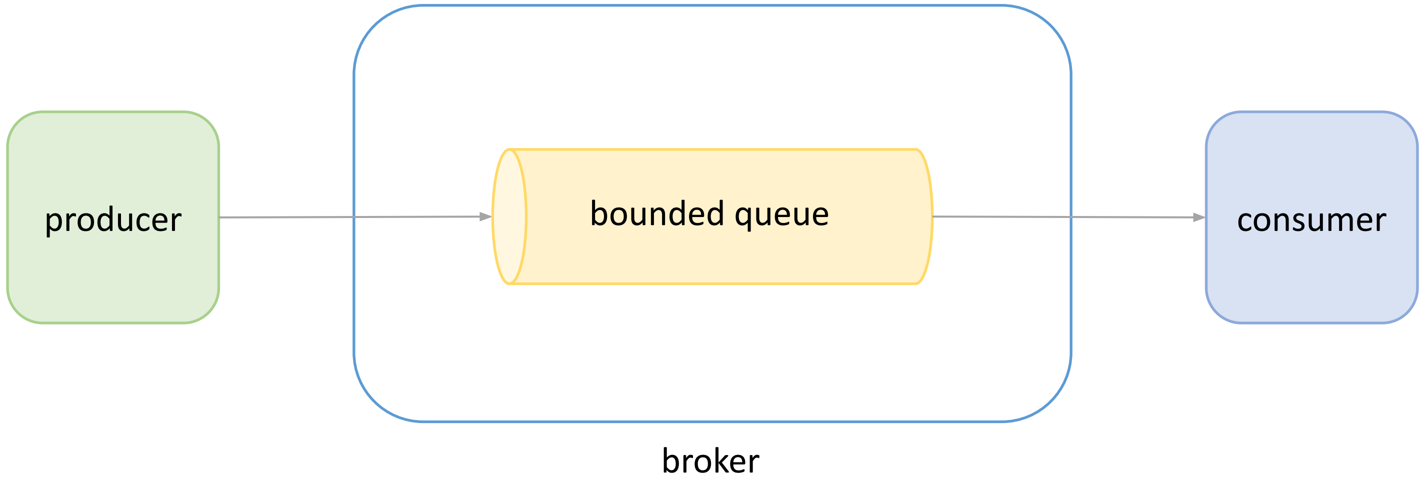
Full and empty queue problems

full

queue

arrival rate > retrieval rate

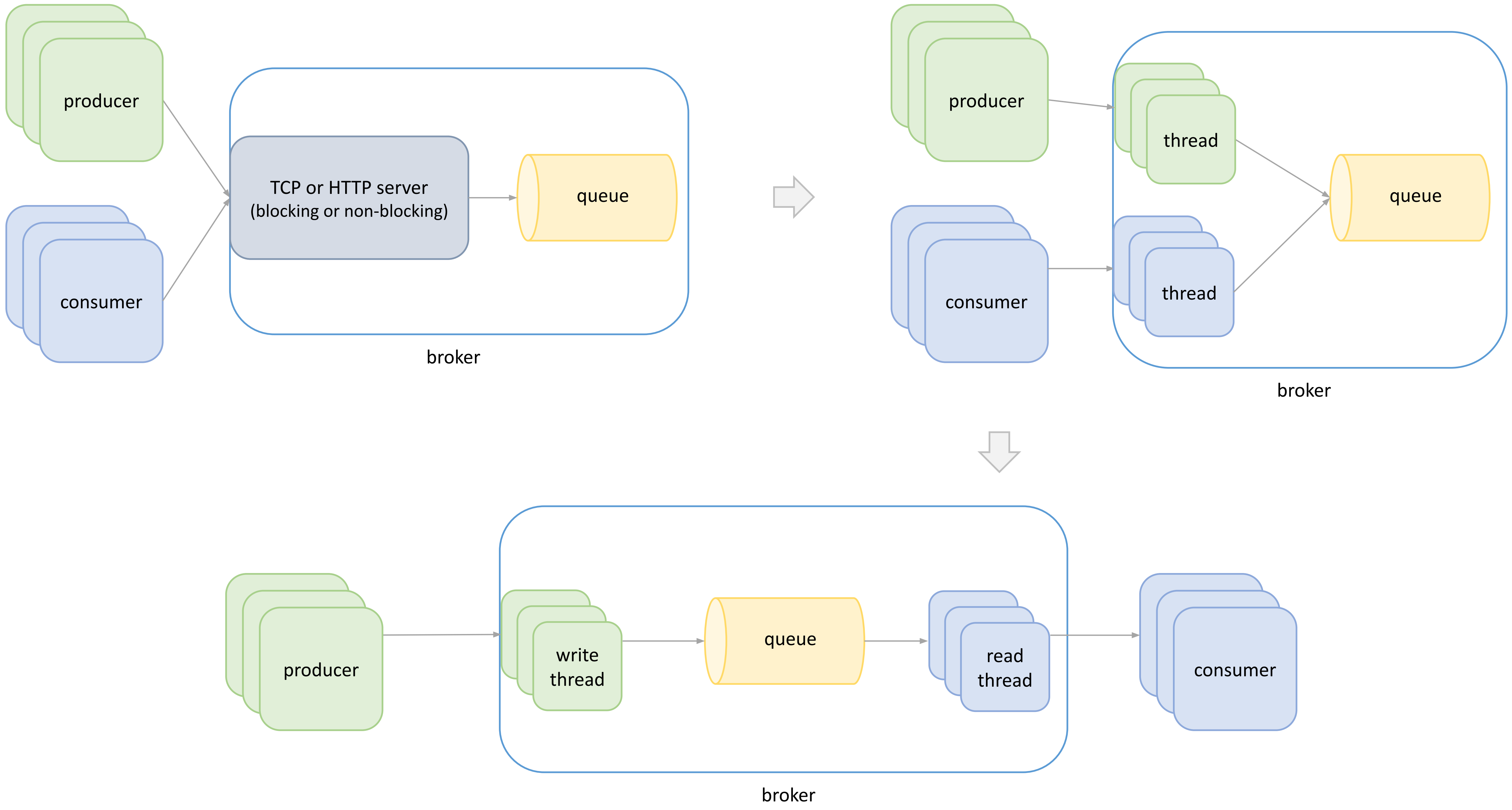


empty

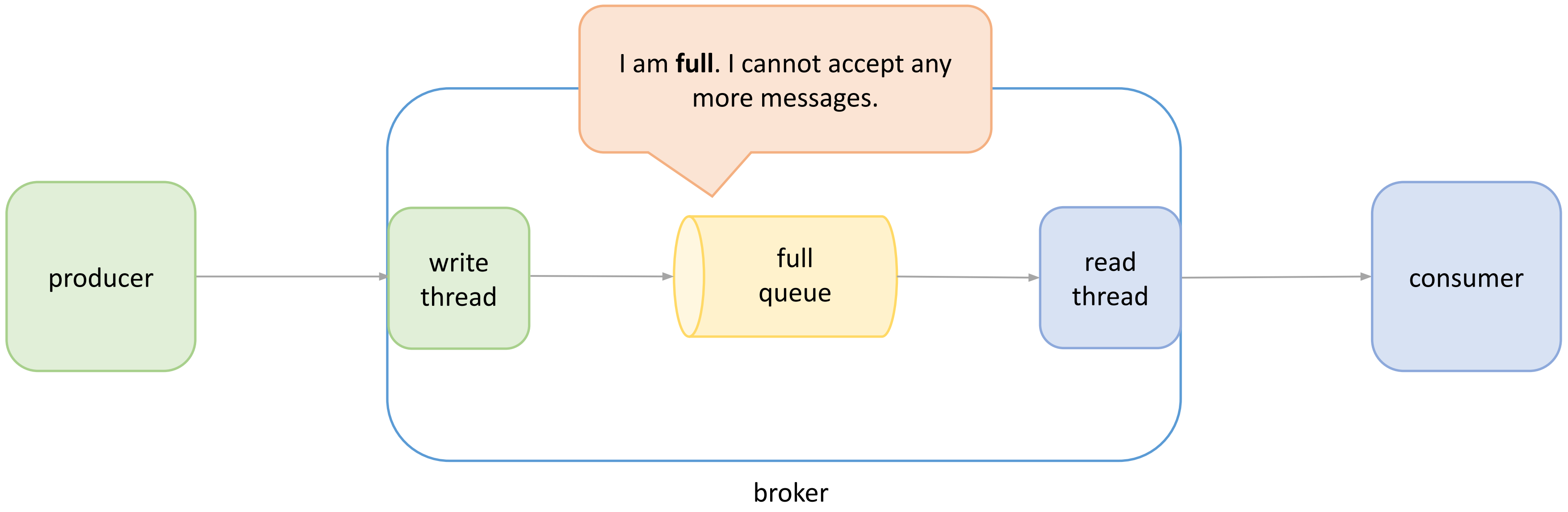
queue

arrival rate < retrieval rate

Full and empty queue problems



Full and empty queue problems

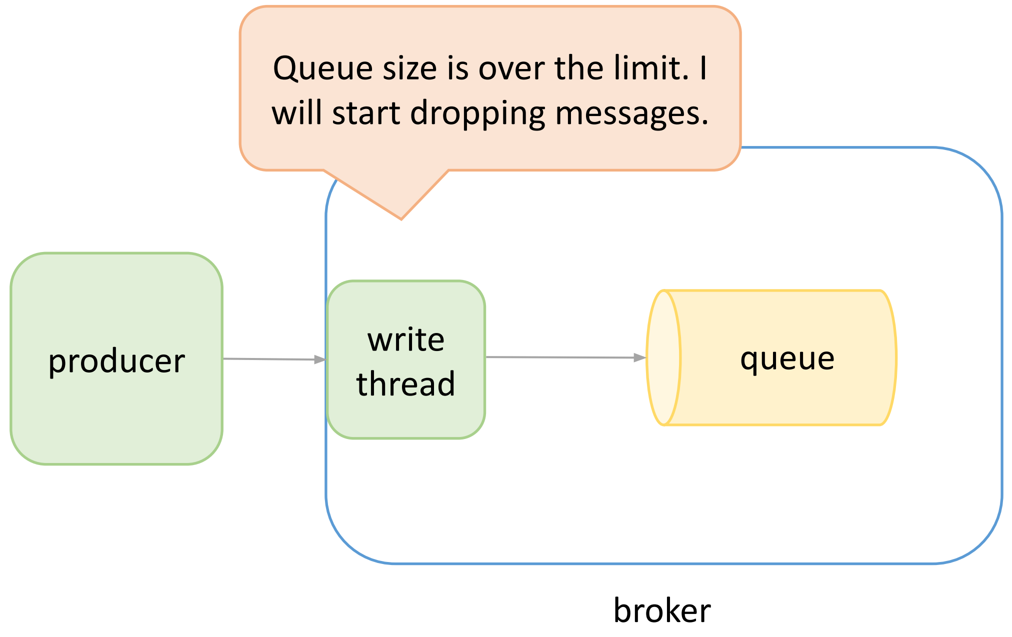


start dropping messages on the floor	→	load shedding, rate limiting
force producers to slow down	→	backpressure
scale consumers up or out	→	elastic scaling (autoscaling)

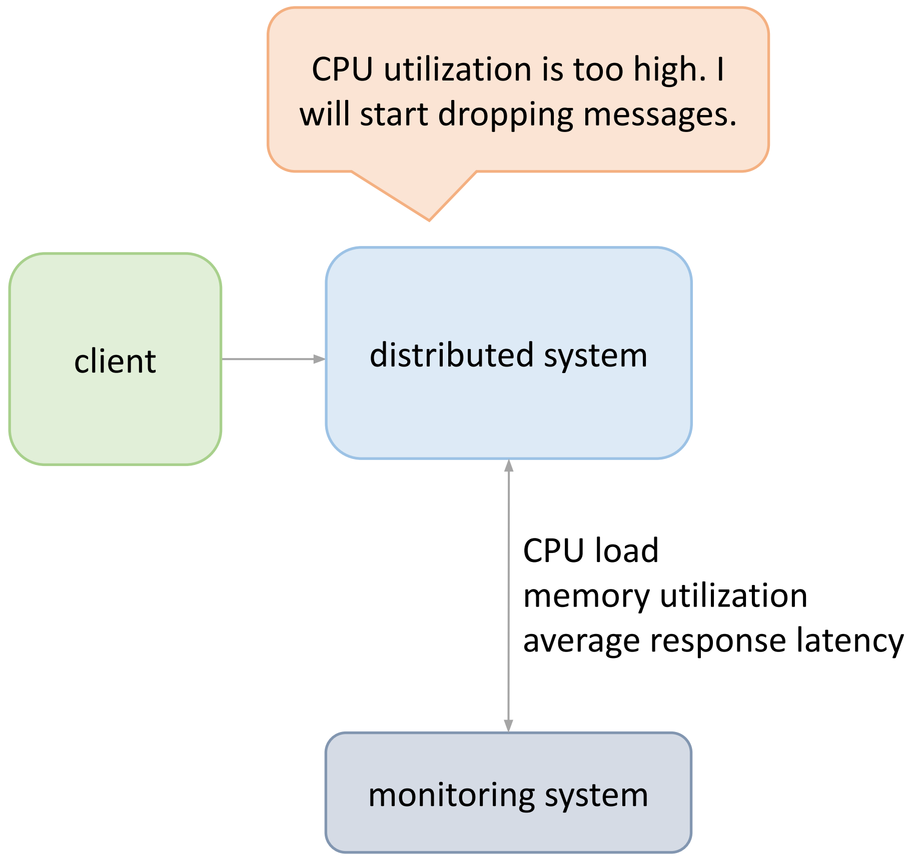
Full and empty queue problems

load shedding

single machine


Queue size is over the limit. I will start dropping messages.
producer
write thread
queue
broker

distributed system (multiple machines)

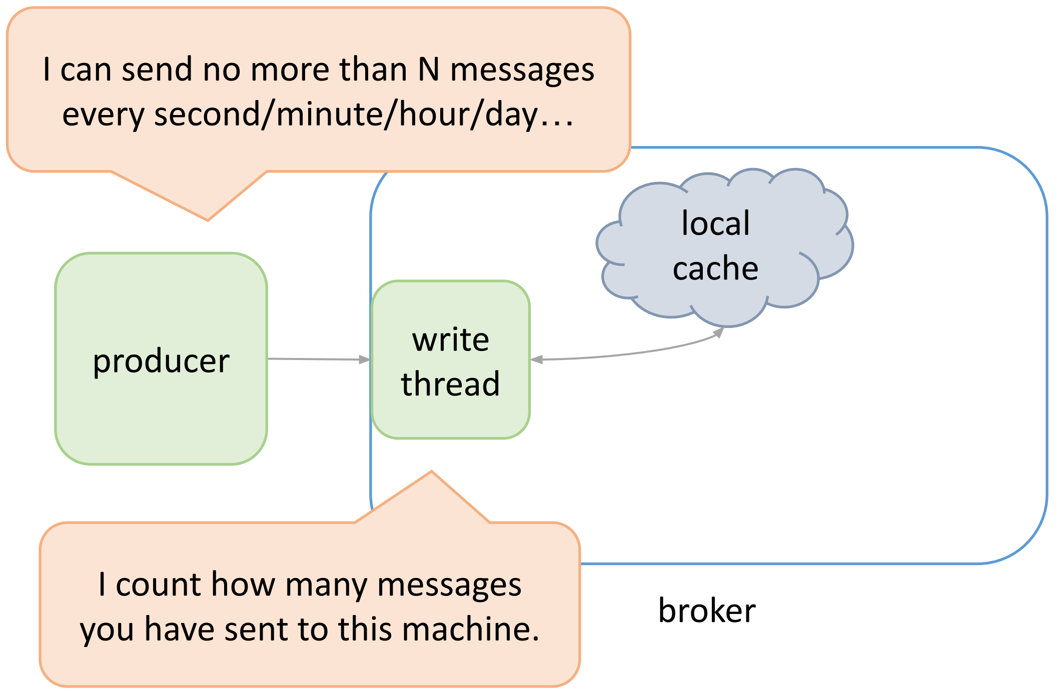


Full and empty queue problems

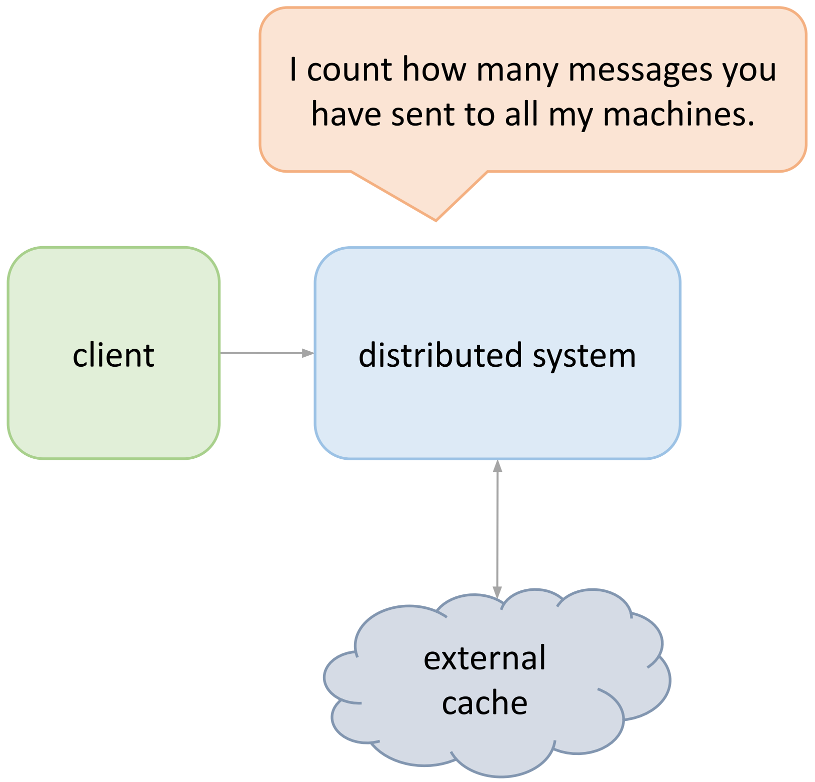
rate limiting

every producer gets a quota limiting how many messages
this producer can send to the broker

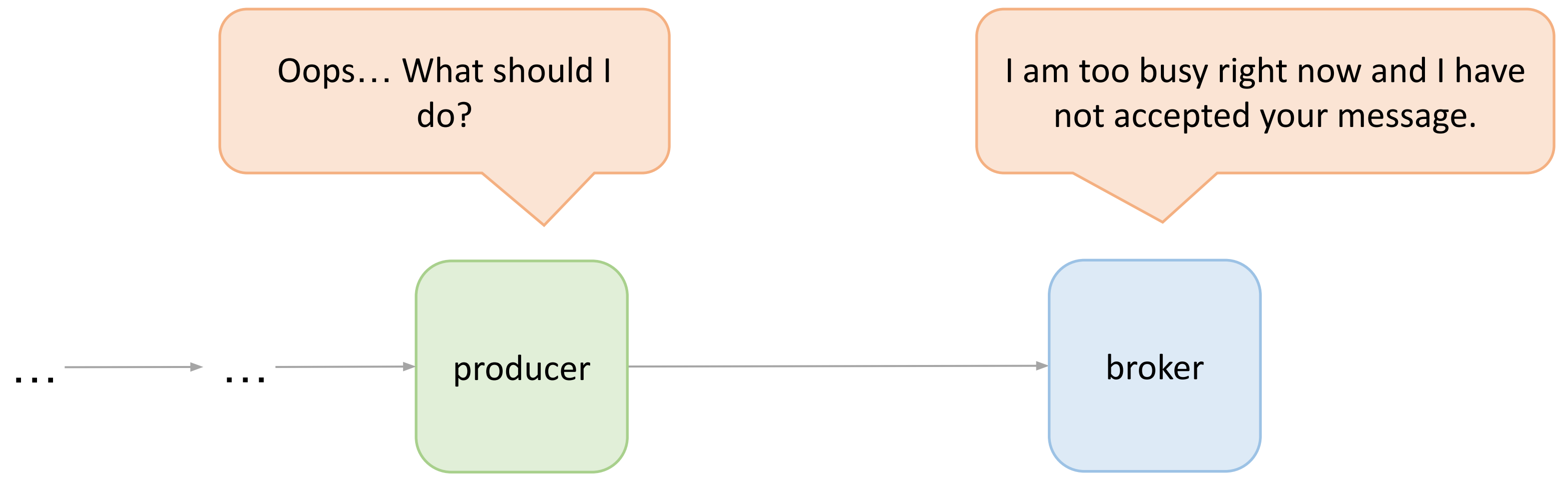
single machine


I can send no more than N messages every second/minute/hour/day…
producer
write thread
local cache
broker
I count how many messages you have sent to this machine.

distributed system (multiple machines)



Full and empty queue problems


Oops… What should I do?
I am too busy right now and I have not accepted your message.
…
…
producer
broker

- do nothing
- buffer messages (in memory or on disk)
- propagate the exception further up the stack
- send messages over the limit to a temporary storage (e.g. another broker or system)

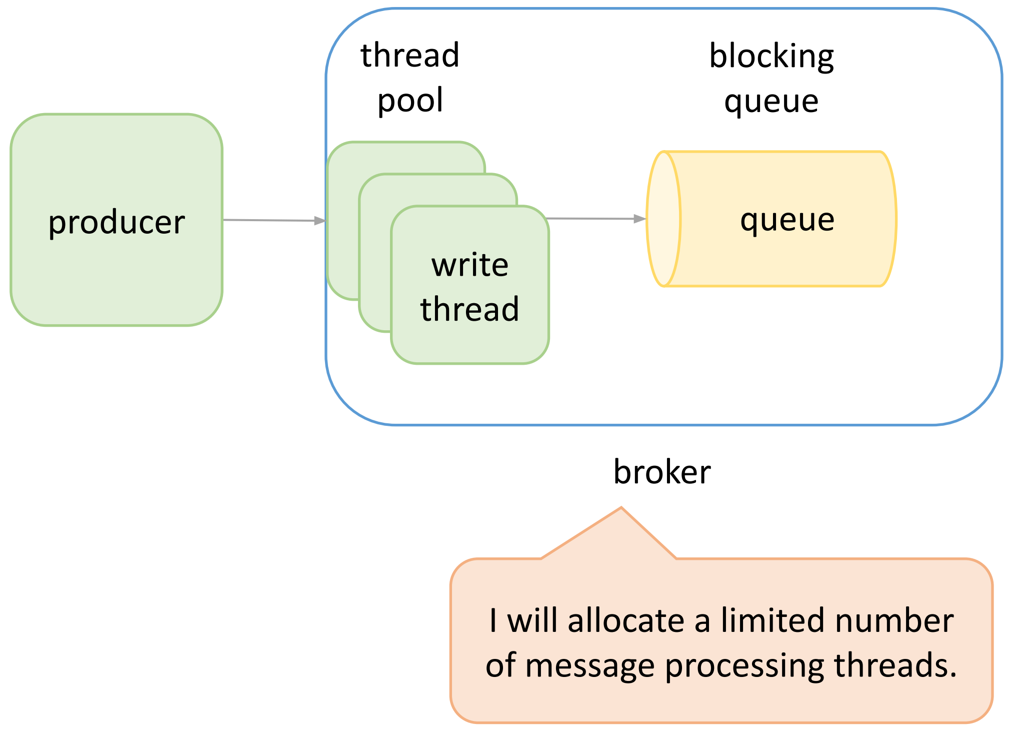
✗ retry immediately

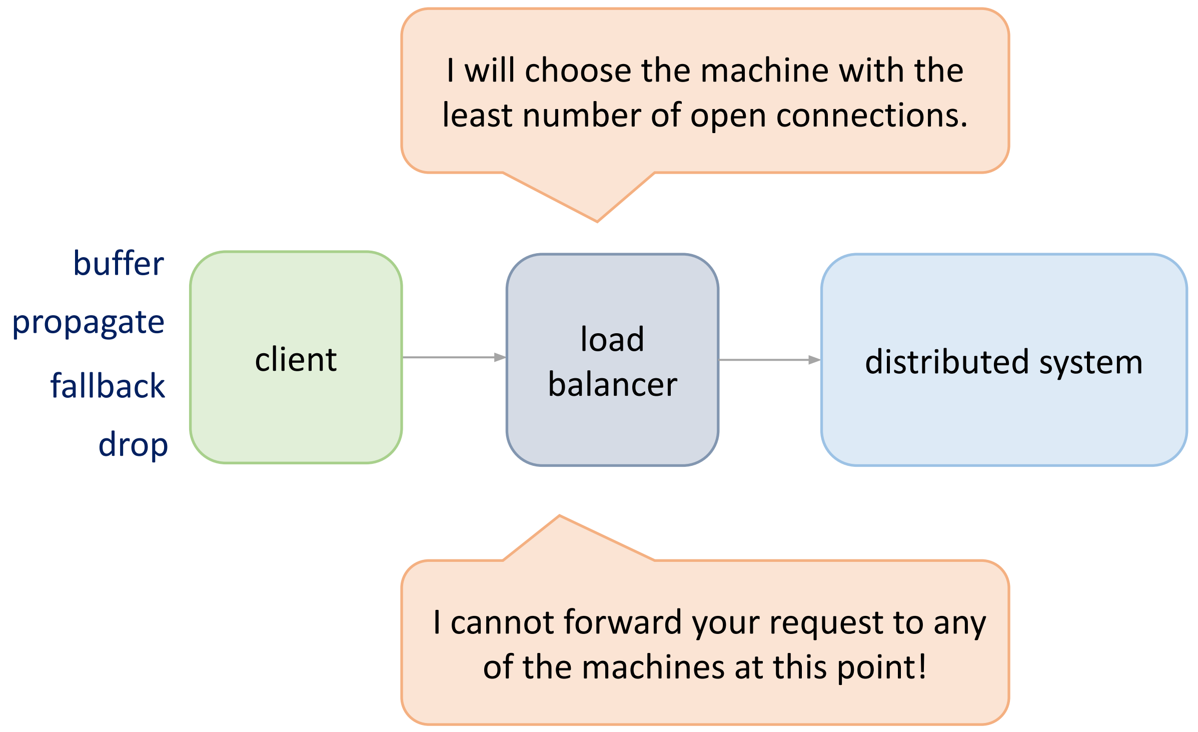
Full and empty queue problems

backpressure

single machine

distributed system
(multiple machines)


thread pool
blocking queue
producer
write thread
queue
broker
I will allocate a limited number of message processing threads.

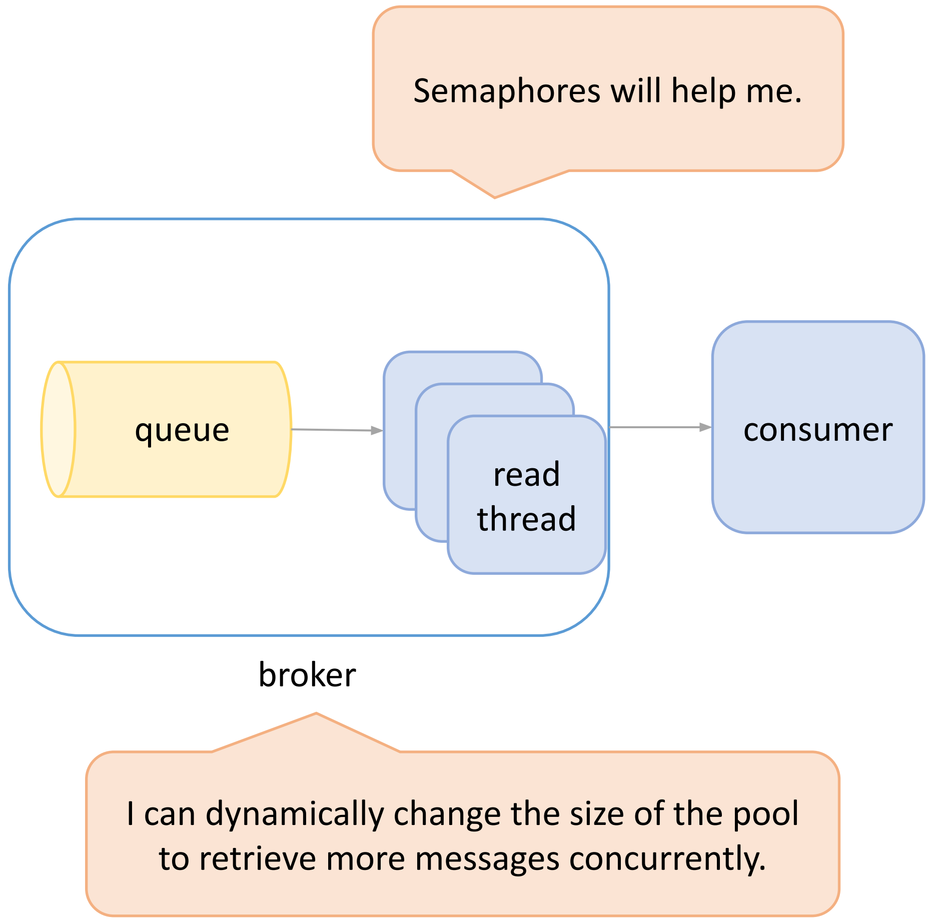

I will choose the machine with the least number of open connections.
buffer
propagate
fallback
drop
client
load balancer
distributed system
I cannot forward your request to any of the machines at this point!

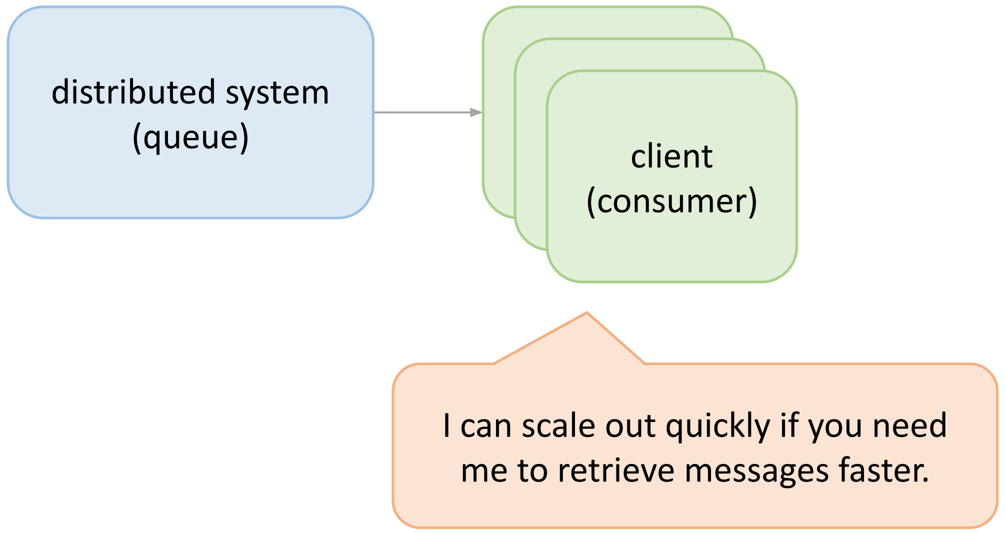
Full and empty queue problems

elastic scaling

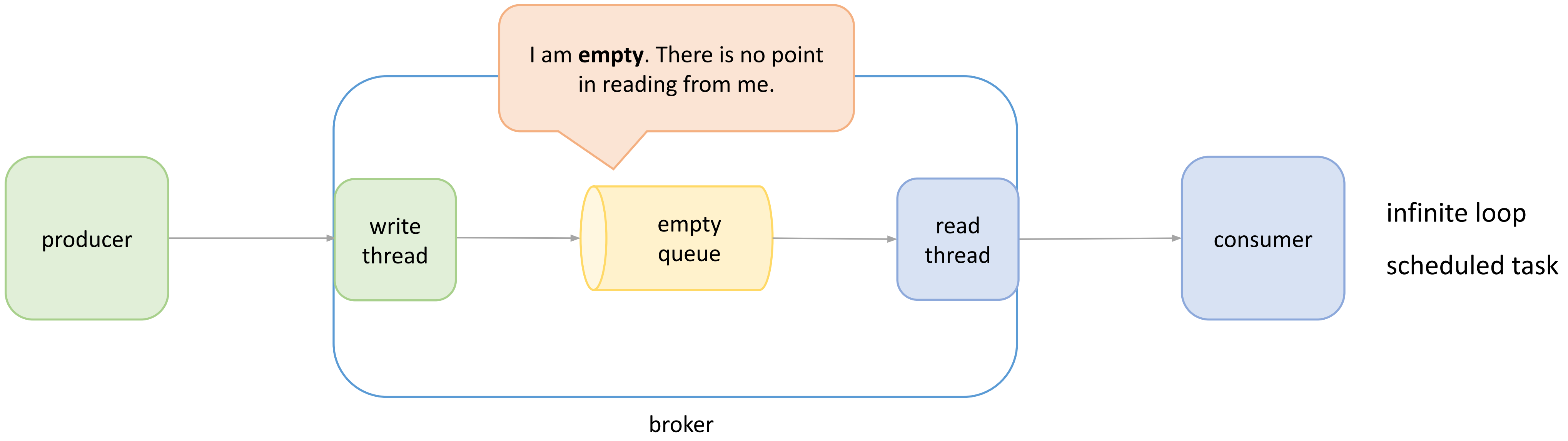
single machine

distributed system
(multiple machines)


Semaphores will help me.
queue
read thread
consumer
broker
I can dynamically change the size of the pool to retrieve more messages concurrently.


distributed system (queue)
client (consumer)
I can scale out quickly if you need me to retrieve messages faster.

Full and empty queue problems



- broker pushes messages when available → WebSocket
- broker blocks the pull request and waits for messages → long polling