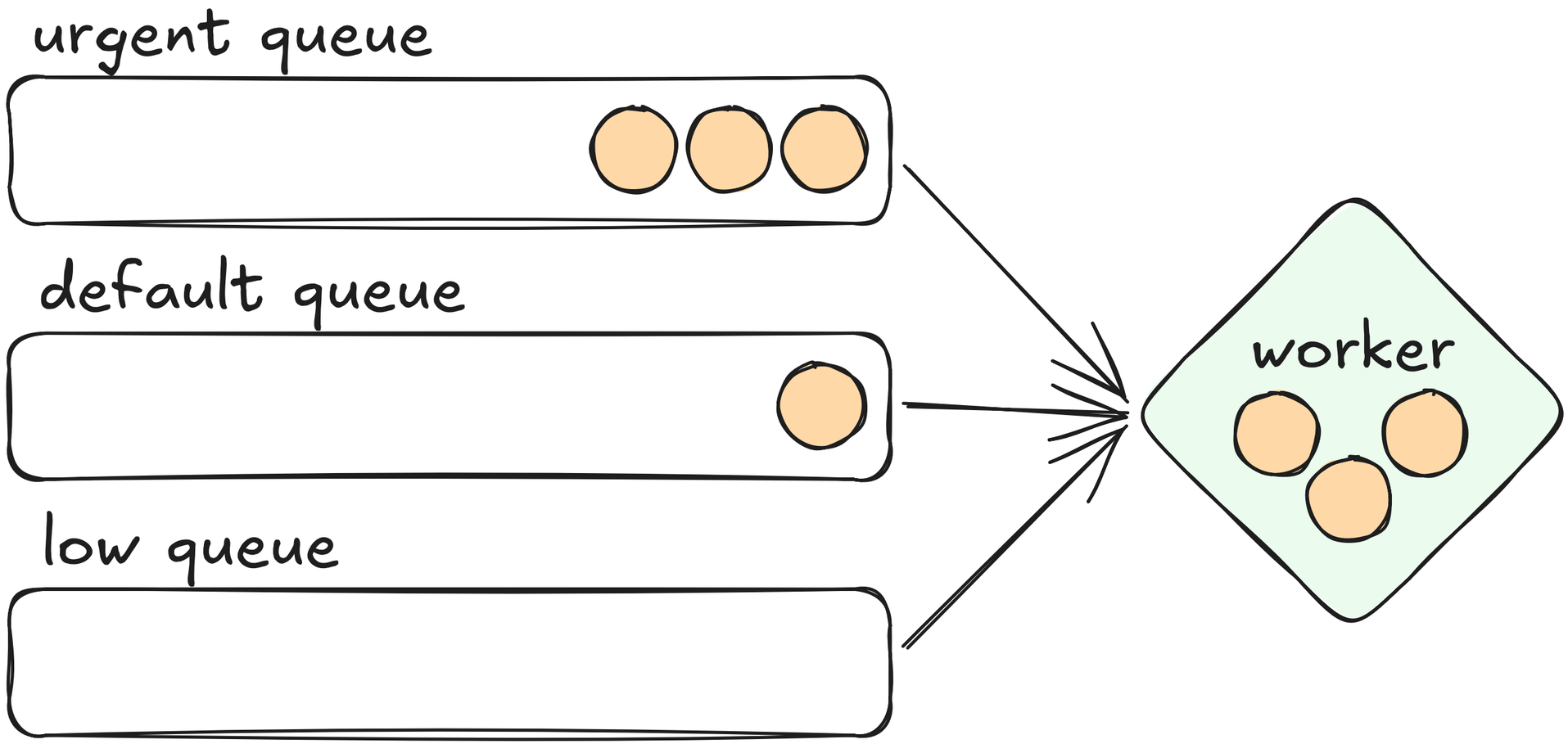
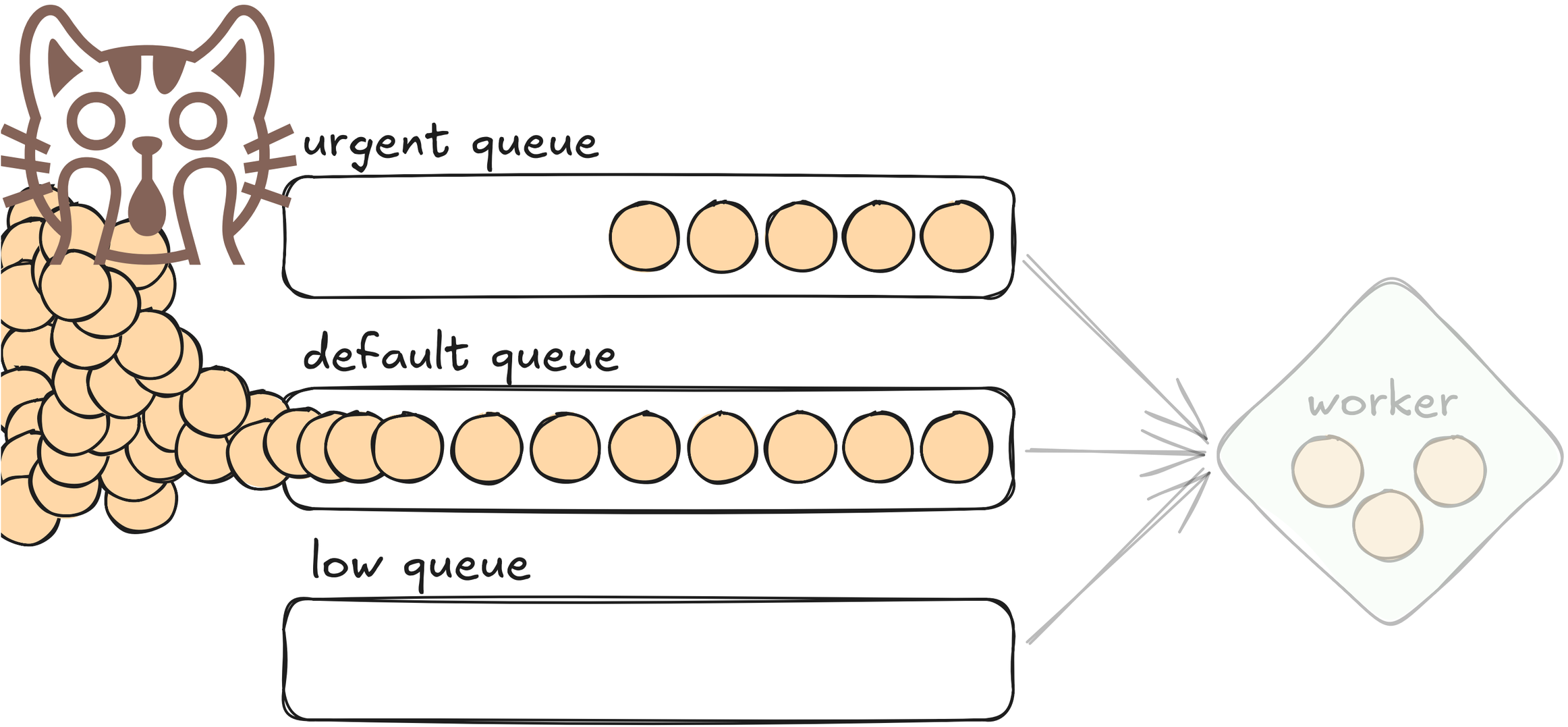
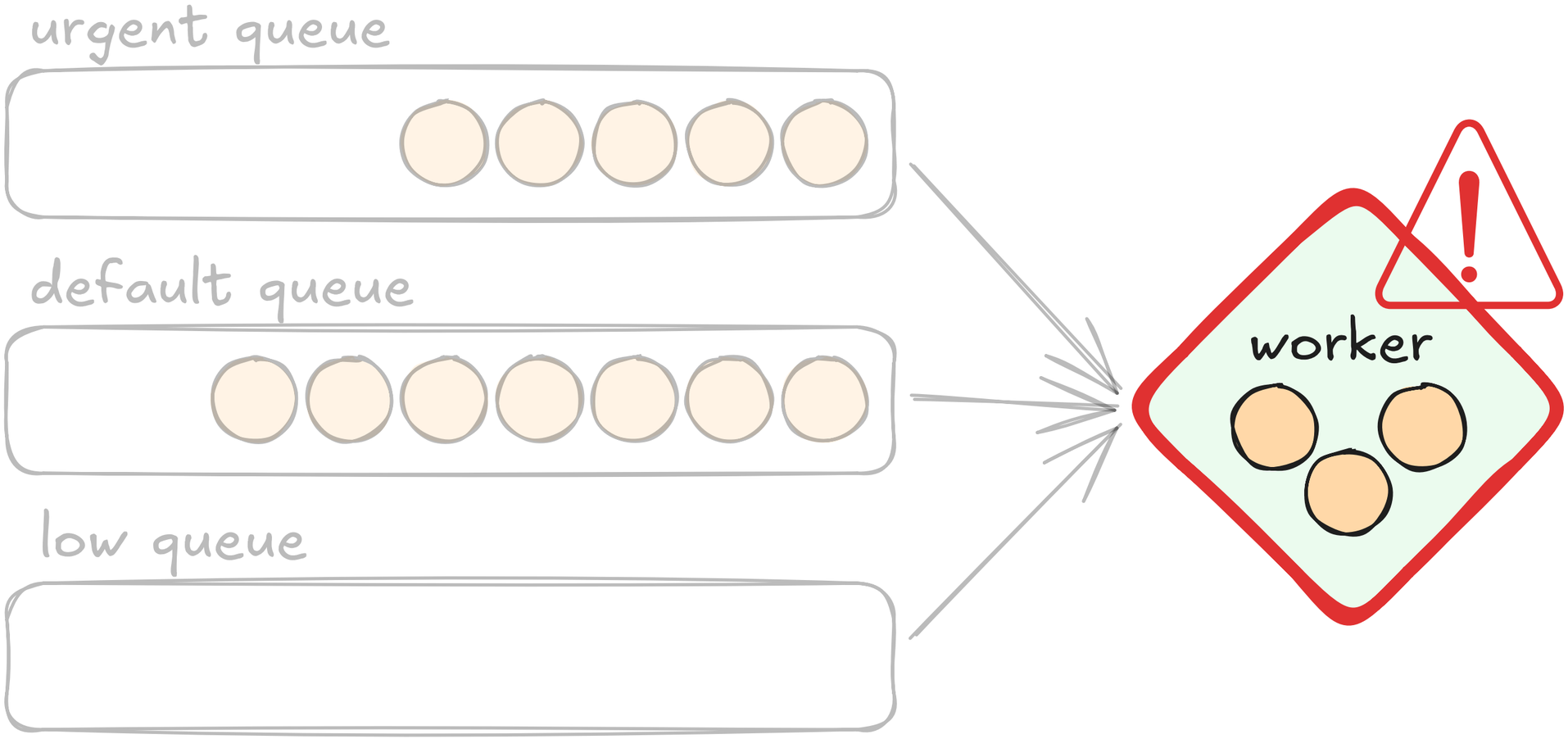
5 Steps to Resilient Job Queues

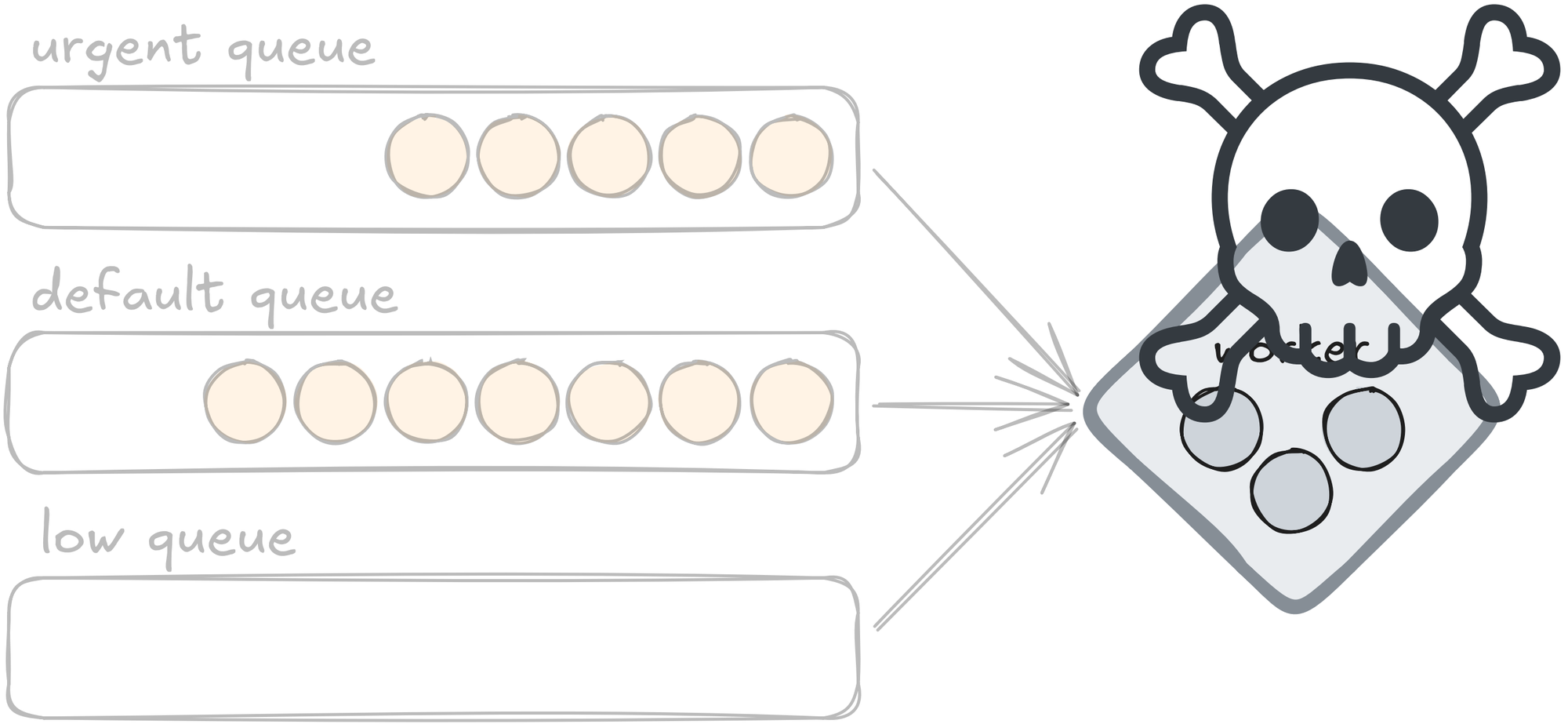
👋 Adam McCrea 👋


urgent queue
default queue
low queue
worker

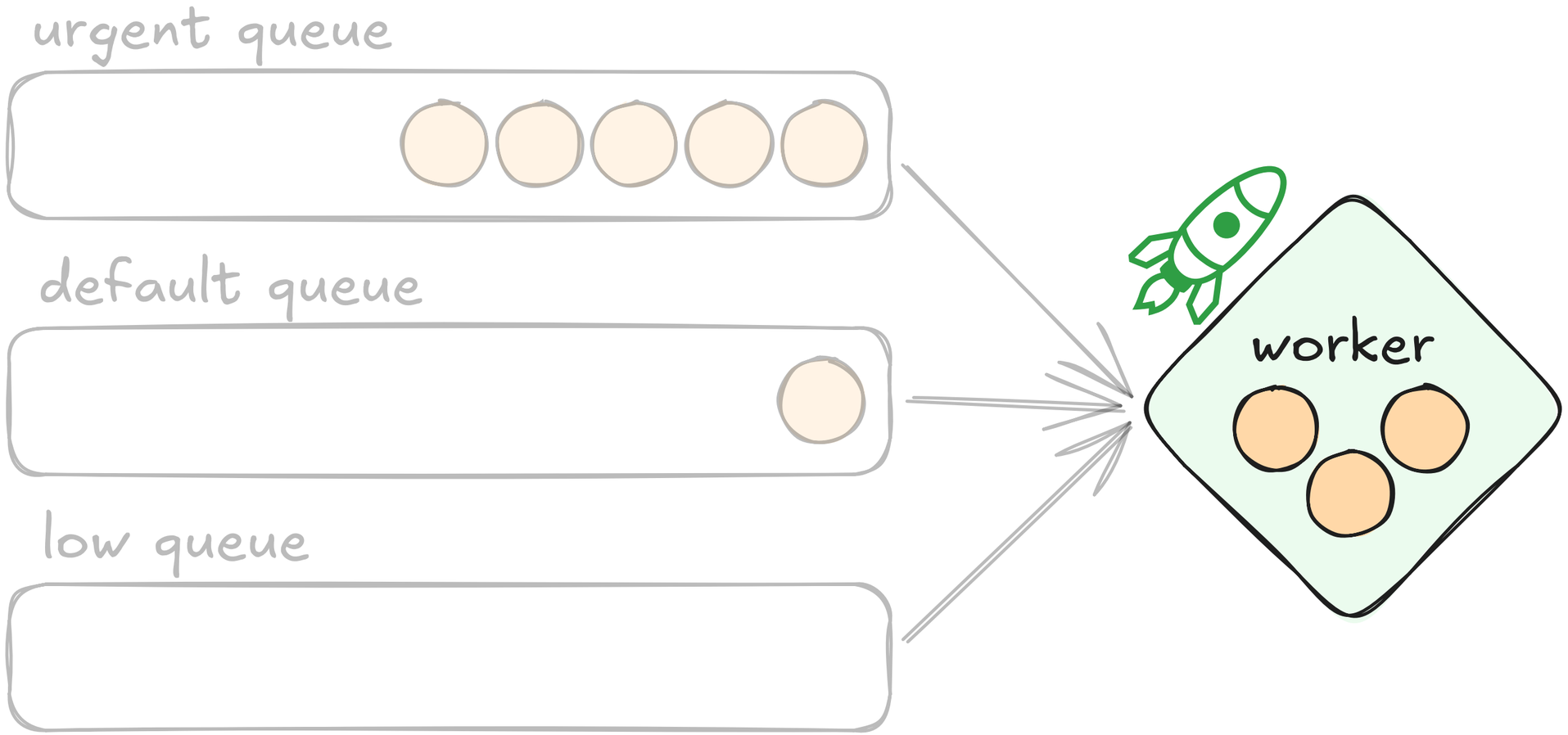
~~Resilient~~ Fragile
job queues...

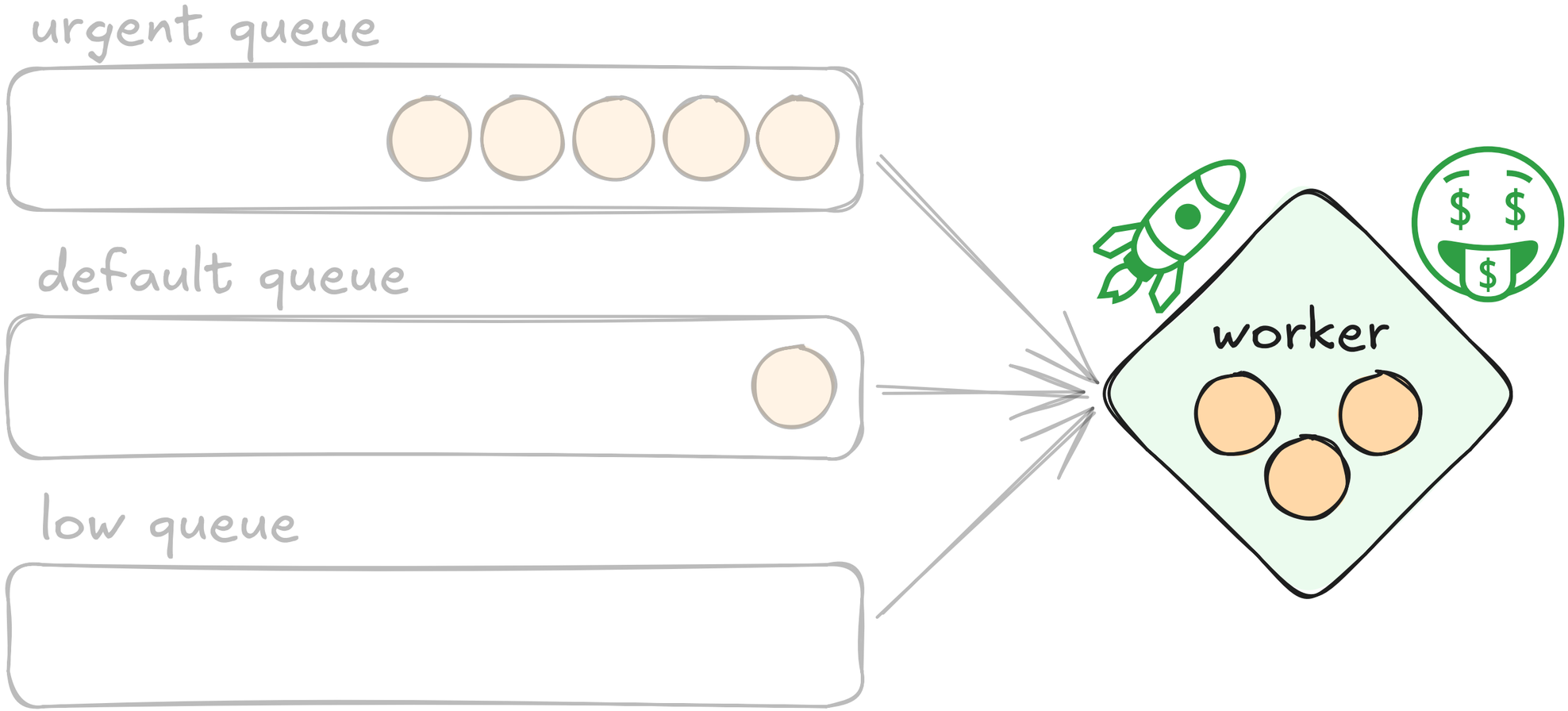

urgent queue
default queue
low queue
worker

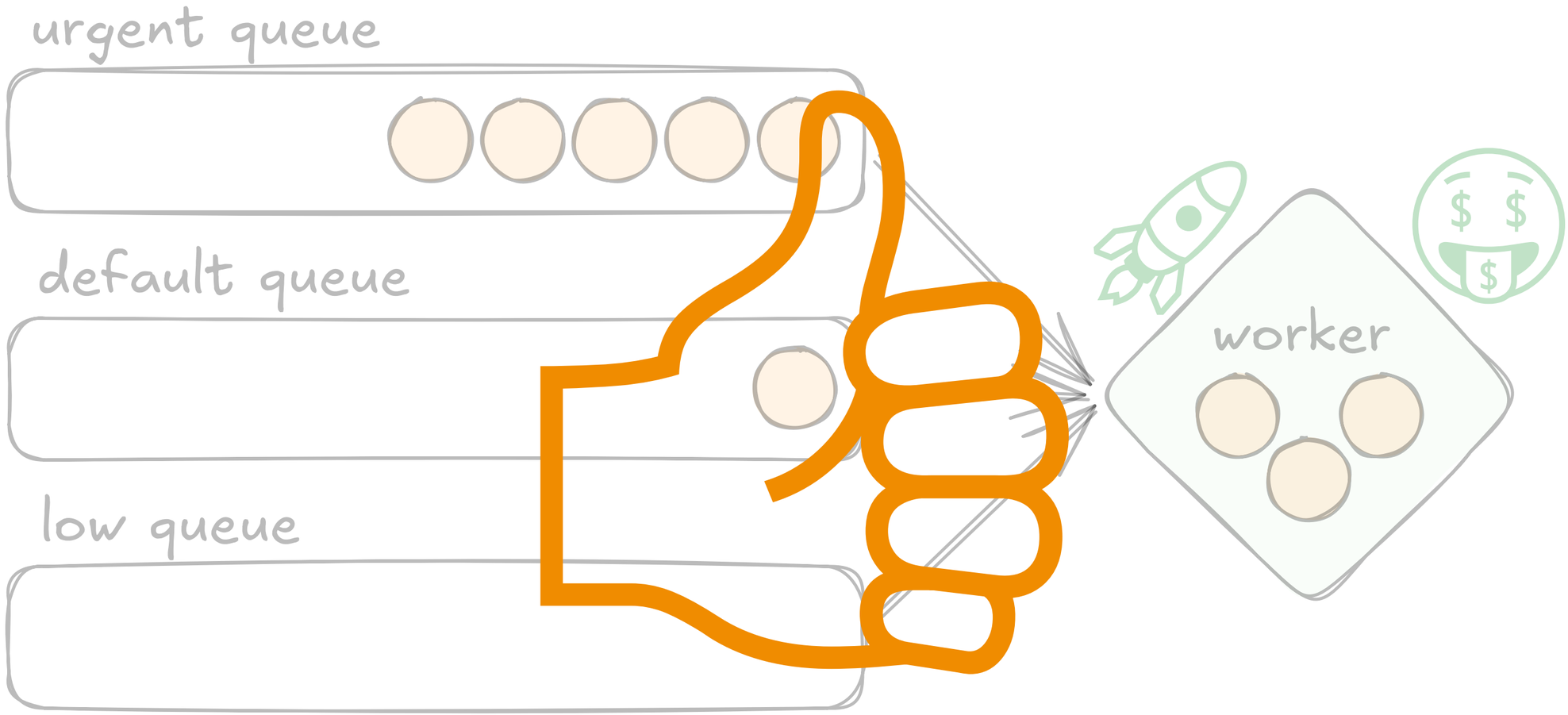

urgent queue
default queue
low queue
worker


urgent queue
default queue
low queue
worker

~~Fragile~~ Resilient
job queues...


urgent queue
default queue
low queue
worker


urgent queue
default queue
low queue
worker
$
$
$


urgent queue
default queue
low queue
worker

Sidekiq

Solid Queue

Good Job

Delayed Job

Job Queues

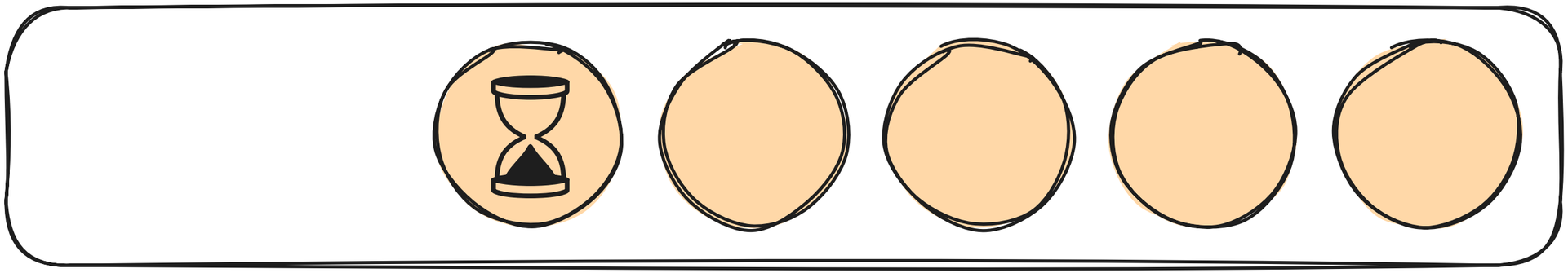
are

Job Queues

Step 1

Explicit latency SLOs

within-5-minutes

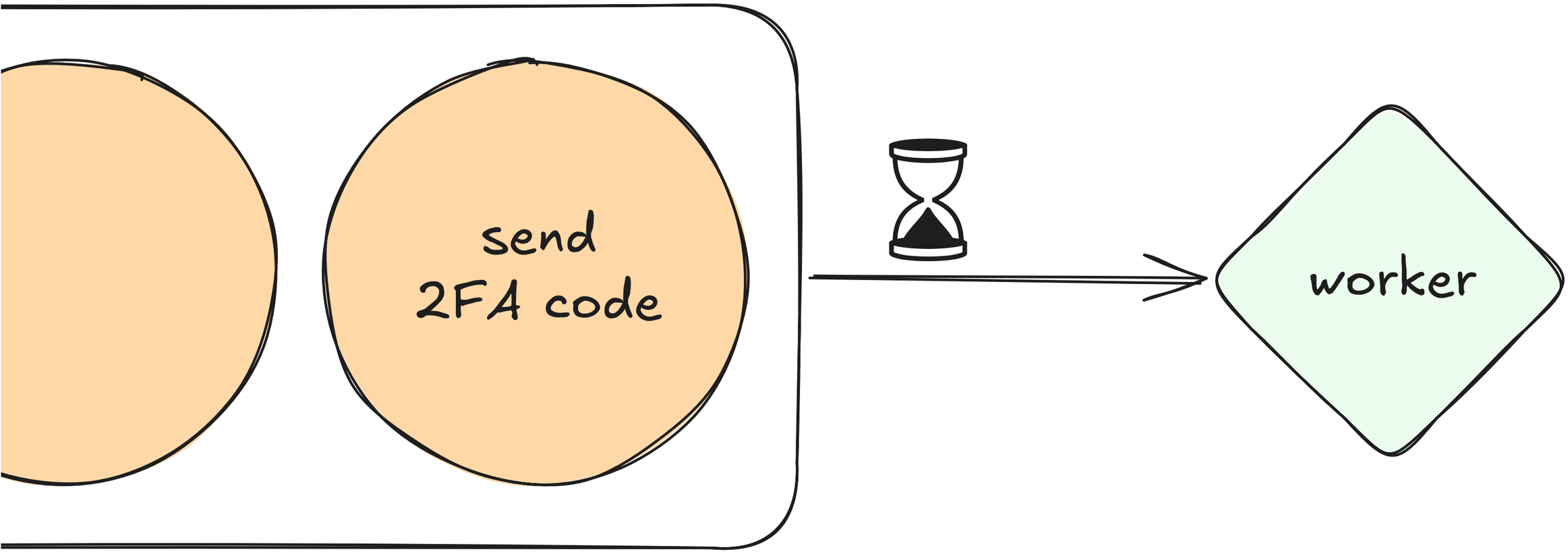


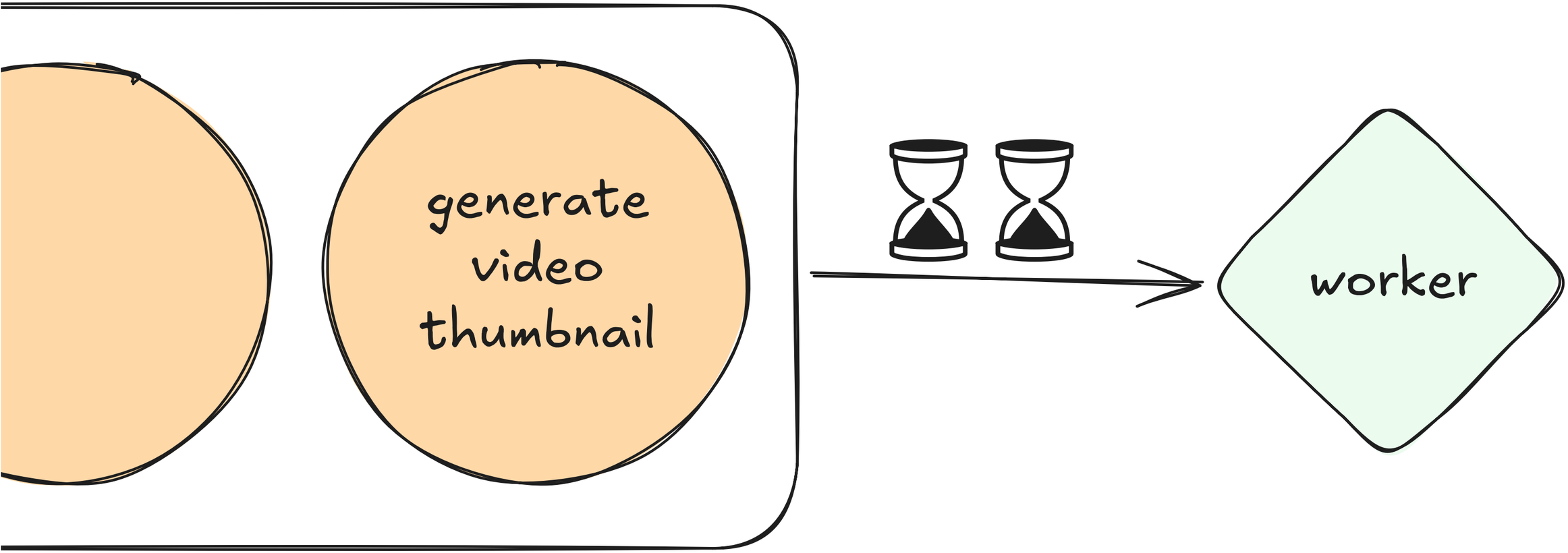
within-5-hours

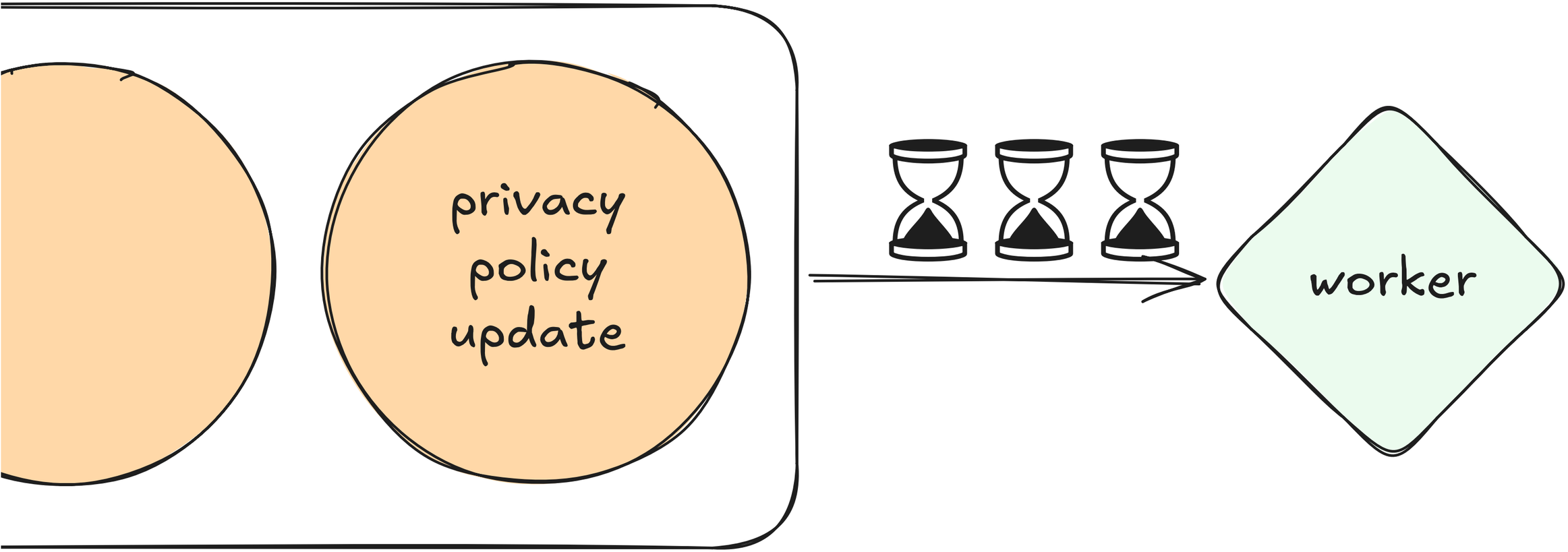
“SLO”

Service Level Objective

Every job in your app has an *implicit* SLO


send
2FA code
worker


generate
video
thumbnail
worker


privacy
policy
update
worker

Don’t make assumptions
Encode SLOs into queue names

default

low

high

urgent

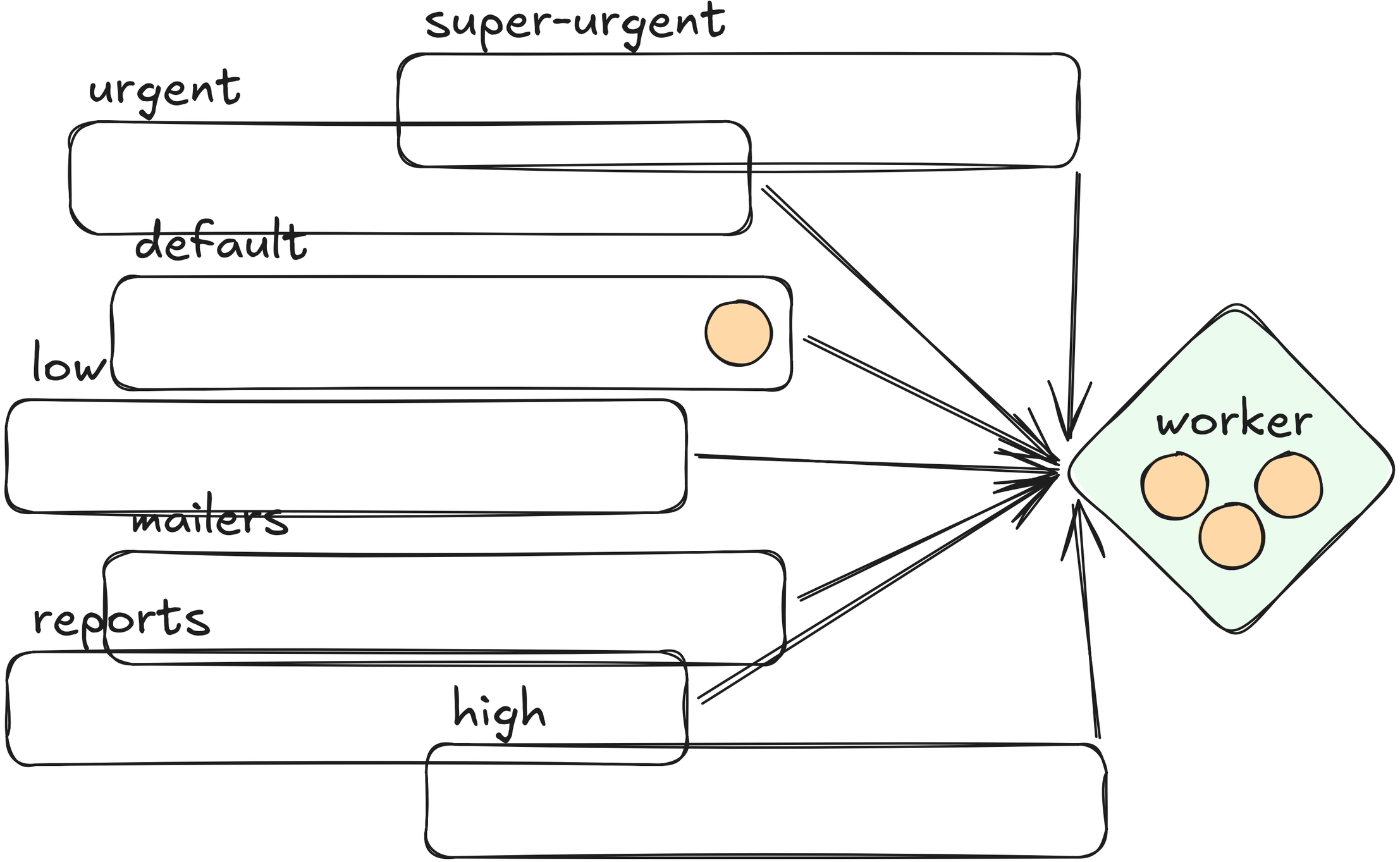
super-urgent

super-urgent-no-really

mailers

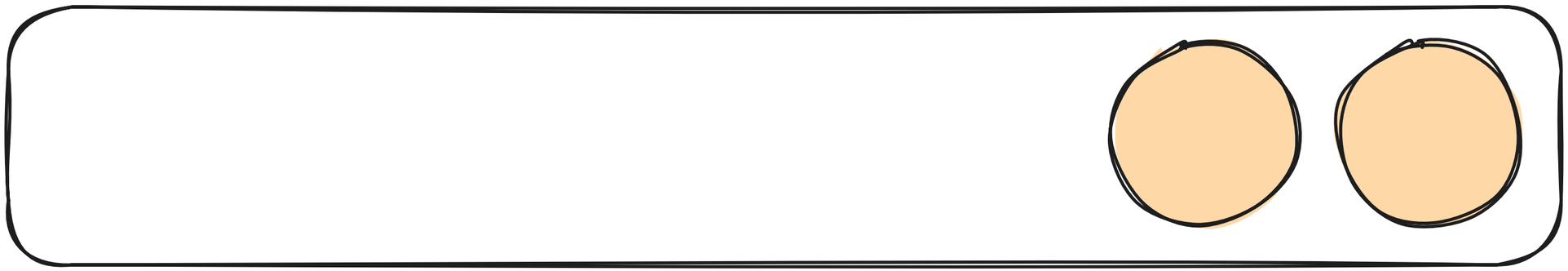
reports

bobs-queue-do-not-use

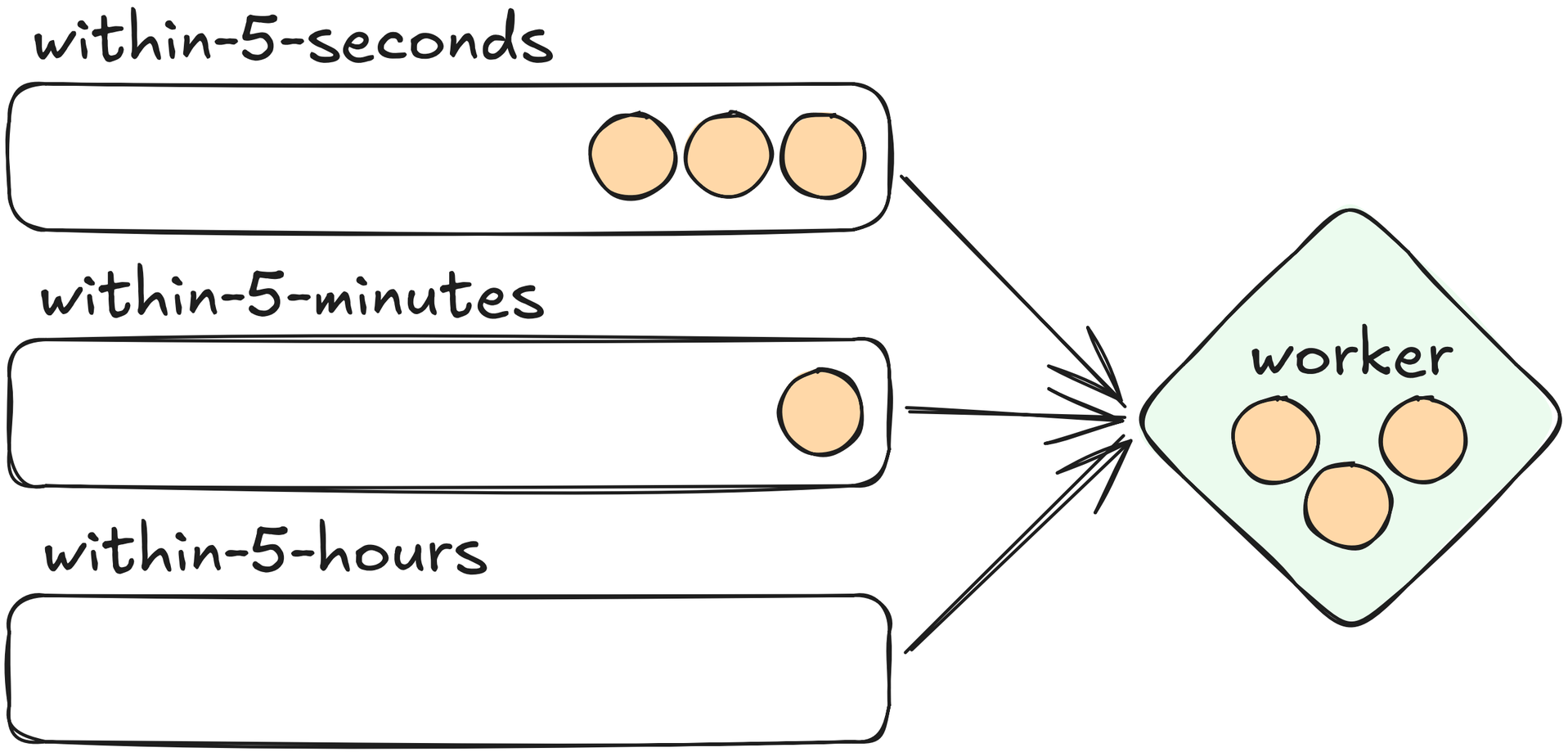

super-urgent
urgent
default
low
mailers
reports
high
worker

urgent??

mailers??



Latency-based queues

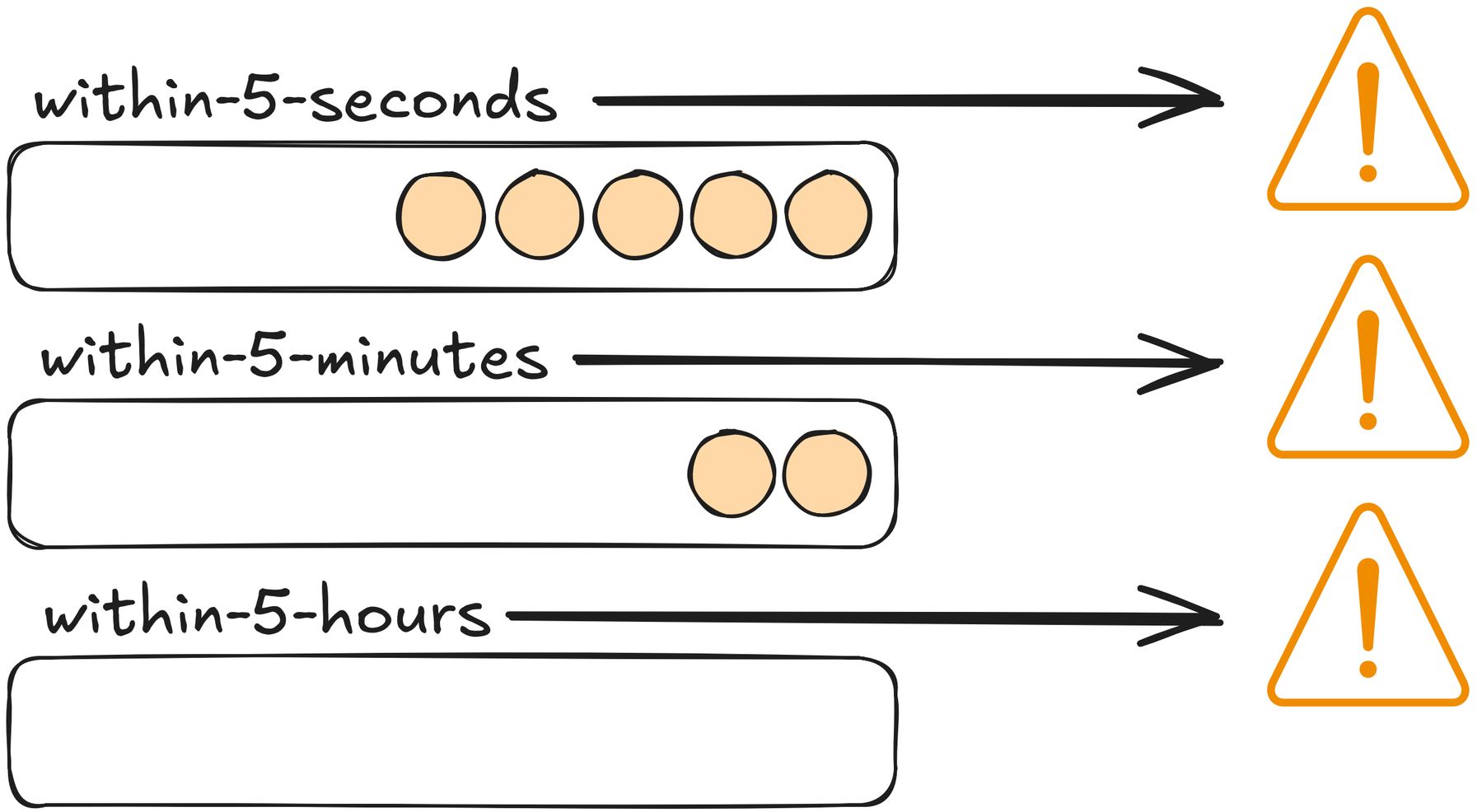

within-5-seconds
within-5-minutes
within-5-hours
worker

Step 2

Latency-based alerts

How do you know when there's a problem?

Queue latency **is the metric that matters**


within-5-seconds
within-5-minutes
within-5-hours

...or avoid alerts entirely

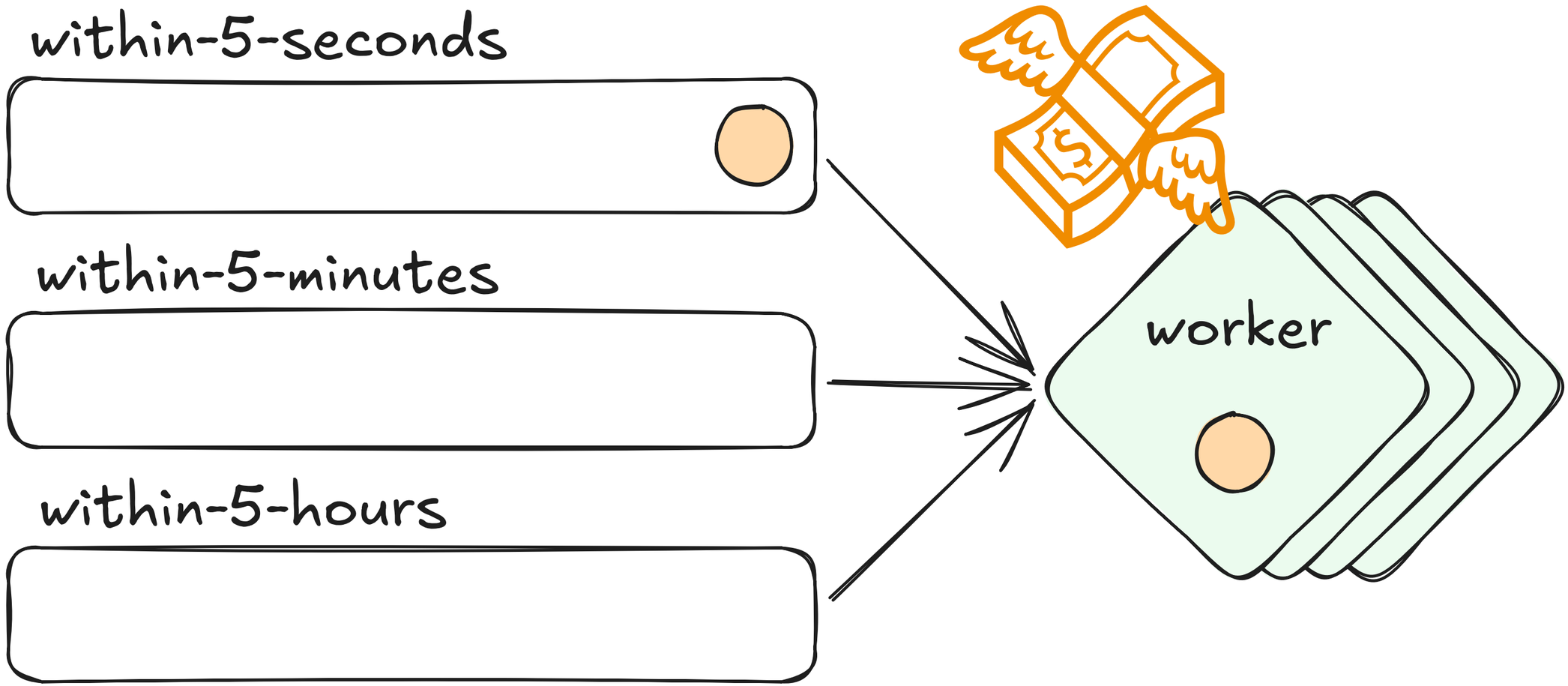
Step 3

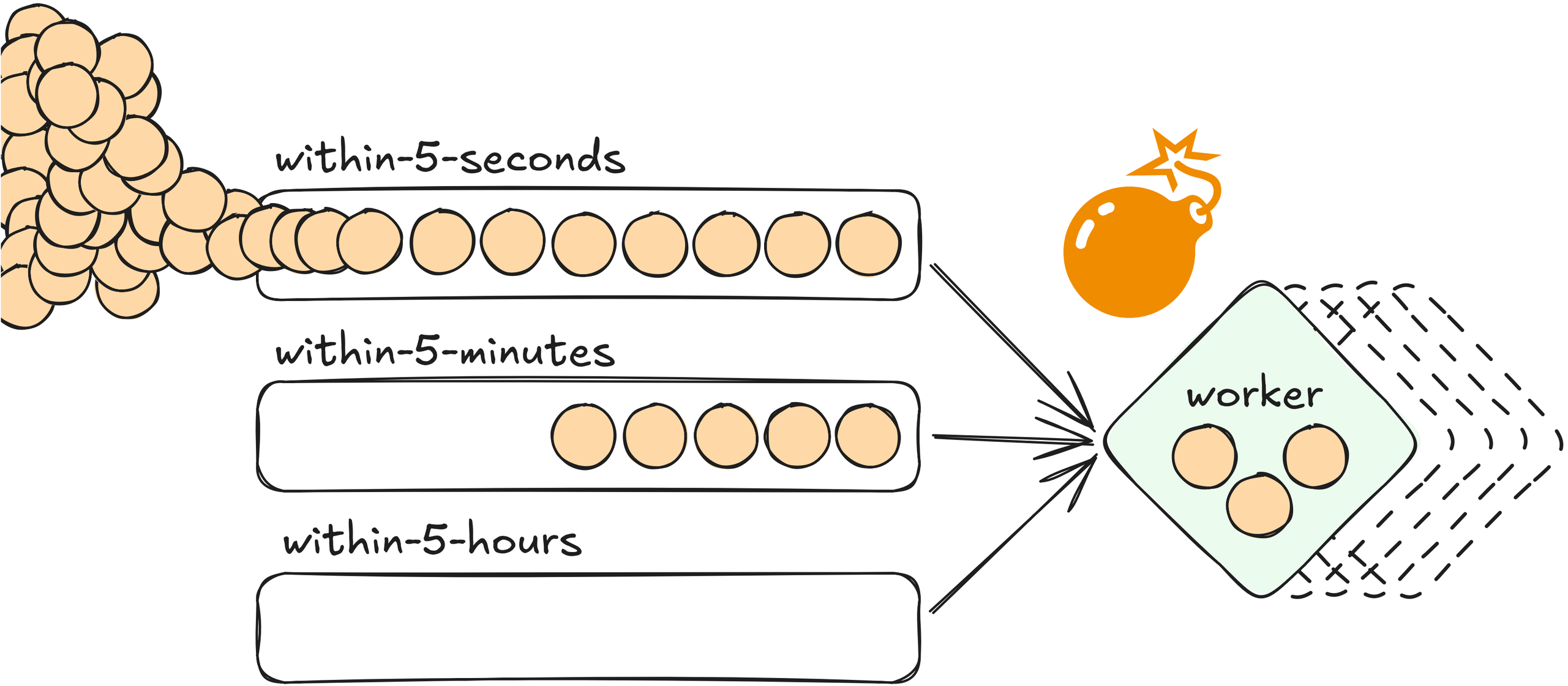
Latency-based autoscaling

Judoscale

Step 3

Latency-based autoscaling


within-5-seconds
within-5-minutes
within-5-hours
worker


within-5-seconds
within-5-minutes
within-5-hours
worker

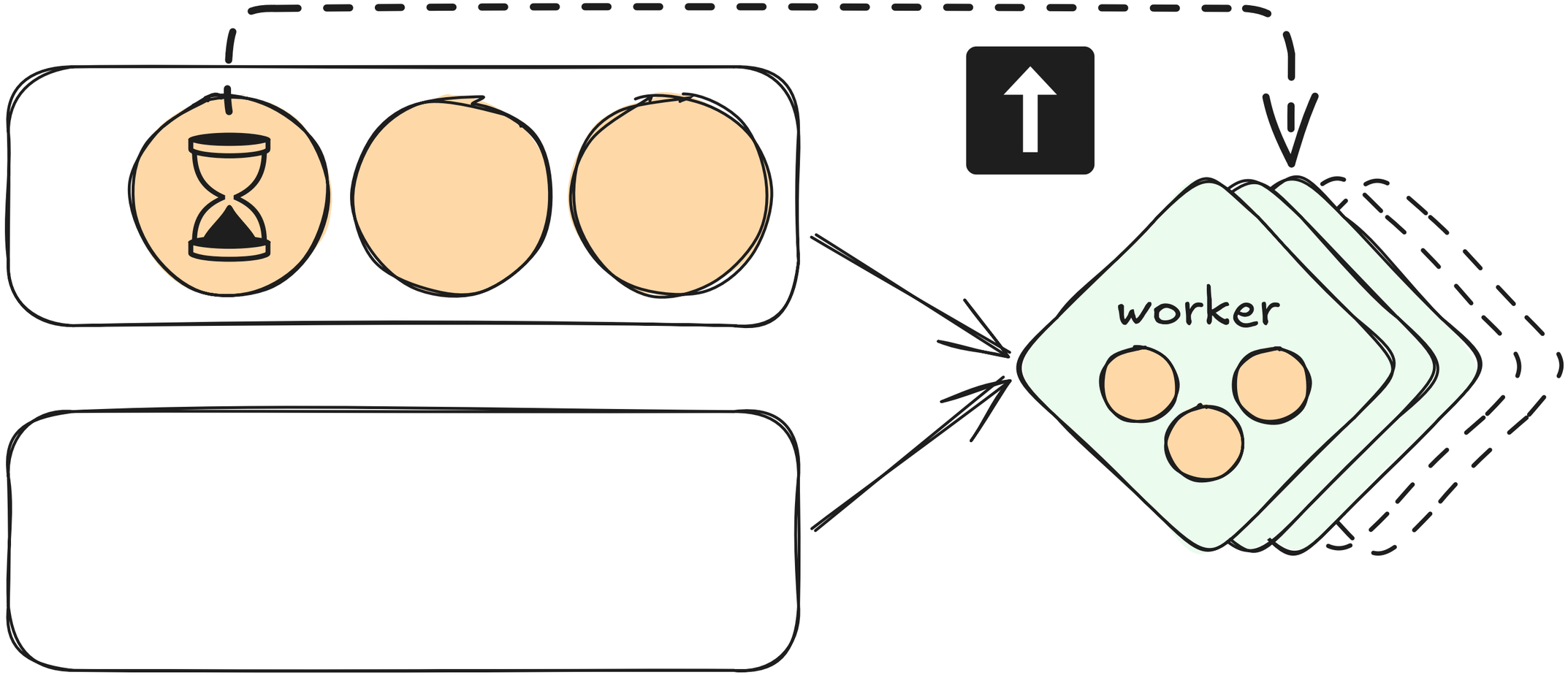
Worker autoscaling 101

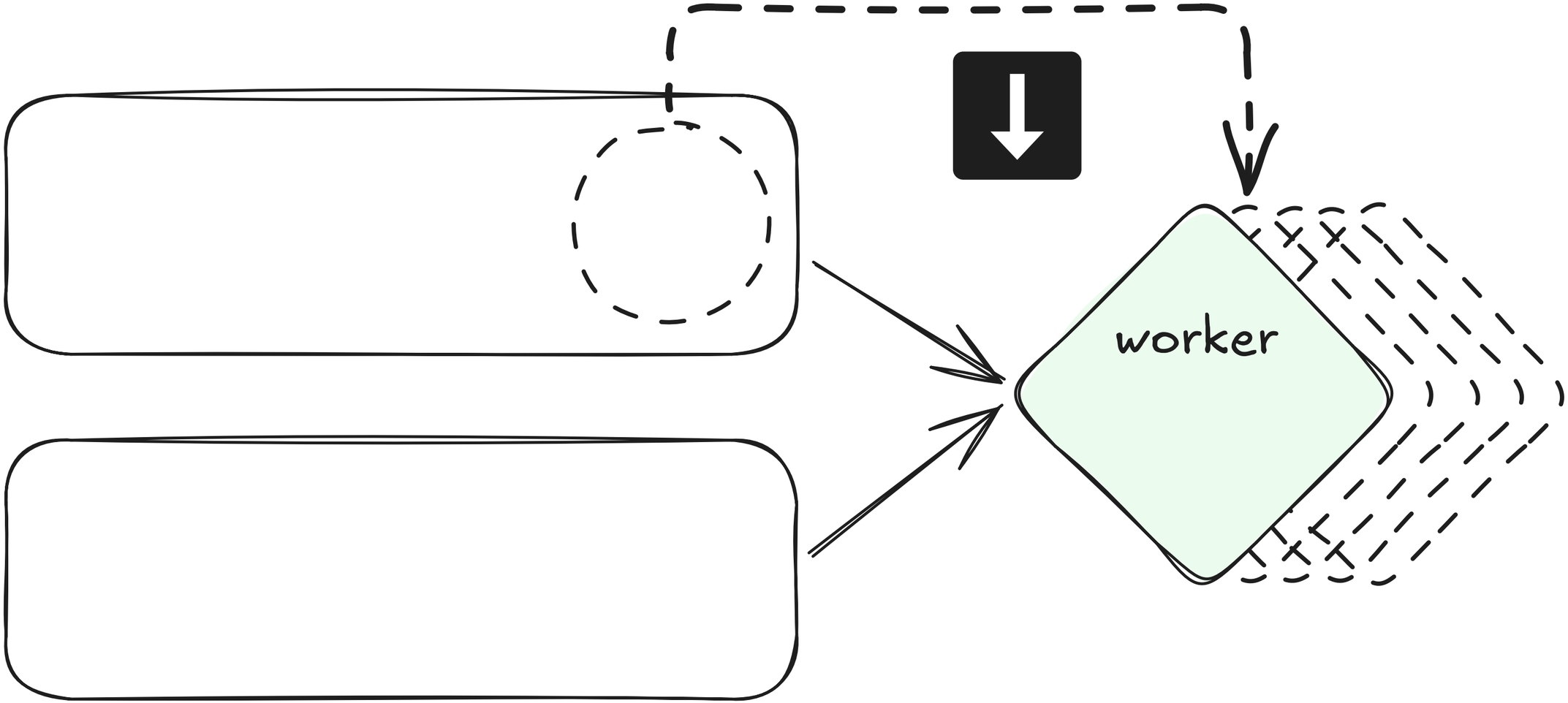
CPU does NOT correlate to queue health

Queue latency **is the metric that matters**

Judoscale

Queue latency **is the metric that matters**

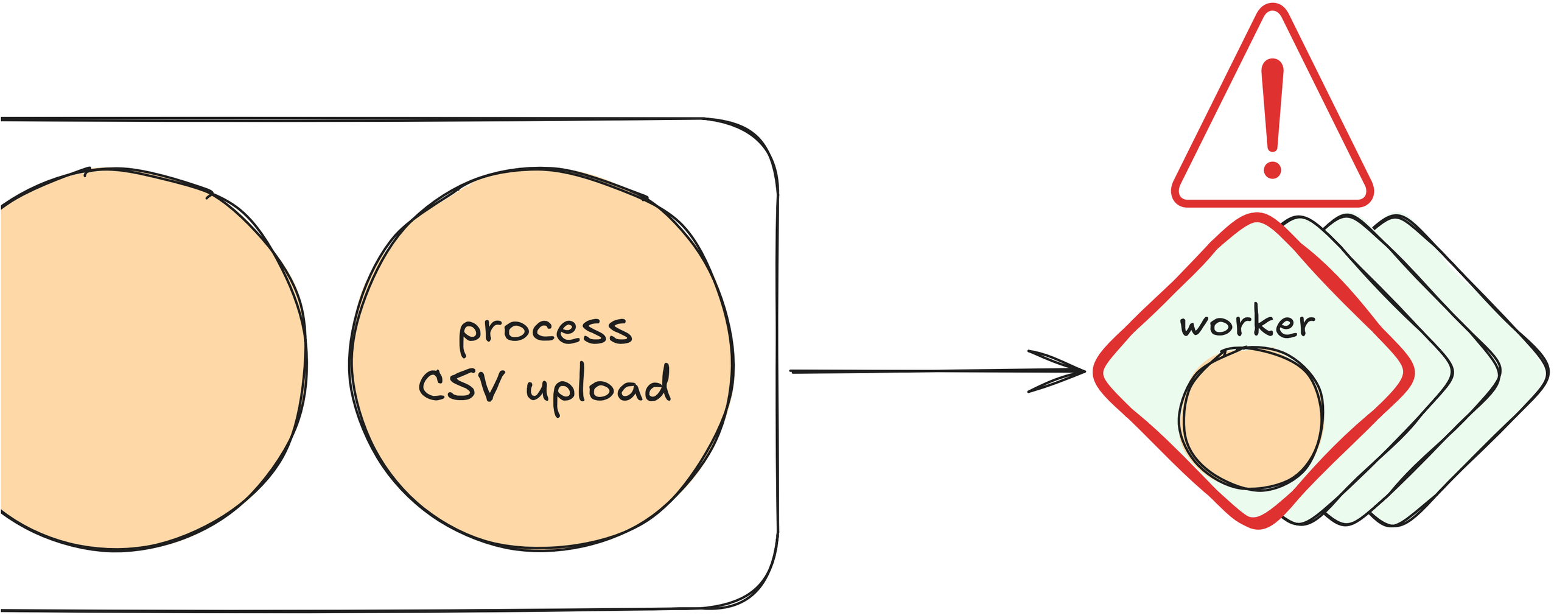

worker

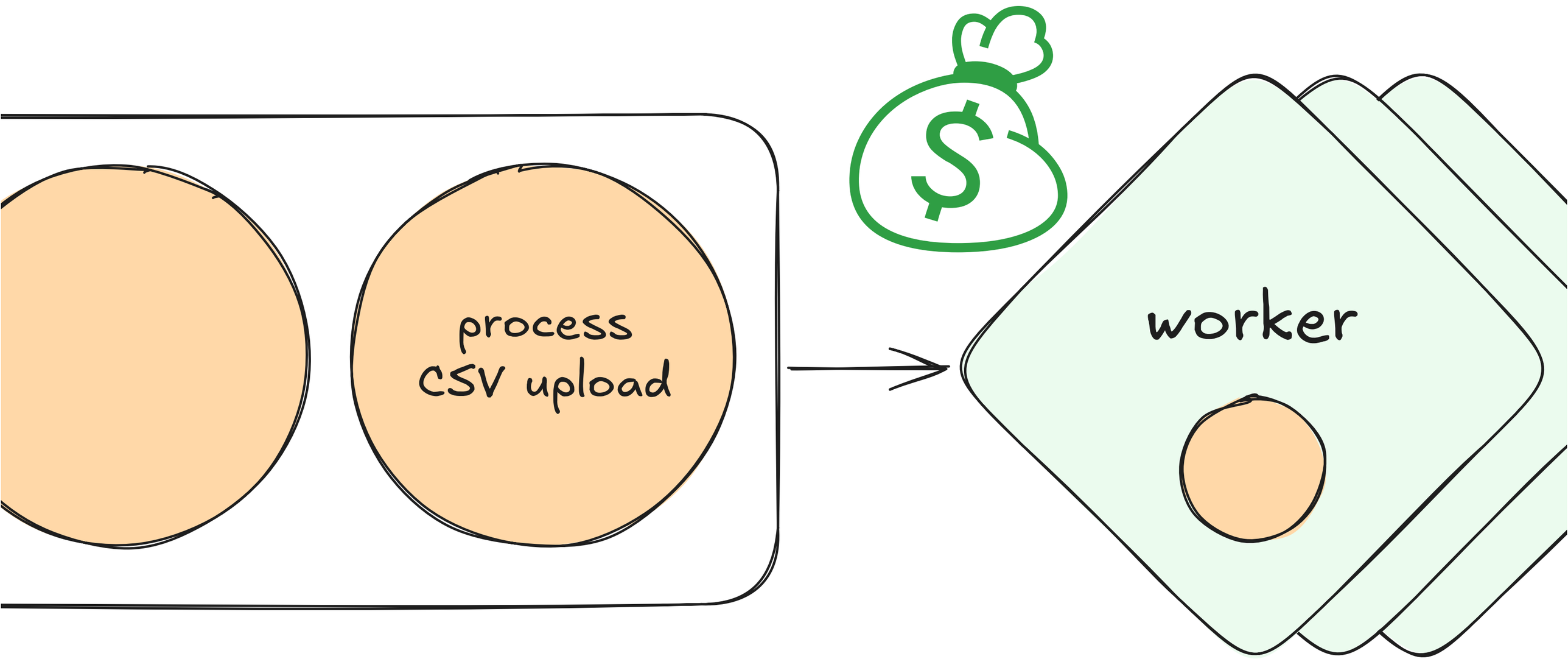

worker

~~One-size-fits-all~~

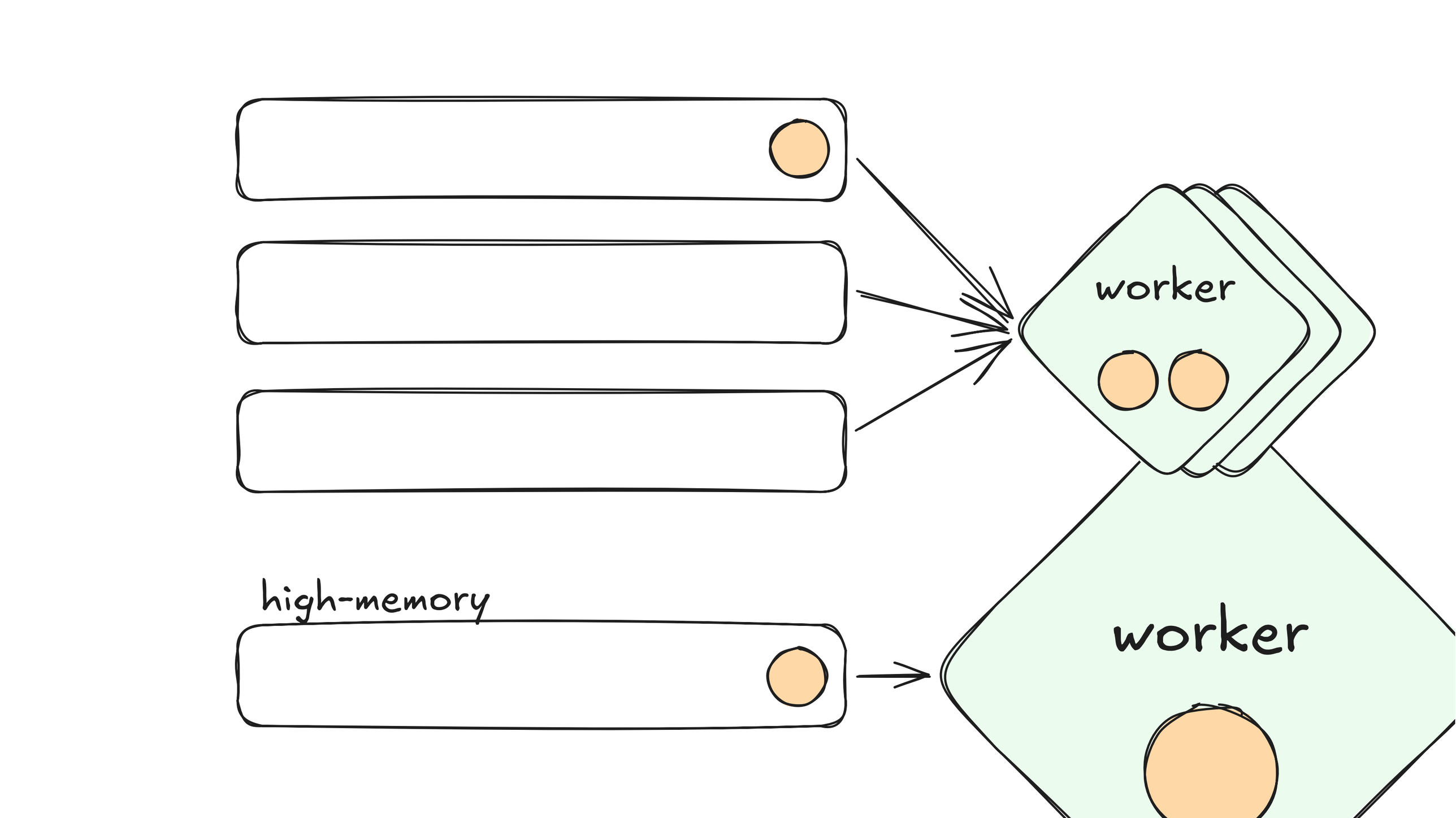
Step 4

Isolate the outliers


process
CSV upload
worker


process
CSV upload
worker

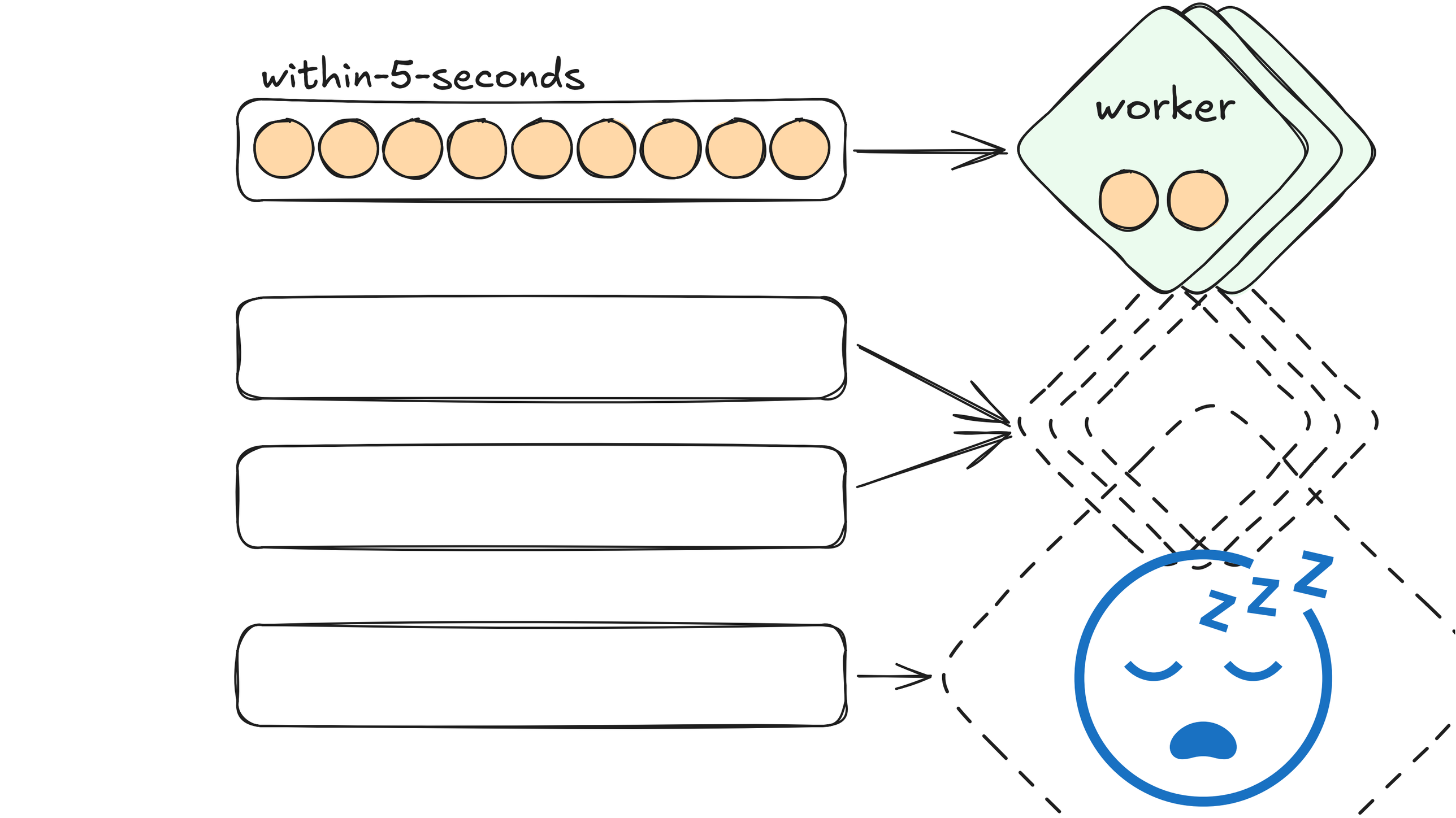
Solution:
dedicated workers
dedicated queues


worker
high-memory
worker

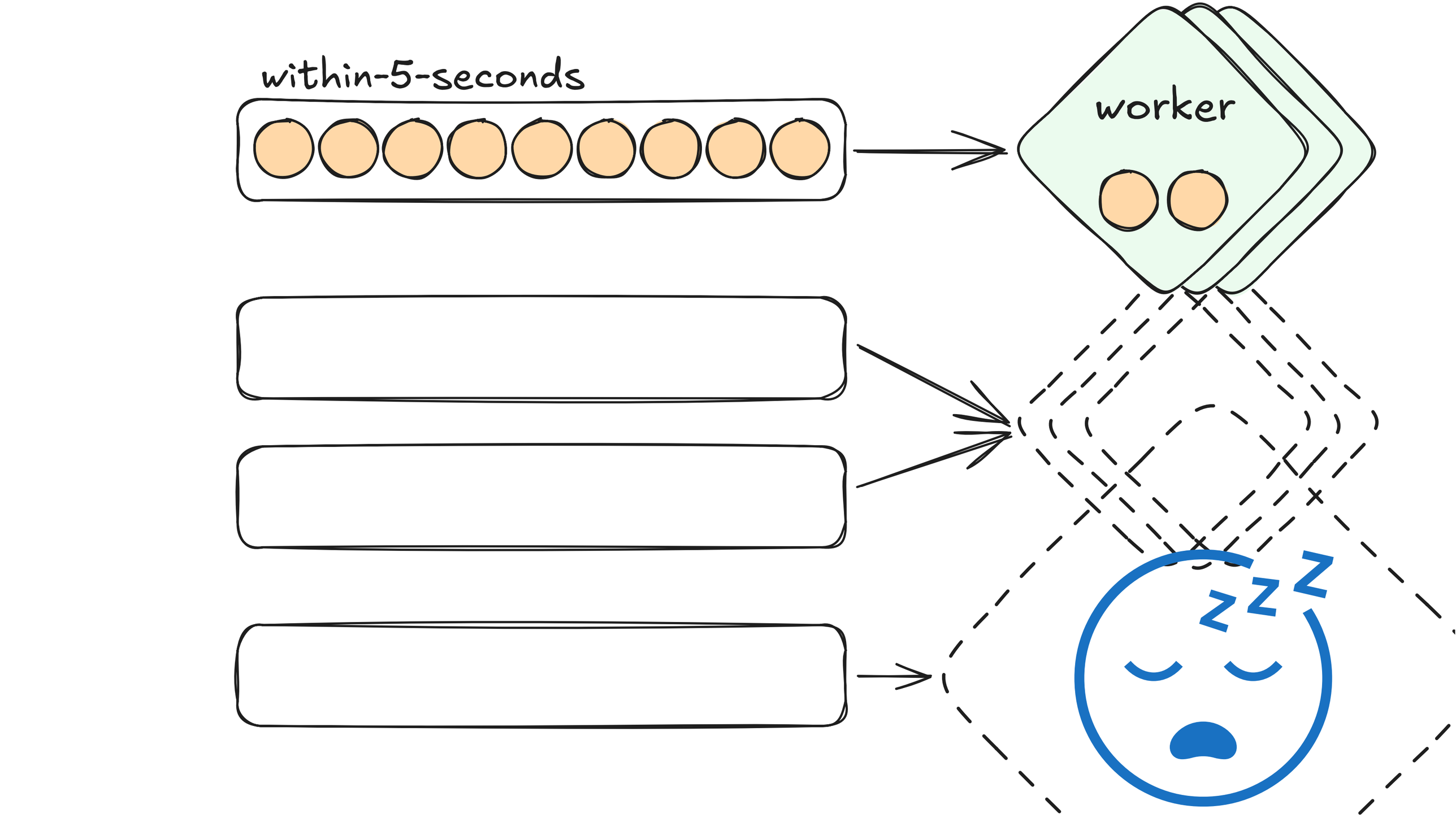
Mo’ workers
Mo’ problems

Step 5

Autoscale to zero


within-5-seconds
worker
Z
Z
Z

Judoscale


within-5-seconds
worker
Z
Z
Z

1. Explicit latency SLOs
2. Latency alerts
3. Latency autoscaling
4. Isolate the outliers
5. Autoscale to zero

💥 SLIDES & SWAG 💥

→ → → → →

5 Steps to Resilient Job Queues

👋 Adam McCrea 👋

💥 SLIDES & SWAG 💥

→ → → → →