]}{2(1-\rho)} + \frac{\mathbb{E}[S]}{2}$$

same for all G/G systems

$$+\frac{\mathbb{E}[I_{off}\,W]}{1-\rho}$$

$$+\frac{\mathbb{E}[I_{setup}\,W]}{1-\rho}$$

$$-\frac{\rho}{1-\rho}\mathbb{E}[I\,R]$$

depend on system & policy

G/G work-decomposition law + WINE in G/G/k/setup

$$\mathbb{E}[W] = \frac{\left(c_A^2 + \rho c_S^2\right)\mathbb{E}[S]}{2(1-\rho)} + \frac{\mathbb{E}[S]}{2}$$

$$+\frac{\mathbb{E}[I_{off}\, W]}{1-\rho}$$

$$+\frac{\mathbb{E}[I_{setup}\, W]}{1-\rho}$$

$$-\frac{\rho}{1-\rho}\mathbb{E}[I\, R]$$

same for all G/G systems

depend on system & policy

WINE

$\leq \ell_{(a)}$ under Gittins

G/G work-decomposition law + WINE in G/G/k/setup

$$\mathbb{E}[W] = \frac{\left(c_A^2 + \rho c_S^2\right)\mathbb{E}[S]}{2(1-\rho)} + \frac{\mathbb{E}[S]}{2}$$

same for all G/G systems

$$+\frac{\mathbb{E}[I_{off}W]}{1-\rho} \xrightarrow{\text{WINE}} \leq \ell_{(a)} \text{ under Gittins}$$

$$+\frac{\mathbb{E}[I_{setup}W]}{1-\rho} \xrightarrow{\text{WINE}} \leq \ell_{(c)} \text{ under Gittins}$$

$$-\frac{\rho}{1-\rho}\mathbb{E}[IR]$$

depend on system & policy

G/G work-decomposition law + WINE in G/G/k/setup

$$\mathbb{E}[W] = \frac{\left(c_A^2 + \rho c_S^2\right)\mathbb{E}[S]}{2(1-\rho)} + \frac{\mathbb{E}[S]}{2}$$

same for all G/G systems

$$+\frac{\mathbb{E}[I_{off}\,W]}{1-\rho} \quad \xrightarrow{\text{WINE}} \quad \le \ell_{(a)} \text{ under Gittins}$$

depend on system & policy

$$+\frac{\mathbb{E}[I_{setup}\,W]}{1-\rho} \quad \xrightarrow{\text{WINE}} \quad \le \ell_{(c)} \text{ under Gittins}$$

$$-\frac{\rho}{1-\rho}\mathbb{E}[I\,R] \quad \xrightarrow{\text{WINE}} \quad \in [\lambda R_{\min}, \lambda R_{\max}] \text{ any policy}$$

G/G work-decomposition law + WINE in G/G/k/setup

$$\mathbb{E}[W] = \frac{\left(c_A^2 + \rho c_S^2\right)\mathbb{E}[S]}{2(1-\rho)} + \frac{\mathbb{E}[S]}{2}$$

same for all G/G systems

$$+\frac{\mathbb{E}[I_{off}\,W]}{1-\rho} \quad \xrightarrow{\text{WINE}} \quad \leq \ell_{(a)} \text{ under Gittins}$$

depend on system & policy

$$+\frac{\mathbb{E}[I_{setup}\,W]}{1-\rho} \quad \xrightarrow{\text{WINE}} \quad \leq \ell_{(c)} \text{ under Gittins}$$

$$-\frac{\rho}{1-\rho}\mathbb{E}[I\,R] \quad \xrightarrow{\text{WINE}} \quad \in [\lambda R_{\min}, \lambda R_{\max}] \text{ any policy}$$

$$\ell_{(b)} = \lambda(R_{\max} - R_{\min})$$

Outline for the rest of the talk

- What is our problem and result? ✓
- How does the G/G/k/setup anaysis work? ✓
- What is the main obstacle and how do we solve it?

Outline for the rest of the talk

- What is our problem and result? ✓
- How does the G/G/k/setup anaysis work? ✓
- What is the main obstacle and how do we solve it?

G/G work decomposition

Work analysis

Work analysis

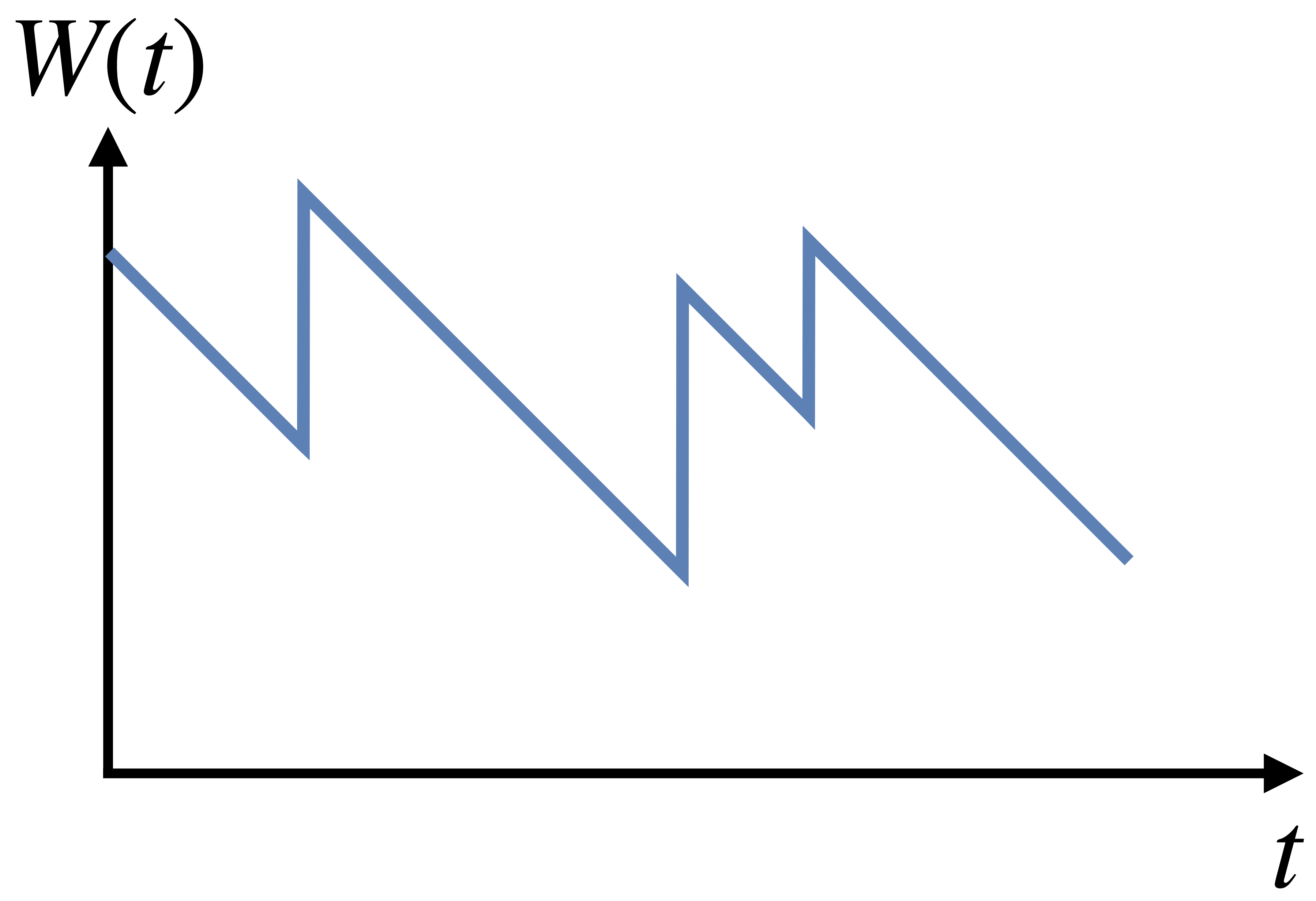
- Why is G/G work analysis harder than M/G?

Work analysis

- Why is G/G work analysis harder than M/G?
- M/G/1: $W(t)$ Markov process ✓

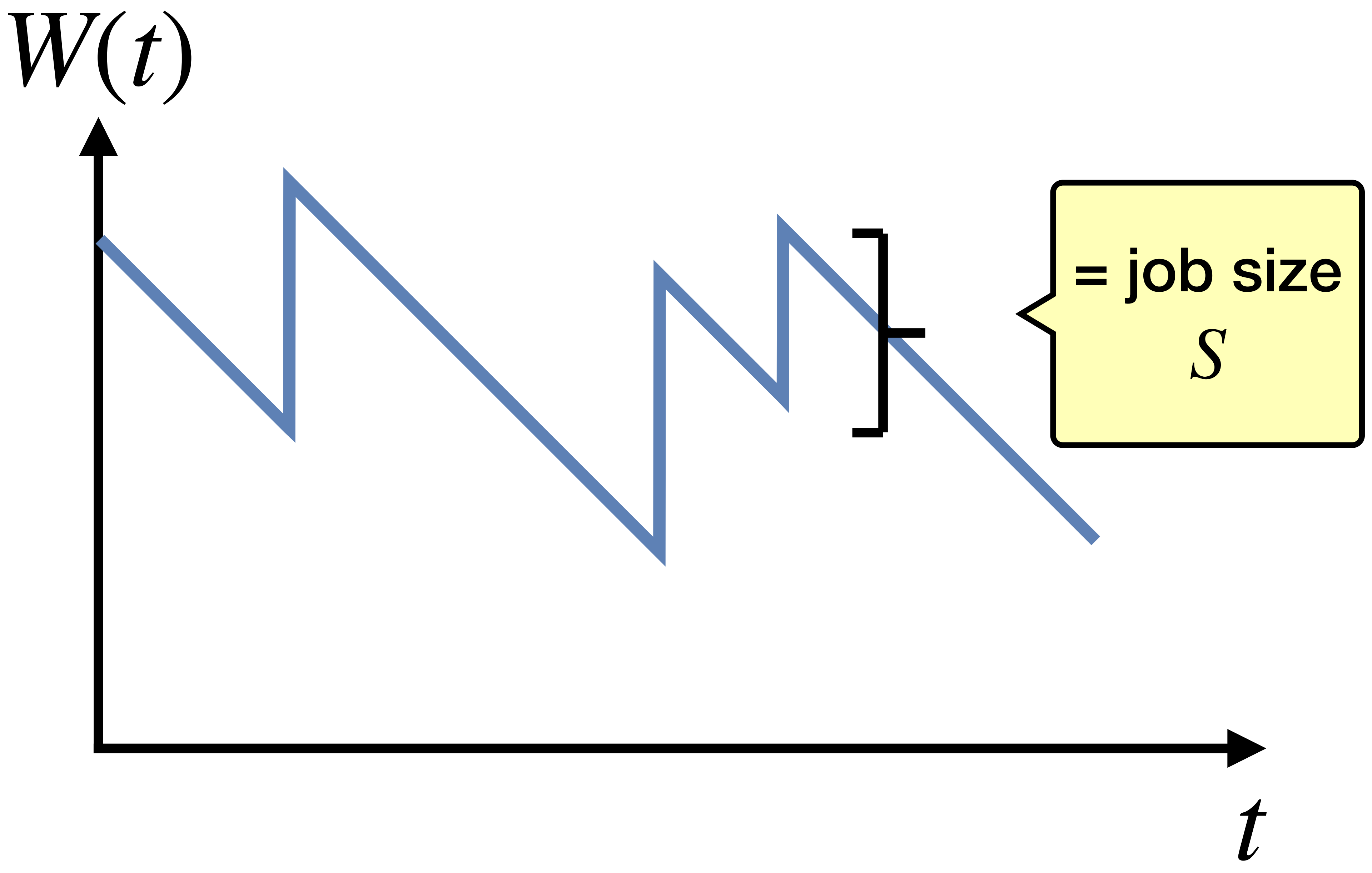
Work analysis

- Why is G/G work analysis harder than M/G?
- M/G/1: $W(t)$ Markov process ✓



Work analysis

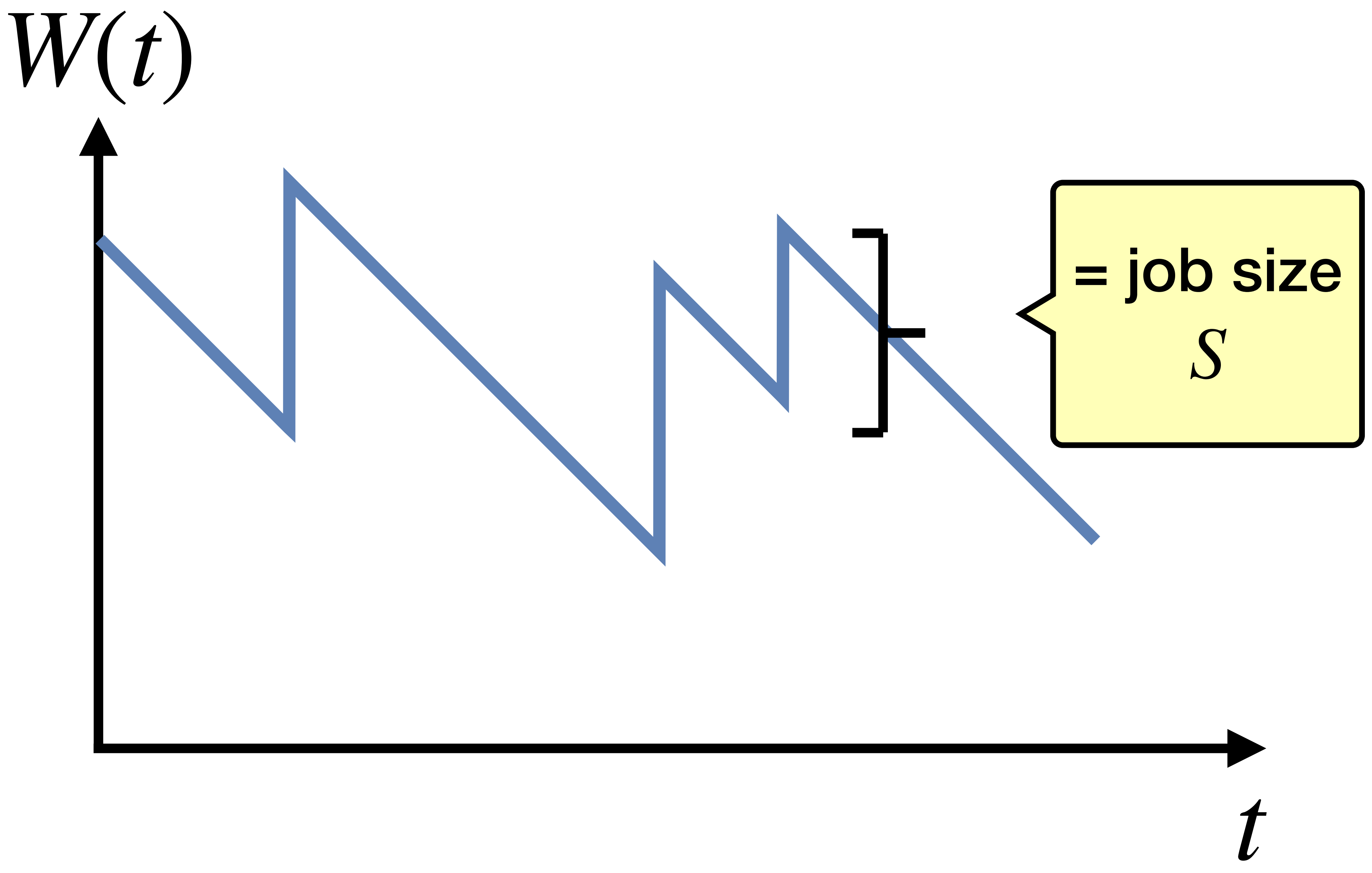
- Why is G/G work analysis harder than M/G?
- M/G/1: $W(t)$ Markov process ✓



= job size S

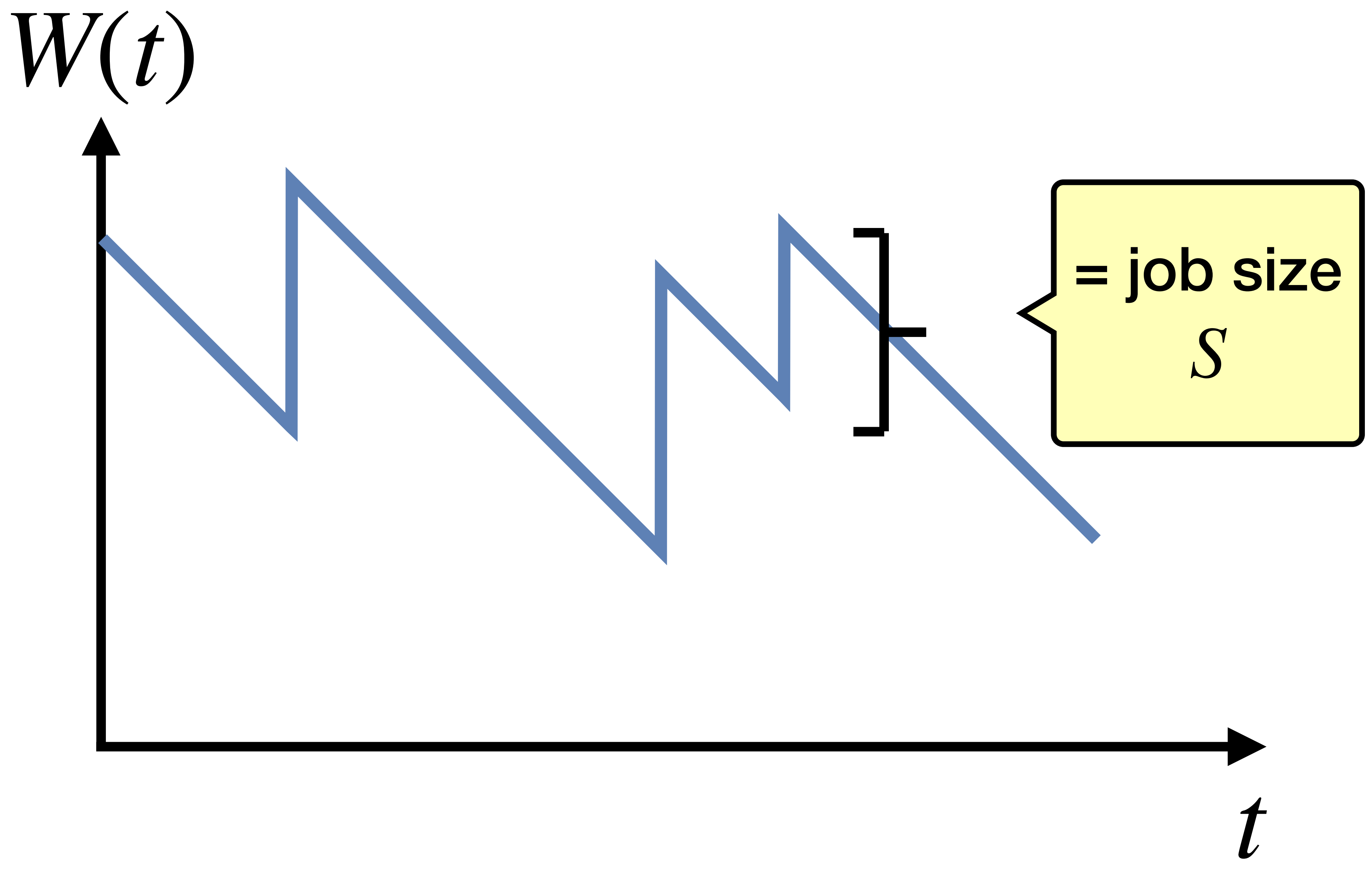
Work analysis

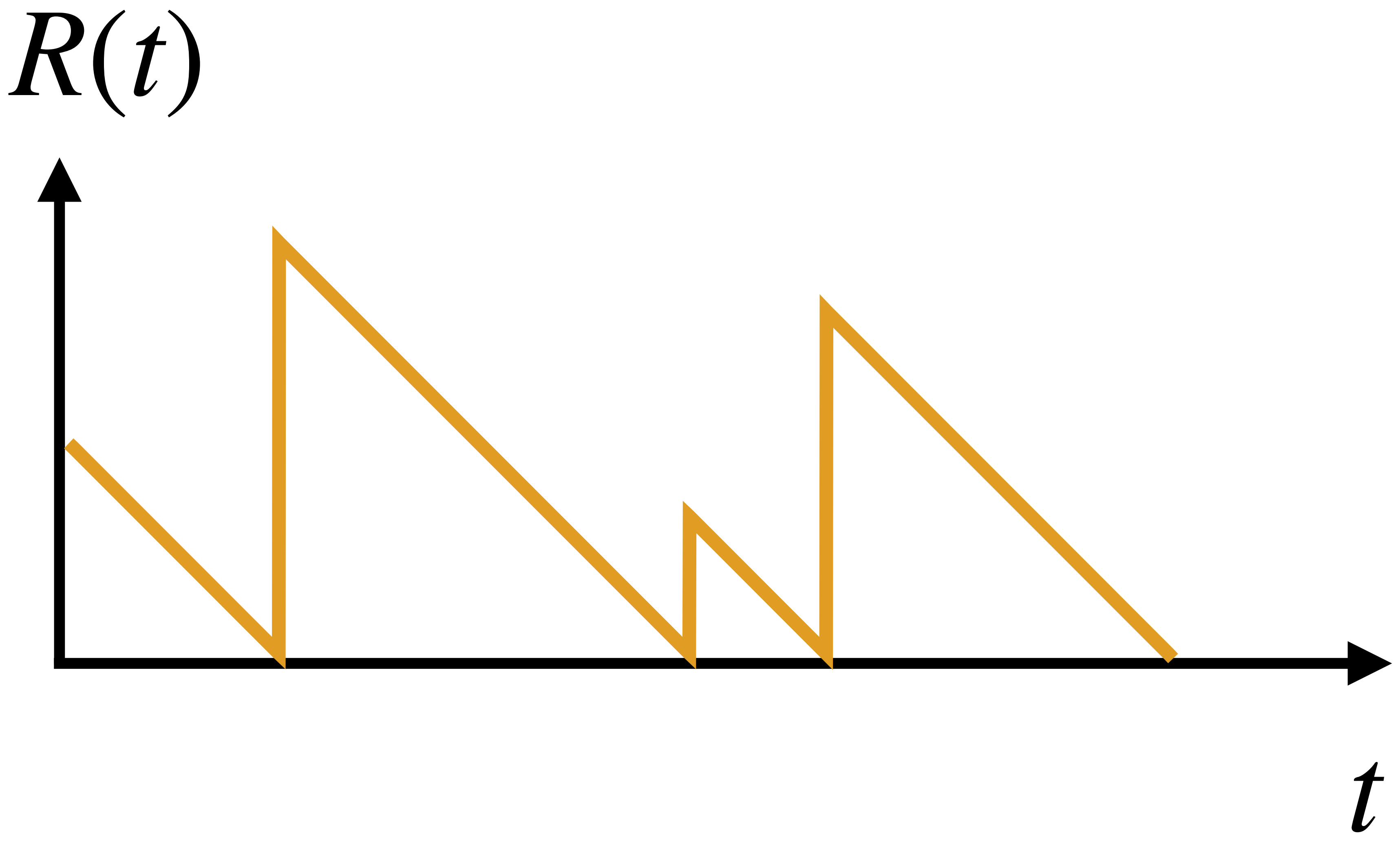
- Why is G/G work analysis harder than M/G?
- M/G/1: $W(t)$ Markov process ✓
- G/G/1: $W(t)$ not a Markov process ✗



Work analysis

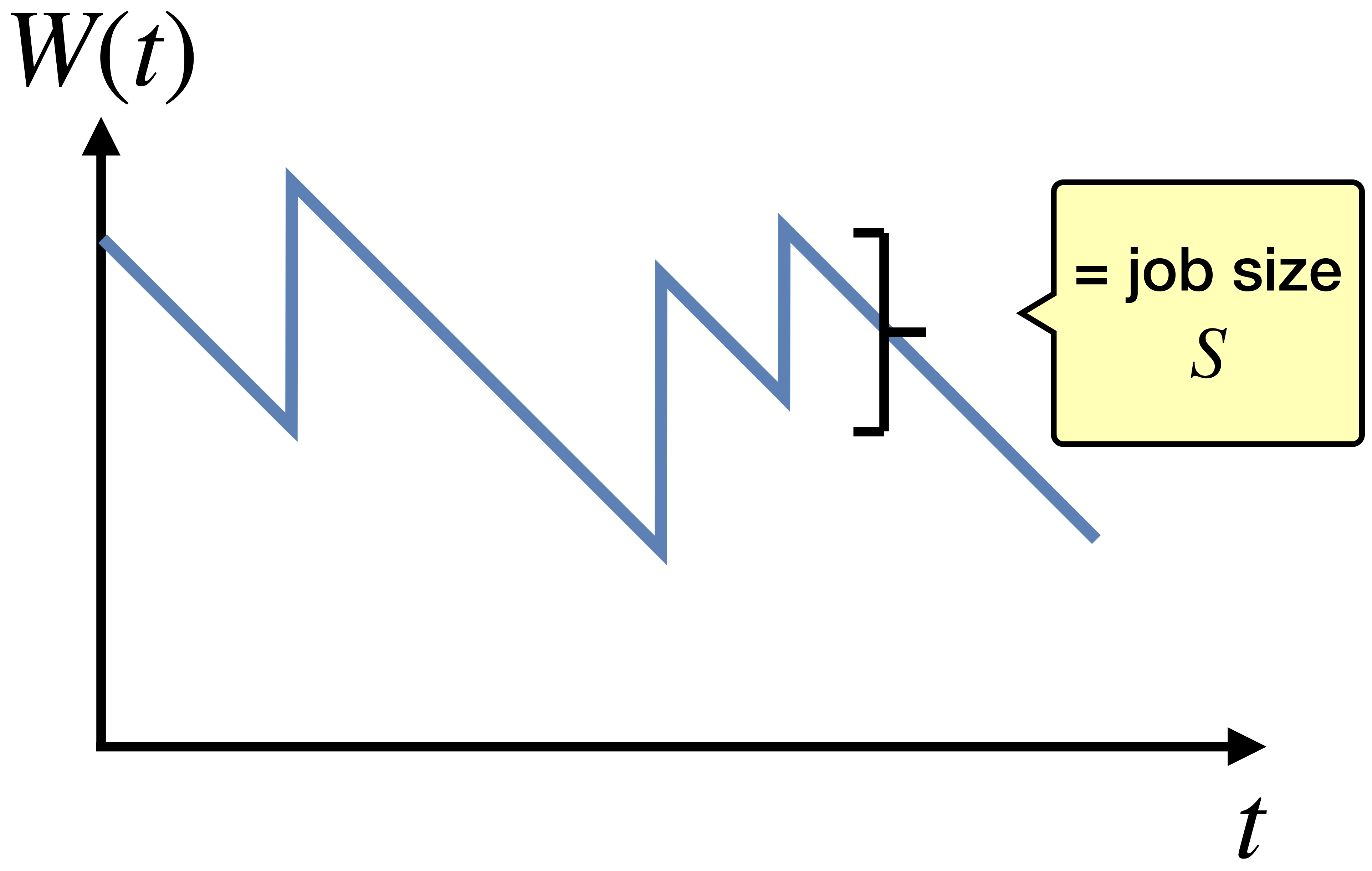
- Why is G/G work analysis harder than M/G?
- M/G/1: $W(t)$ Markov process ✓
- G/G/1: $W(t)$ not a Markov process ✗





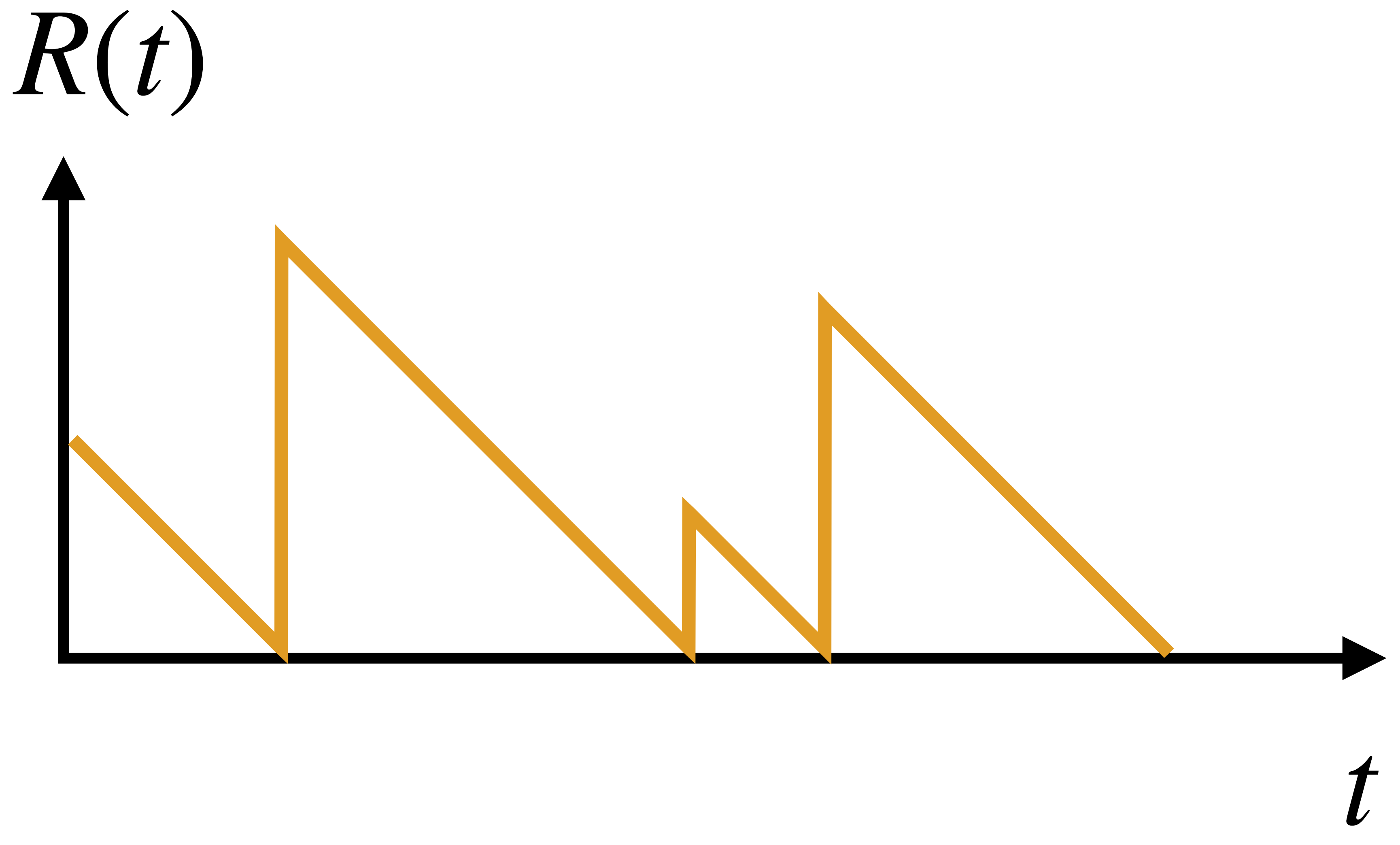
Work analysis

- Why is G/G work analysis harder than M/G?
- M/G/1: $W(t)$ Markov process ✓
- G/G/1: $W(t)$ not a Markov process ✗



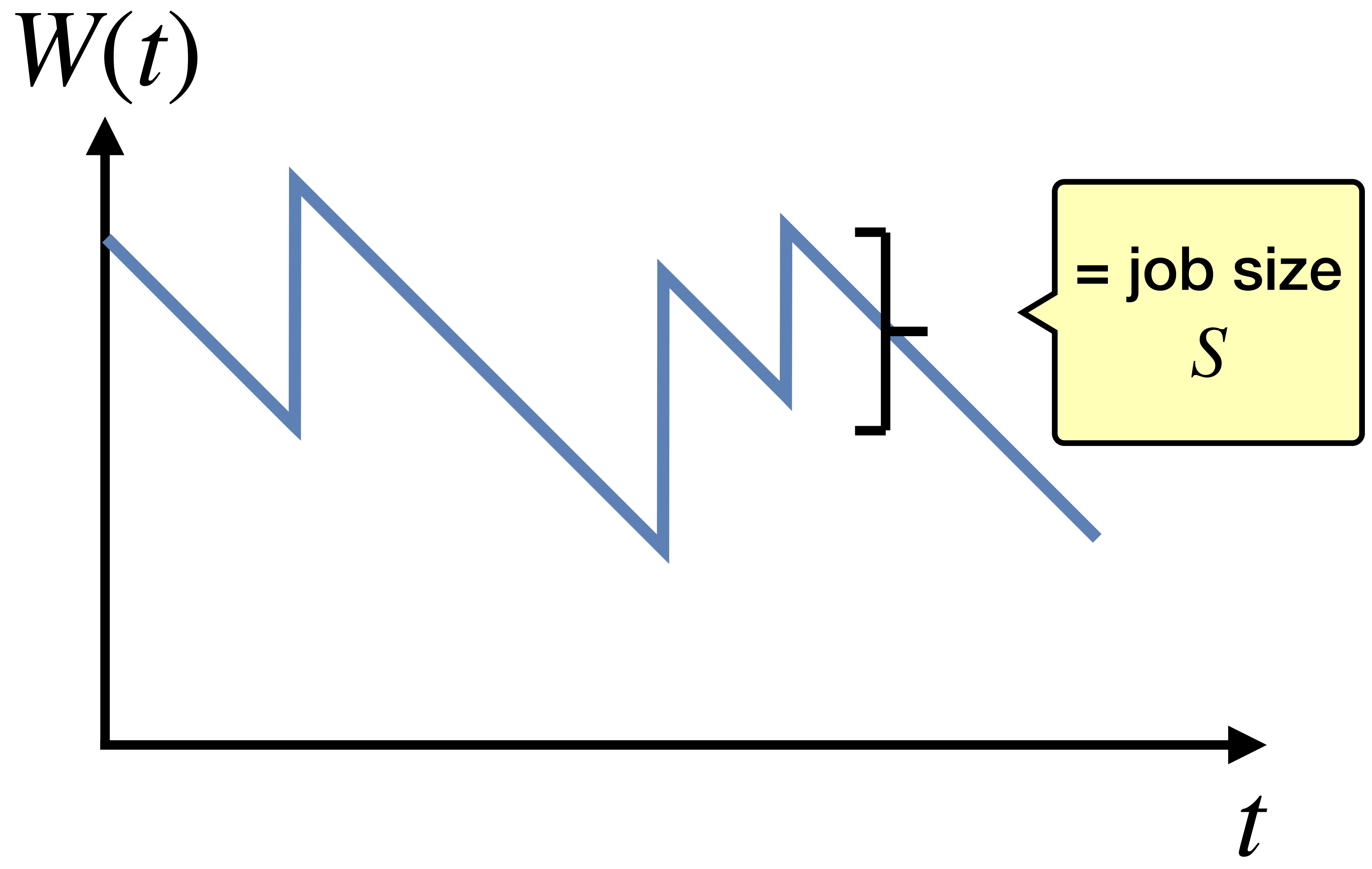
= job size S

residual arrival time

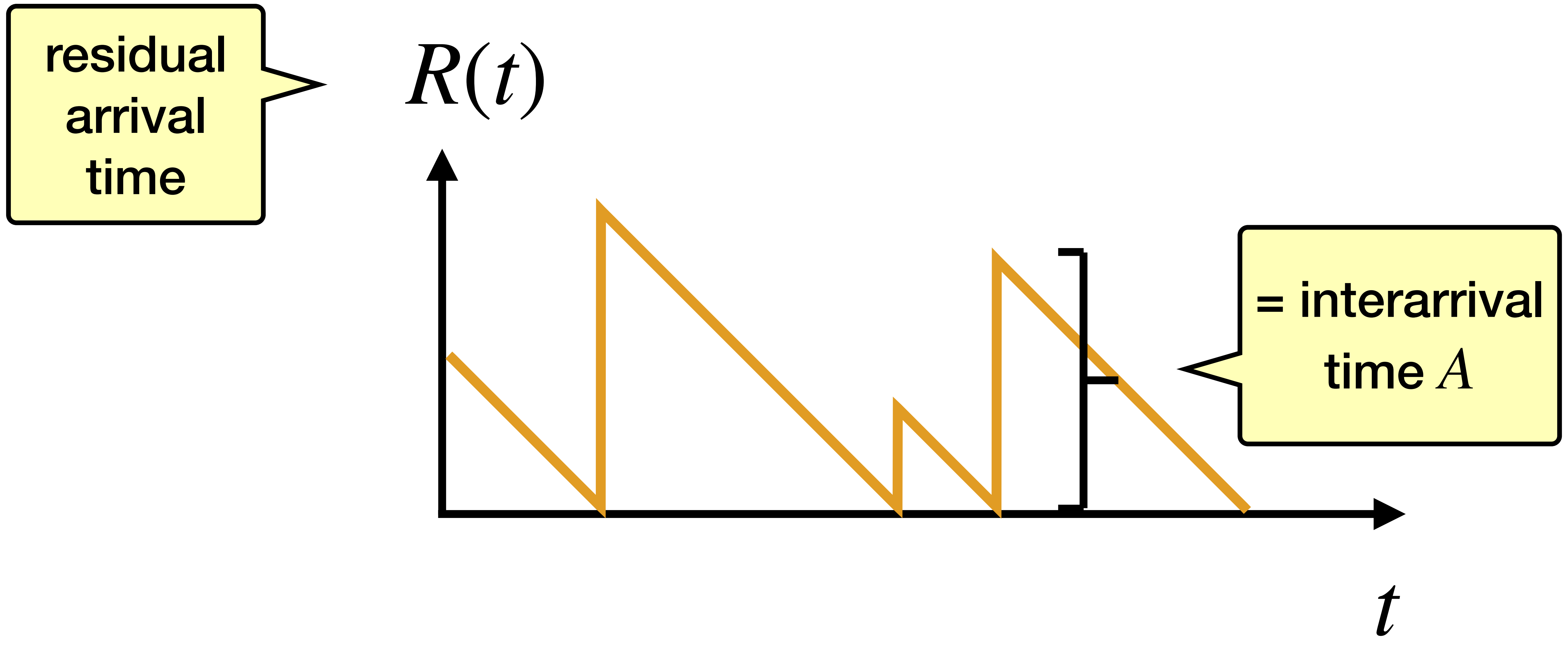


Work analysis

- Why is G/G work analysis harder than M/G?
- M/G/1: $W(t)$ Markov process ✓
- G/G/1: $W(t)$ not a Markov process ✗



= job size S

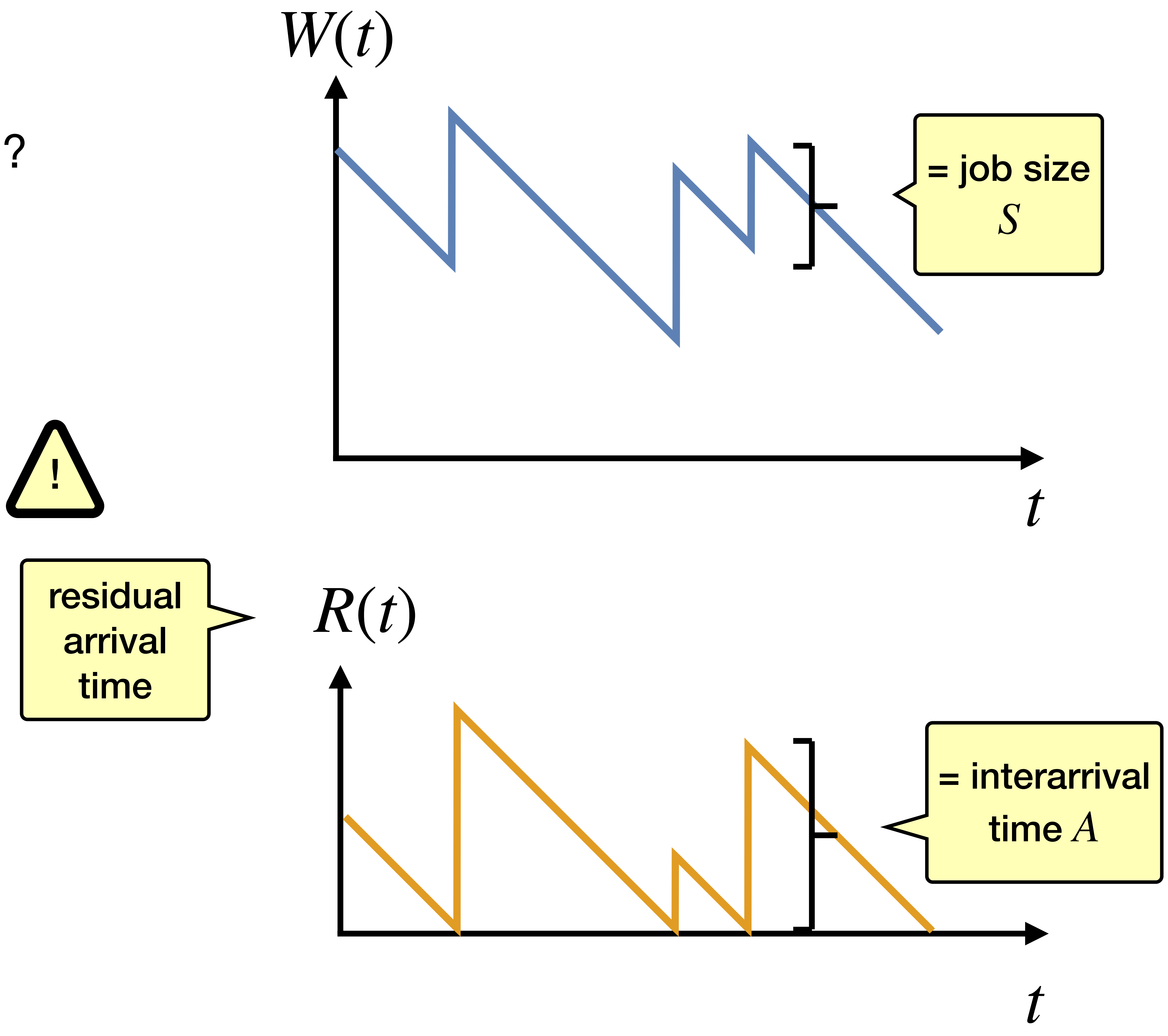


residual arrival time

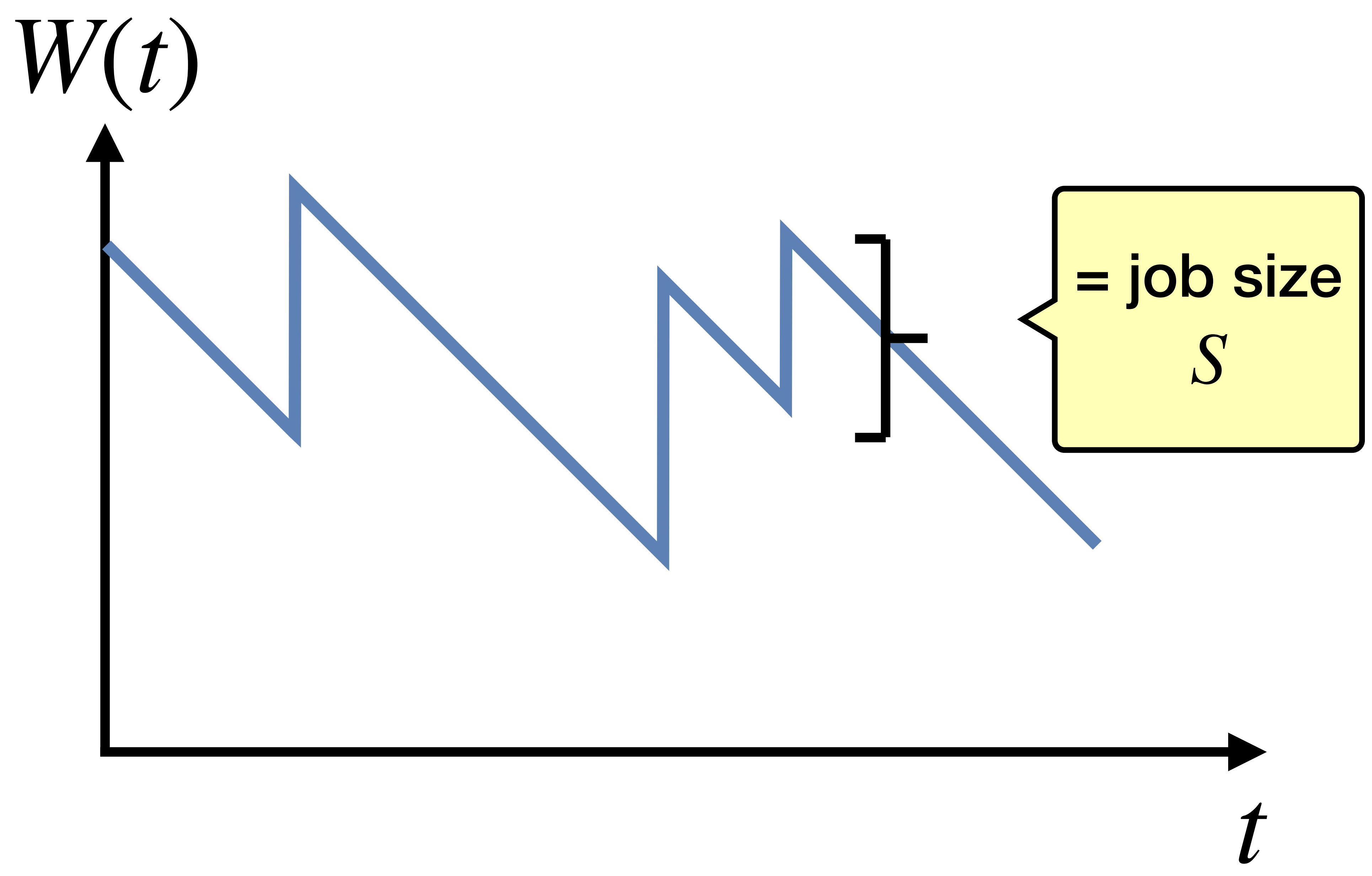
= interarrival time A

Work analysis

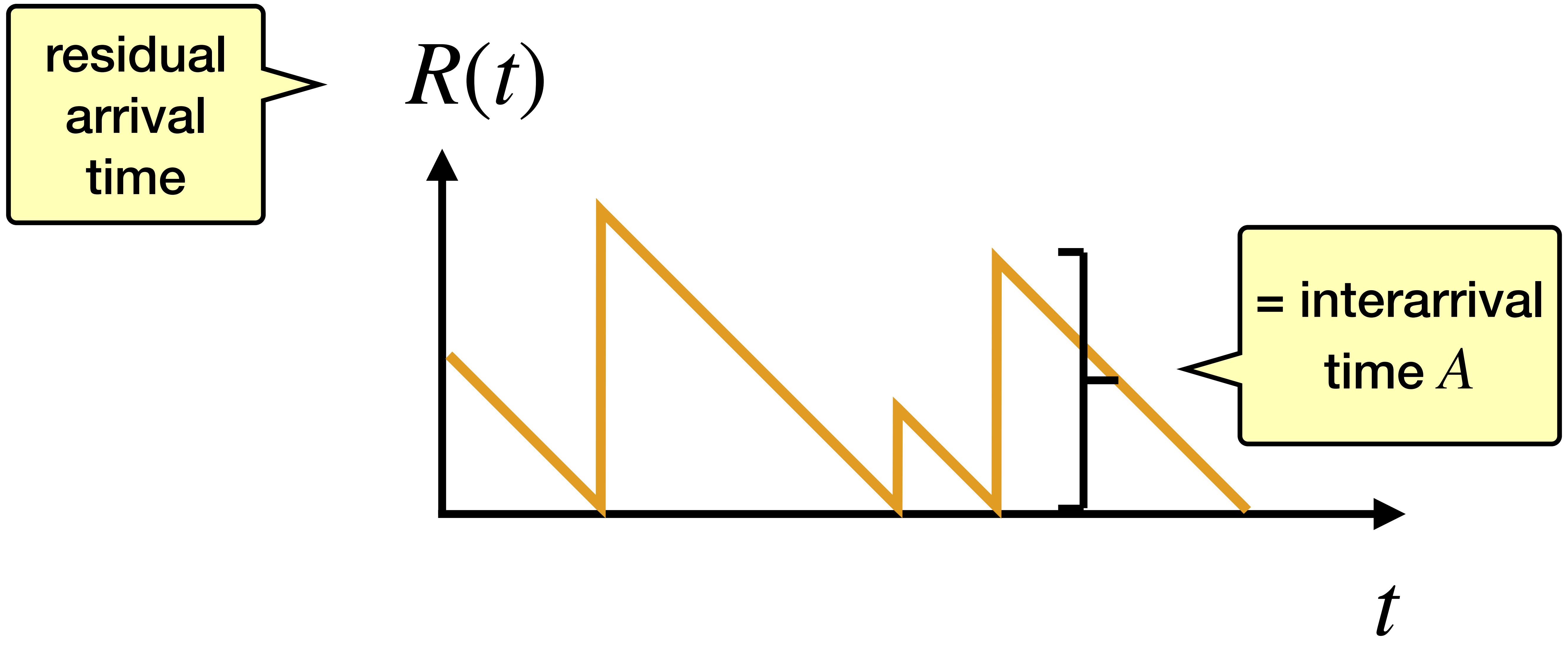
- Why is G/G work analysis harder than M/G?
- M/G/1: $W(t)$ Markov process ✓
- G/G/1: $W(t)$ not a Markov process ✗
- $(W(t), R(t))$ Markov, but two dimensional ⚠



Work analysis



= job size S

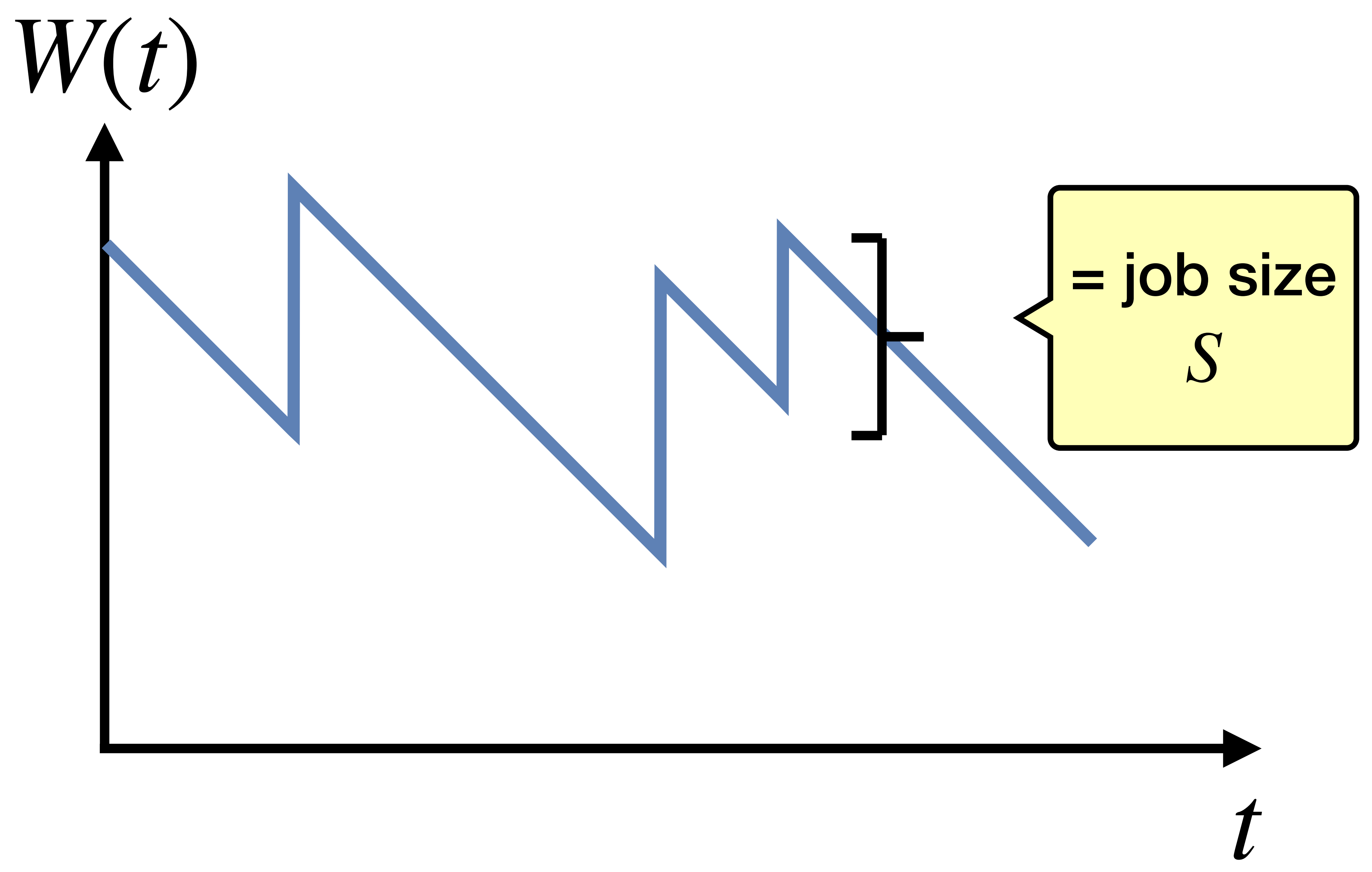


residual arrival time

= interarrival time A

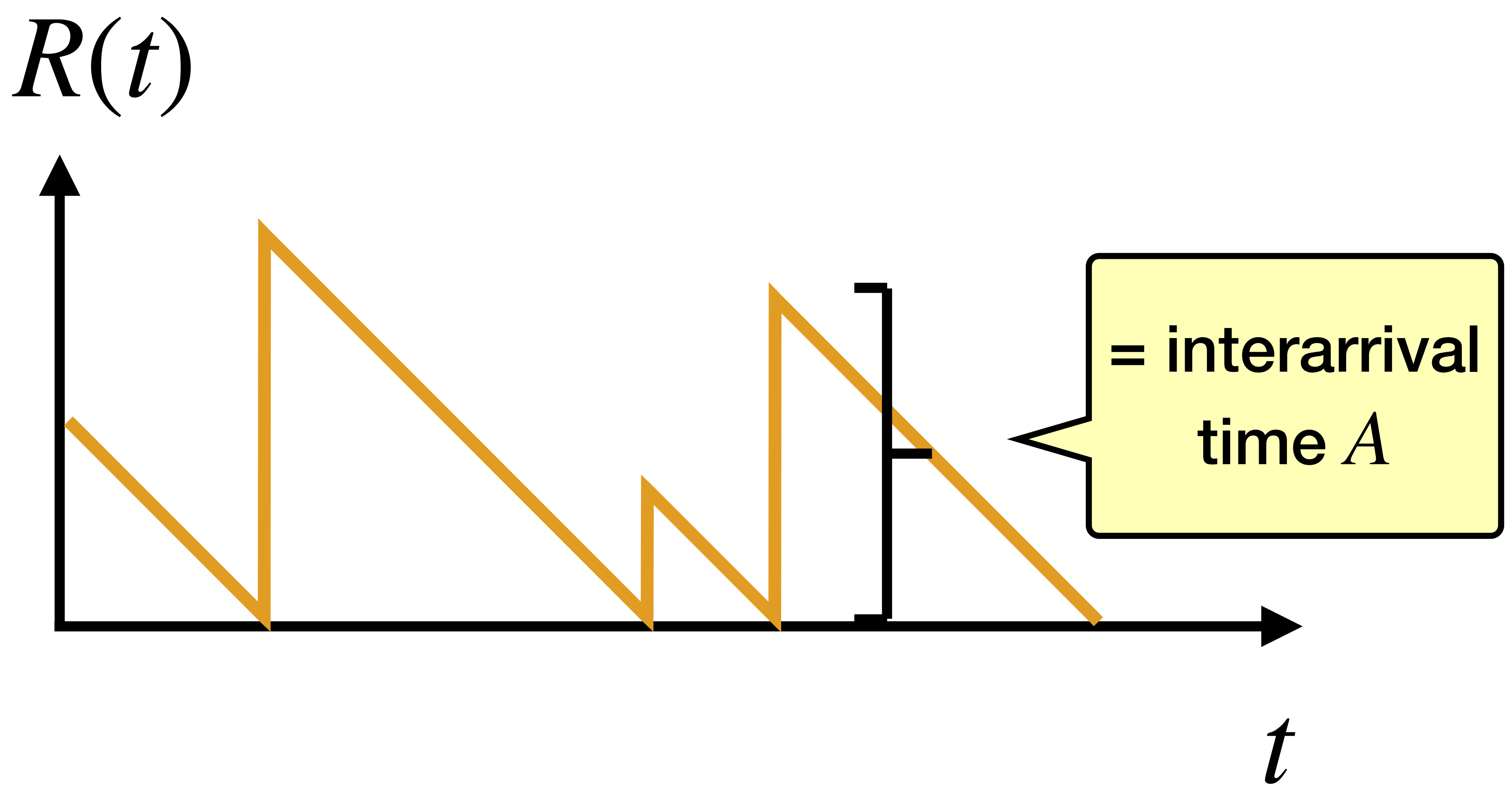
Work analysis

Idea: consider $W(t) - \rho R(t)$



= job size S

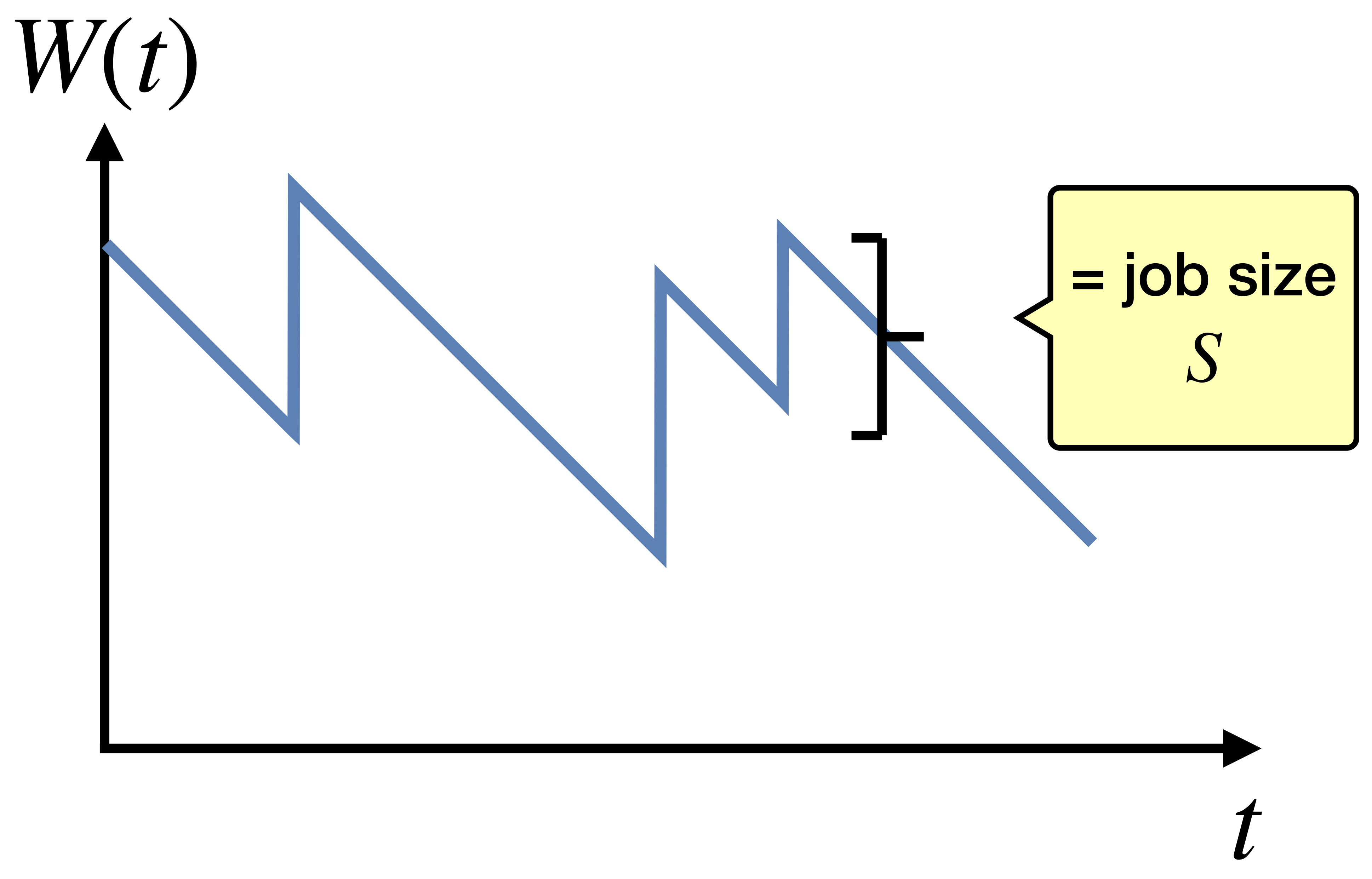
residual arrival time

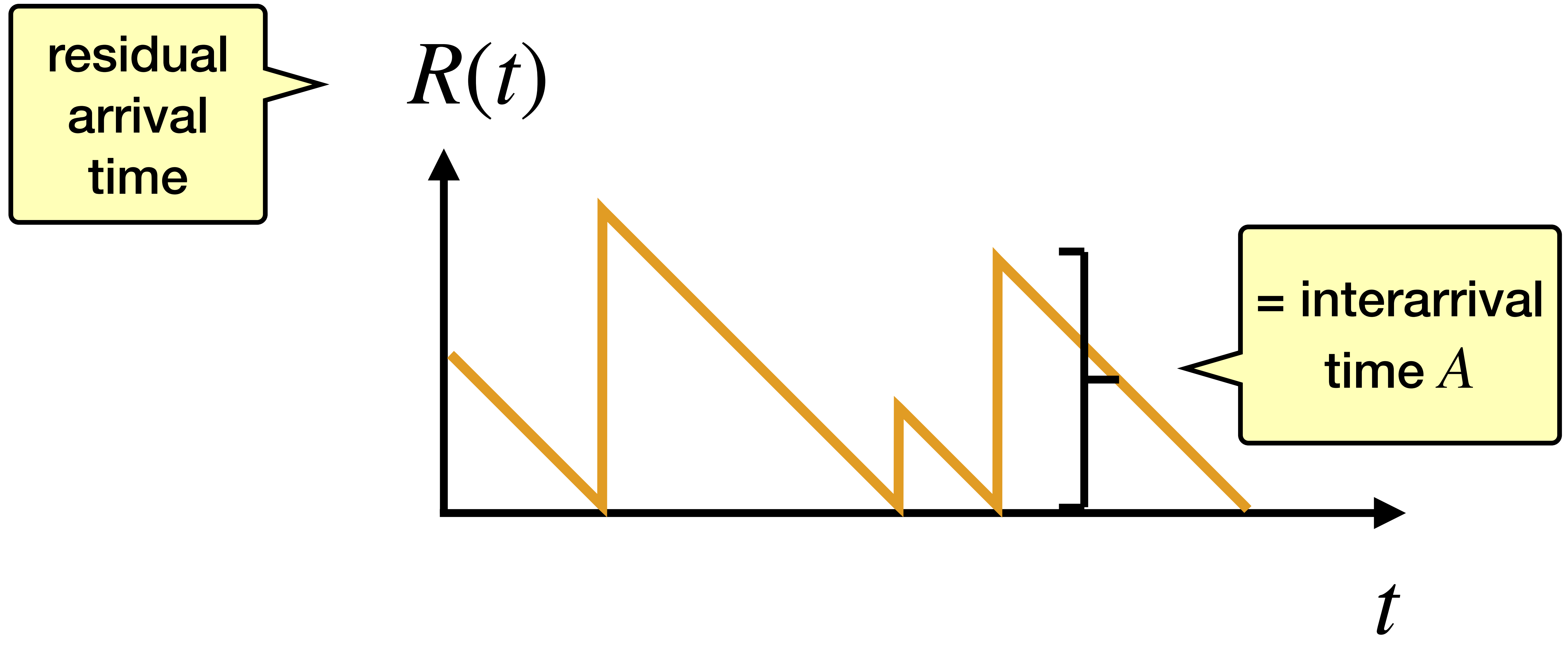


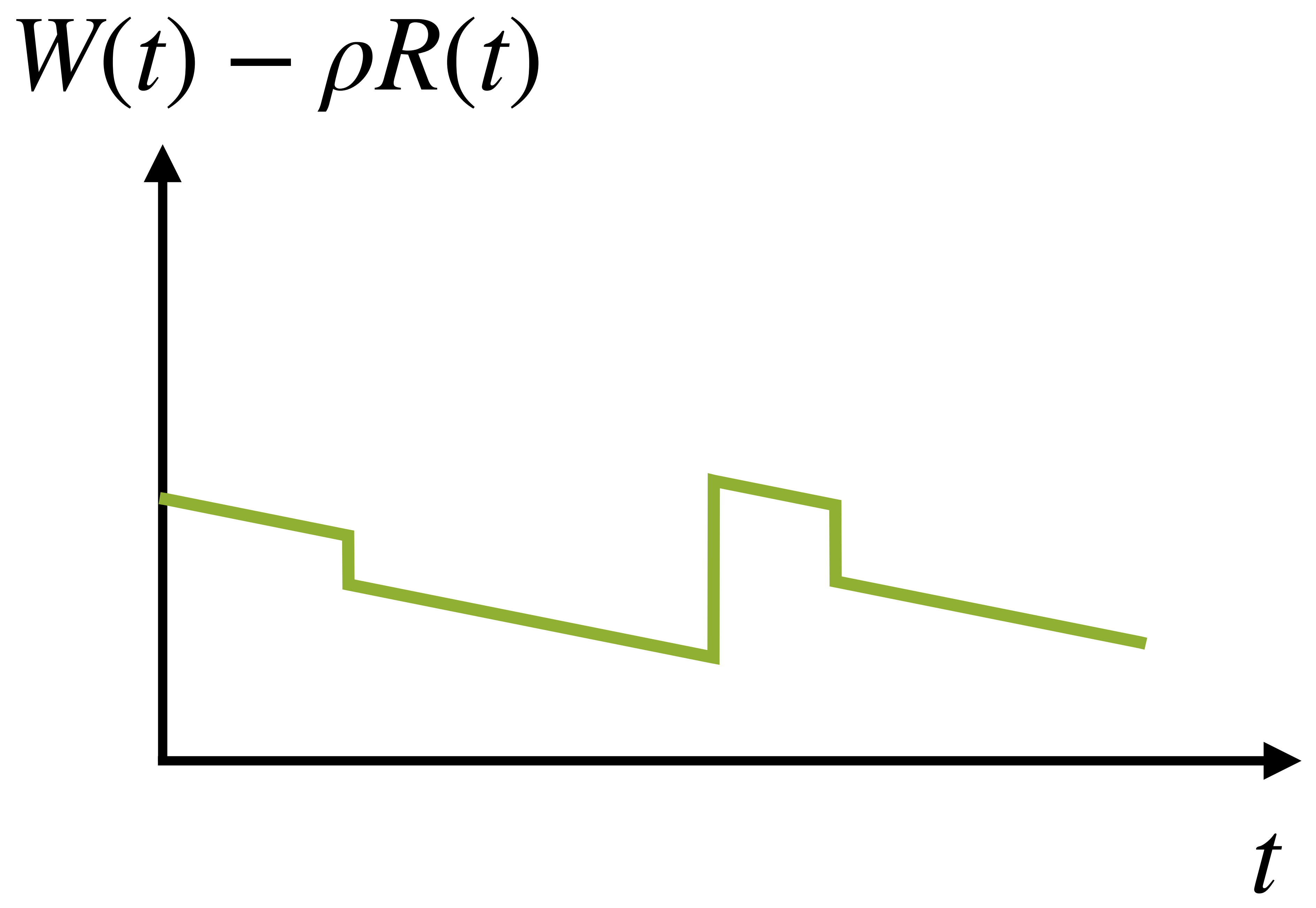
= interarrival time A

Work analysis

Idea: consider $W(t) - \rho R(t)$

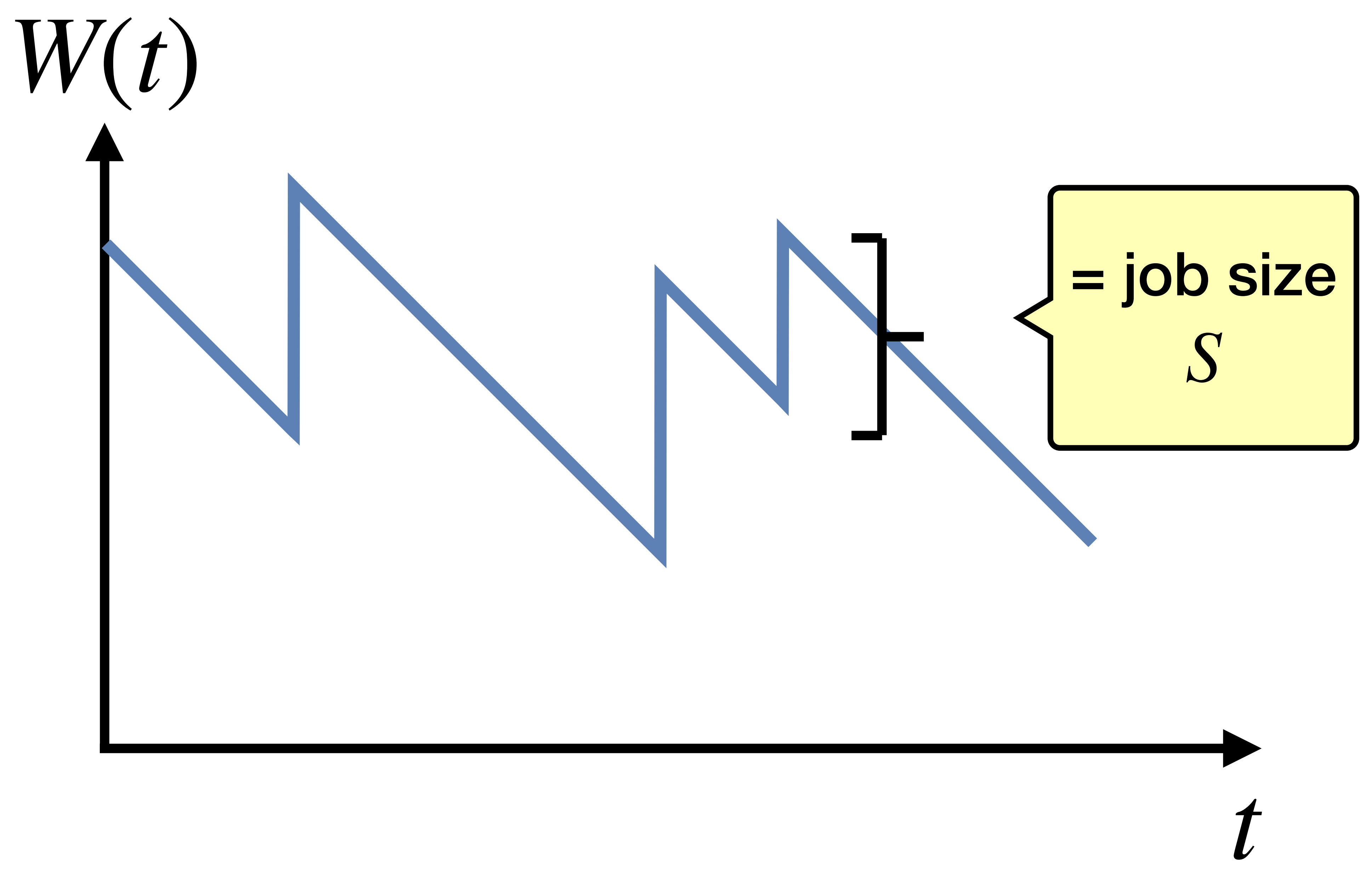




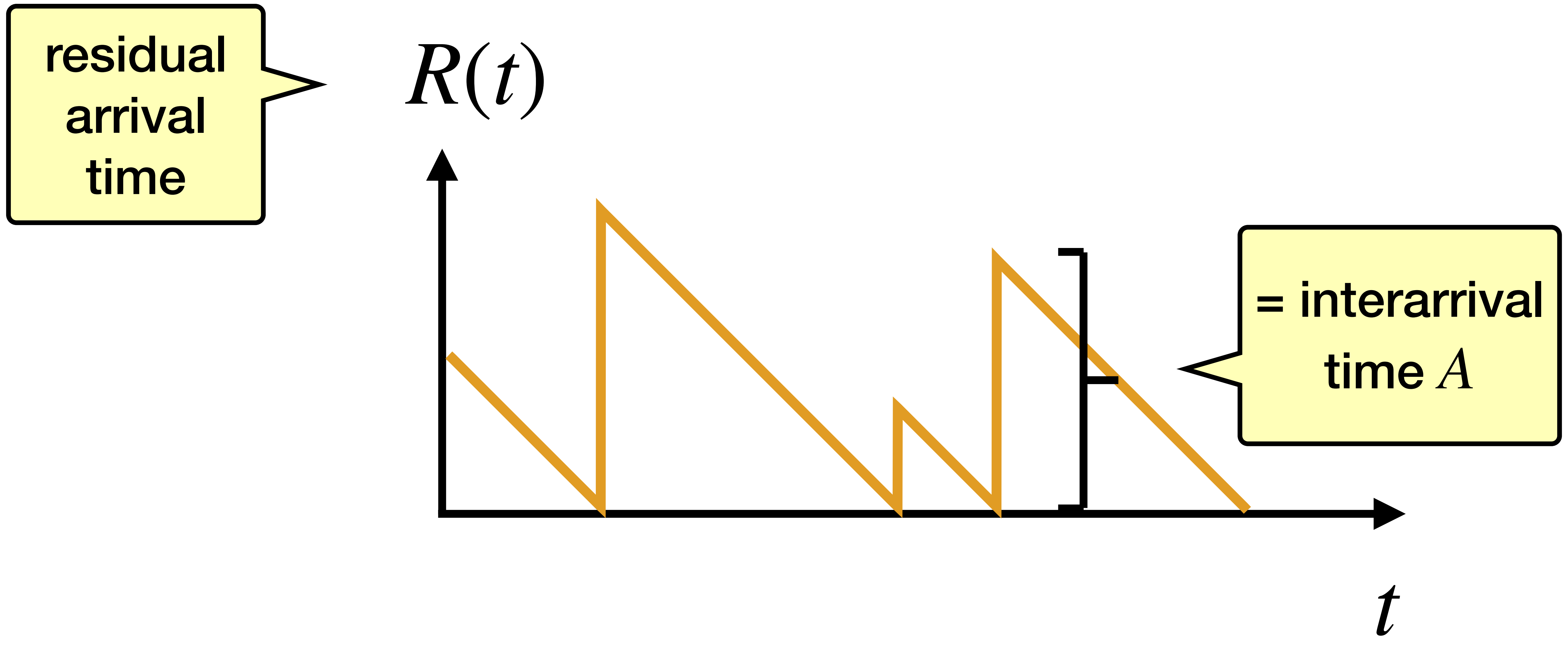


Work analysis

Idea: consider $W(t) - \rho R(t)$

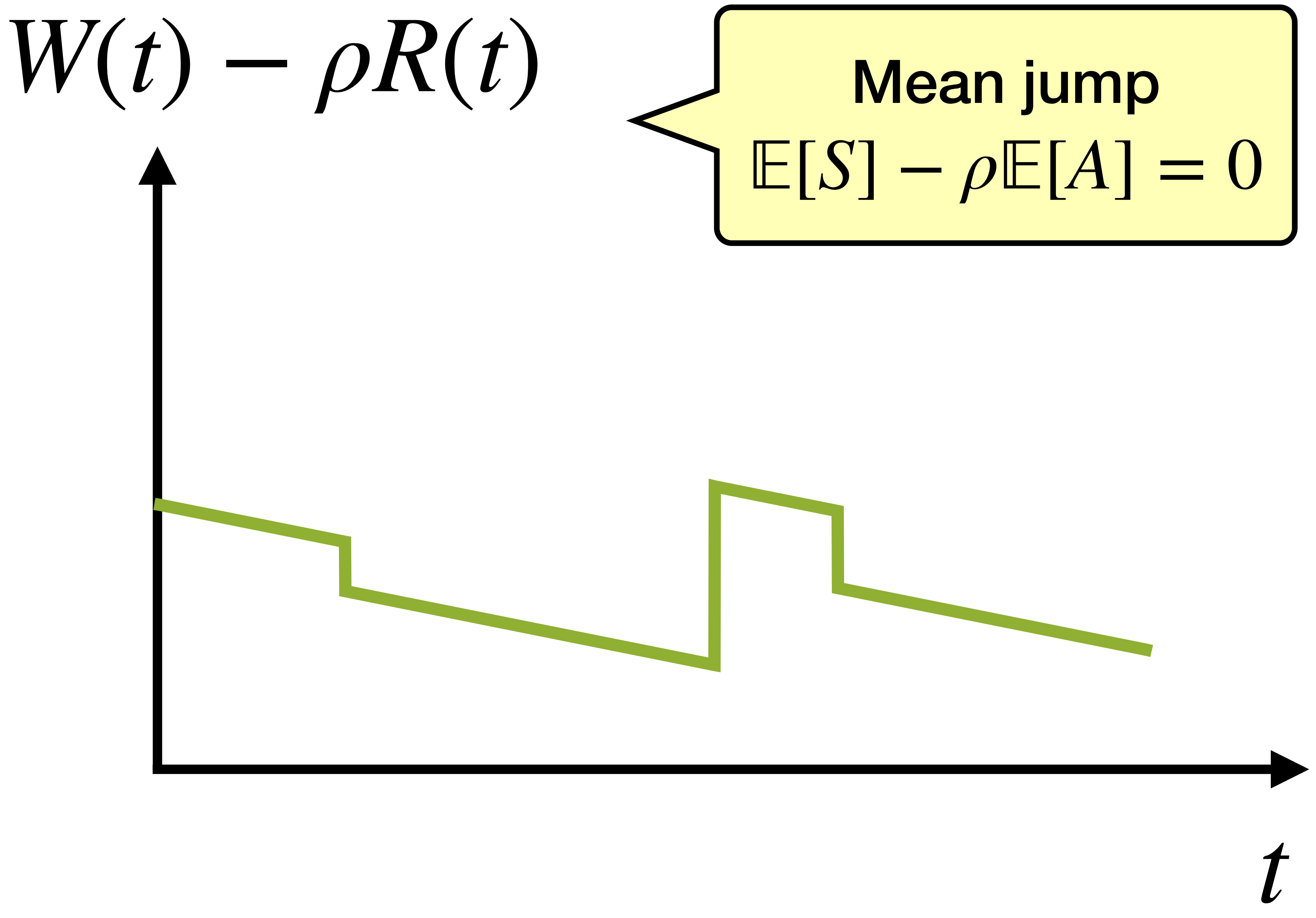


= job size S



residual arrival time

= interarrival time A

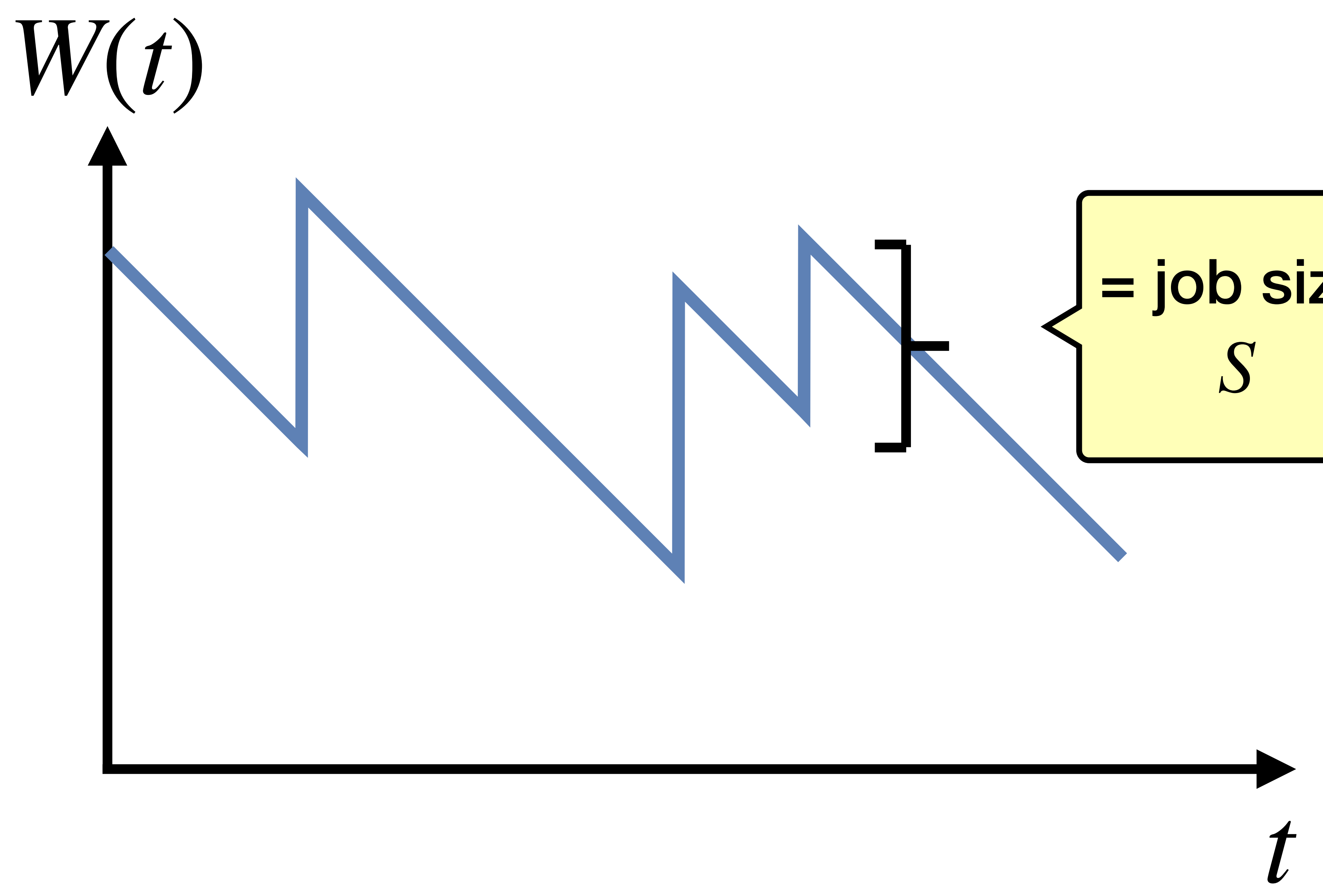


Mean jump $\mathbb{E}[S] - \rho\mathbb{E}[A] = 0$

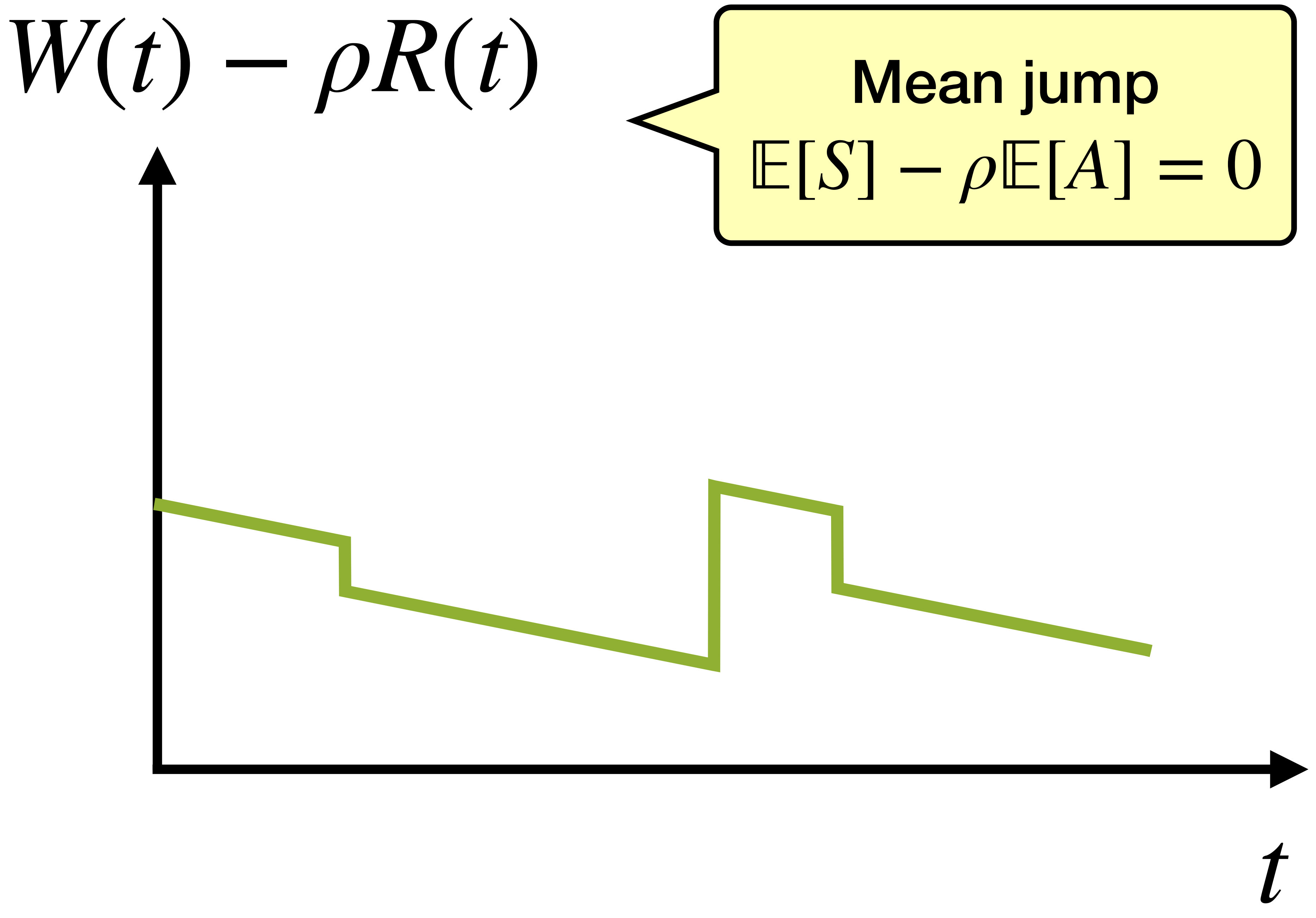
Work analysis

Idea: consider $W(t) - \rho R(t)$

$W(t) - \rho R(t)$ decreases at const rate + noise

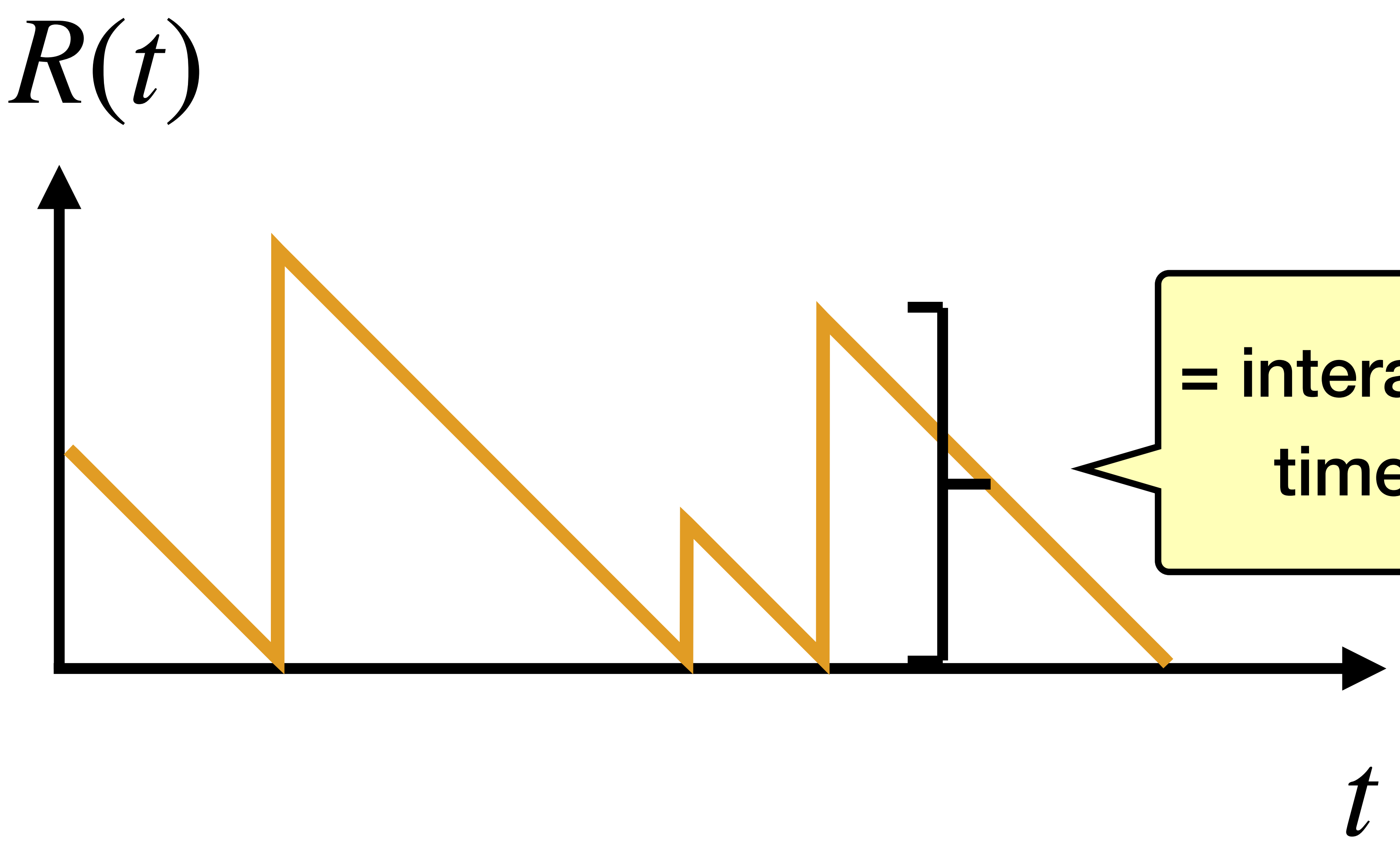


= job size S



Mean jump
$\mathbb{E}[S] - \rho\mathbb{E}[A] = 0$

residual arrival time



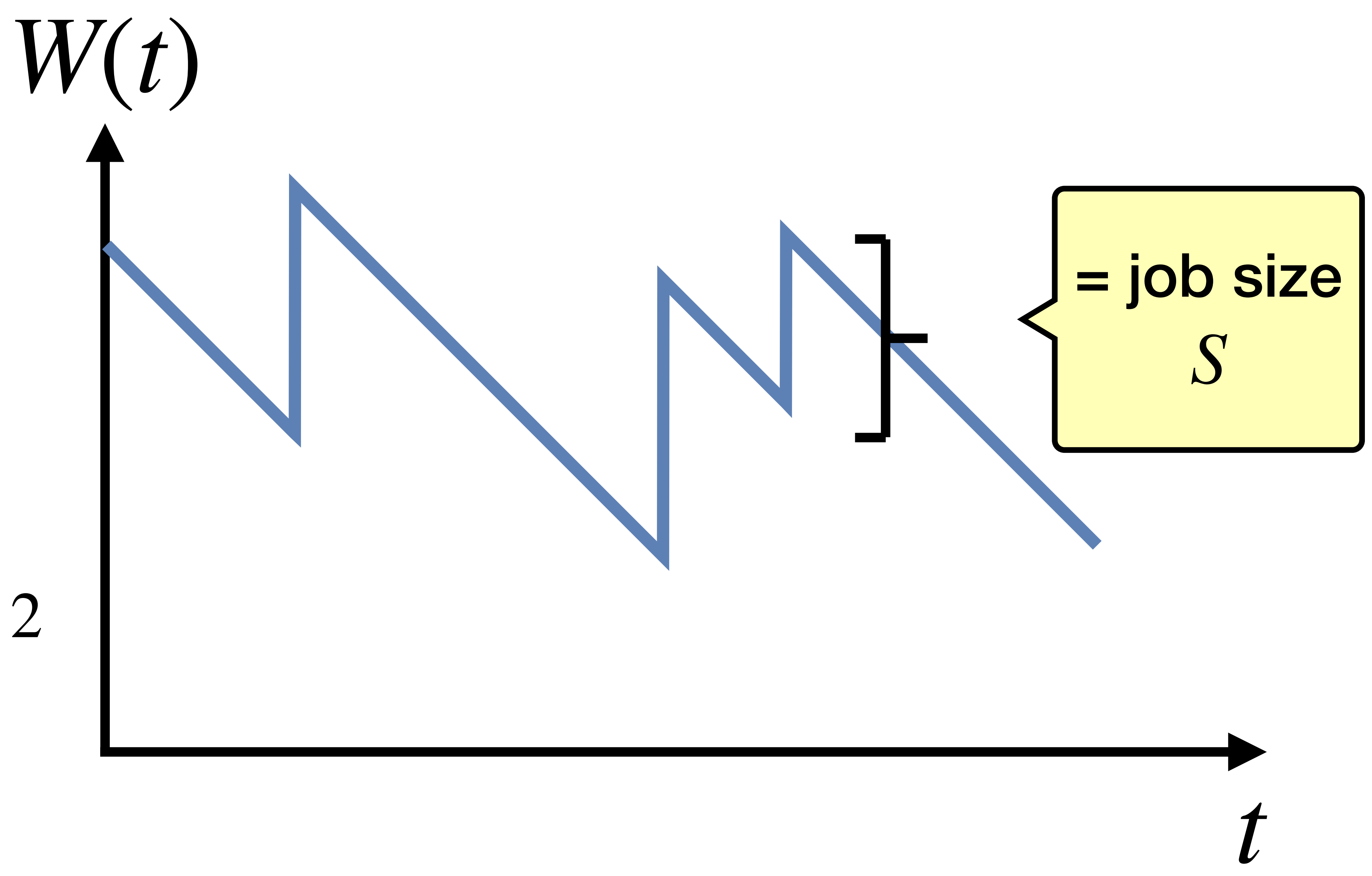
= interarrival time A

Work analysis

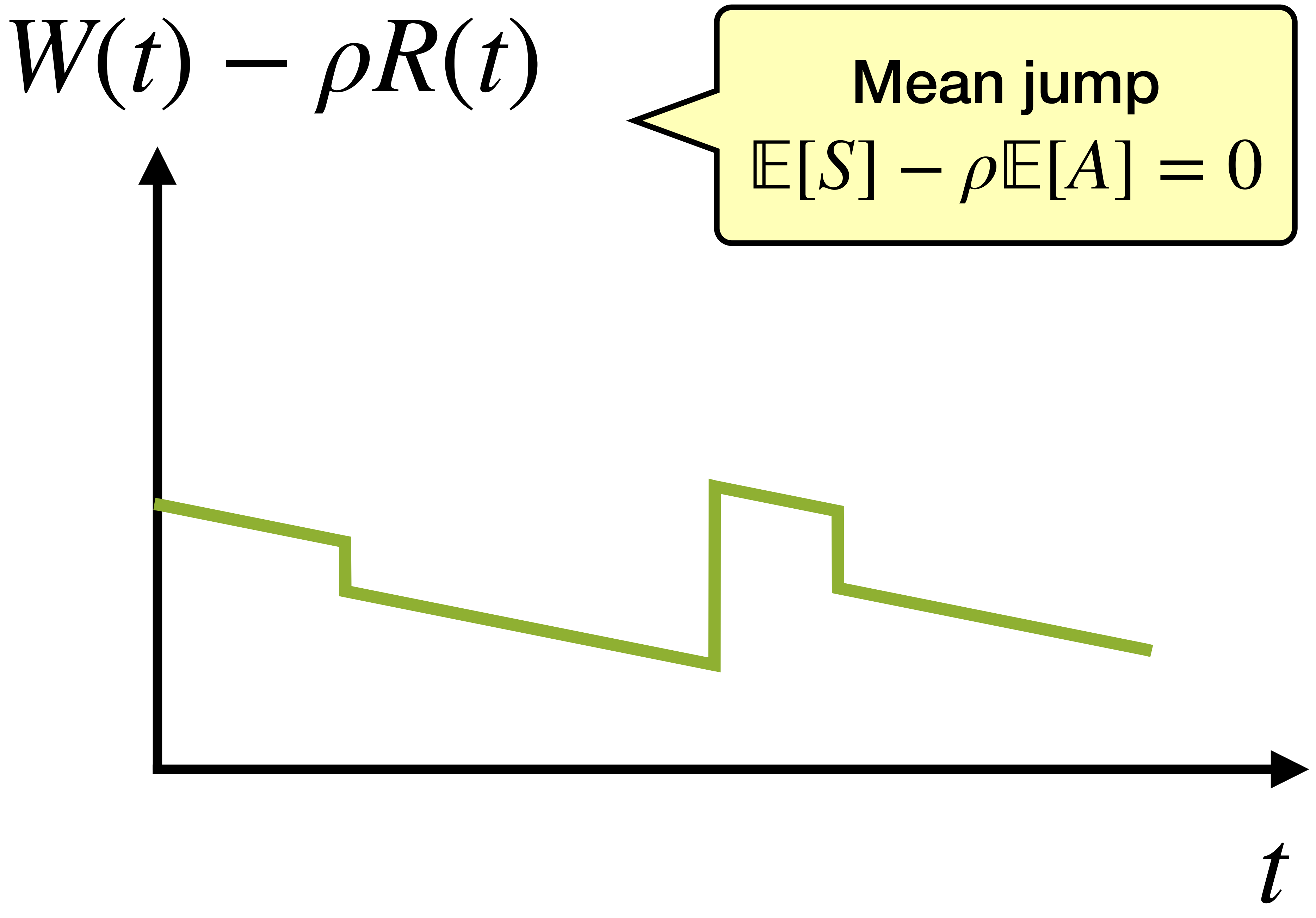
Idea: consider $W(t) - \rho R(t)$

$W(t) - \rho R(t)$ decreases at const rate + noise

Formally, apply Rate Conservation Law to $\big(W(t) - \rho R(t)\big)^2$

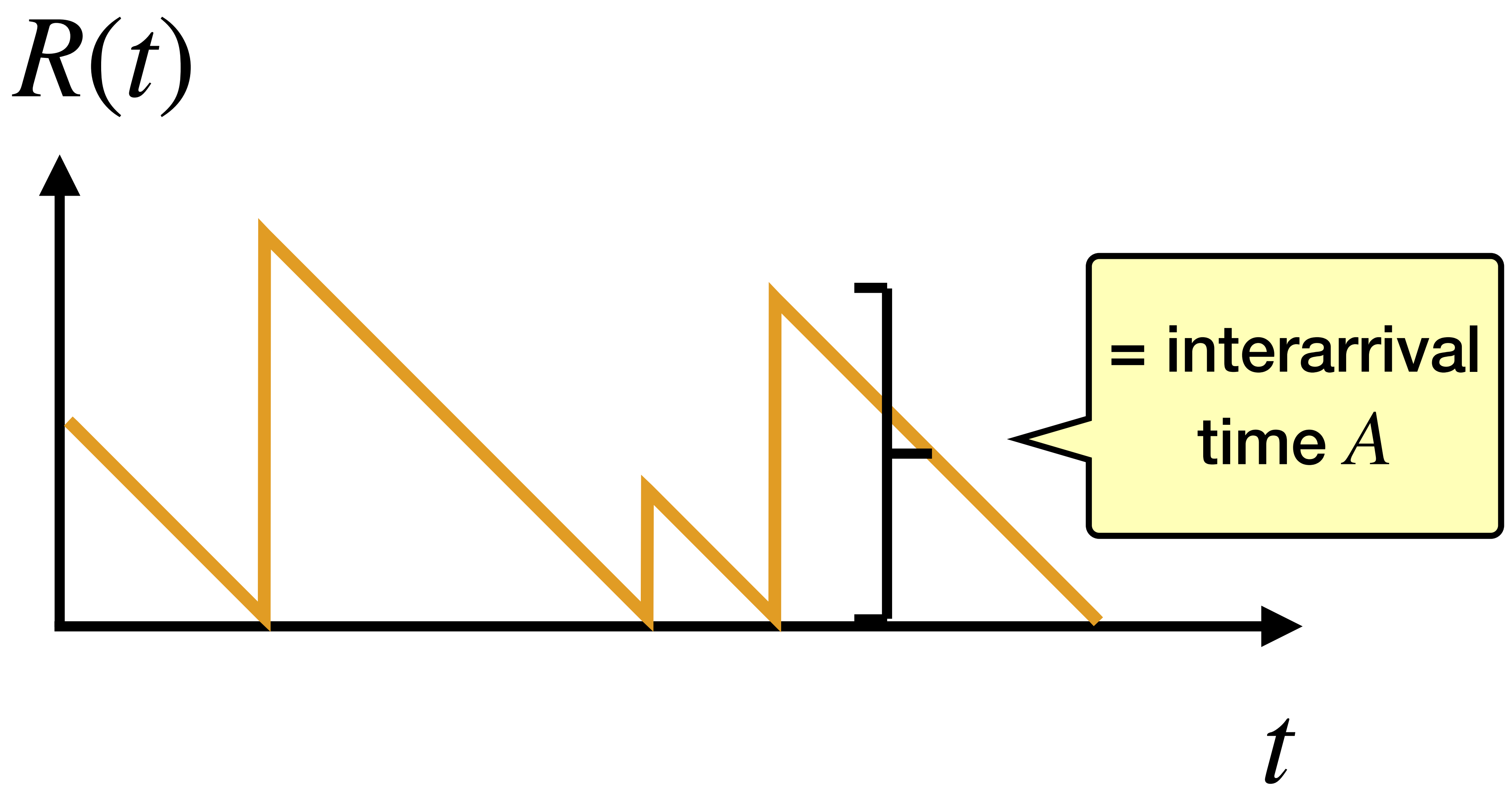


= job size S



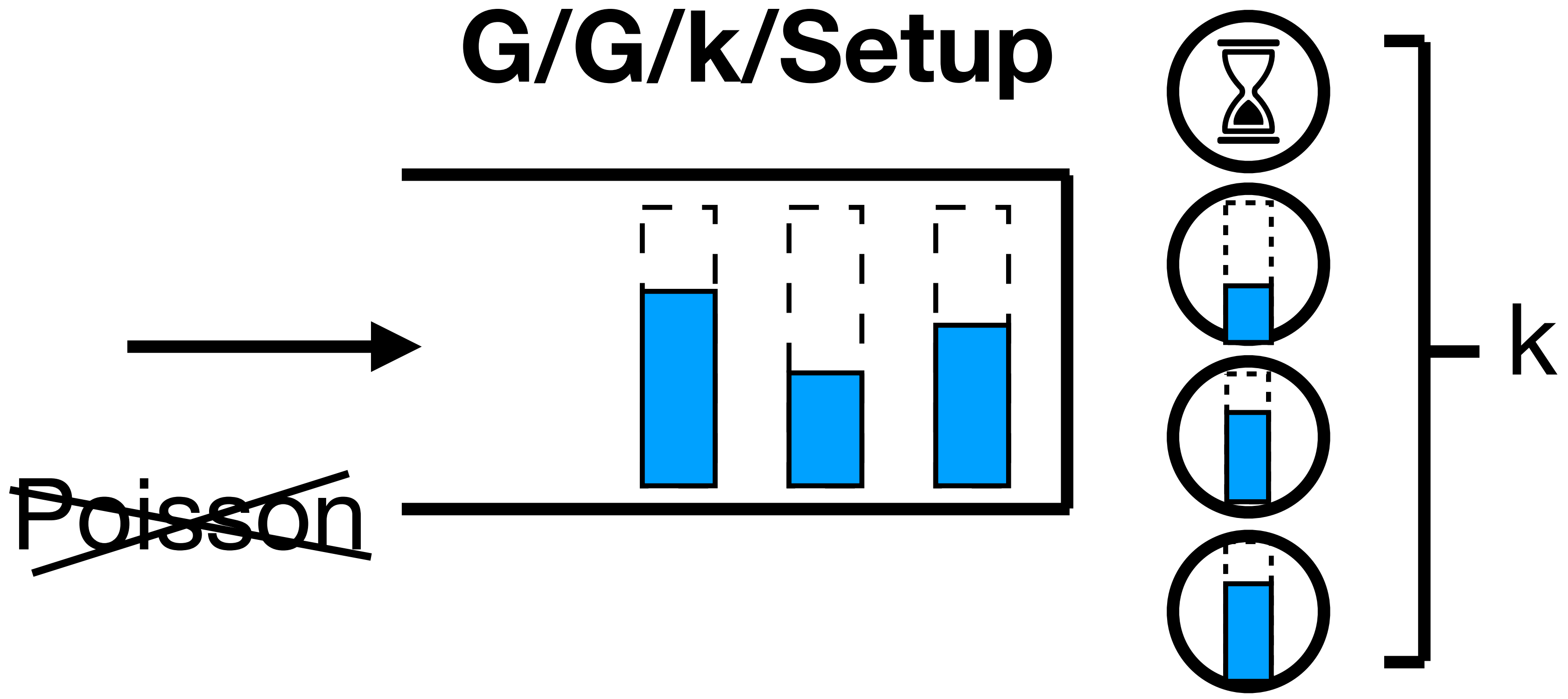
Mean jump $\mathbb{E}[S] - \rho\mathbb{E}[A] = 0$

residual arrival time

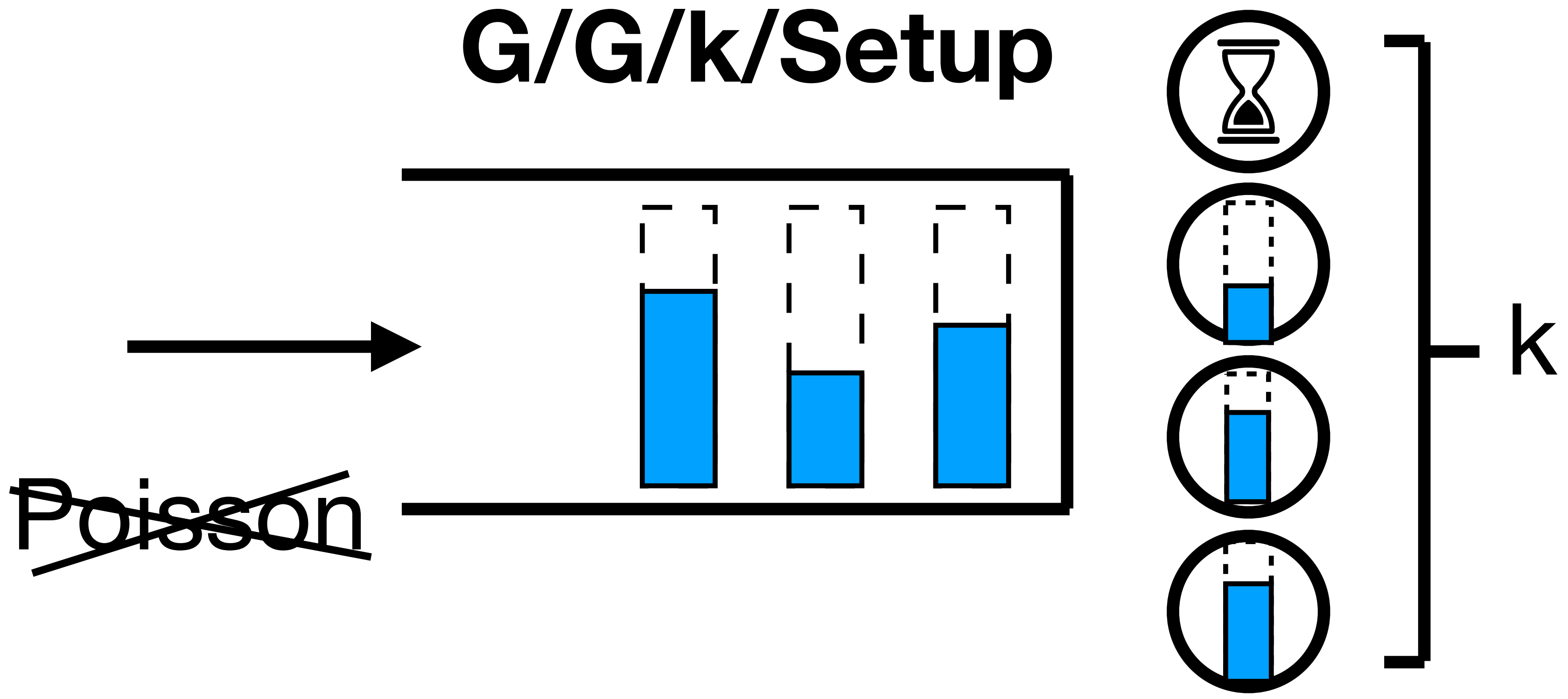


= interarrival time A

Conclusion

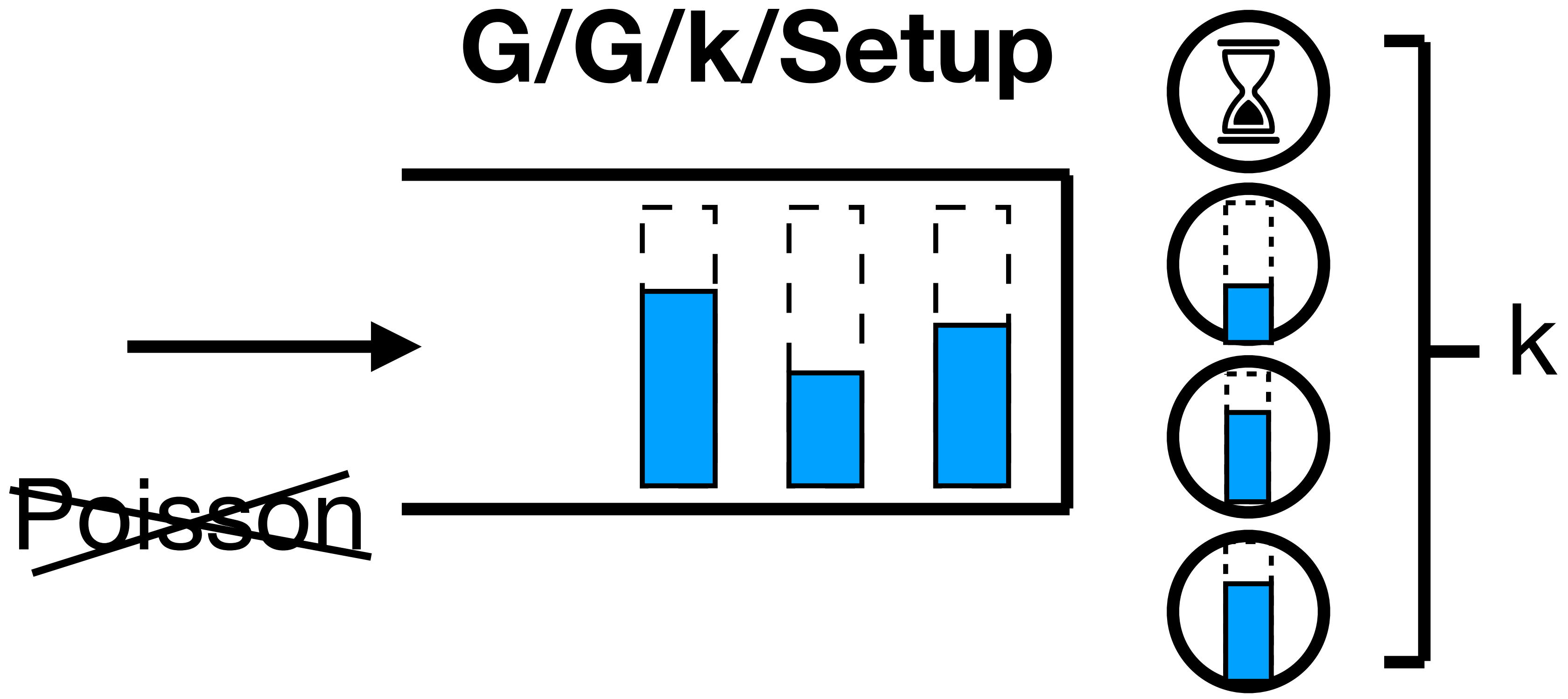


Conclusion



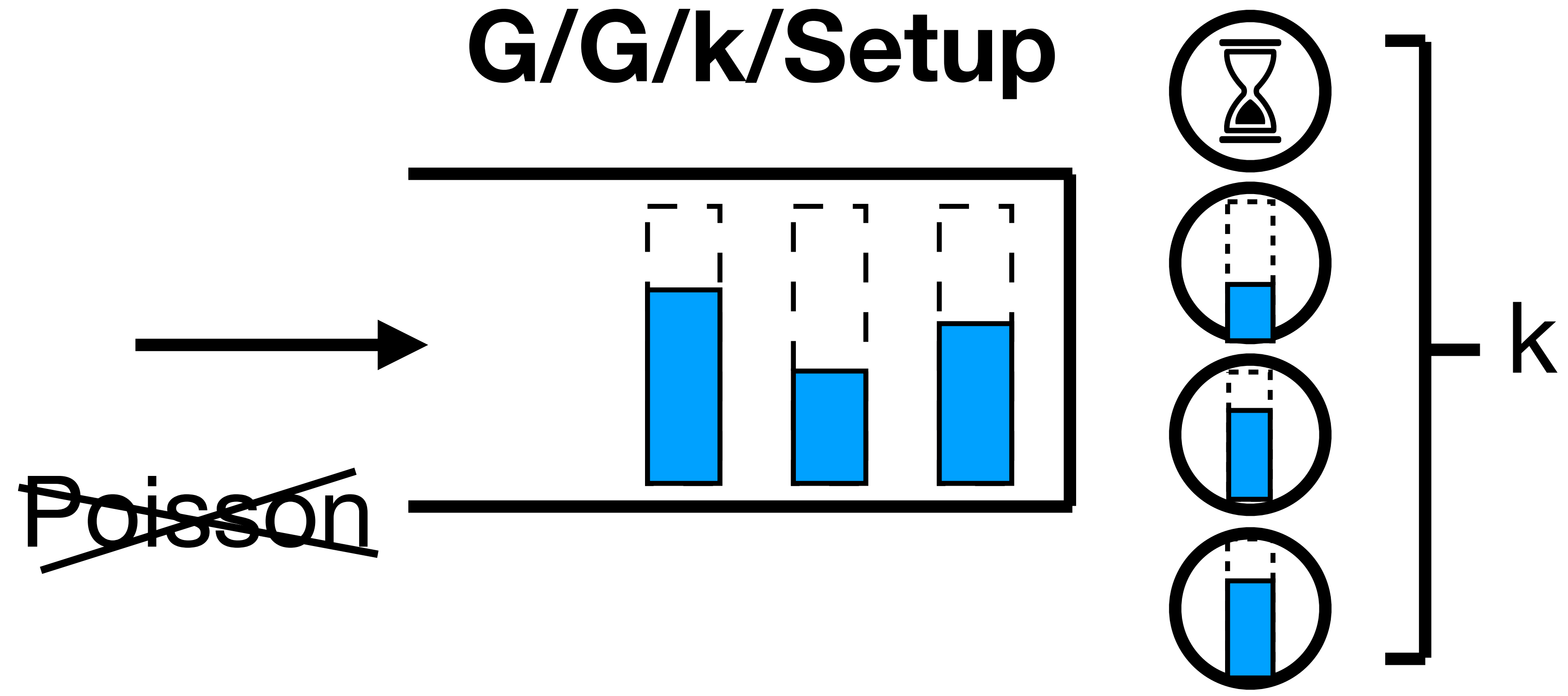
- **Problem:** scheduling in G/G/k/setup

Conclusion



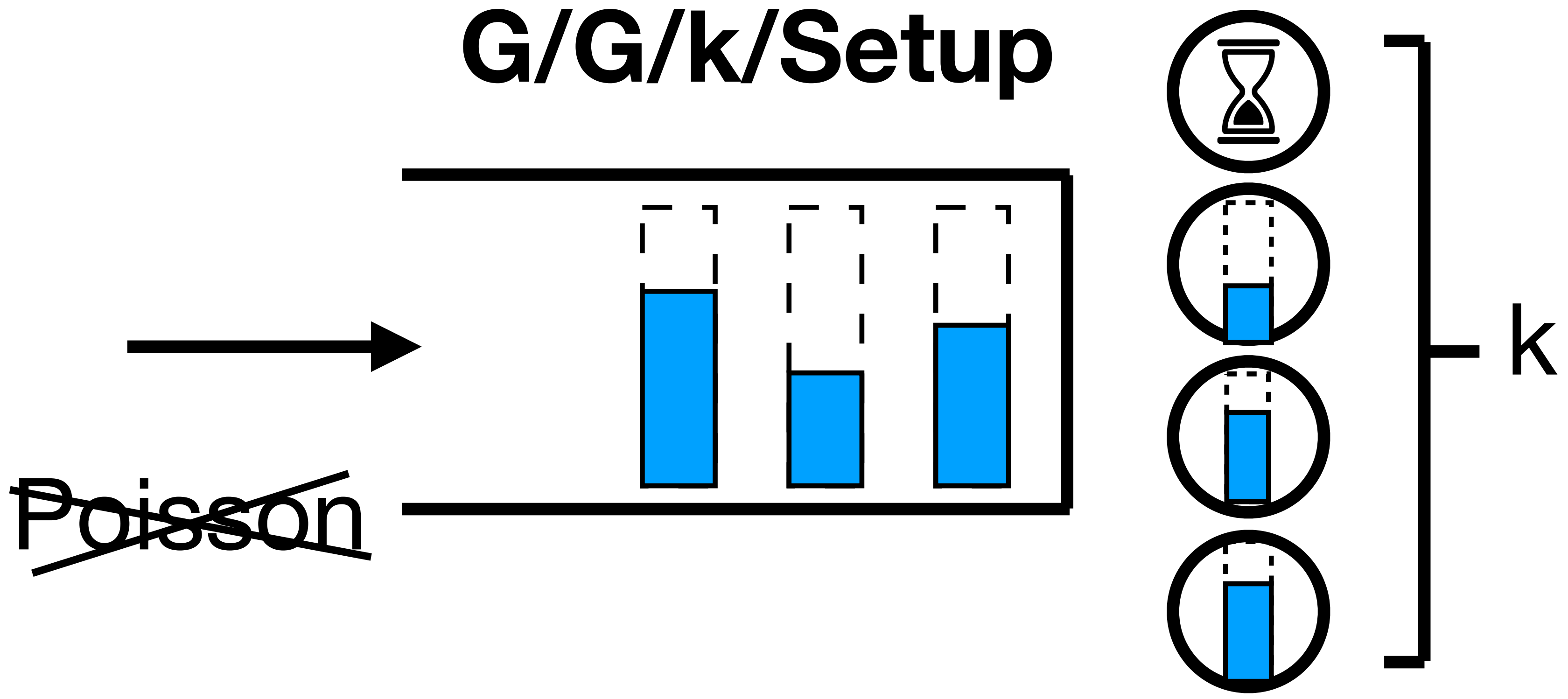
- **Problem:** scheduling in G/G/k/setup
- **Question:** Is Gittins good?

Conclusion



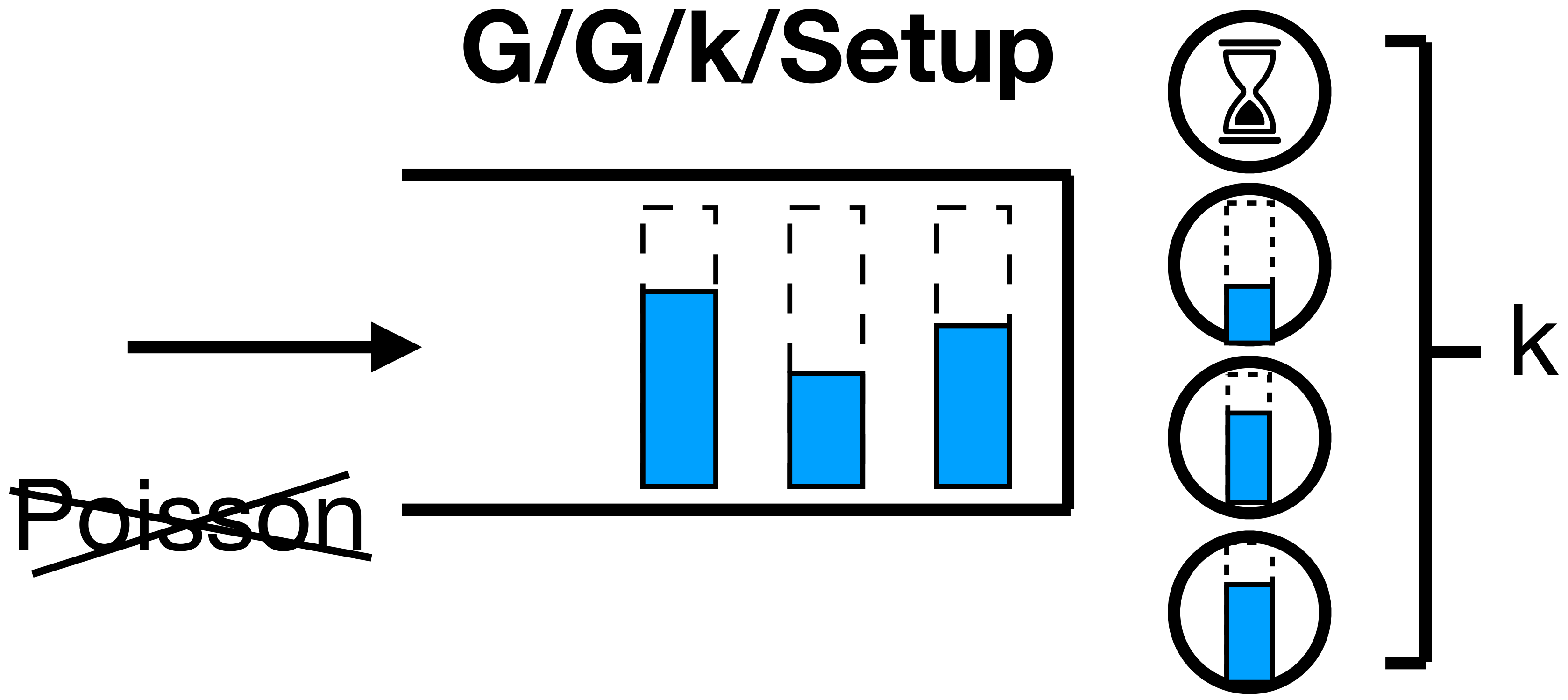
- **Problem:** scheduling in G/G/k/setup
- **Question:** Is Gittins good?
- **Result:** gap $\leq \ell_{(a)} + \ell_{(b)} + \ell_{(c)}$

Conclusion



- **Problem:** scheduling in G/G/k/setup
- **Question:** Is Gittins good?
- **Result:** gap $\leq \ell_{(a)} + \ell_{(b)} + \ell_{(c)}$
- **Corollary:** heavy-traffic optimal if $\mathbb{E}[S^2 \log S] < \infty$

Conclusion

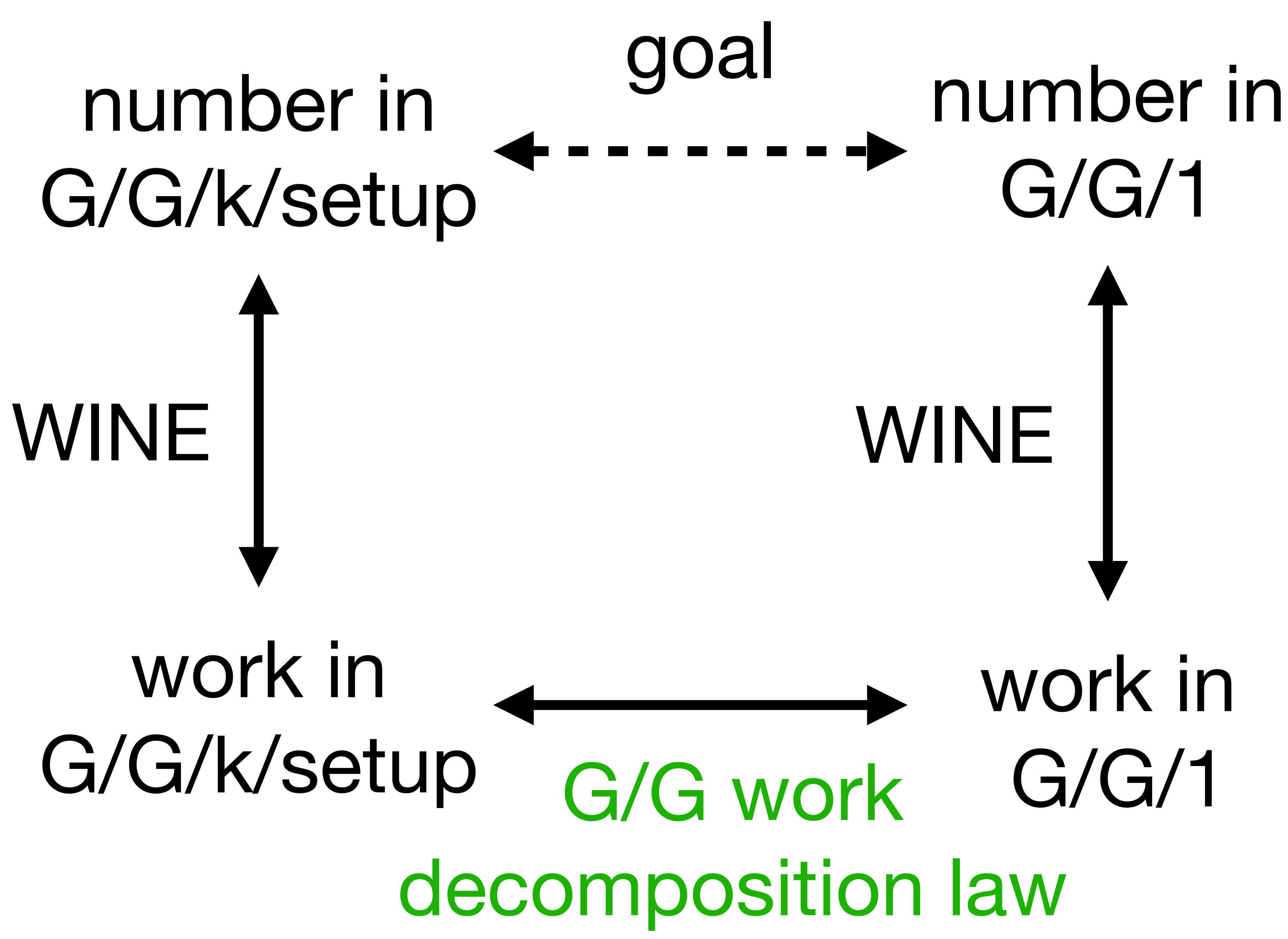


- **Problem:** scheduling in G/G/k/setup
- **Question:** Is Gittins good?
- **Result:** gap $\leq \ell_{(a)} + \ell_{(b)} + \ell_{(c)}$
- **Corollary:** heavy-traffic optimal if $\mathbb{E}[S^2 \log S] < \infty$
- **Key tool:** new G/G work-decomposition law

Conclusion

G/G/k/Setup

- **Problem:** scheduling in G/G/k/setup
- **Question:** Is Gittins good?
- **Result:** gap $\leq \ell_{(a)} + \ell_{(b)} + \ell_{(c)}$
- **Corollary:** heavy-traffic optimal if $\mathbb{E}[S^2 \log S] < \infty$
- **Key tool:** new G/G work-decomposition law



number in G/G/k/setup

goal

number in G/G/1

WINE

WINE

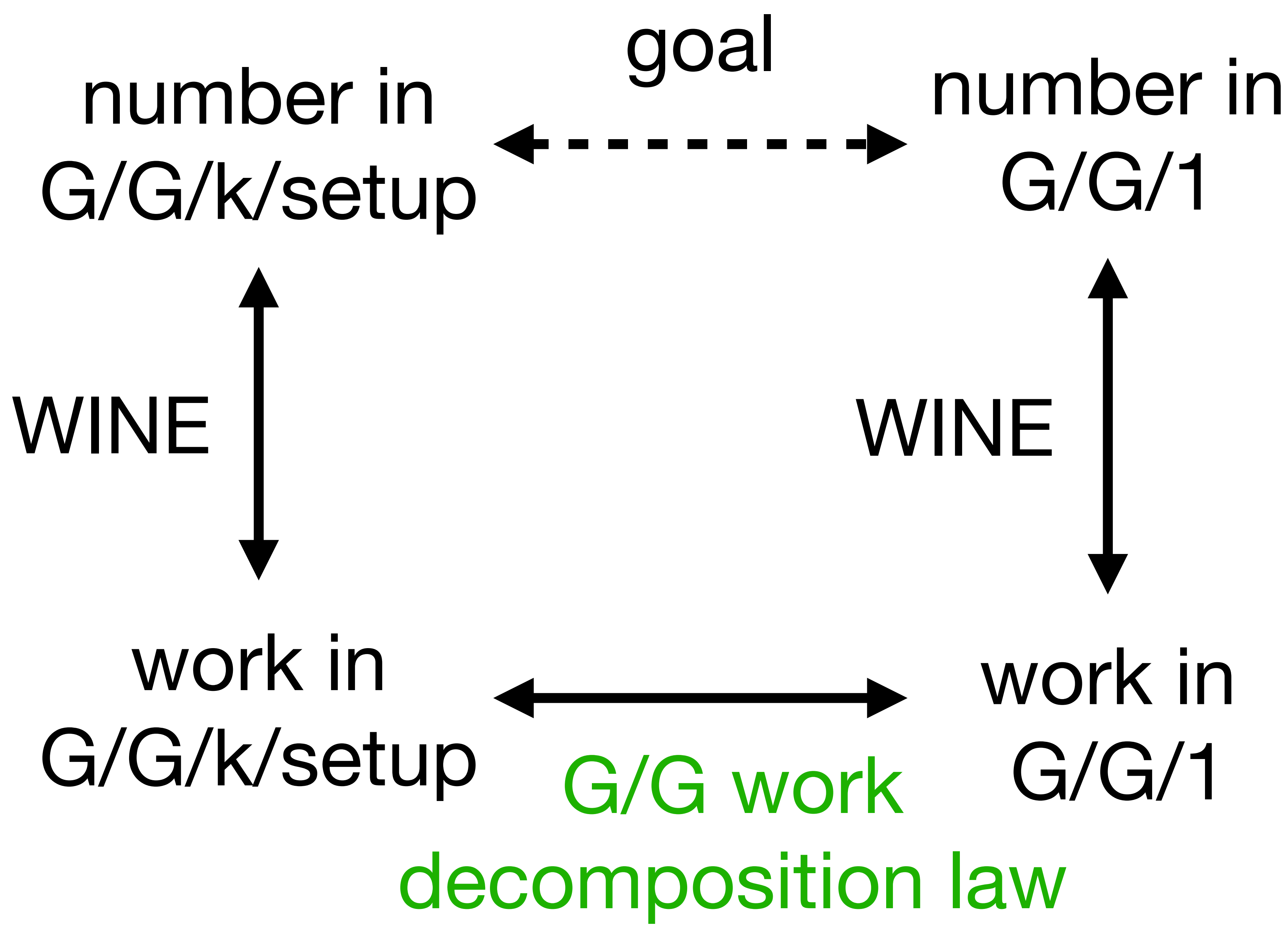
work in G/G/k/setup

work in G/G/1

G/G work decomposition law

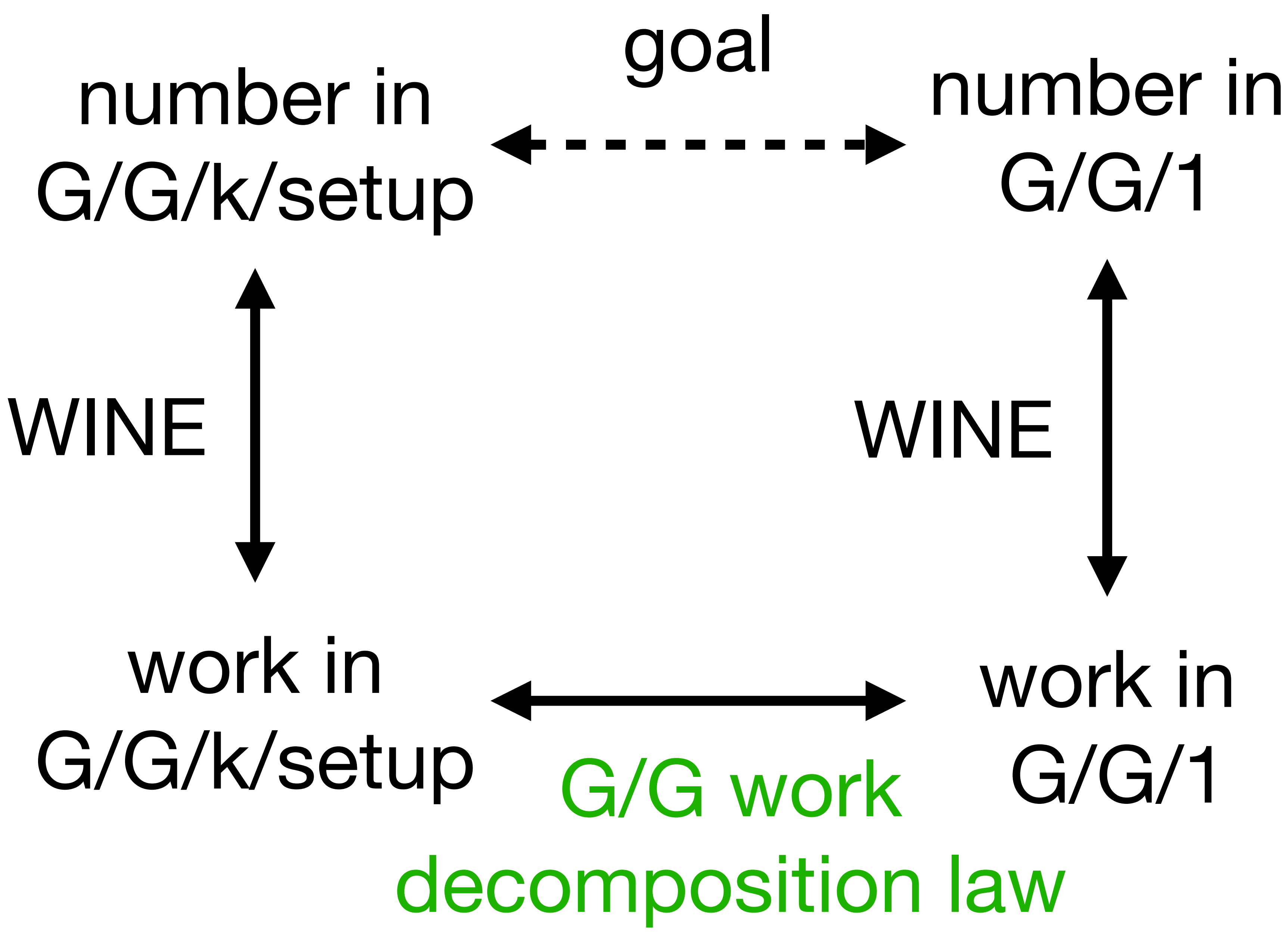
Conclusion

- **Problem:** scheduling in G/G/k/setup
- **Question:** Is Gittins good?
- **Result:** gap $\leq \ell_{(a)} + \ell_{(b)} + \ell_{(c)}$
- **Corollary:** heavy-traffic optimal if $\mathbb{E}[S^2 \log S] < \infty$
- **Key tool:** new G/G work-decomposition law
- **Takeaway:** Gittins applicable far beyond M/G/1



Conclusion

- **Problem:** scheduling in G/G/k/setup
- **Question:** Is Gittins good?
- **Result:** gap $\leq \ell_{(a)} + \ell_{(b)} + \ell_{(c)}$
- **Corollary:** heavy-traffic optimal if $\mathbb{E}[S^2 \log S] < \infty$
- **Key tool:** new G/G work-decomposition law
- **Takeaway:** Gittins applicable far beyond M/G/1



number in G/G/k/setup

goal

number in G/G/1

WINE

WINE

work in G/G/k/setup

work in G/G/1

G/G work decomposition law

Thank you!