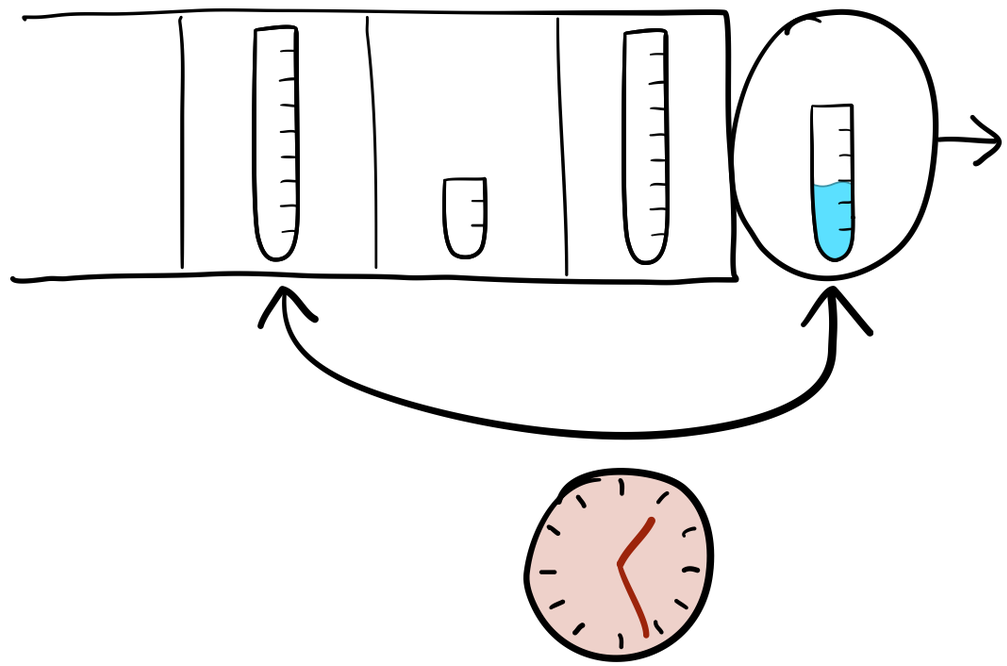
Queueing with

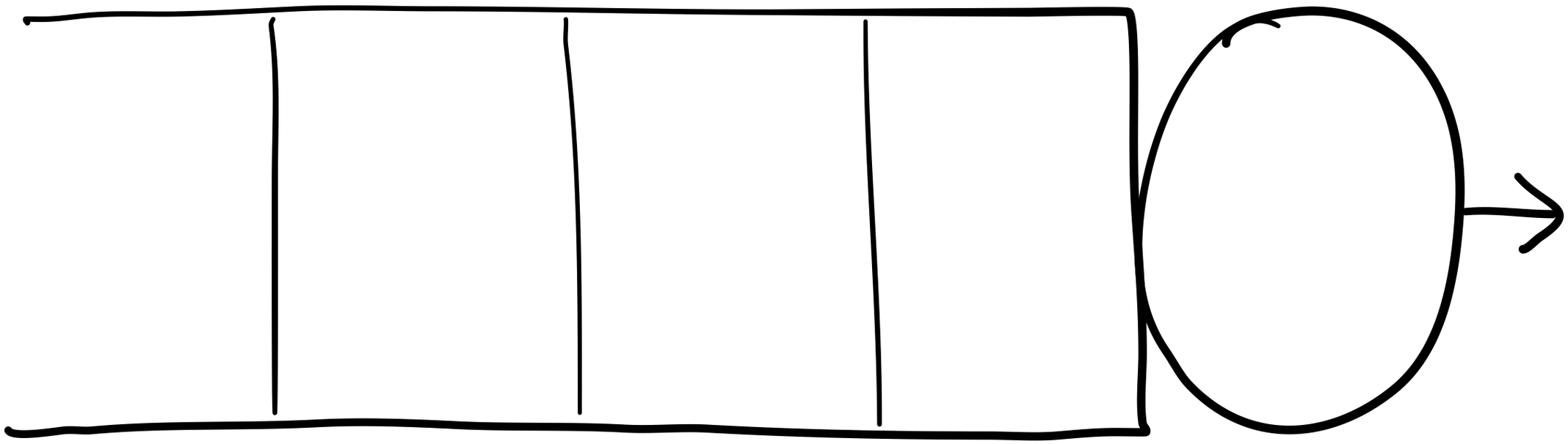
Preemption Costs

Ziv Scully

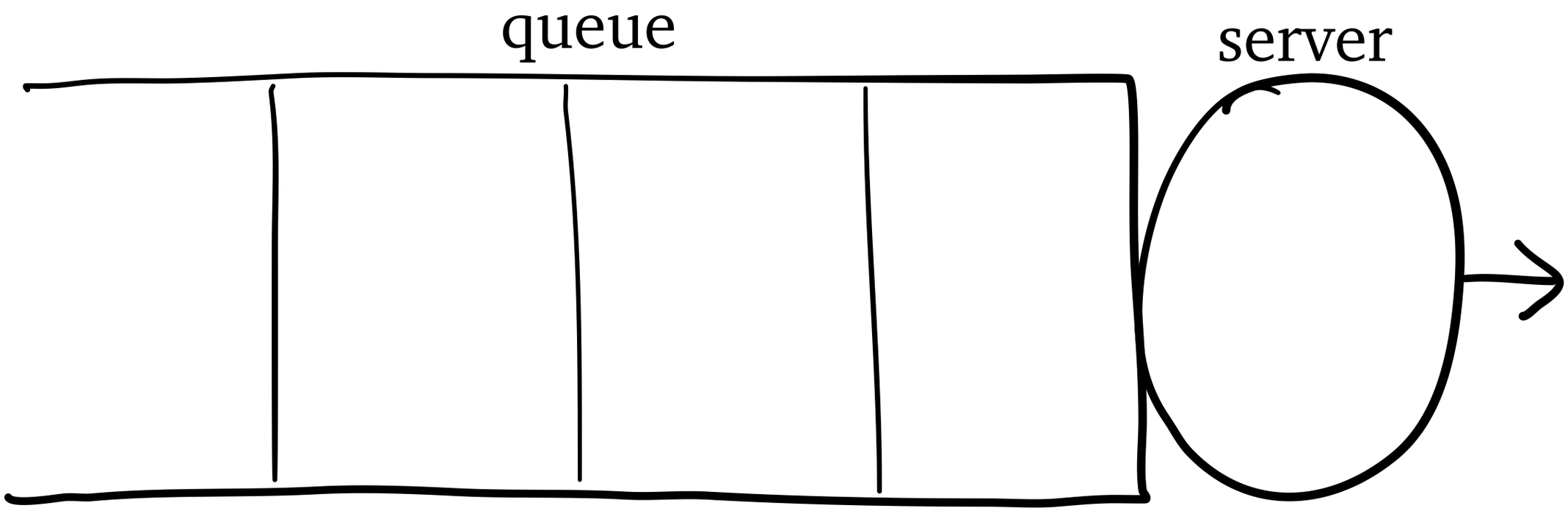
Carnegie Mellon University



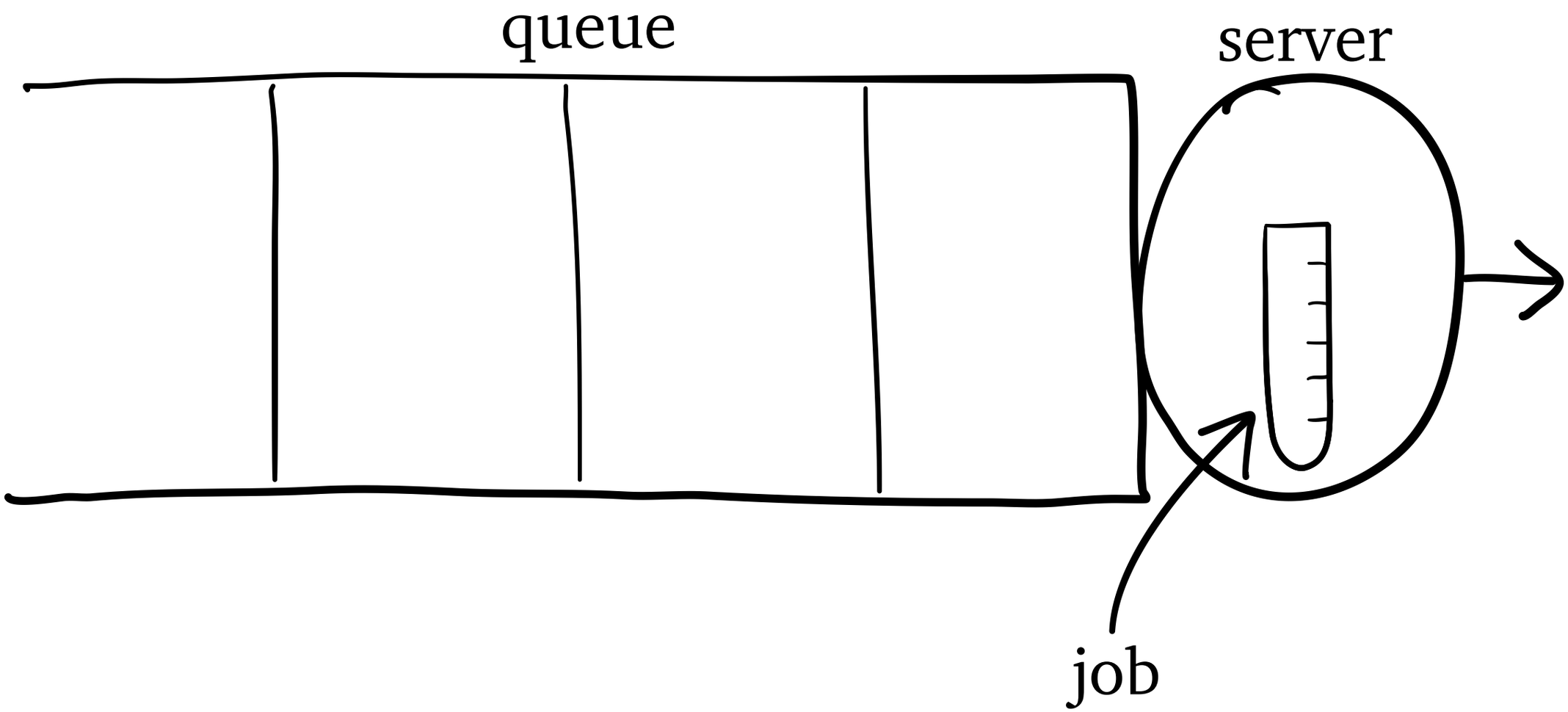
M/G/1 Queue



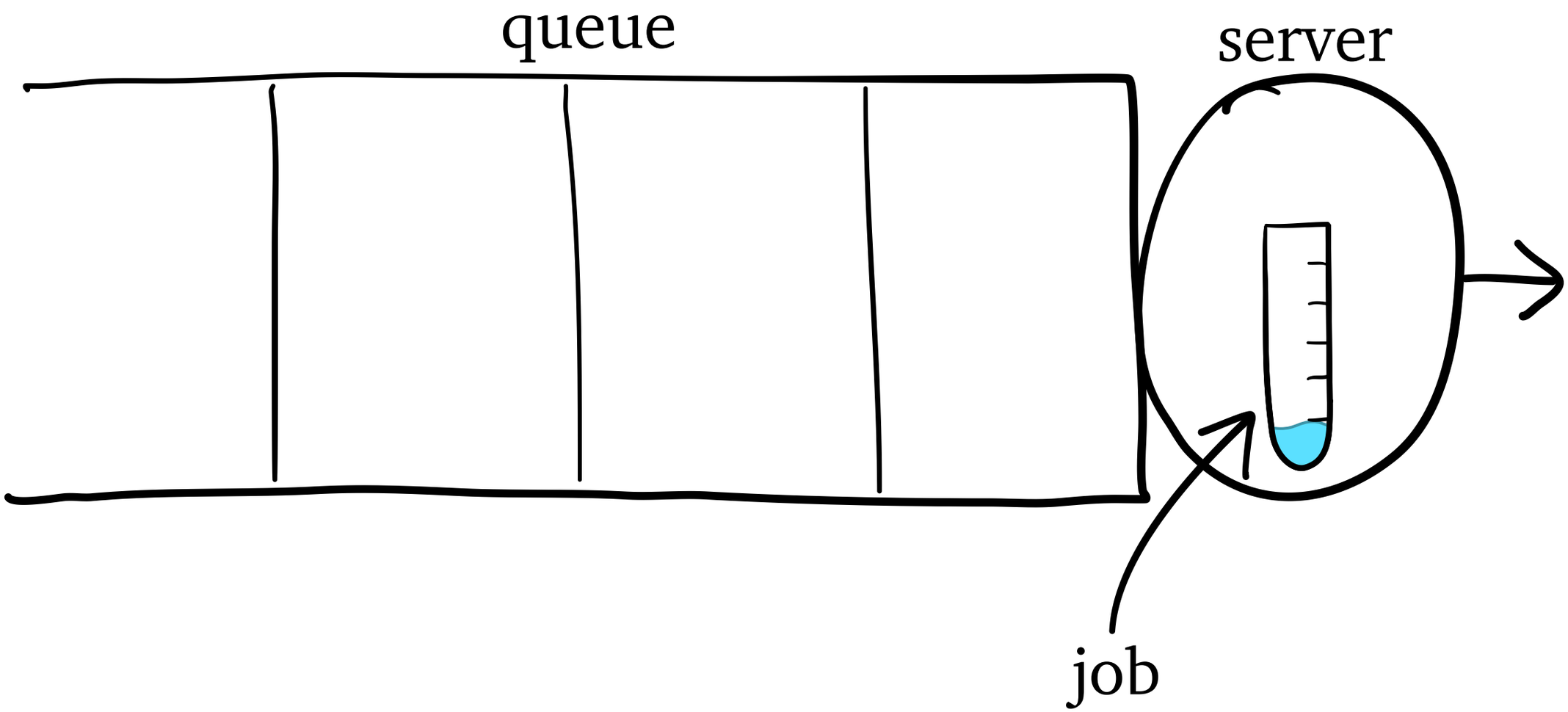
M/G/1 Queue



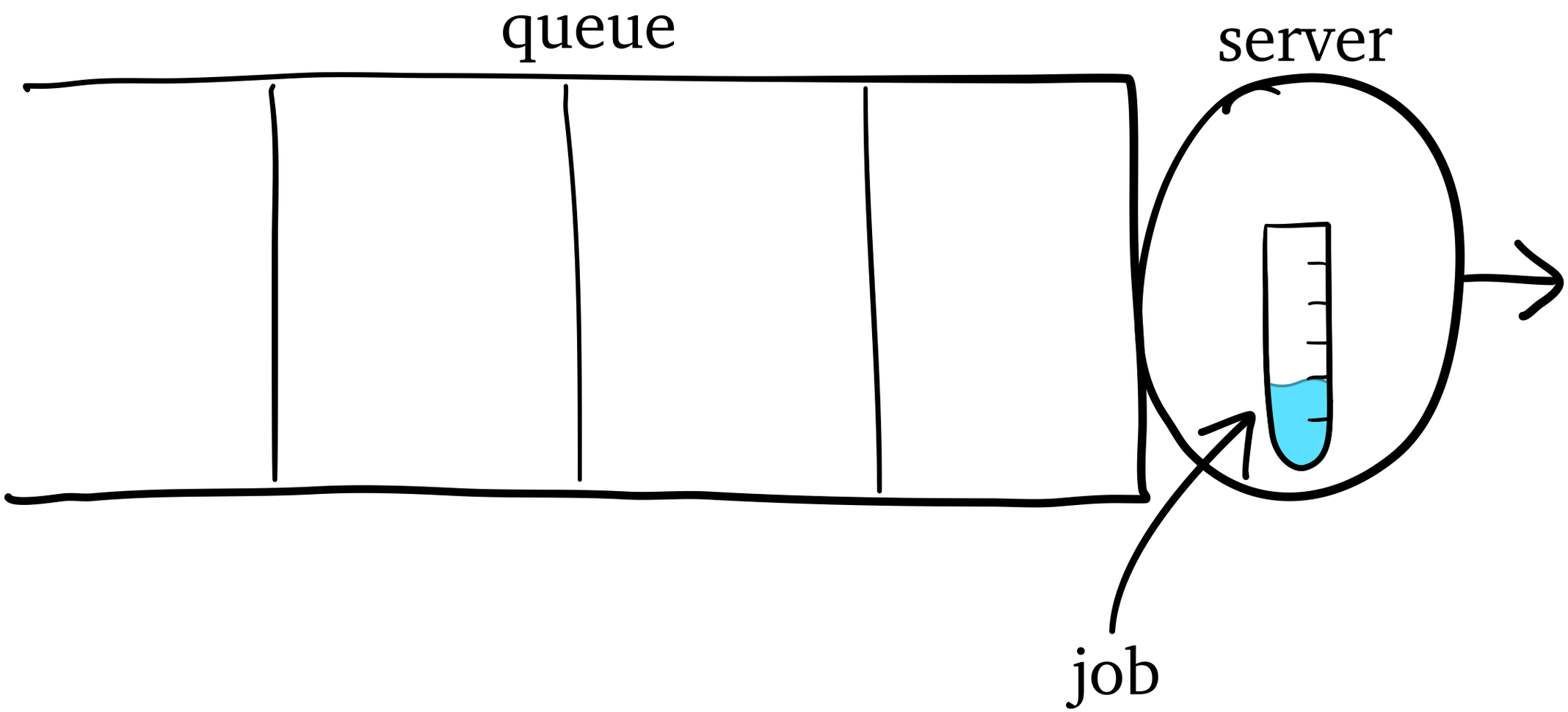
M/G/1 Queue



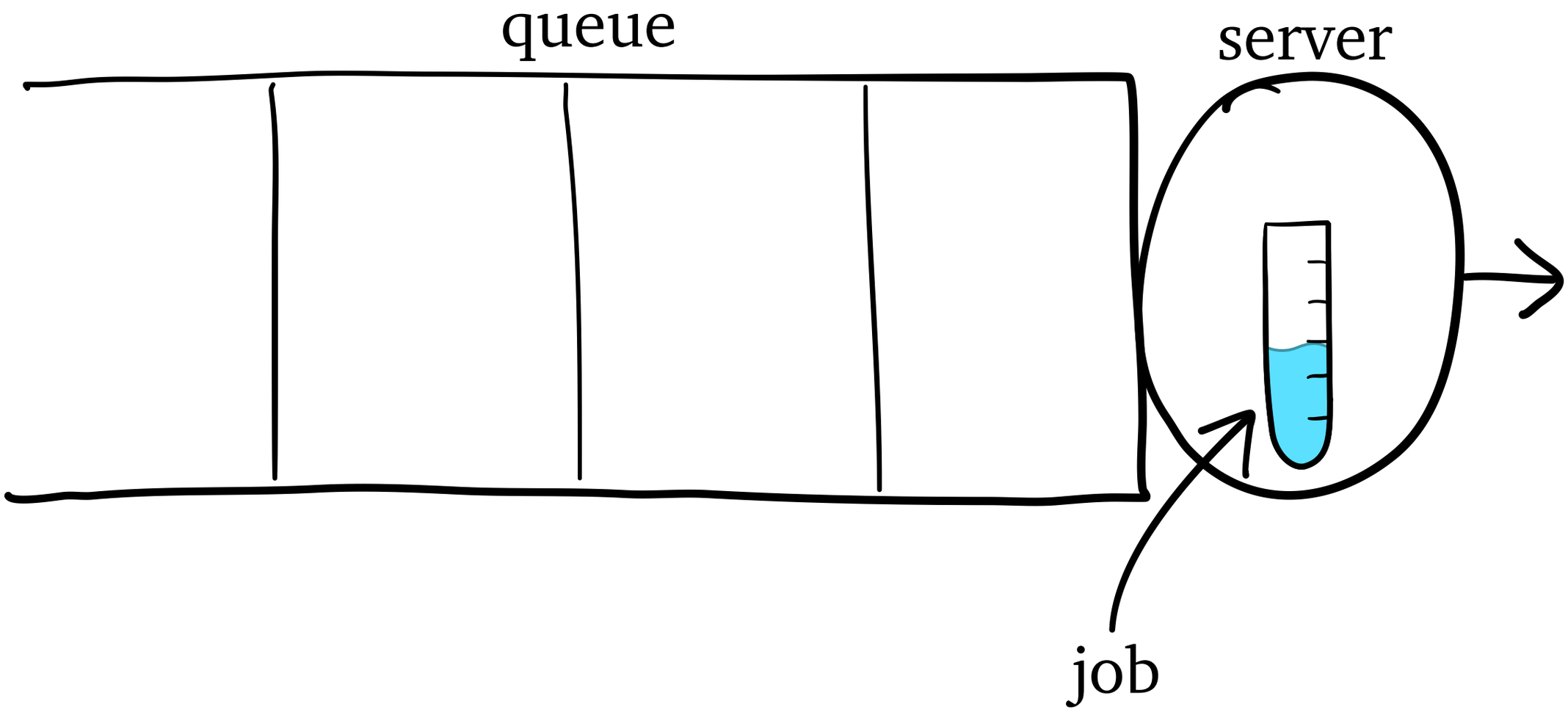
M/G/1 Queue



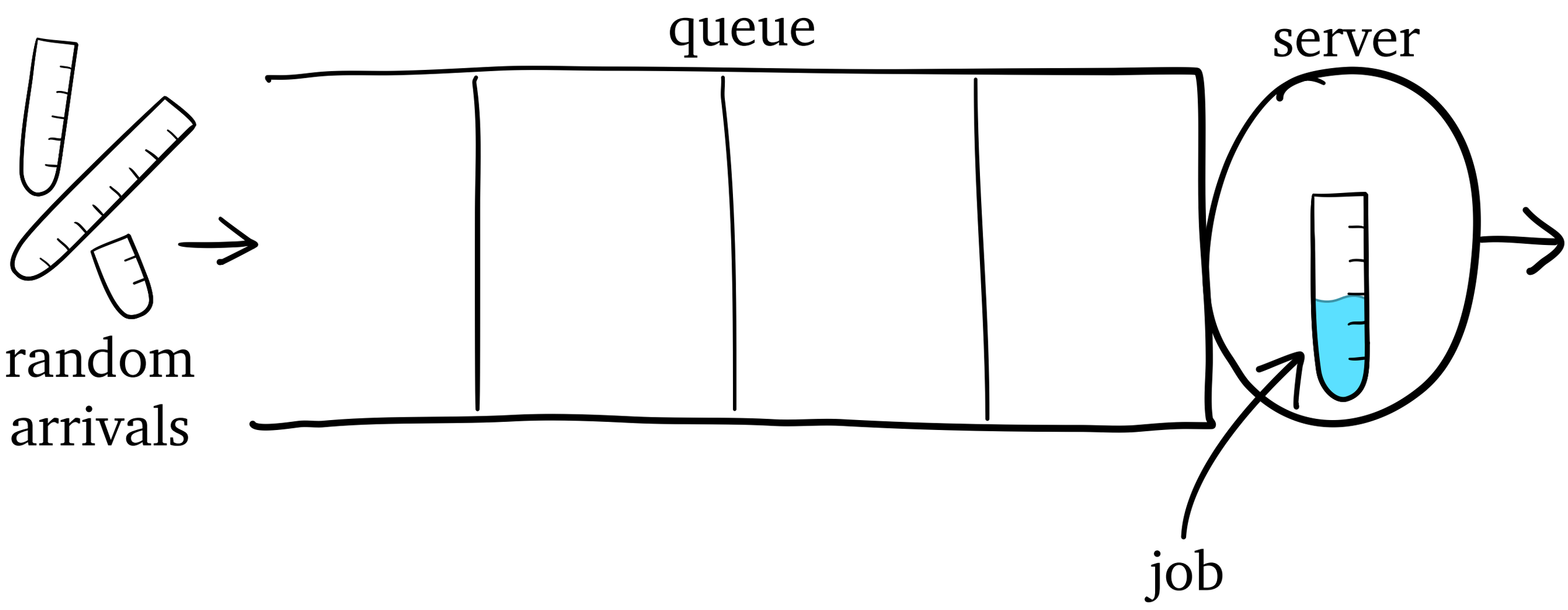
M/G/1 Queue



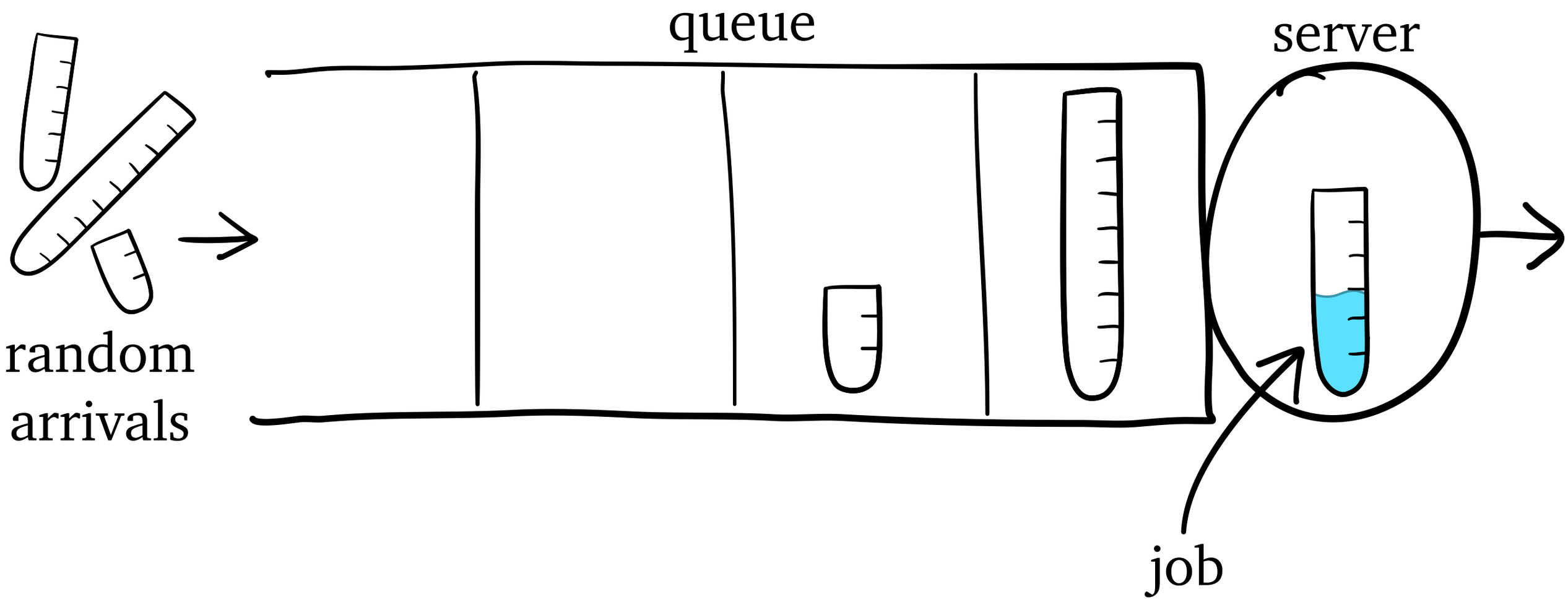
M/G/1 Queue



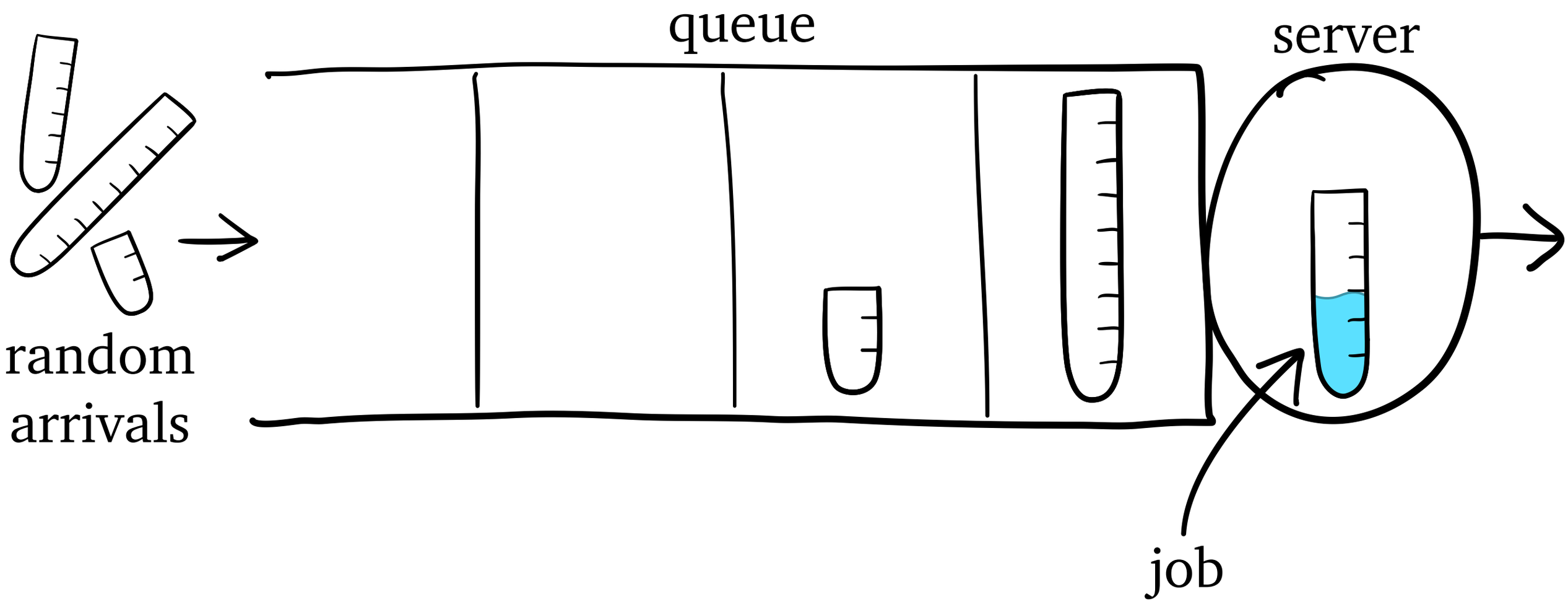
M/G/1 Queue



M/G/1 Queue

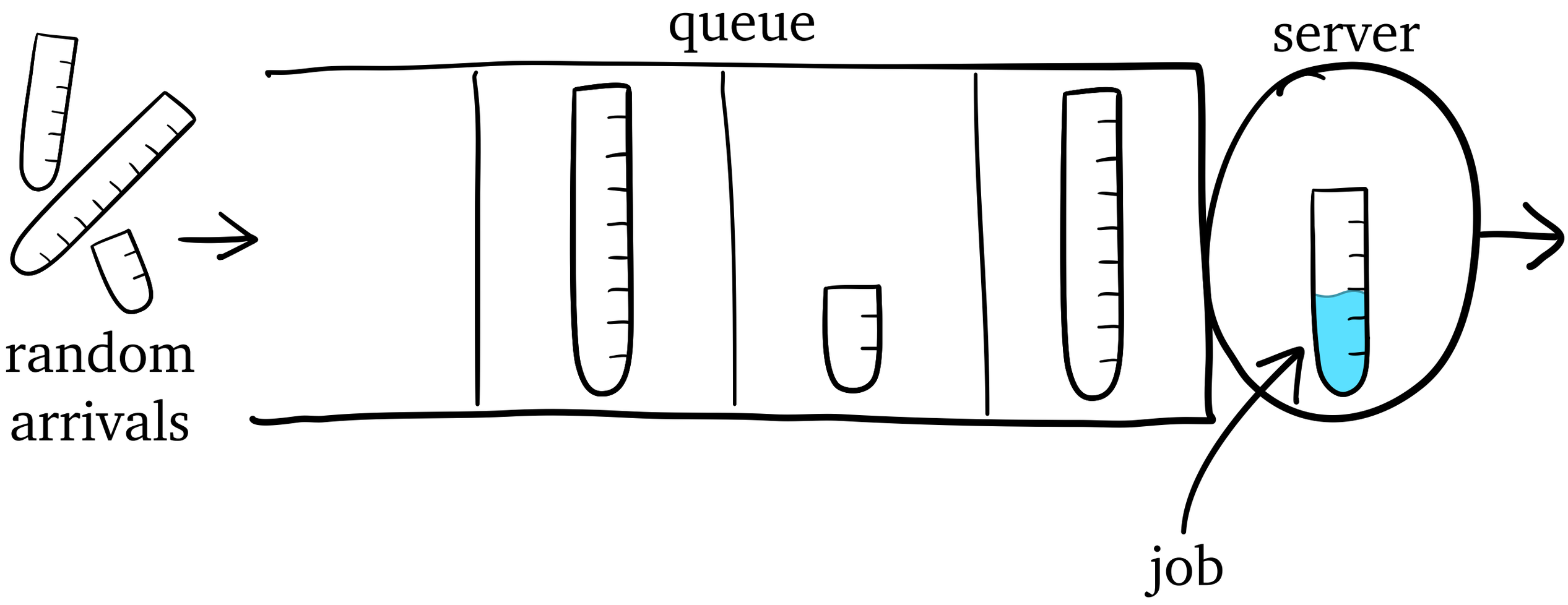


M/G/1 Queue



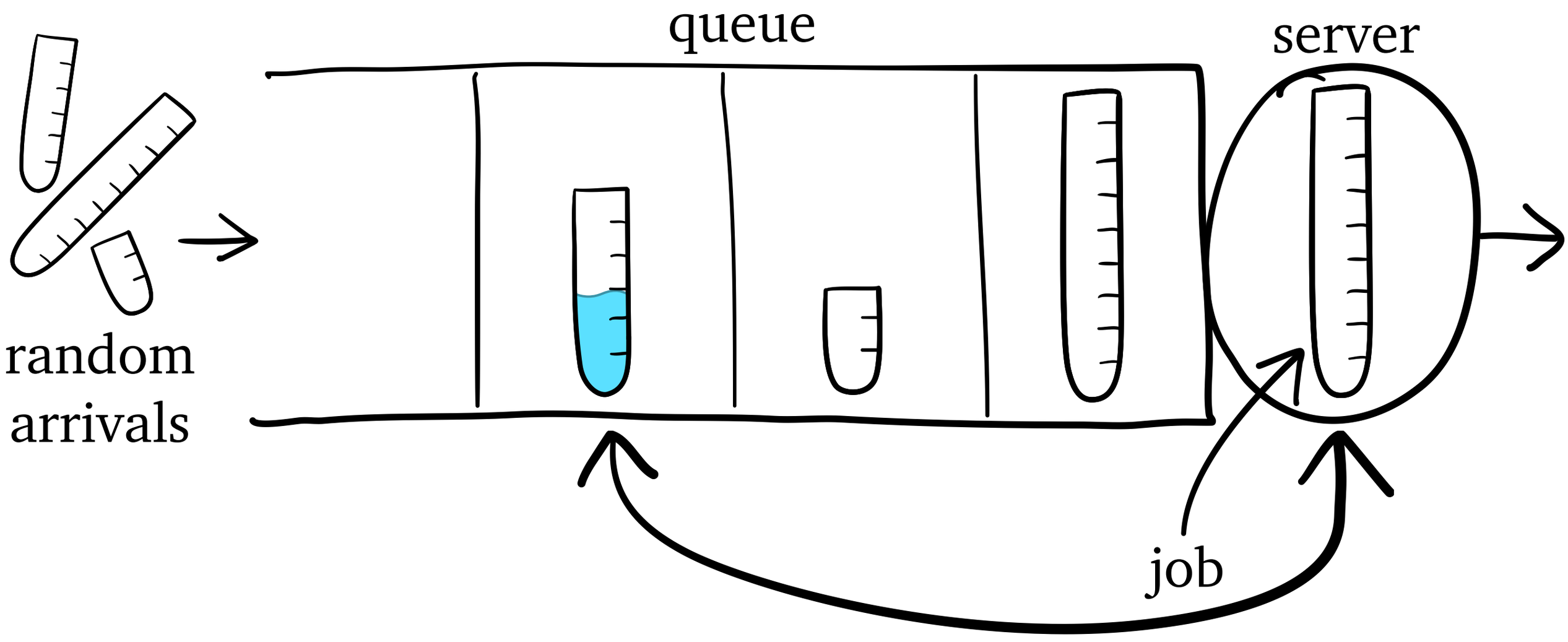
Scheduling policy: picks which job to serve

M/G/1 Queue



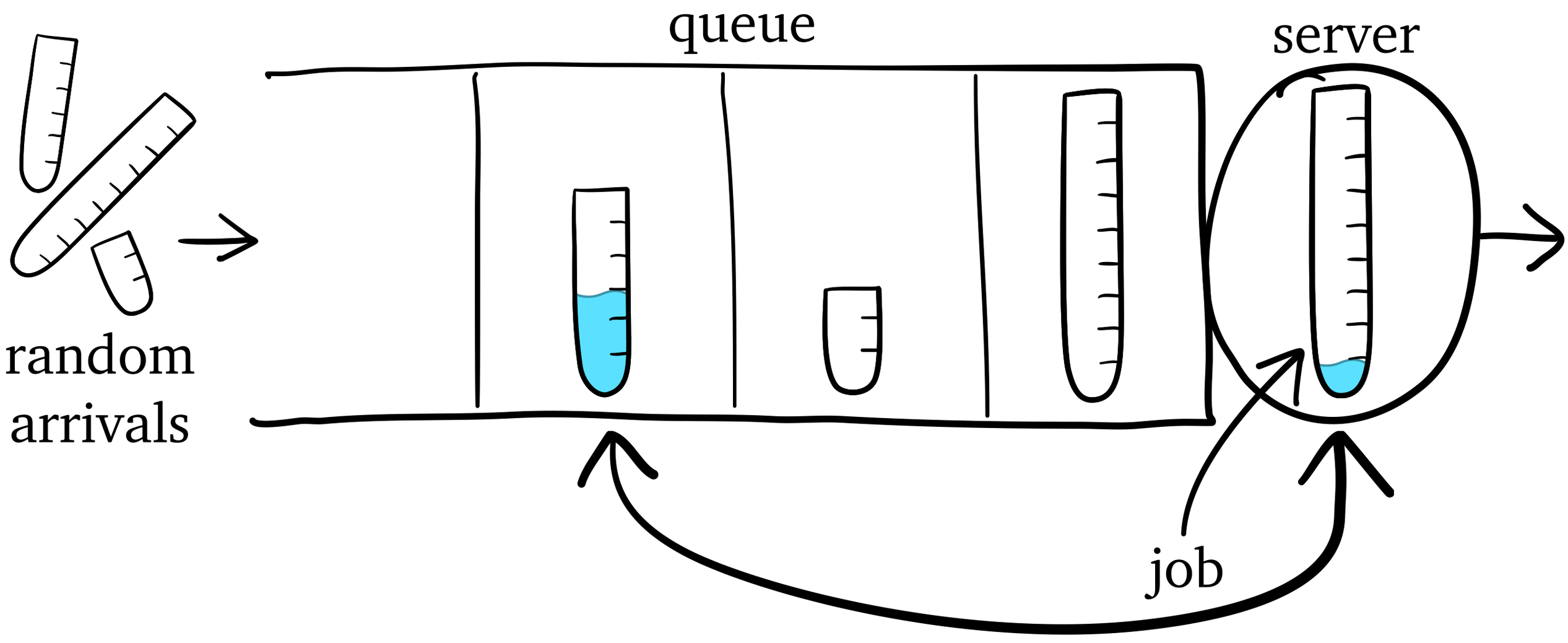
Scheduling policy: picks which job to serve

M/G/1 Queue



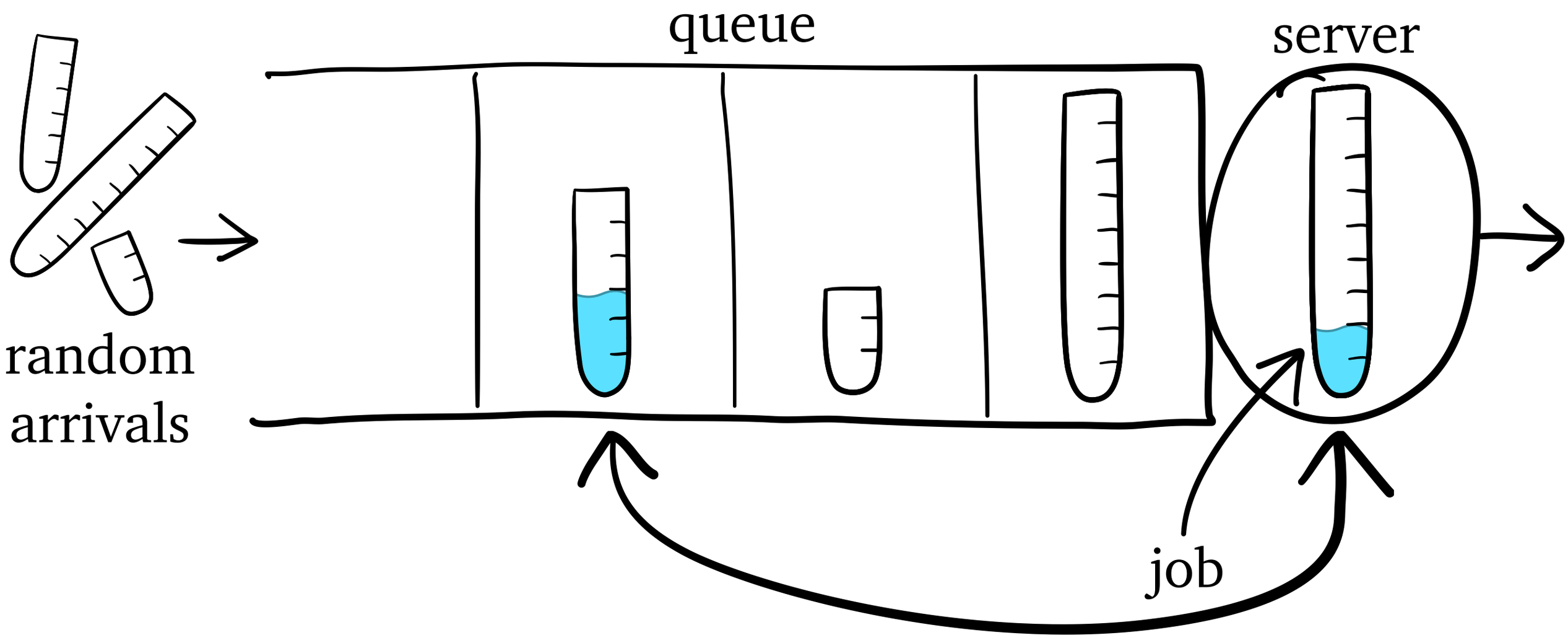
Scheduling policy: picks which job to serve

M/G/1 Queue



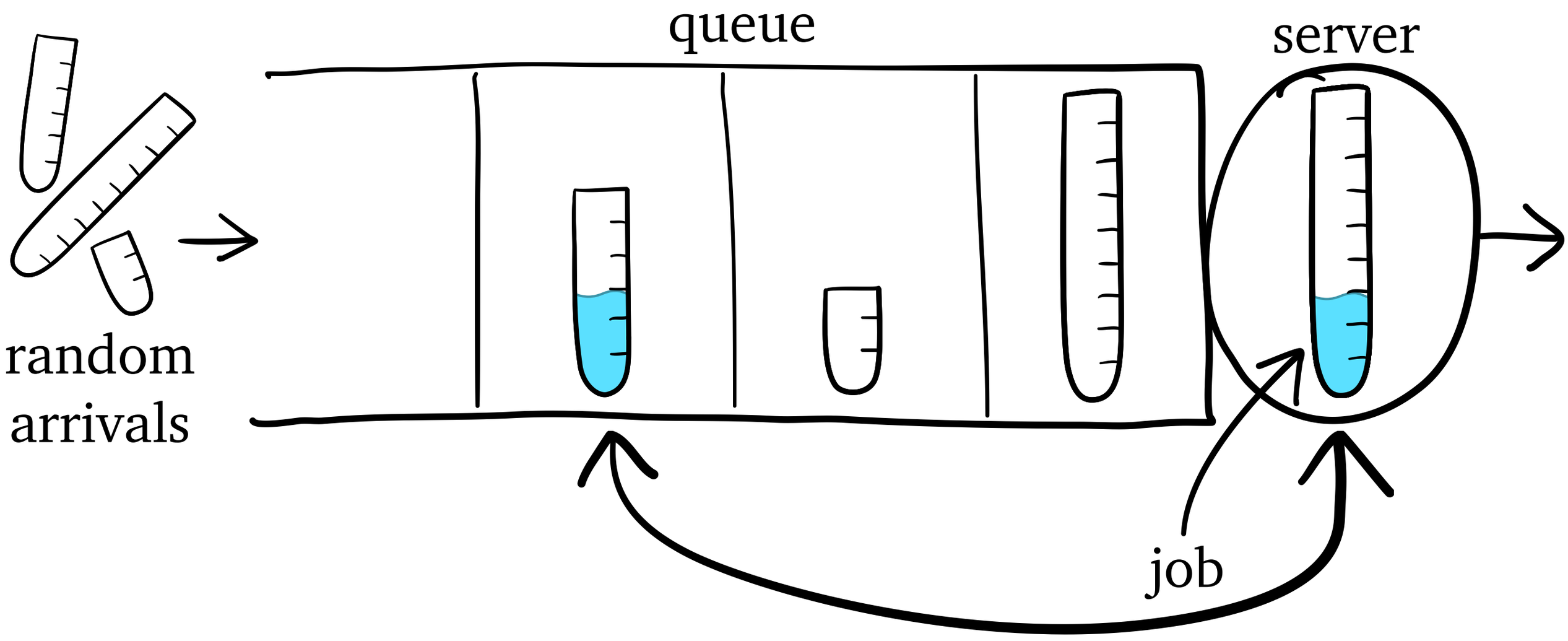
Scheduling policy: picks which job to serve

M/G/1 Queue



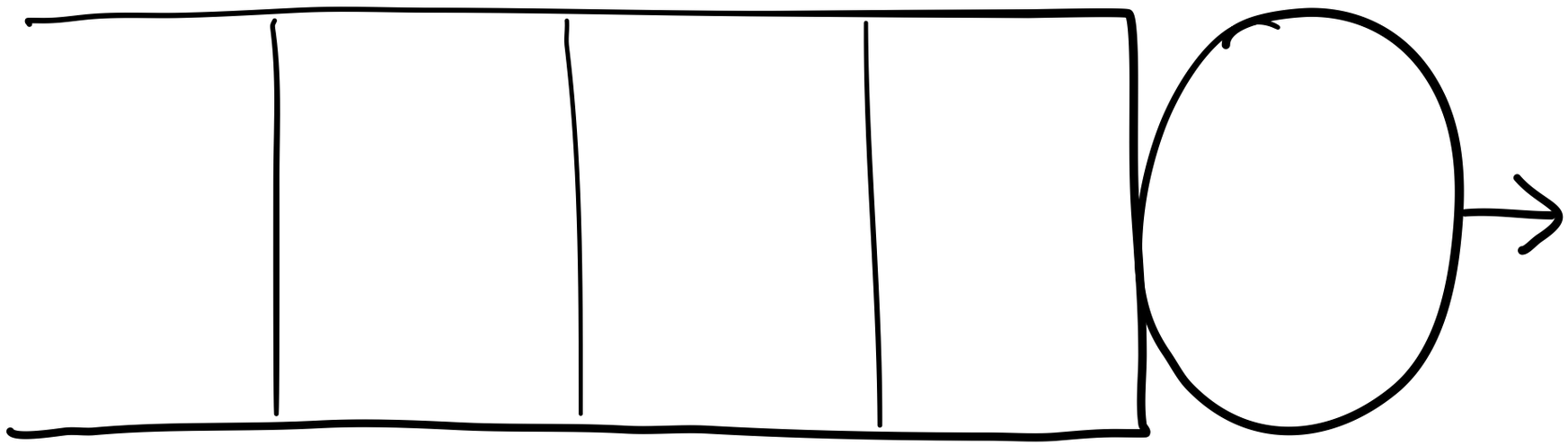
Scheduling policy: picks which job to serve

M/G/1 Queue

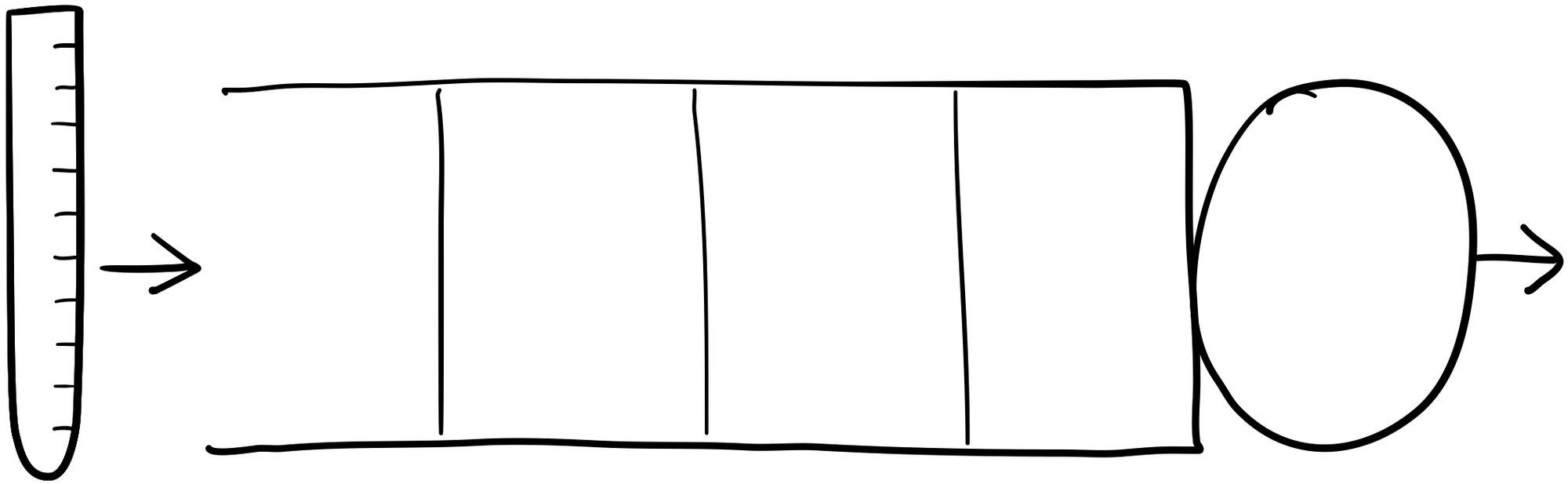


Scheduling policy: picks which job to serve

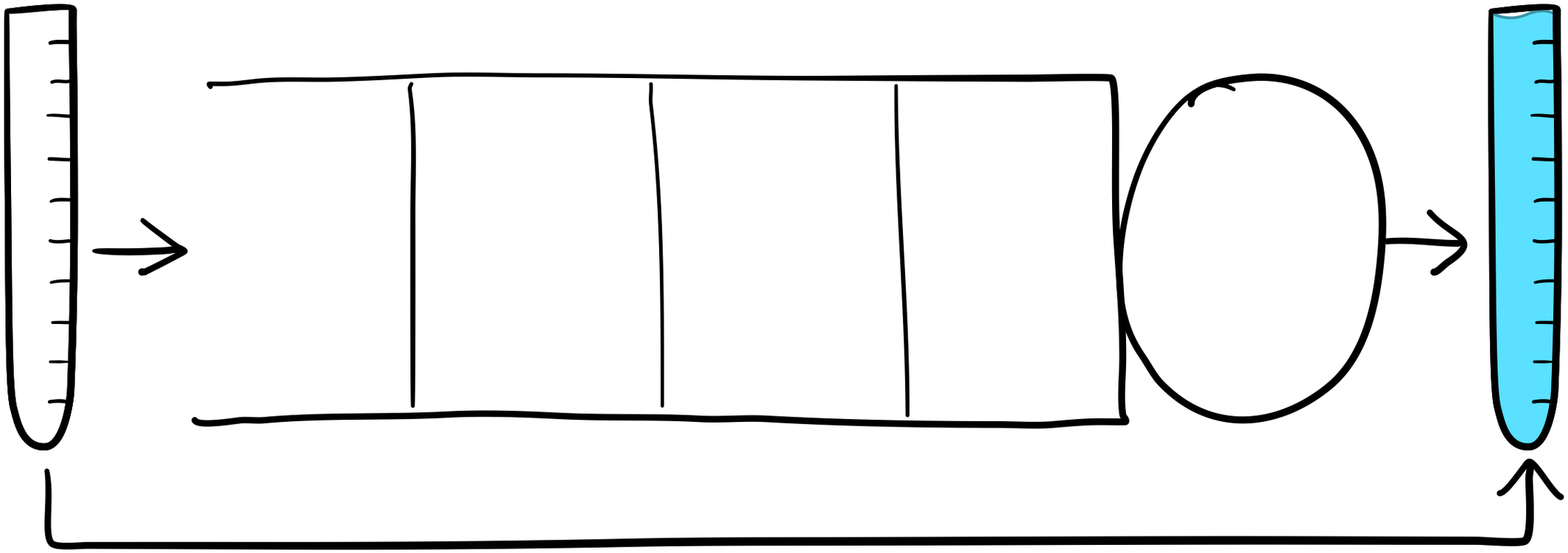
Response Time



Response Time

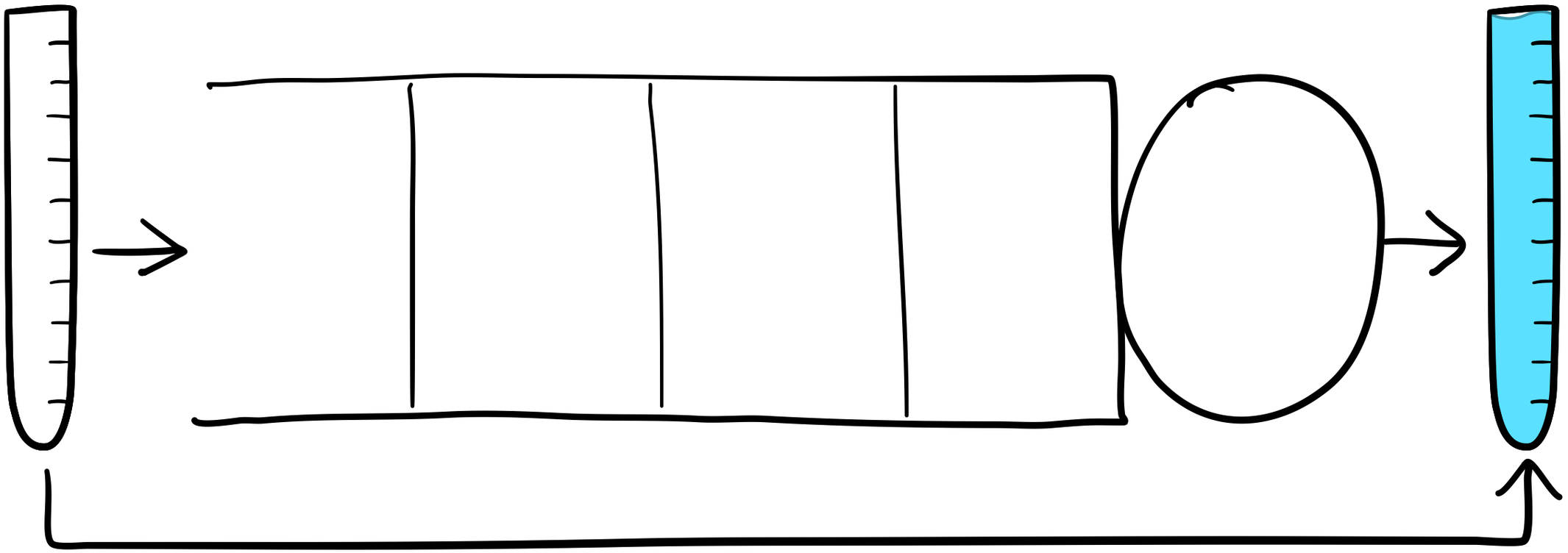


Response Time



$T = response\ time$

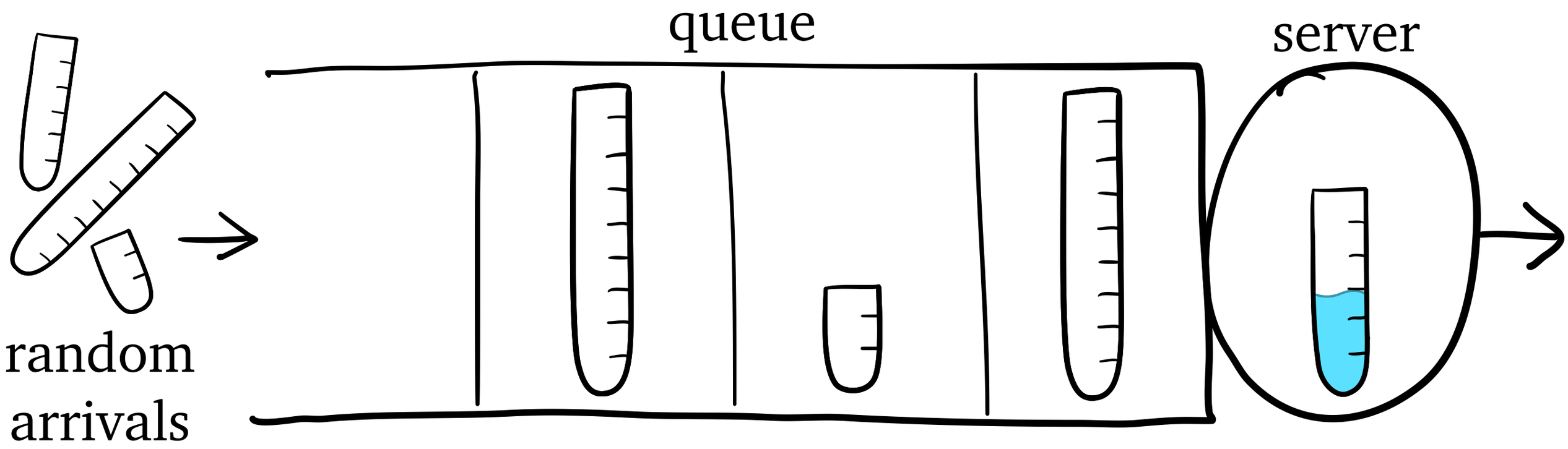
Response Time



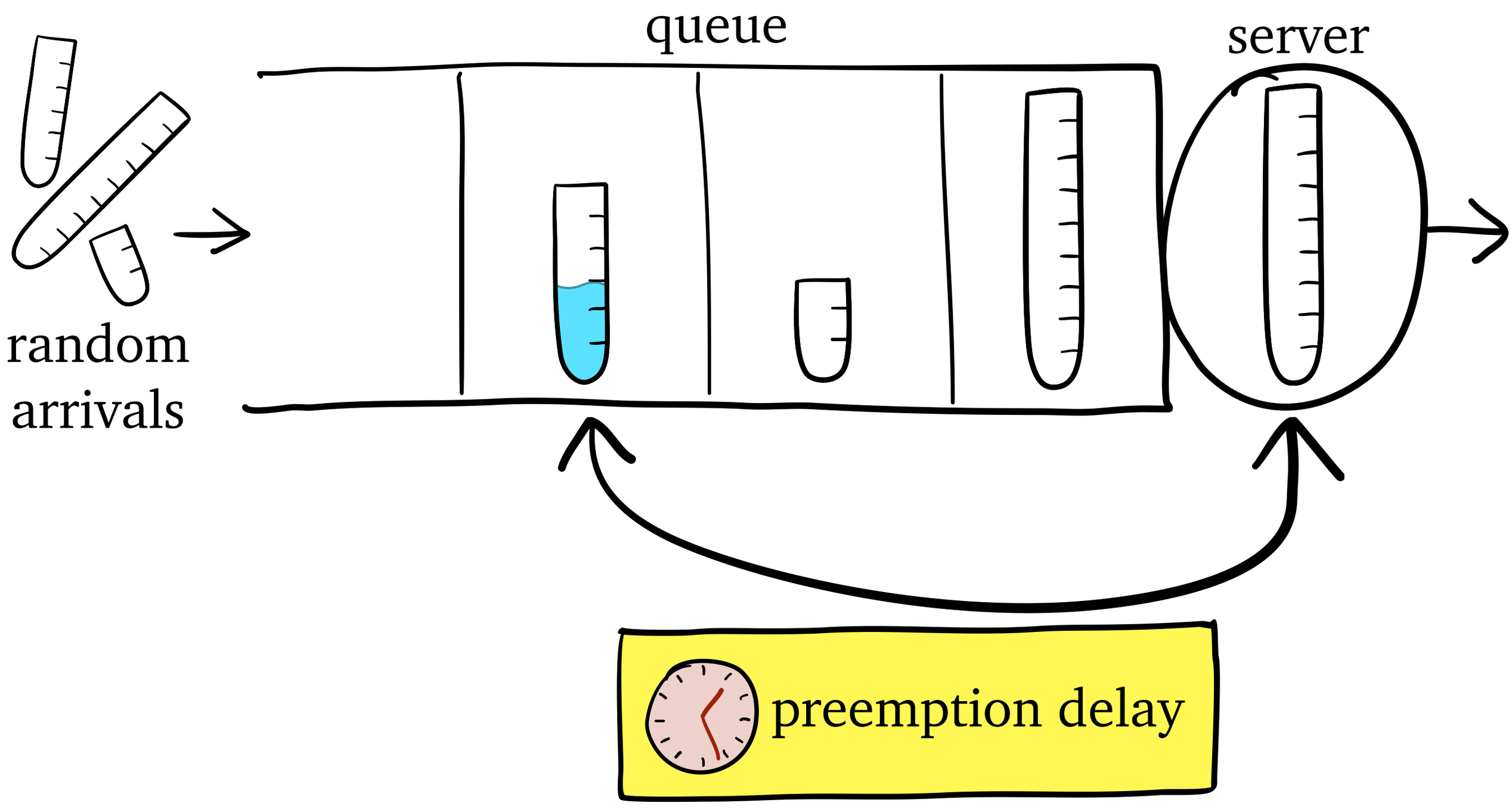
T = *response time*

Goal: analyze and minimize *mean response time* $\mathbf{E}[T]$

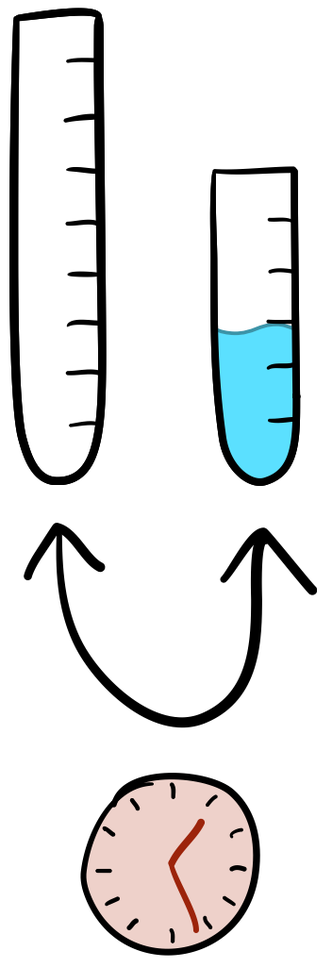
M/G/1 Queue



M/G/1 Queue

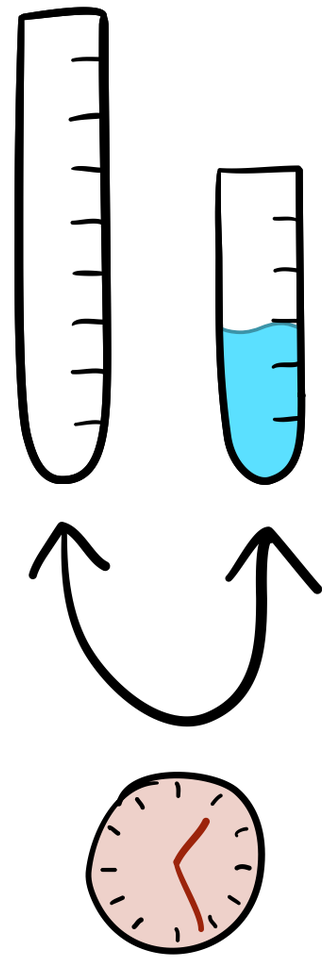


What is $\mathbf{E}[T]$ with Delays?



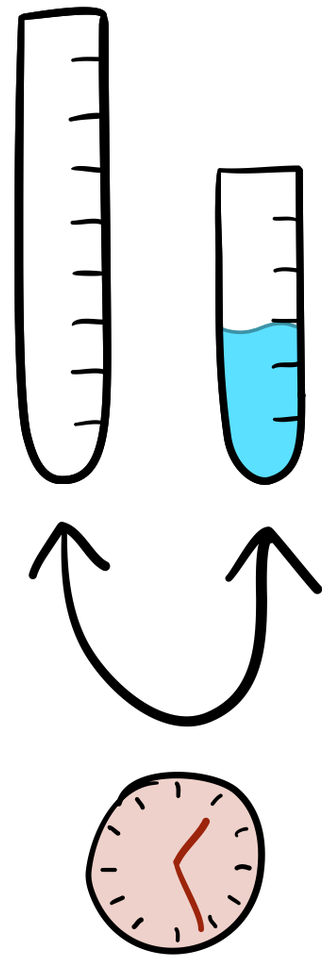
What is $\mathbf{E}[T]$ with Delays?

- Any nonpreemptive policy: add delay to job size



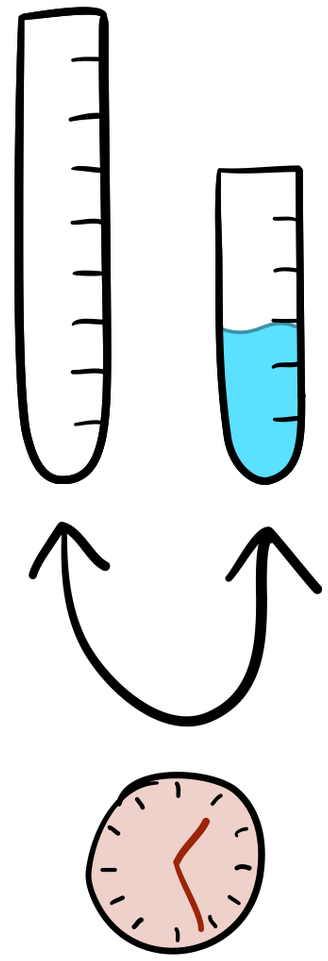
What is $\mathbf{E}[T]$ with Delays?

- Any nonpreemptive policy:
  add delay to job size
- Any preemptive policy:
  even *stability* hard to know

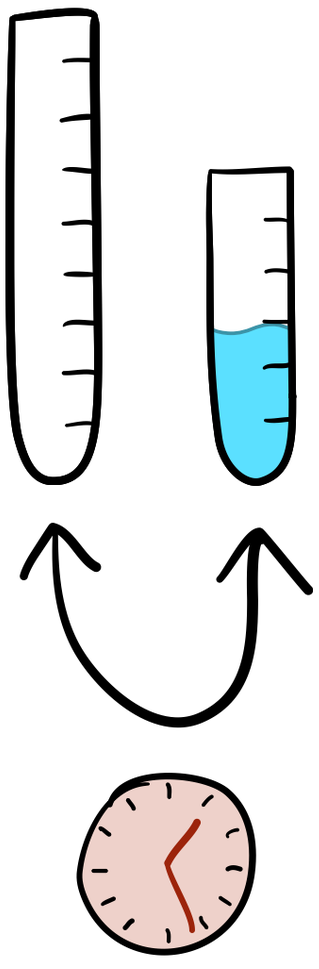


What is $\mathbf{E}[T]$ with Delays?

- Any nonpreemptive policy:
  add delay to job size
- Any preemptive policy:
  even *stability* hard to know
  - Only known analysis: SRPT [Goerg ’86]

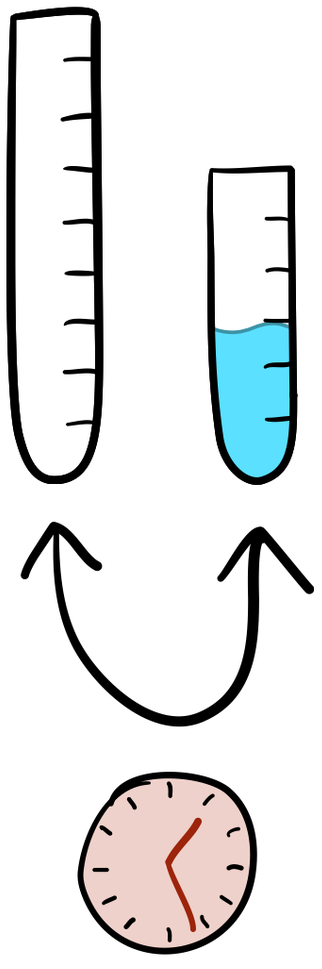


Is There a Stable Policy?



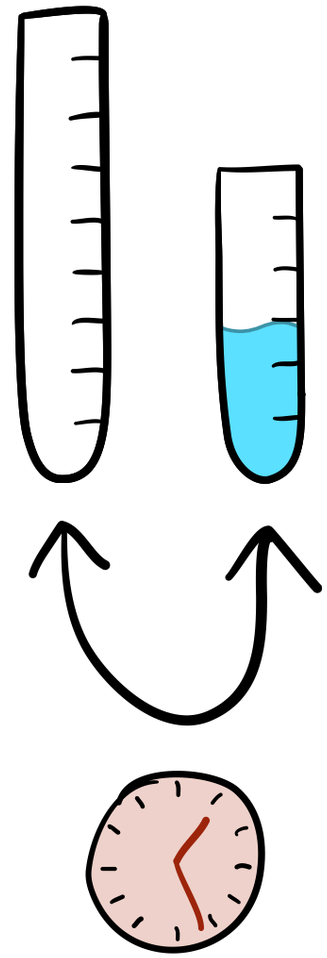
Is There a Stable Policy?

- When load is near 1, *preemptive* policies can raise effective load to over 1



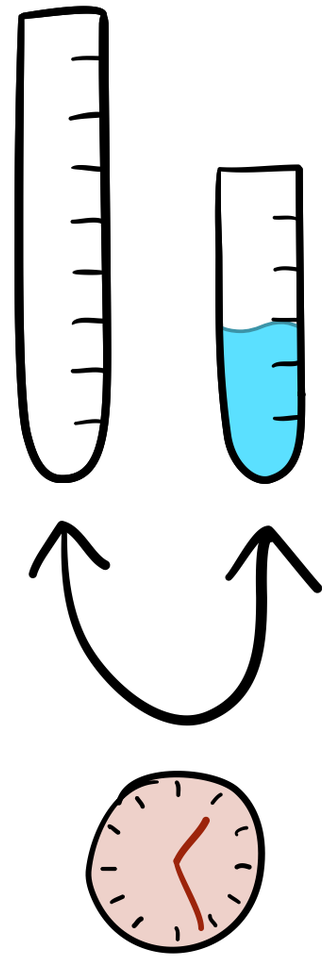
Is There a Stable Policy?

- When load is near 1, *preemptive* policies can raise effective load to over 1
  - *load instability*



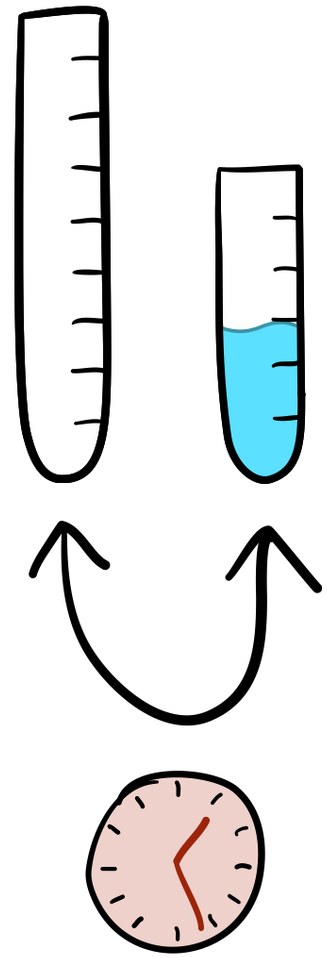
Is There a Stable Policy?

- When load is near 1, *preemptive* policies can raise effective load to over 1
  - *load instability*
- When $\mathbf{E}[X^2]$ is infinite, *nonpreemptive* policies have infinite $\mathbf{E}[T]$



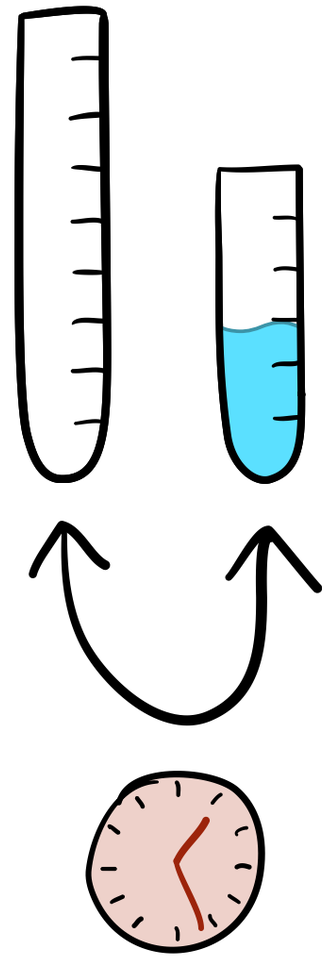
Is There a Stable Policy?

- When load is near 1, *preemptive* policies can raise effective load to over 1
  - *load instability*
- When $\mathbf{E}[X^2]$ is infinite, *nonpreemptive* policies have infinite $\mathbf{E}[T]$
  - *response time instability*



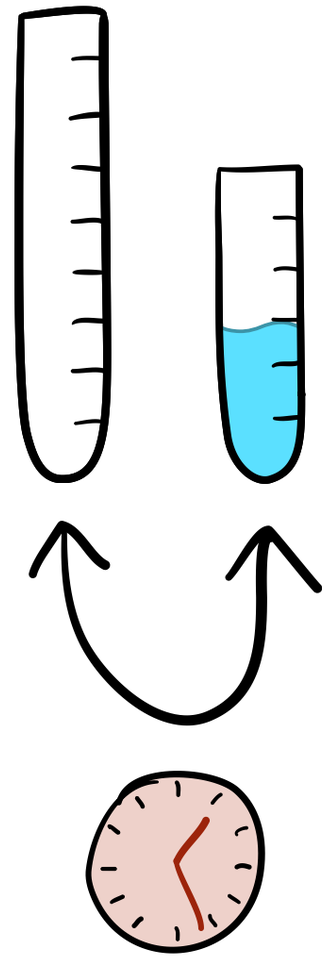
Is There a Stable Policy?

- When load is near 1, *preemptive* policies can raise effective load to over 1
  - *load instability*
- When $\mathbf{E}[X^2]$ is infinite, *nonpreemptive* policies have infinite $\mathbf{E}[T]$
  - *response time instability*
- Is there a policy that has both?

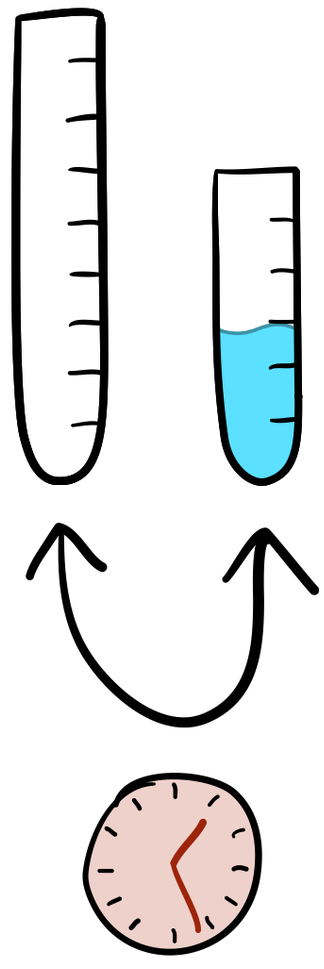


Is There a Stable Policy?

- When load is near 1, *preemptive* policies can raise effective load to over 1
  - *load instability*
- When $\mathbf{E}[X^2]$ is infinite, *nonpreemptive* policies have infinite $\mathbf{E}[T]$
  - *response time instability*
- Is there a policy that has both?
  - Probably a variant of PLCFS?

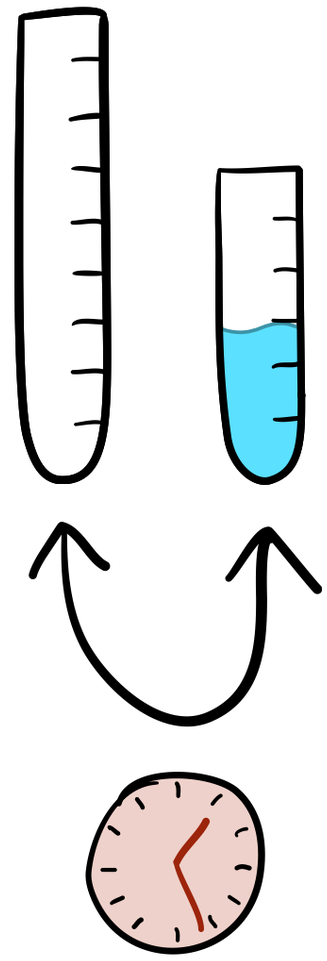


How do we Minimize $\mathbf{E}[T]$?



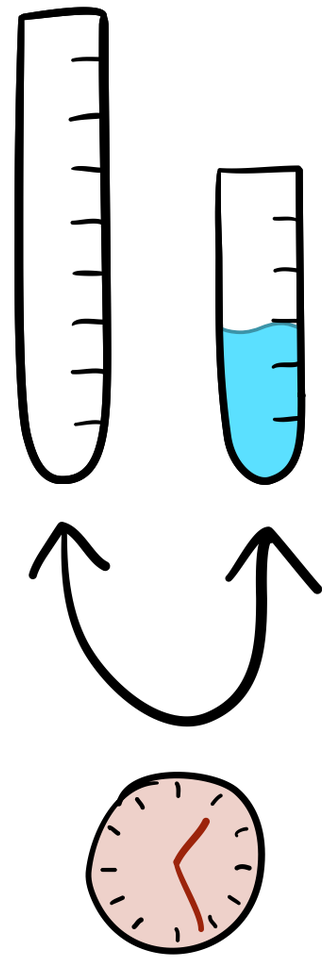
How do we Minimize $\mathbf{E}[T]$?

- Policy for no preemption delay: *Gittins policy*



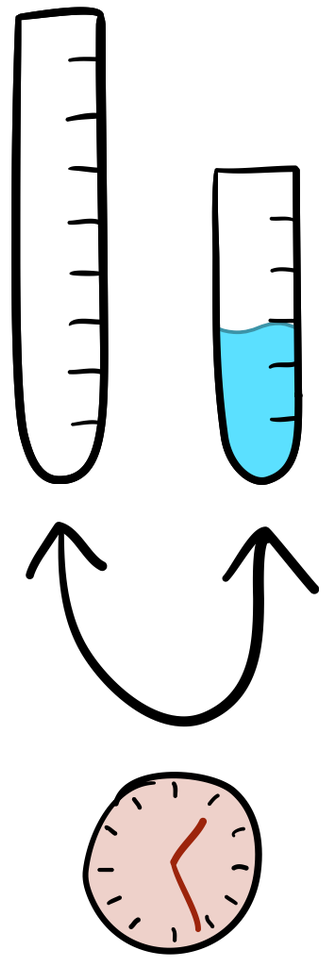
How do we Minimize $\mathbf{E}[T]$?

- Policy for no preemption delay: *Gittins policy*
- Comes from multiarmed bandits



How do we Minimize $\mathbf{E}[T]$?

- Policy for no preemption delay: *Gittins policy*
  - Comes from multiarmed bandits
- Probably related: multiarmed bandits with swicthing costs [Asawa and Teneketzis ’96]



How do we Minimize $\mathbf{E}[T]$?

- Policy for no preemption delay: *Gittins policy*
  - Comes from multiarmed bandits
- Probably related: multiarmed bandits with swicthing costs [Asawa and Teneketzis ’96]
  - … but *cost* and *delay* are different