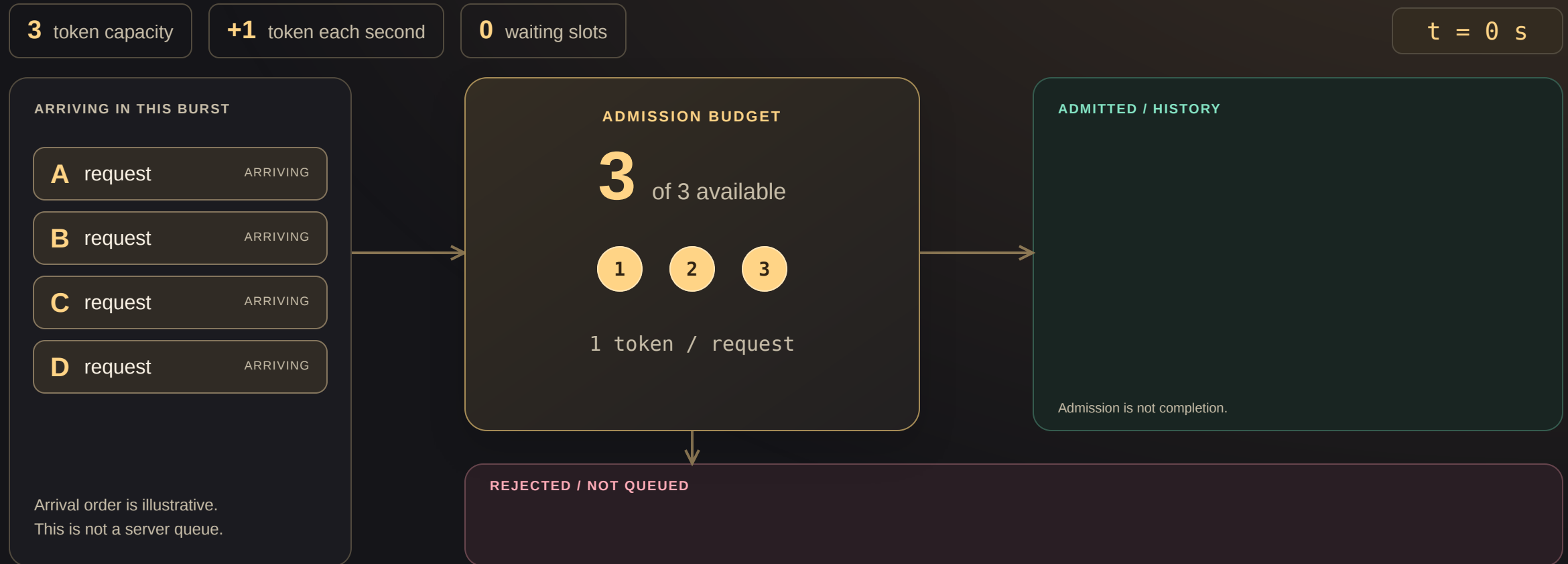


Throttle the burst.

A small bucket absorbs a burst. One token buys admission for one request.



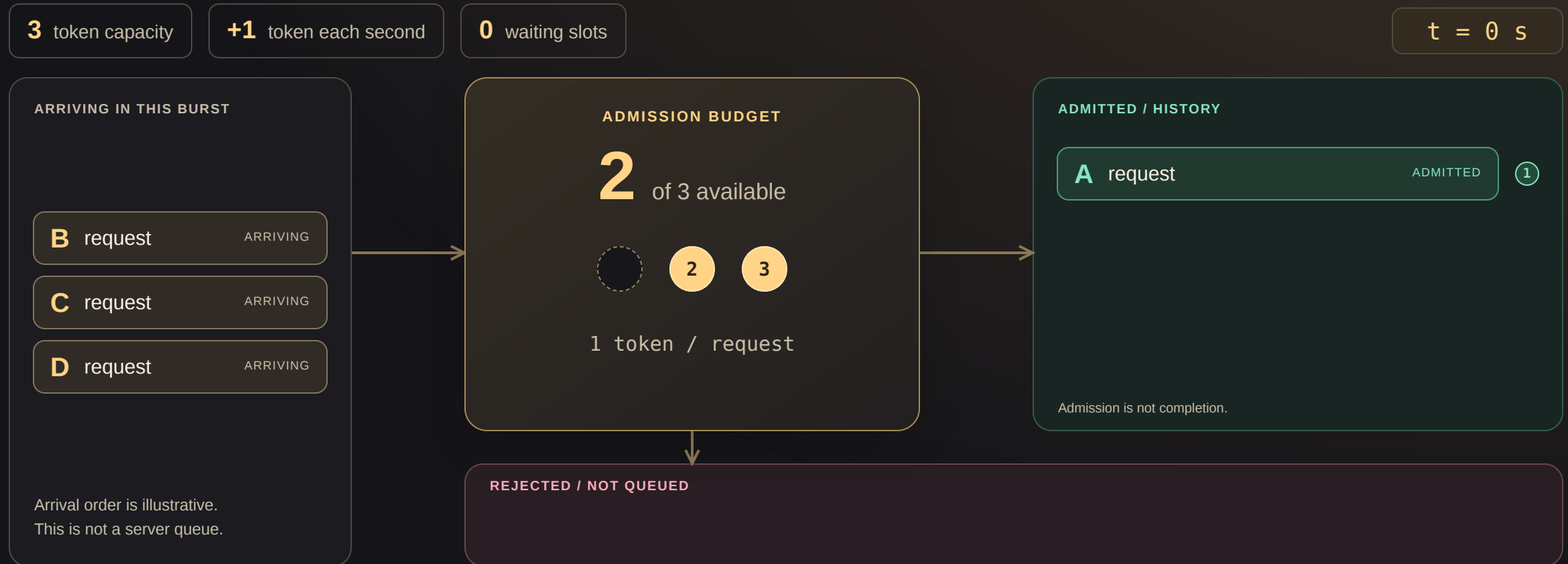
START FULL. Four arrivals meet a budget of three.

ADMITTED **0** / REJECTED **0**

Use the arrow keys. Each slide is a complete state; the transition supplies the motion.

One request. One token.

A is admitted. Its token moves out of the available budget and into the spent-token history.



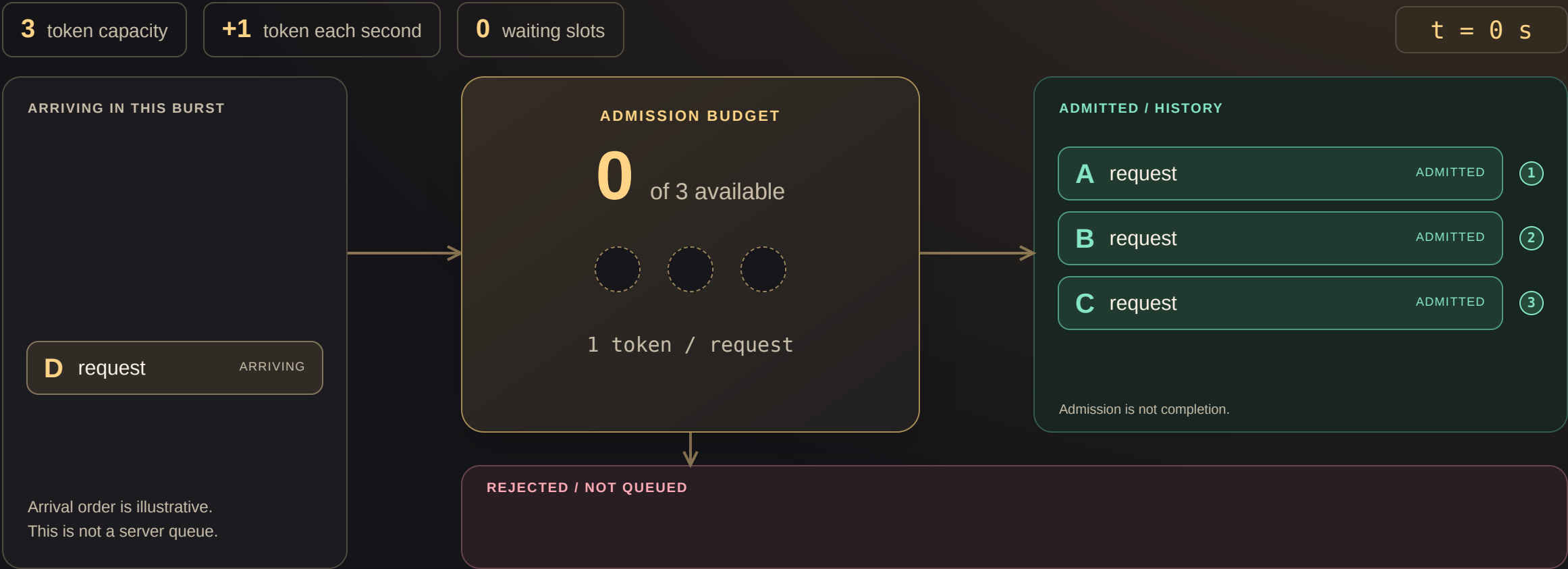
3 - 1 = 2. Admission spends budget immediately.

ADMITTED **1** / REJECTED **0**

The small numbered dots beside admitted requests record tokens already spent.

Three tokens. Three admissions.

B and C use the remaining tokens. The bucket is empty before the next clock tick.



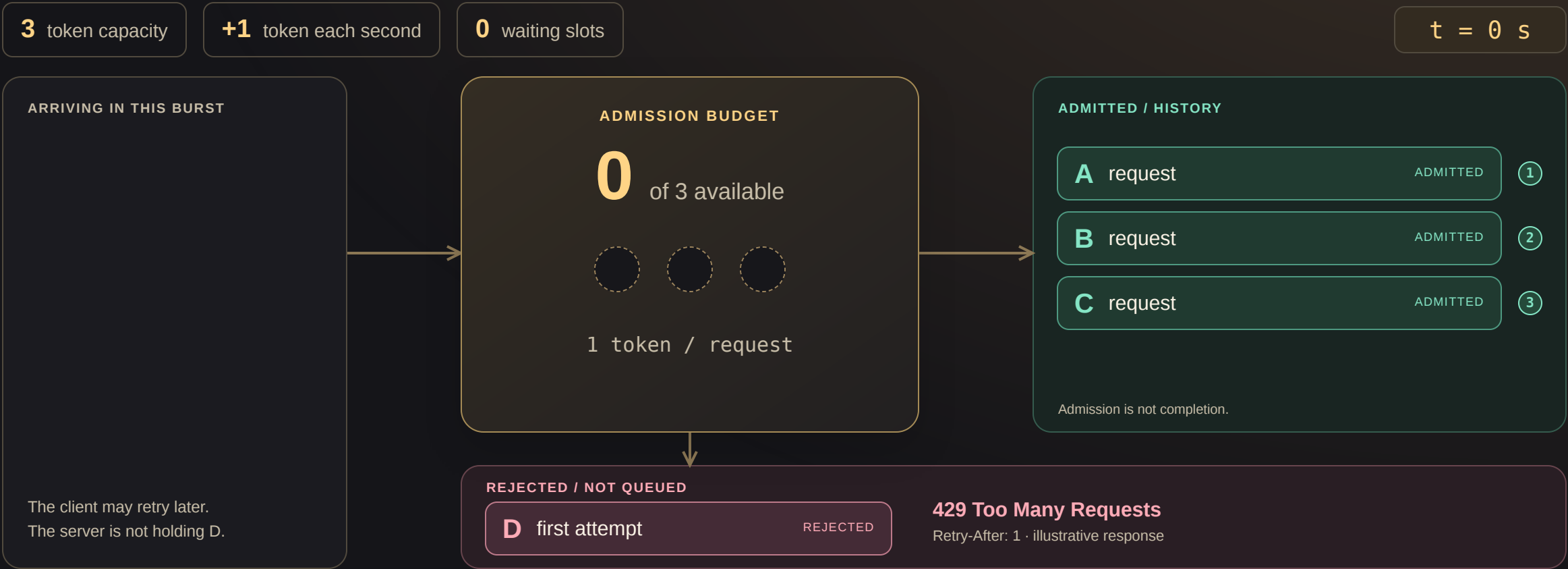
2 - 2 = 0. The initial burst consumed the stored budget.

ADMITTED 3 / REJECTED 0

A refill rate of one token per second can still allow an initial burst of three.

No token? Reject. Don't queue.

D cannot be admitted. This example returns HTTP 429 and tells the client when to try again.



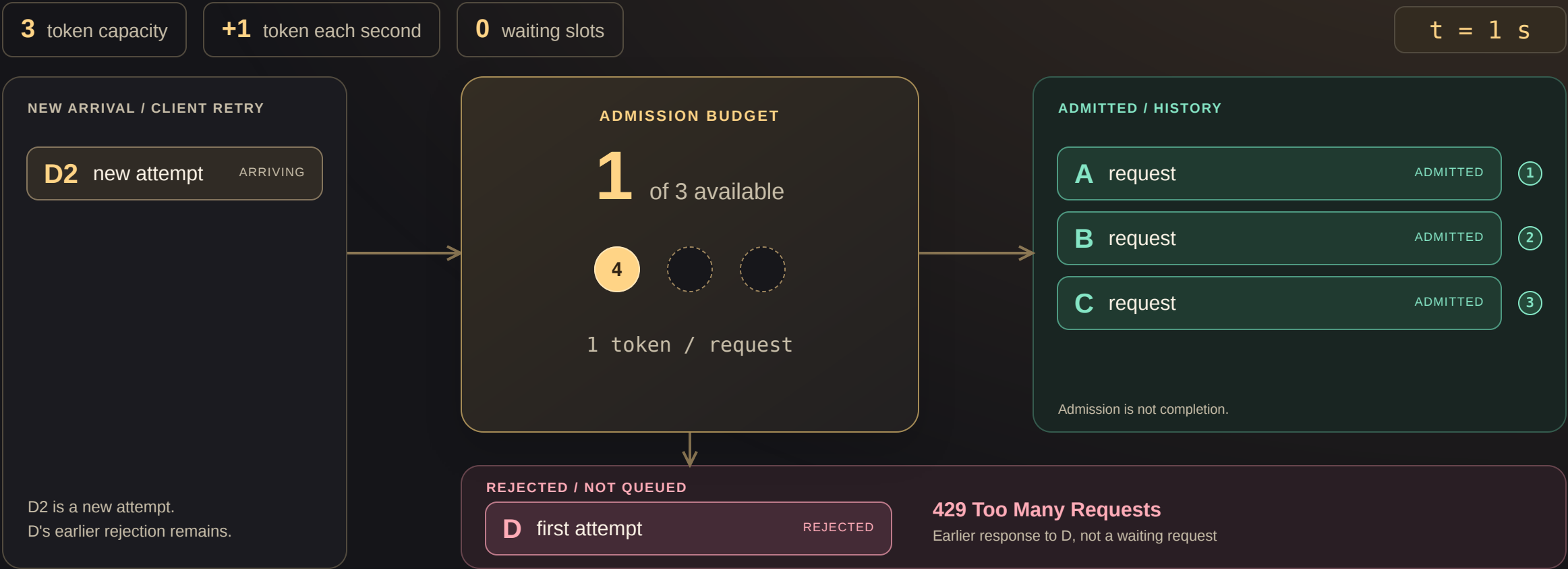
0 TOKENS. Rejecting D does not spend a token.

ADMITTED **3** / REJECTED **1**

429 and Retry-After are policy choices here, not universal defaults for every rate limiter.

The clock creates the next token.

One second passes. A new token appears, and the client sends a new attempt: D2.



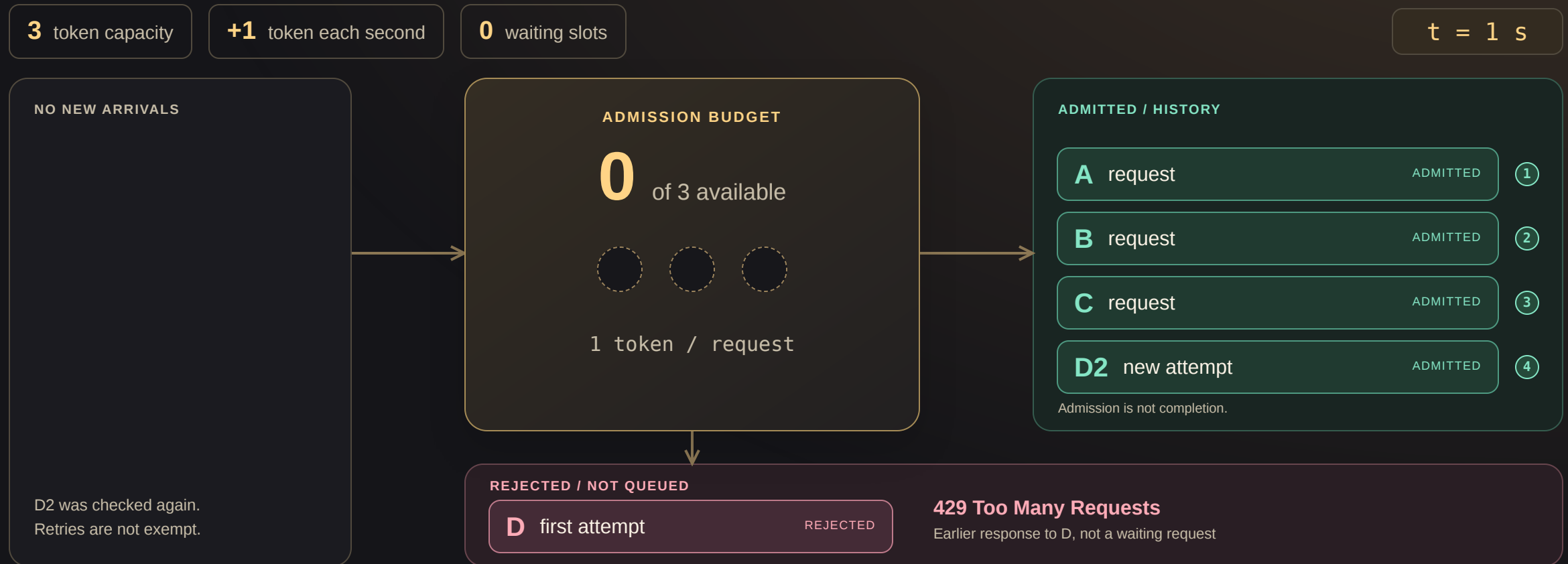
0 + 1 = 1. Request completion did not refill this bucket.

ADMITTED 3 / REJECTED 1

A rate limit budgets admission over time. It does not directly limit concurrent in-flight work.

The retry earns the next token.

D2 is admitted this time. The original rejected attempt is still part of the history.



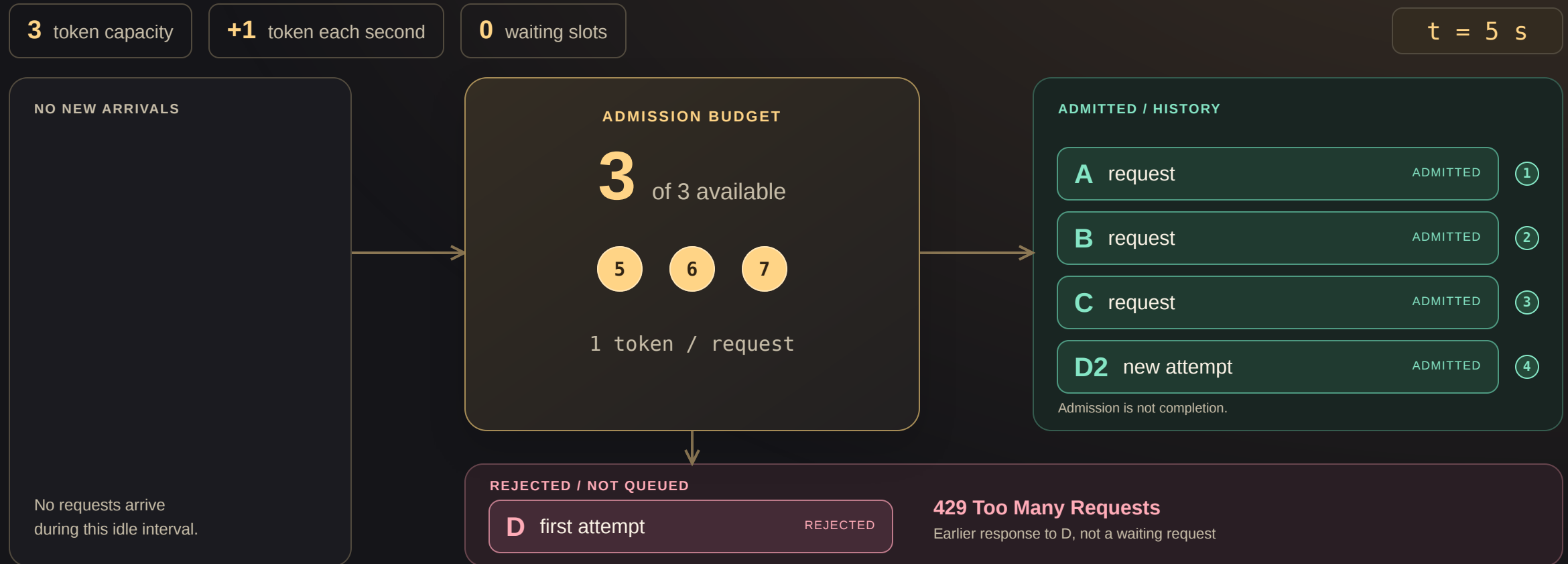
1 - 1 = 0. Four admissions used three initial tokens plus one refill.

ADMITTED **4** / REJECTED **1**

There are no competing arrivals in this example. Production retries still need backoff and careful policy handling.

Idle time never makes an infinite bucket.

After several quiet seconds, the bucket is full again. Additional refill is discarded at the capacity limit.



$\text{MIN}(3, 0 + 4) = 3$. The cap bounds the next burst.

ADMITTED **4** / REJECTED **1**