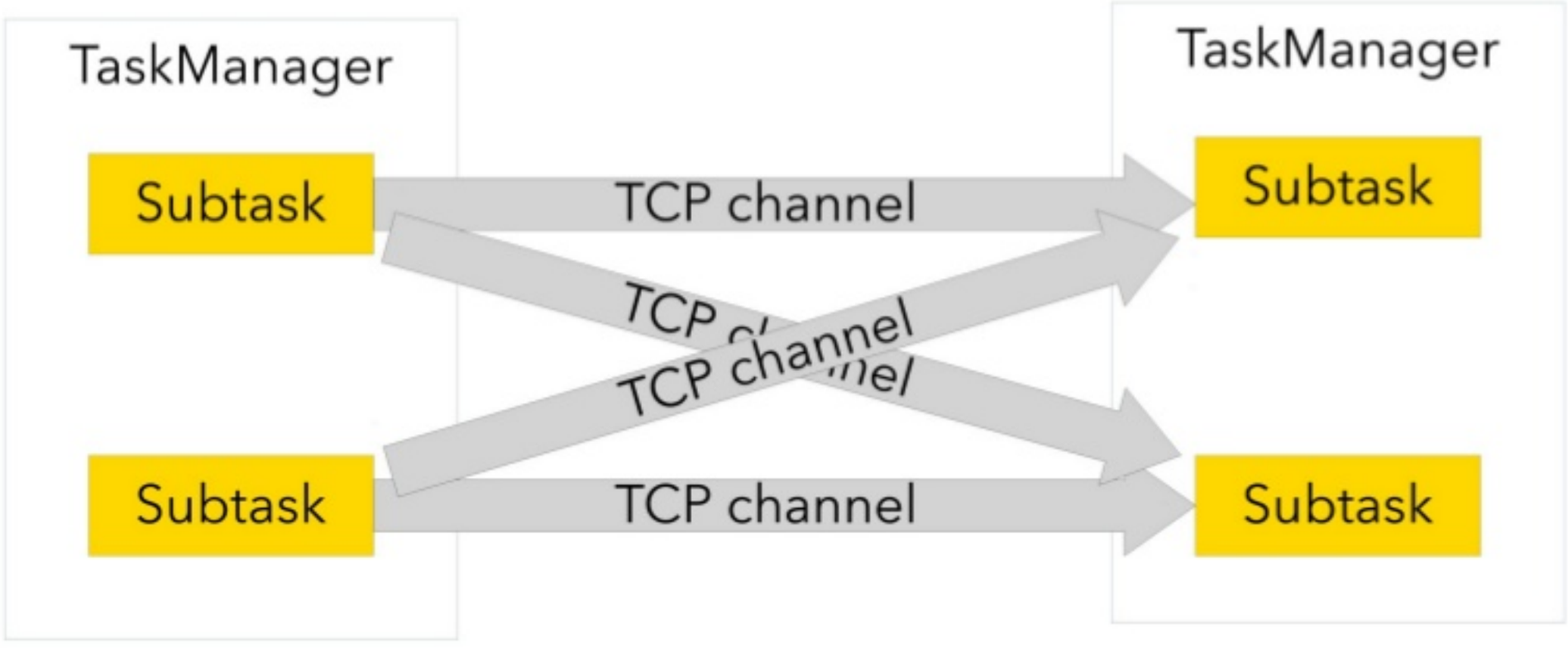
Building a network stack for optimal throughput / low-latency trade-offs

Nico Kruber
nico@data-artisans.com

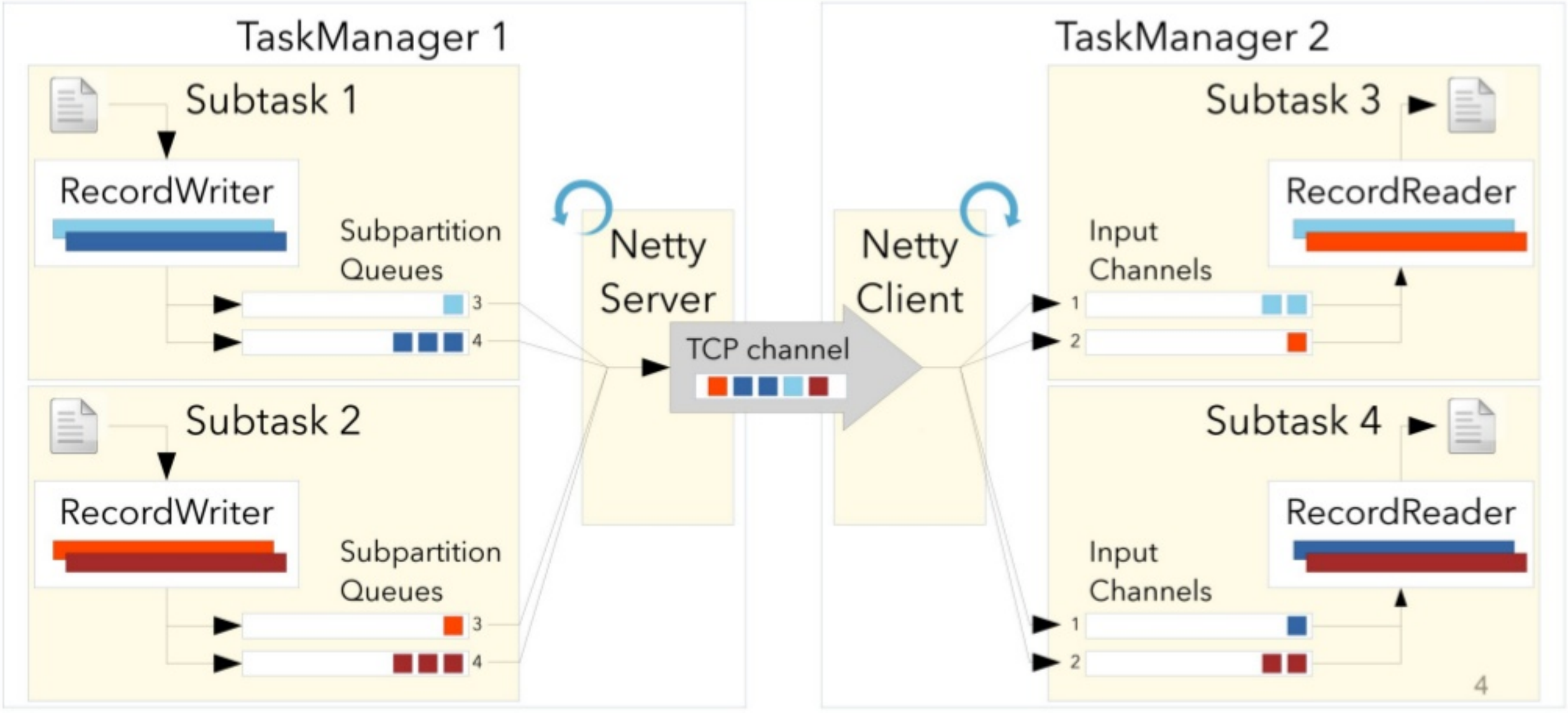
dataArtisans

Flink 1.3 Network Stack

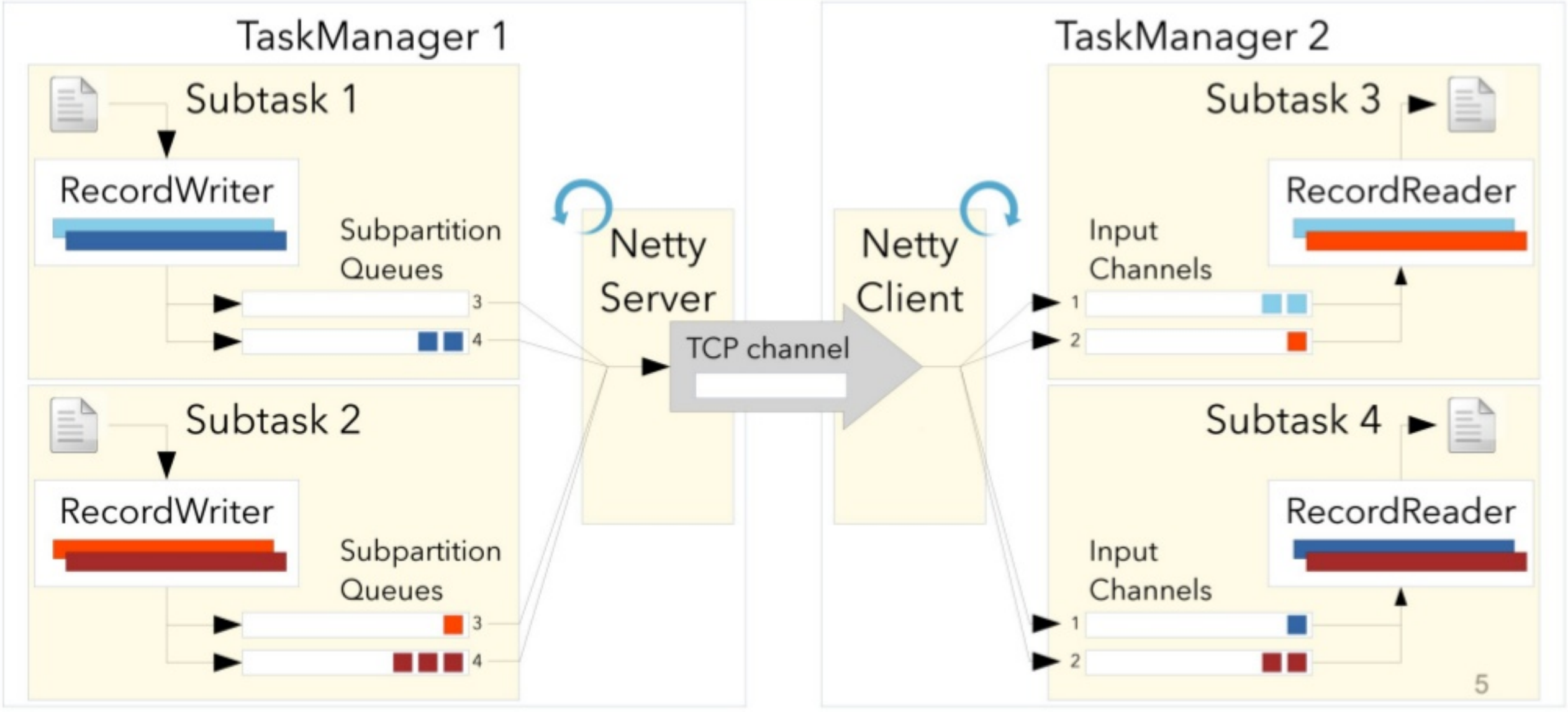
What Operators Expect...



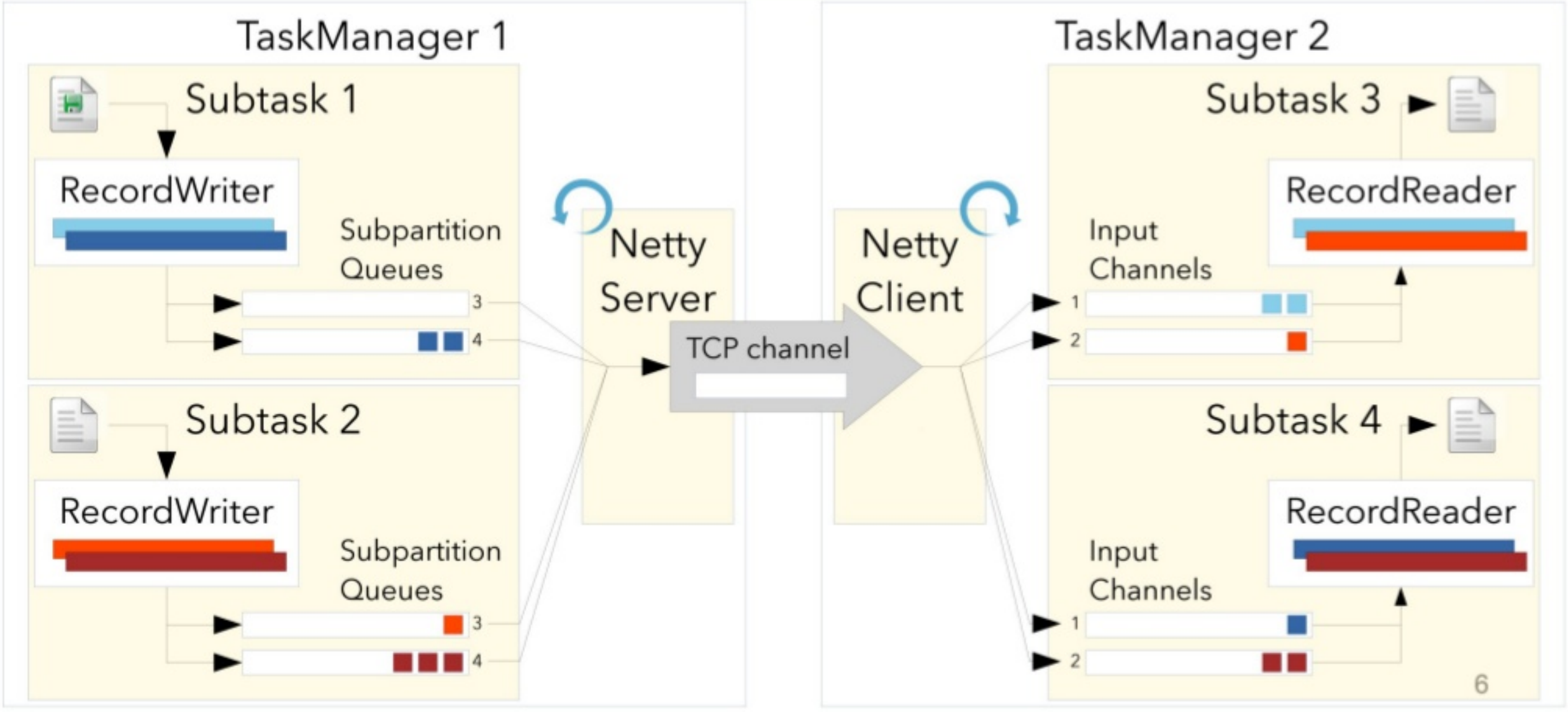
...and what's really going on!



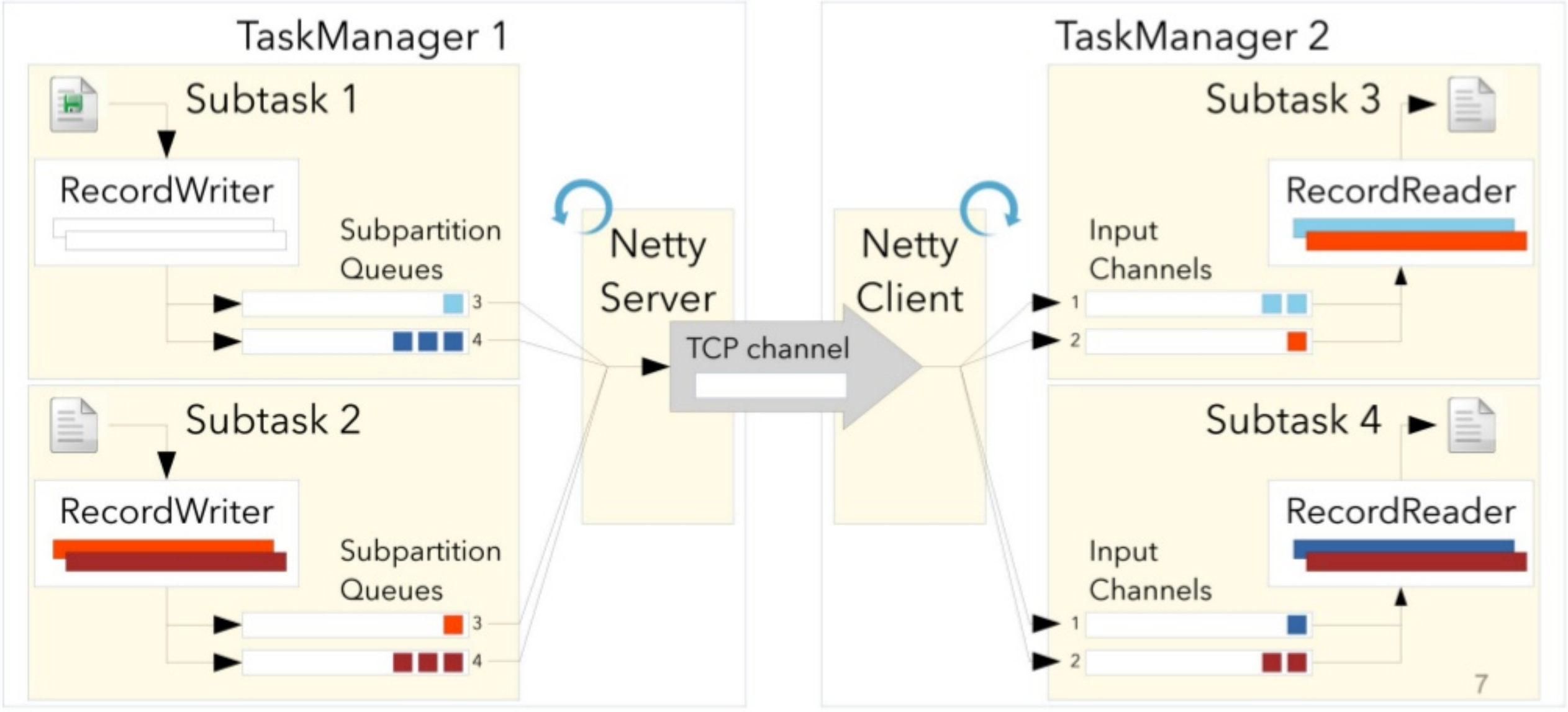
Record Flow with Checkpointing



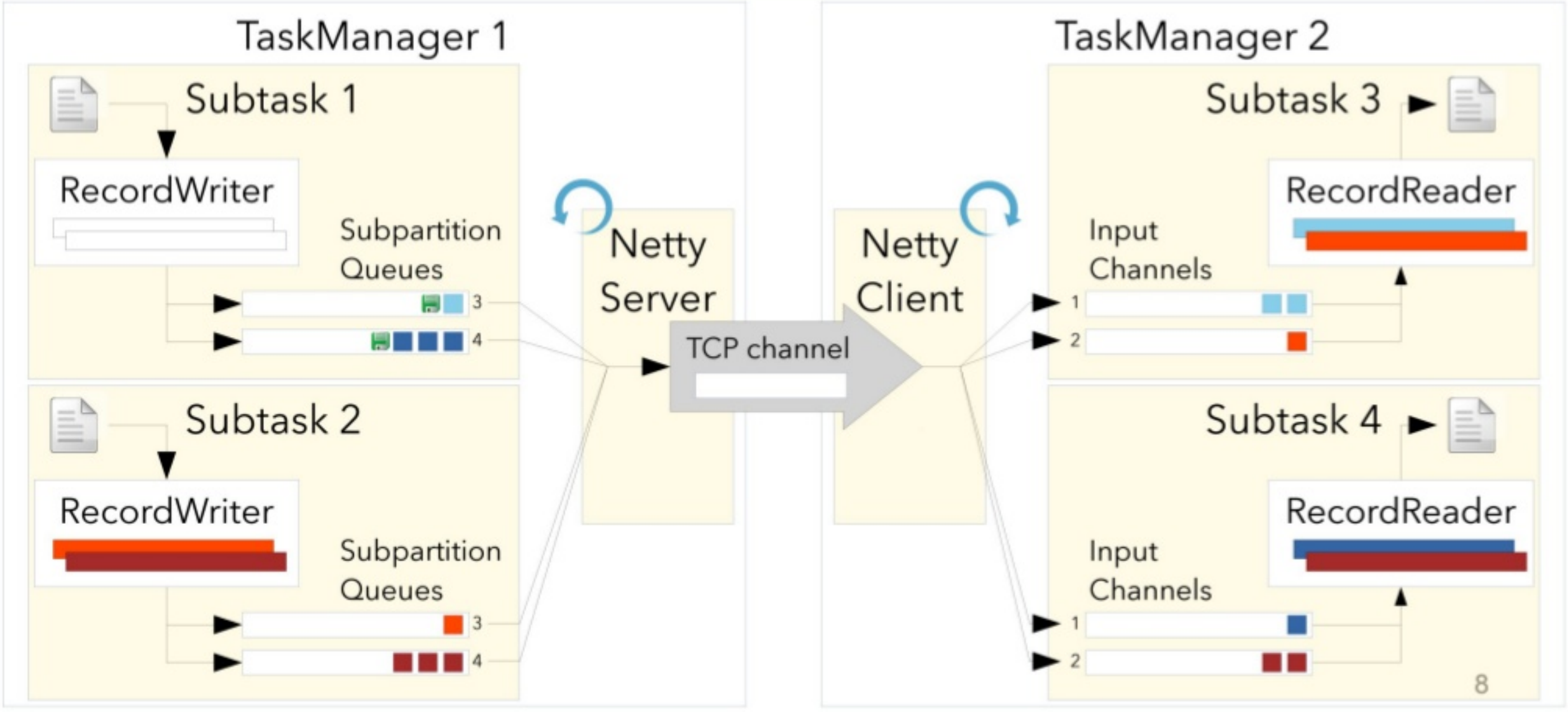
Record Flow with Checkpointing



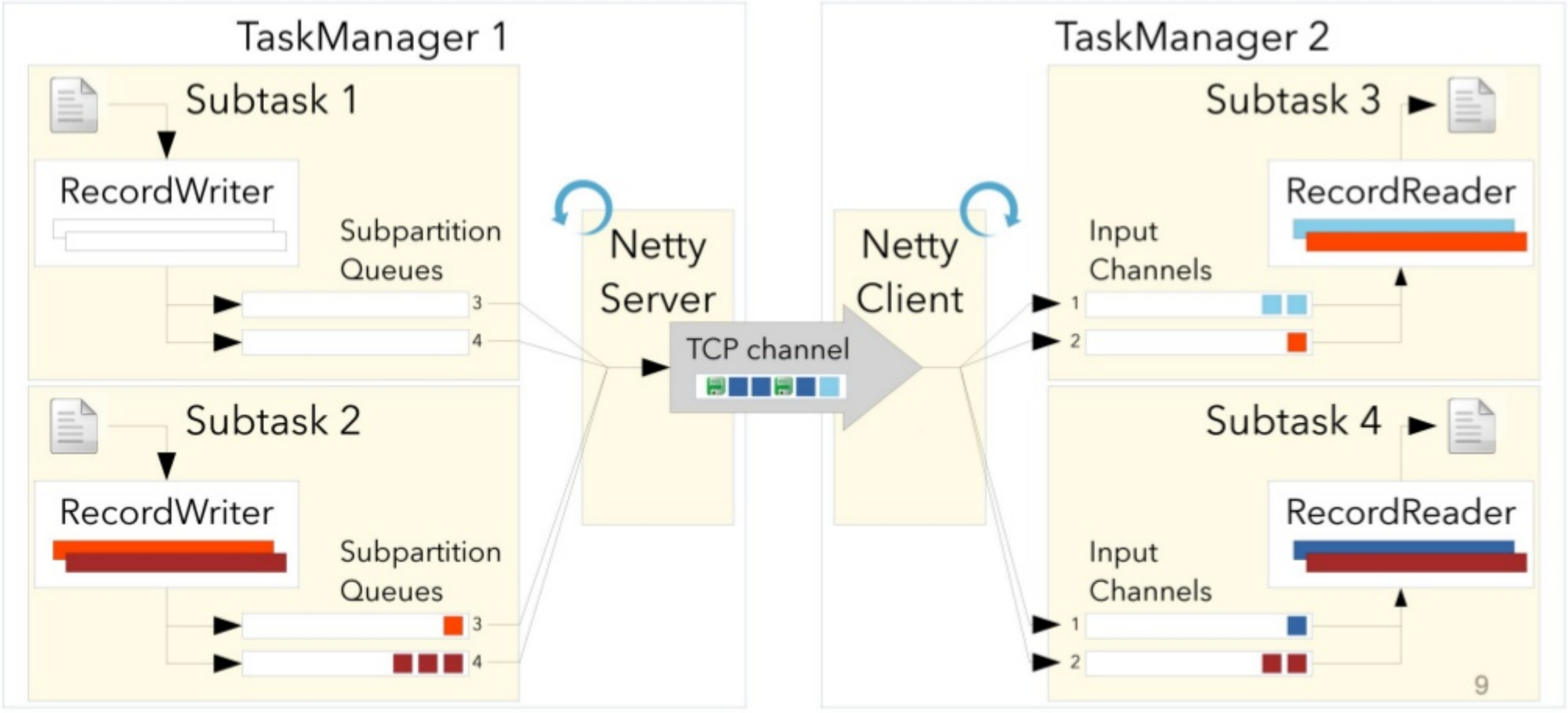
Record Flow with Checkpointing



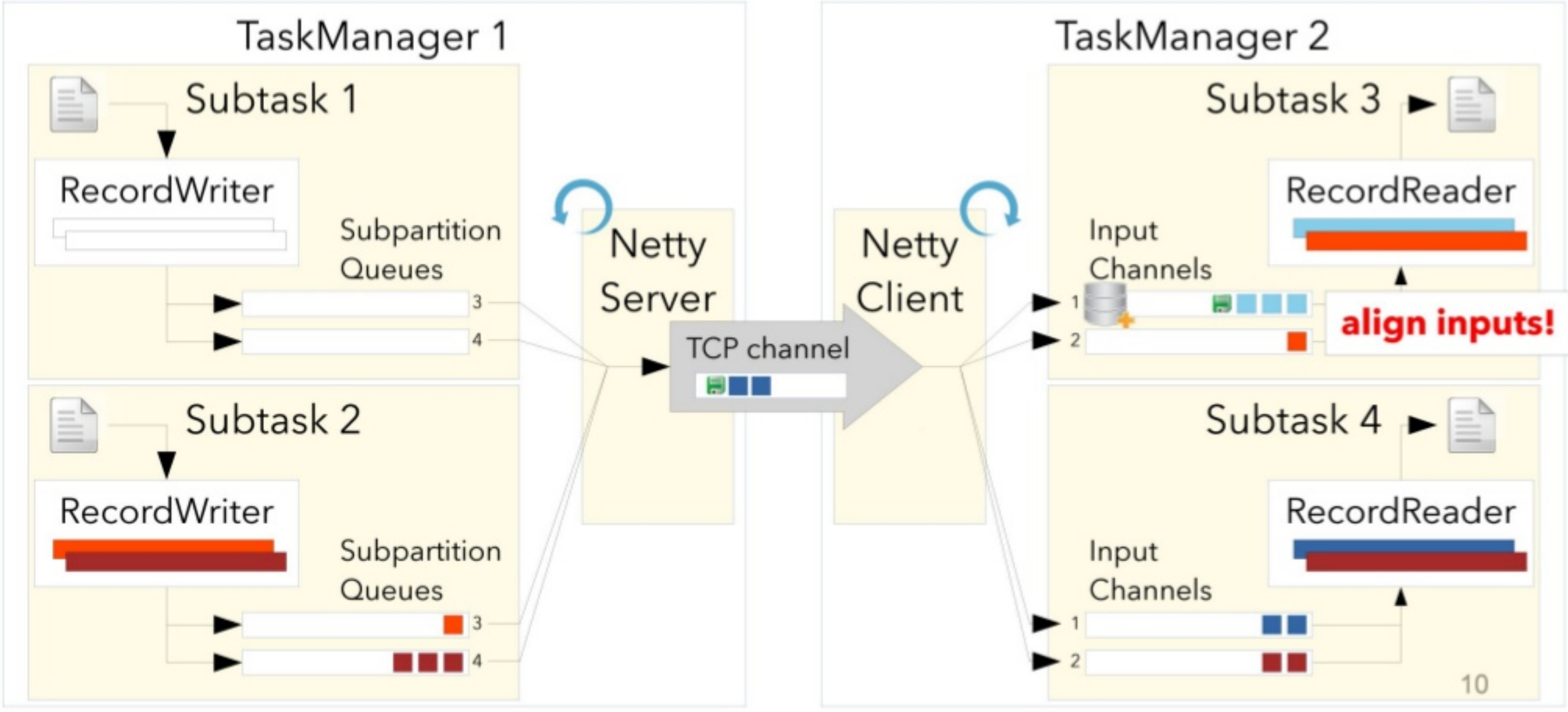
Record Flow with Checkpointing



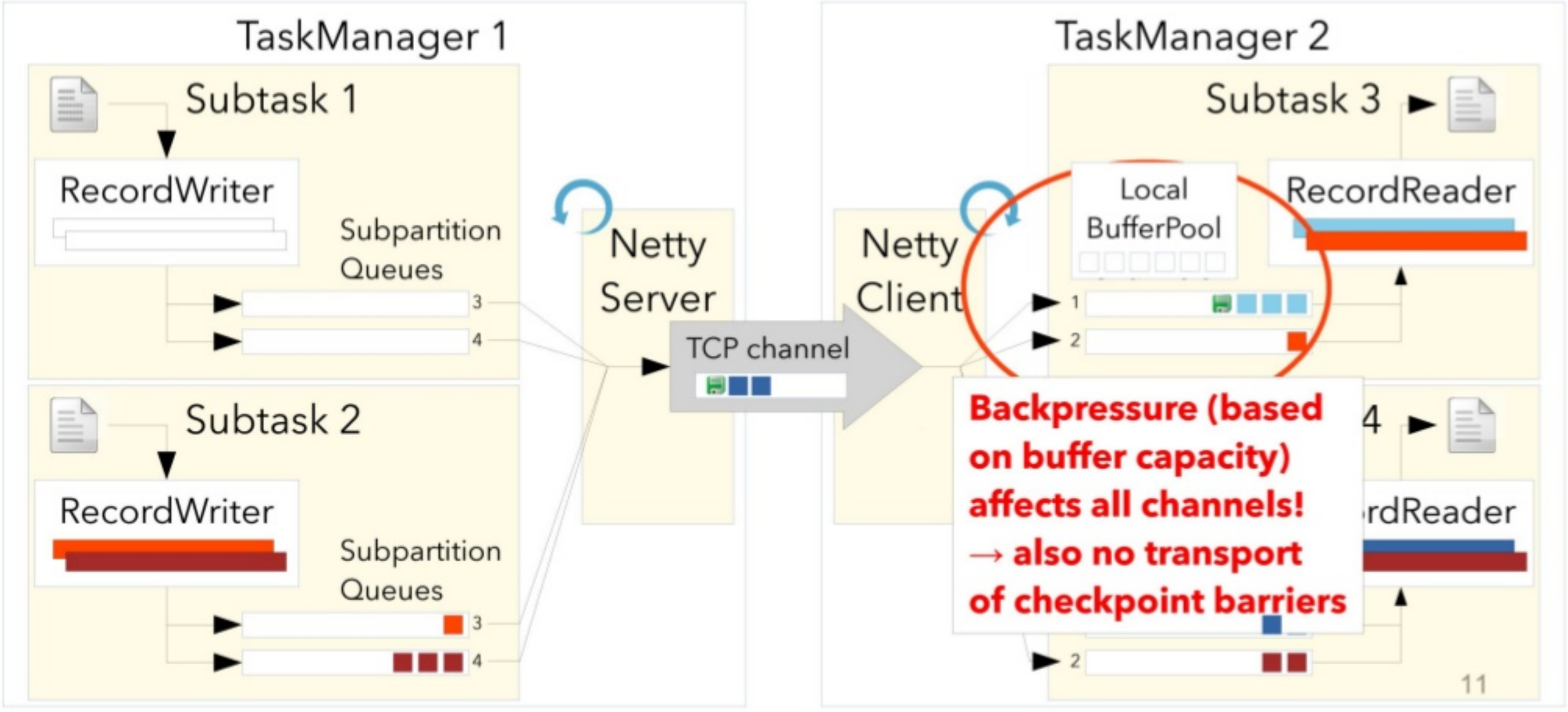
Record Flow with Checkpointing



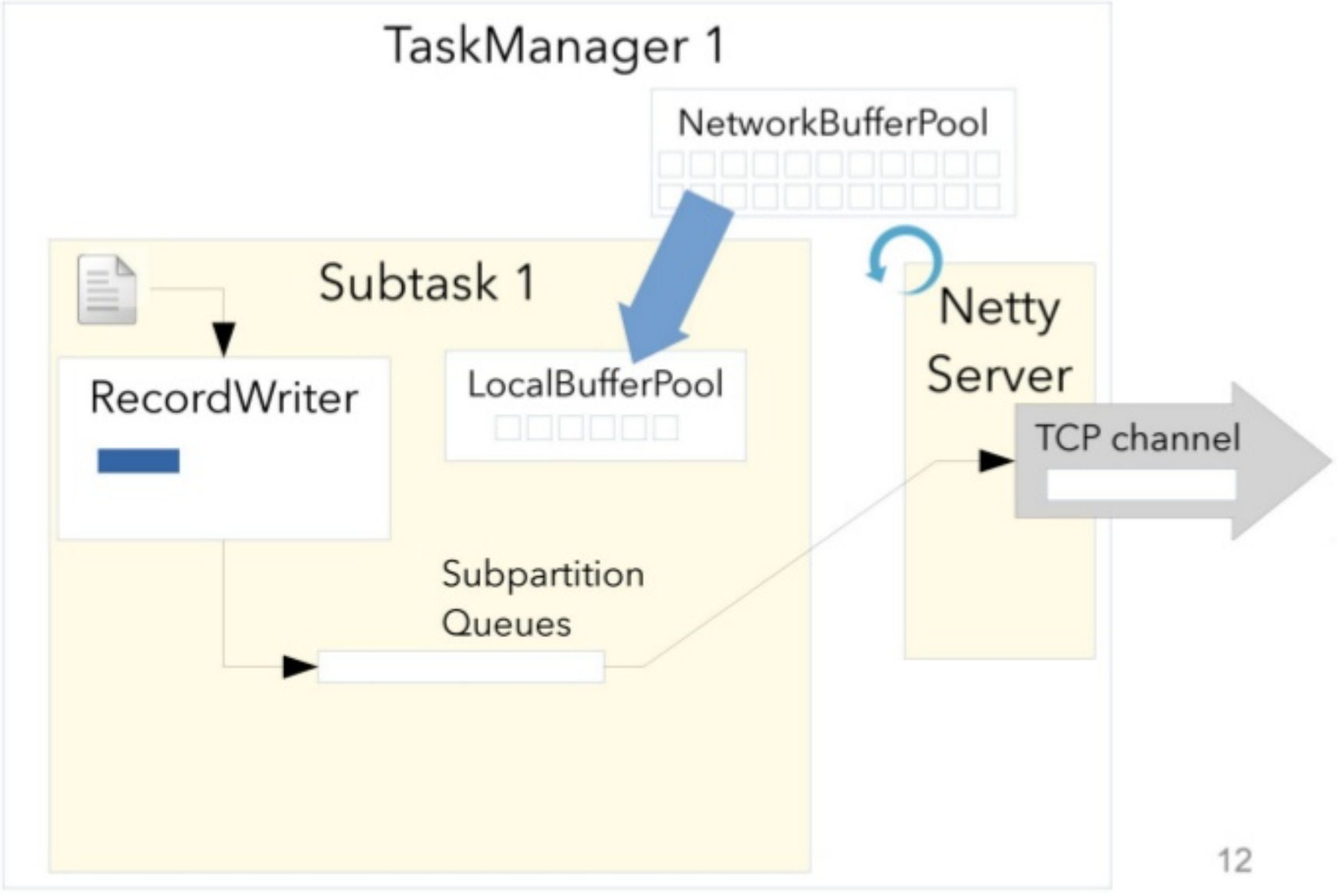
Record Flow with Checkpointing



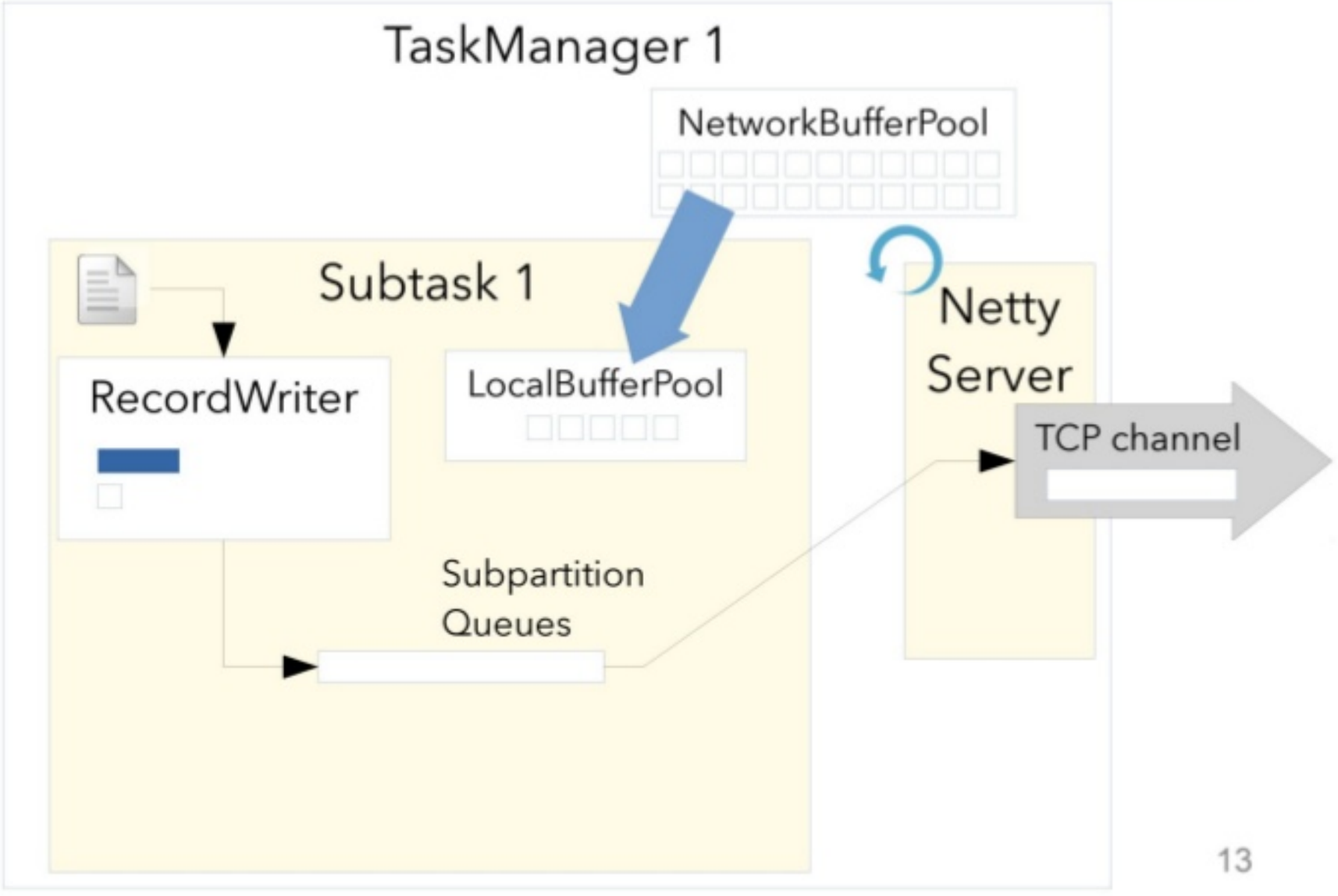
Record Flow with Checkpointing



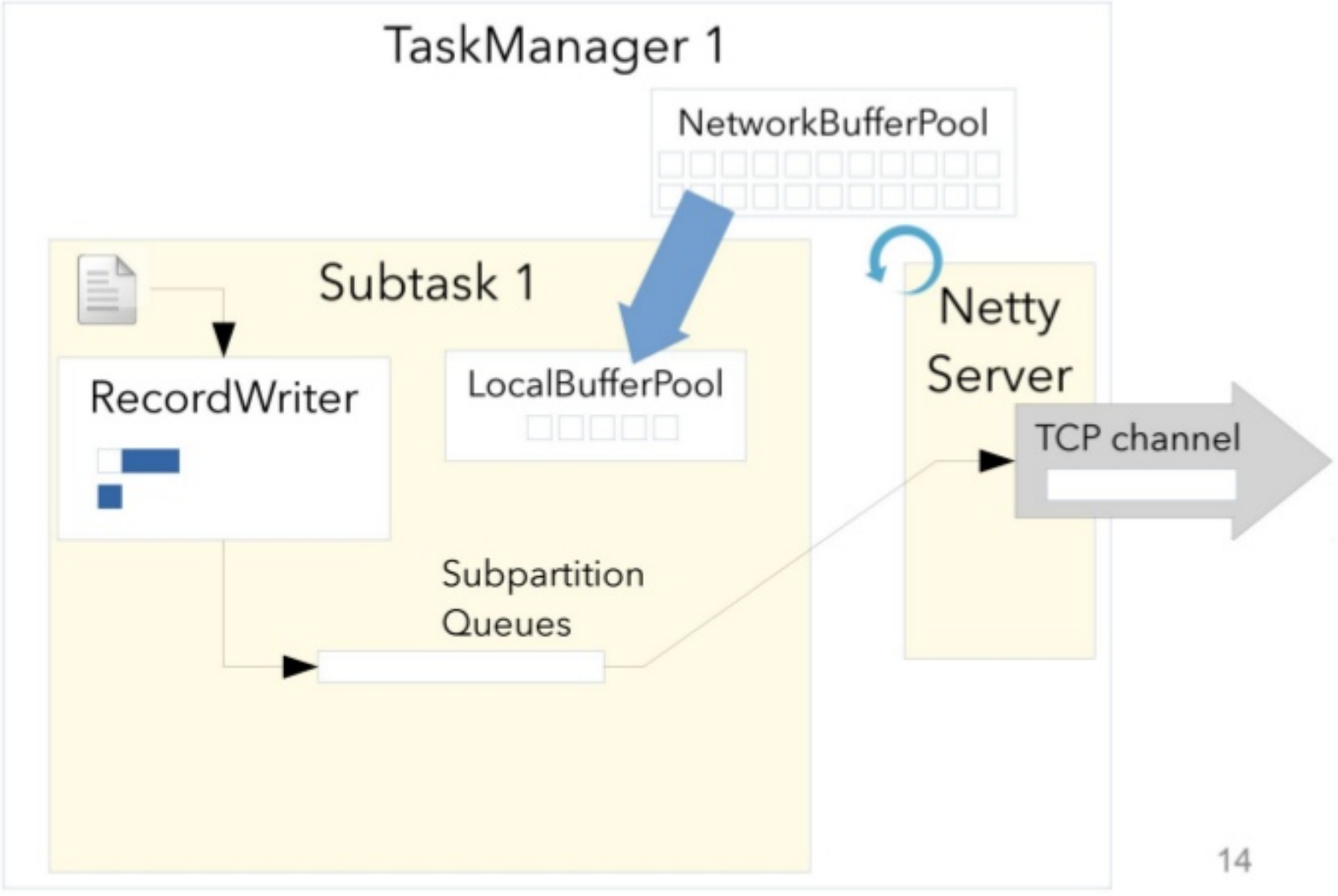
Network Buffers and Output Flusher



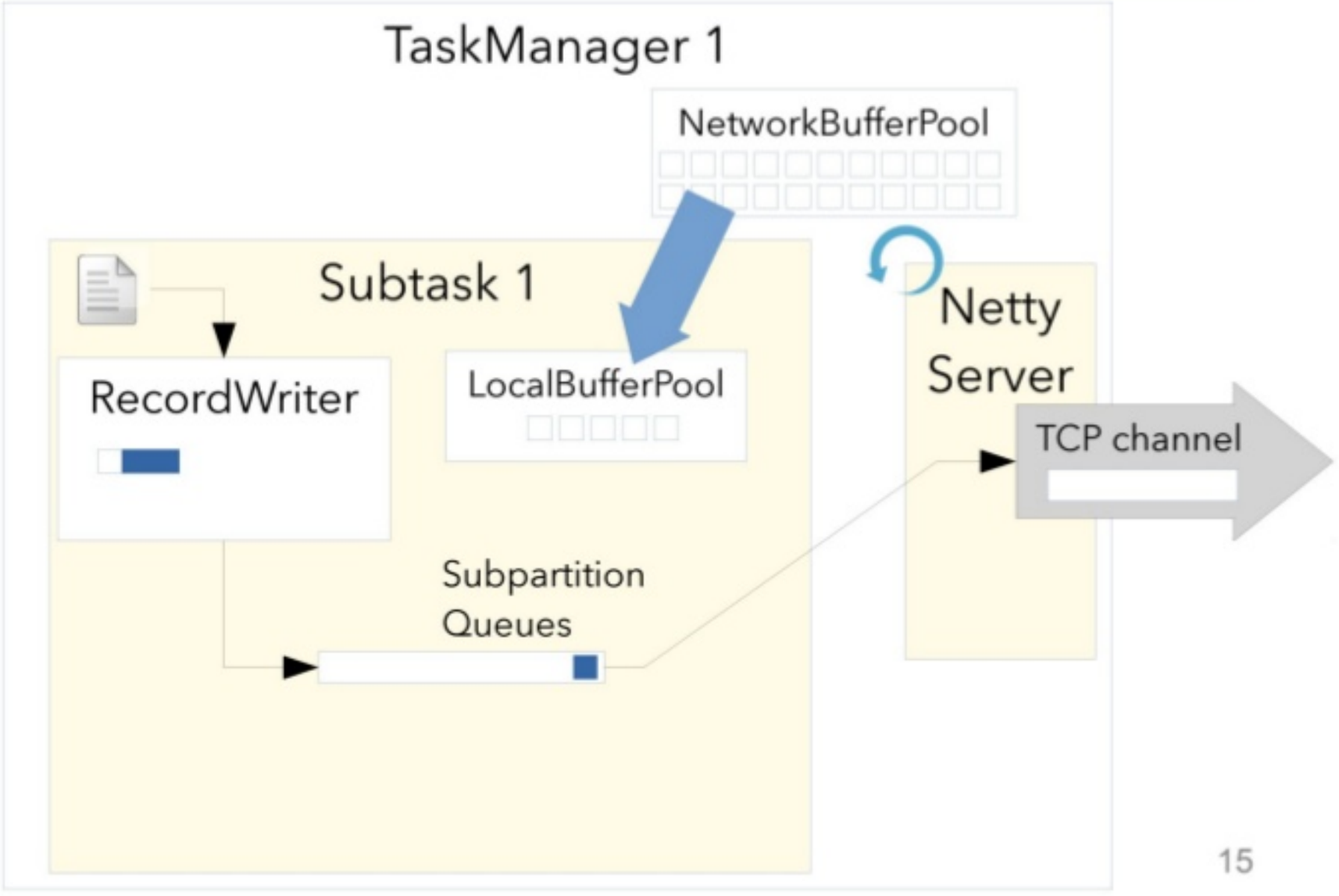
Network Buffers and Output Flusher



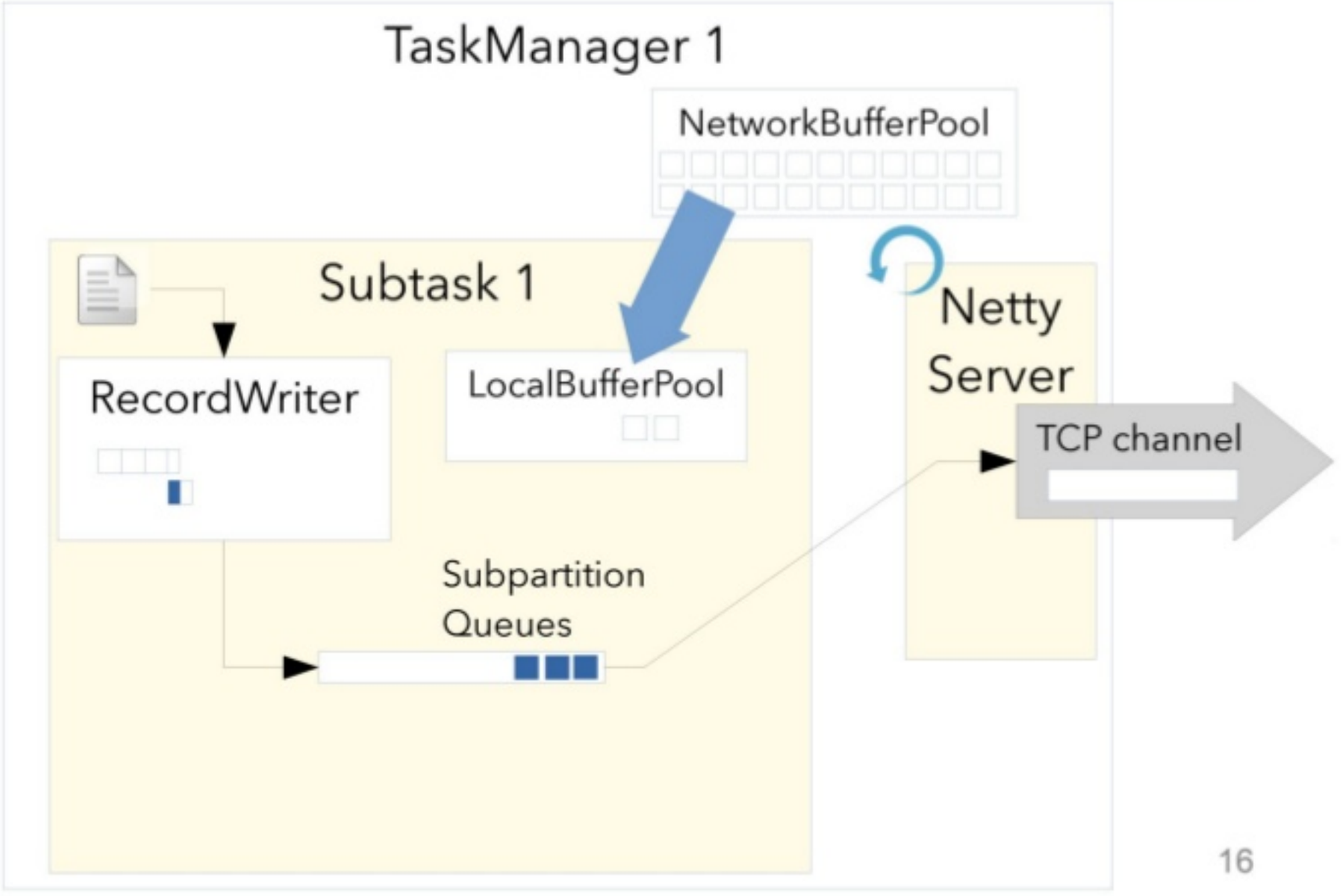
Network Buffers and Output Flusher



Network Buffers and Output Flusher

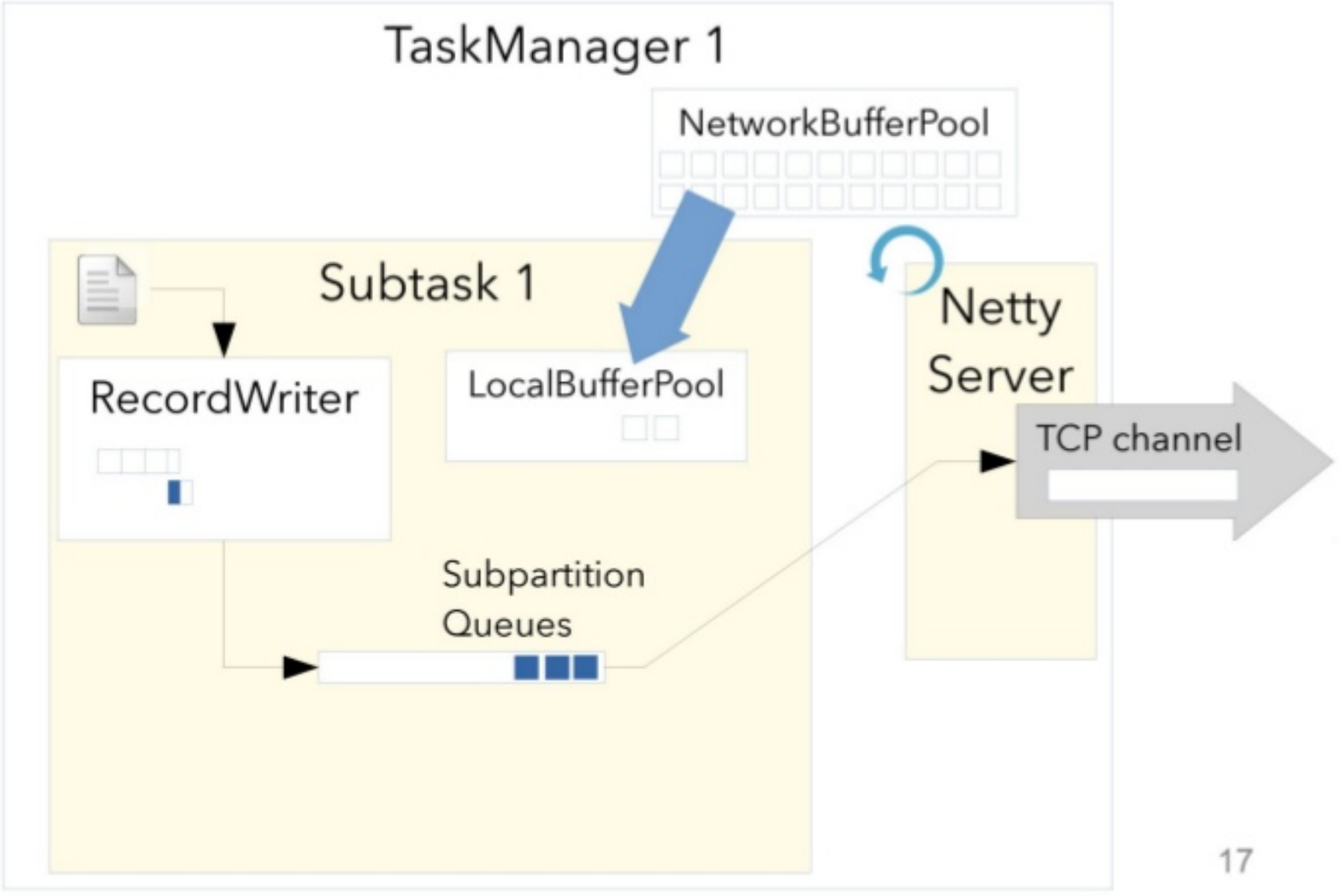


Network Buffers and Output Flusher



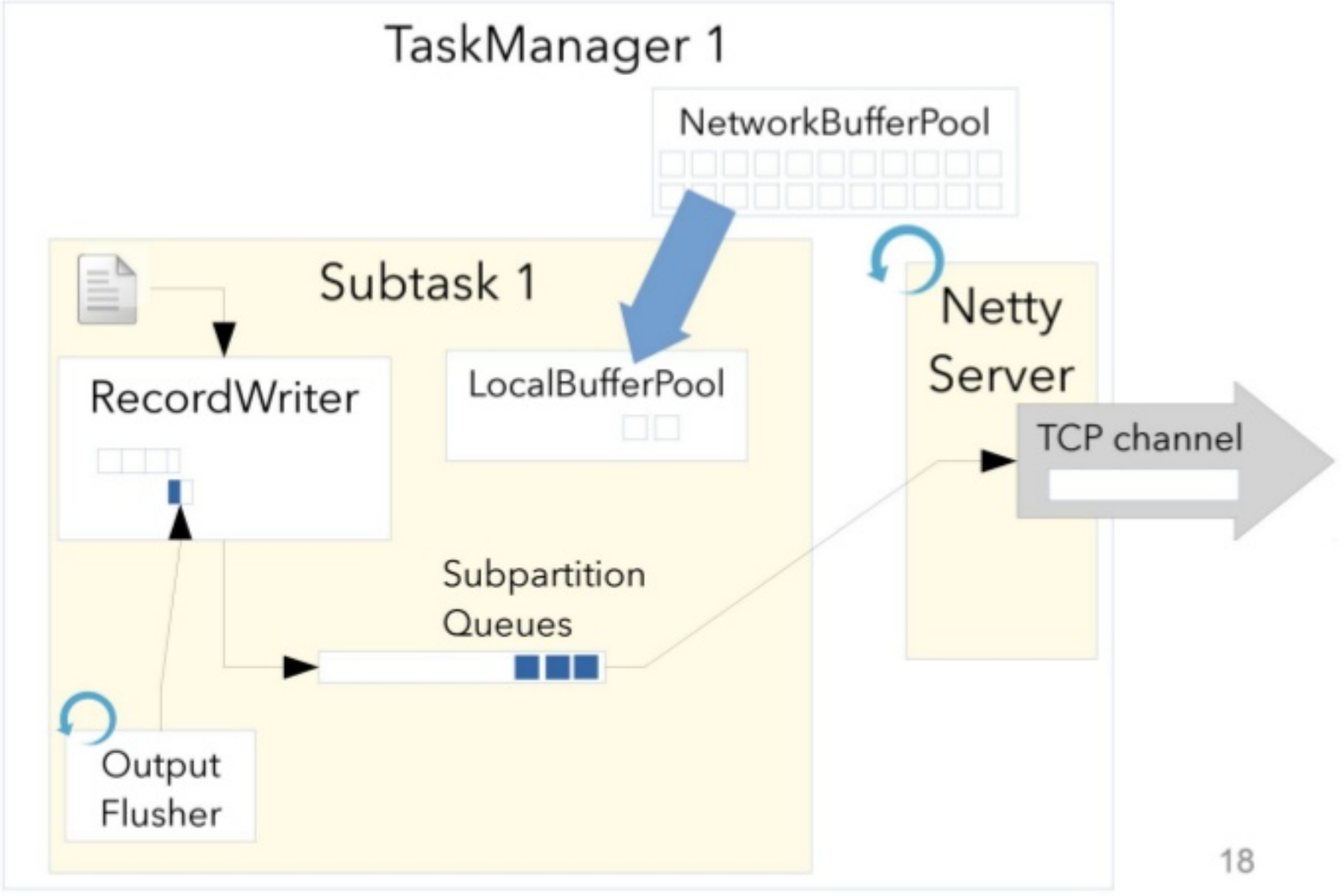
Network Buffers and Output Flusher

- Variant 1:
  send partial buffer
  right away
- Variant 2:
  wait for buffer to
  become full



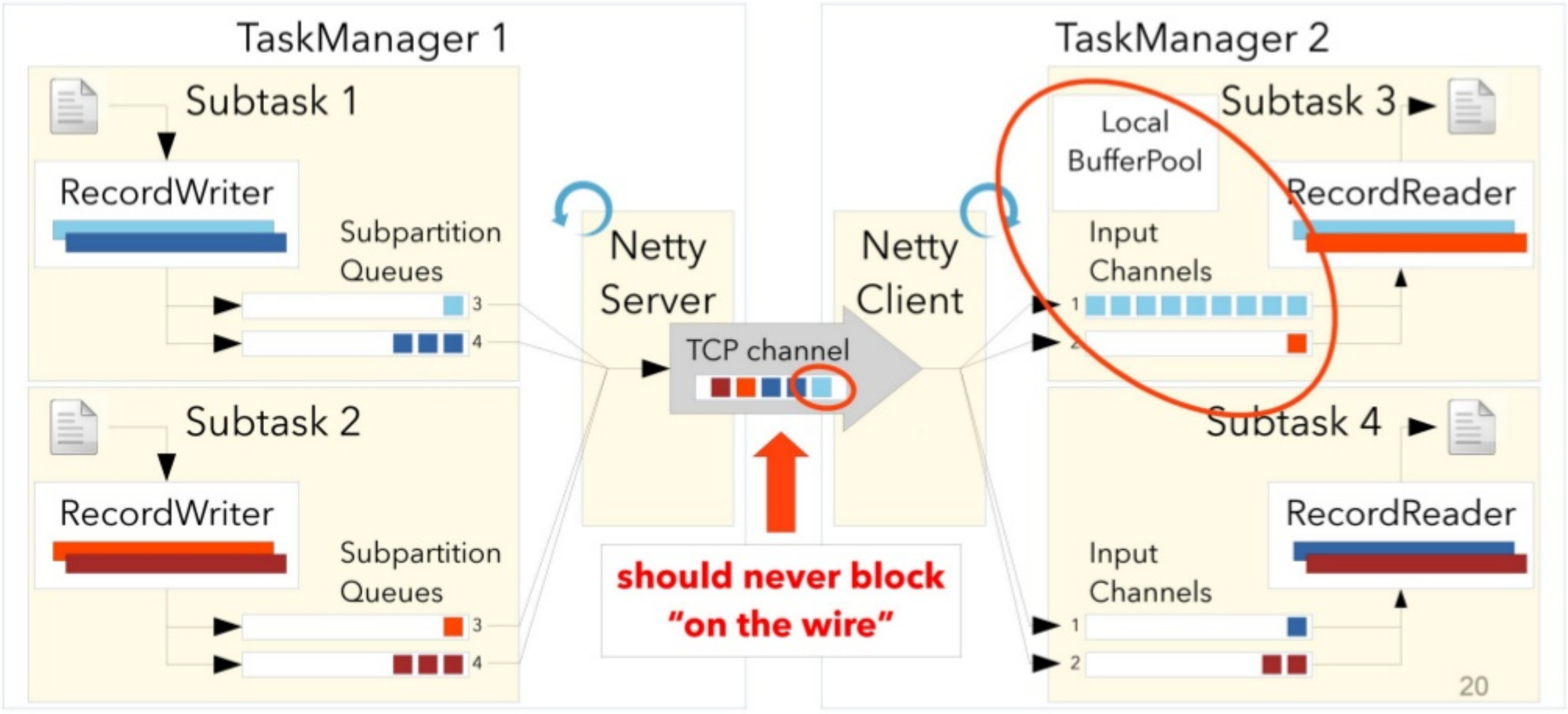
Network Buffers and Output Flusher

- Variant 1:
  send p… …ffer
  ri… …ay
  low throughput!
- Variant 2:
  wait fo… … to
  be… …e full
  high latency!
- Variant 3:
  flush buffers when full or after timeout (default: 100ms)

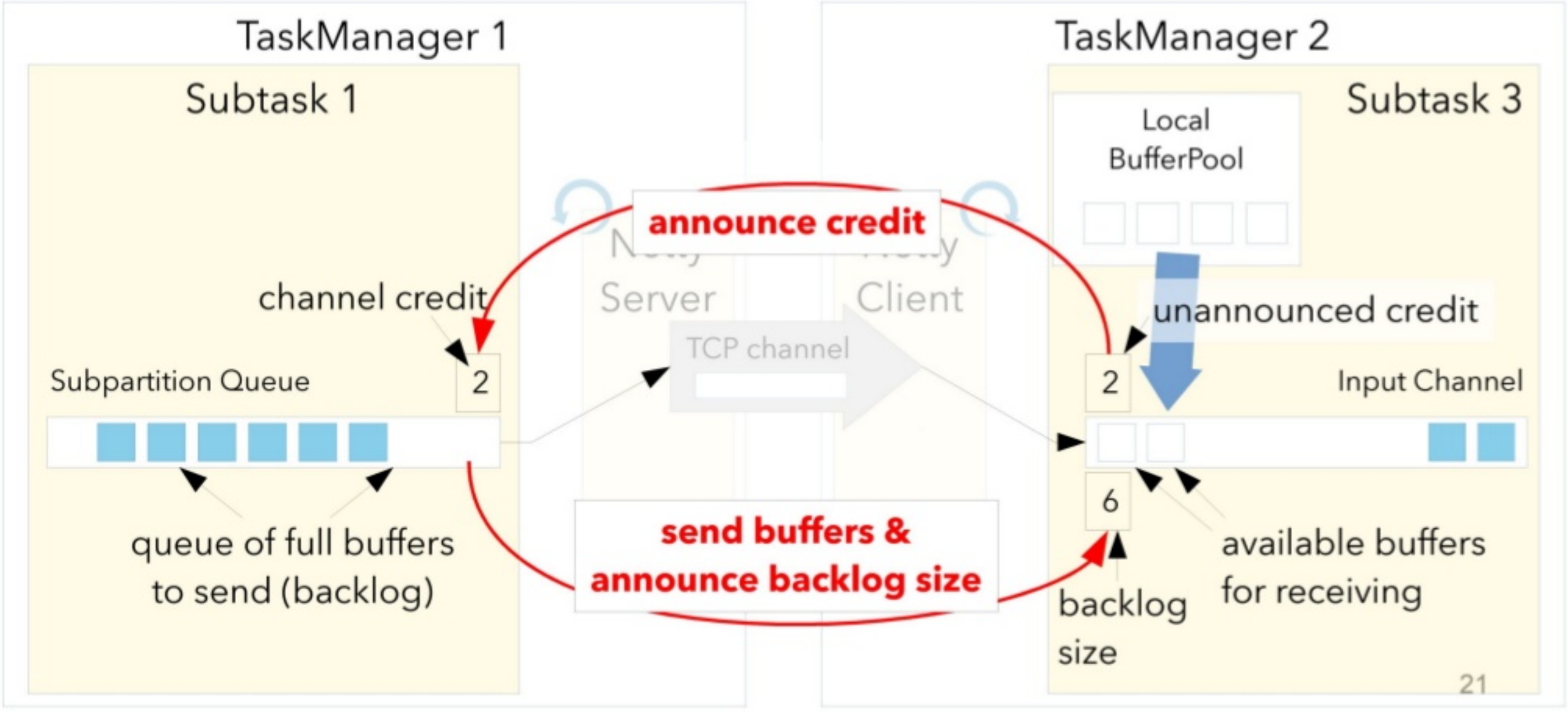


Adding Flow Control

Backpressure and Flow Control



Credit-Based Flow Control (FLINK-7282)



Credit-Based Flow Control (FLINK-7282)

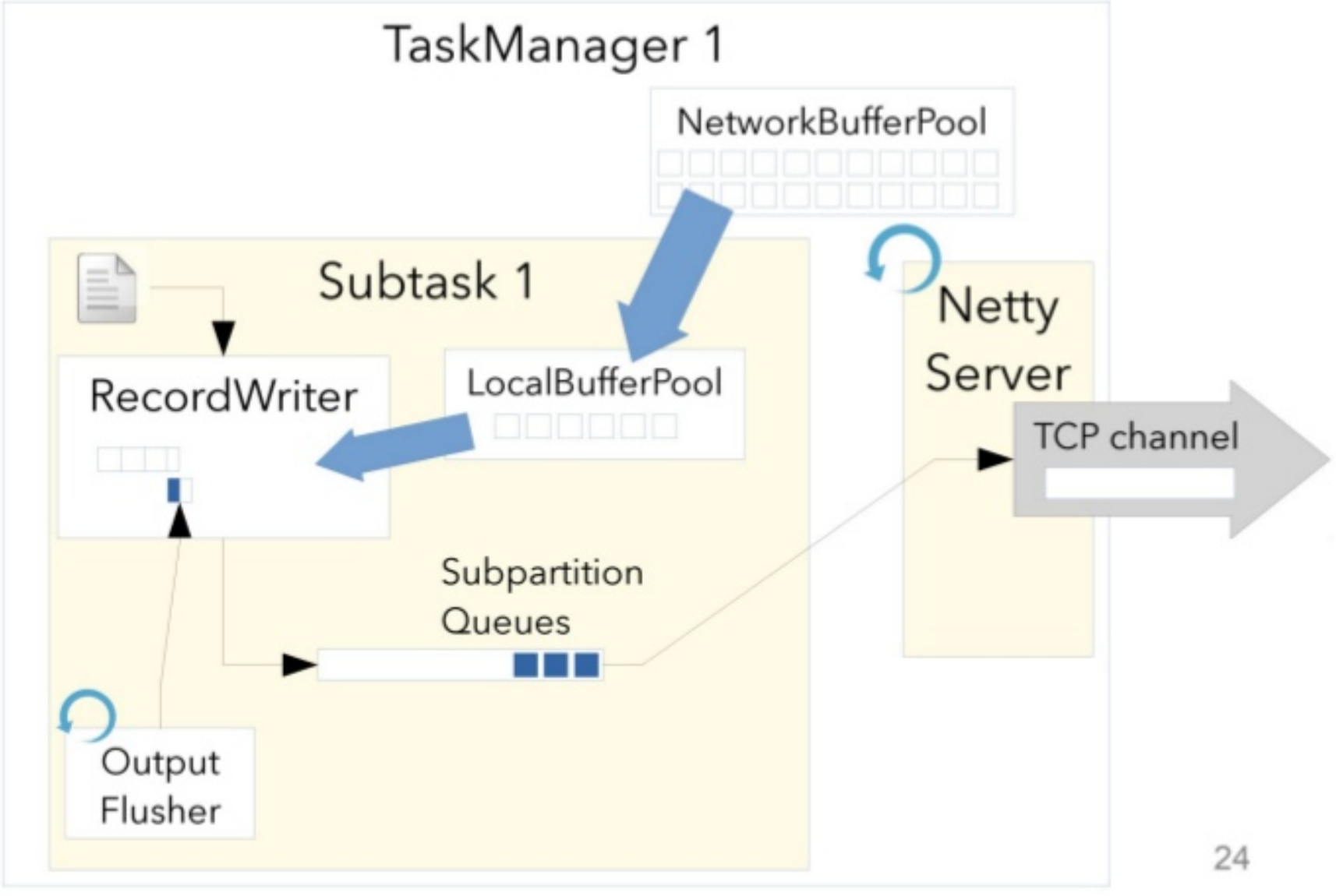
- no messages on the wire for which the receiver does not have buffers
  → single channel no longer stalls multiplexed input
- fine-grained backpressure control
- Improves checkpoint alignments
- cost: additional announce messages → piggy-backed into BufferRequest and BufferResponse messages

Event-Driven Network IO

How to remove the Output Flusher

Event-Driven Flushes (FLINK-7612)

- Variant 3:
  flush buffers when full
  or after timeout
  (default: 100ms)

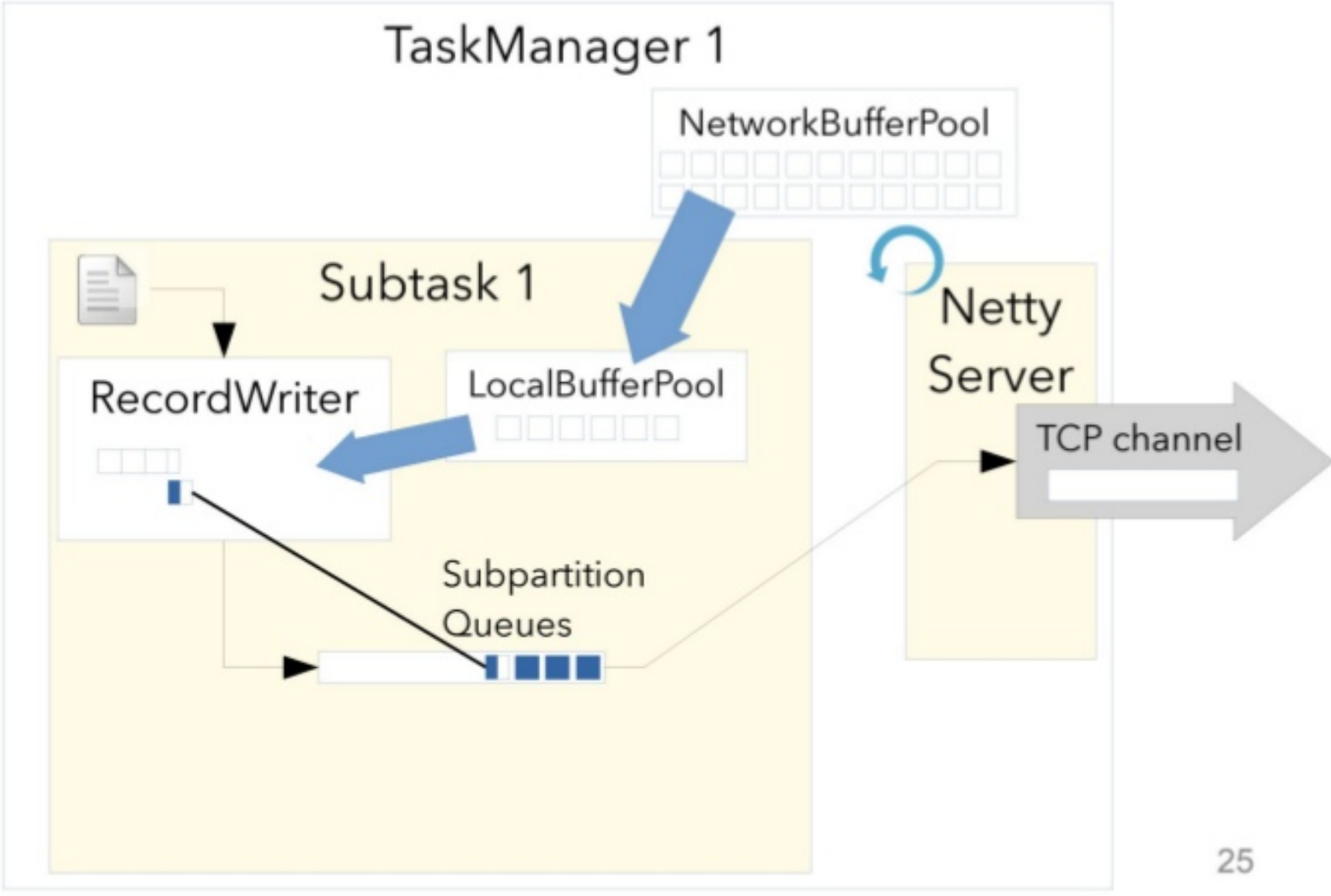


Event-Driven Flushes (FLINK-7612)

- Variant 3:
  flush buff[illegible]n full or after [illegible](default: 100ms)

Flink 1.3

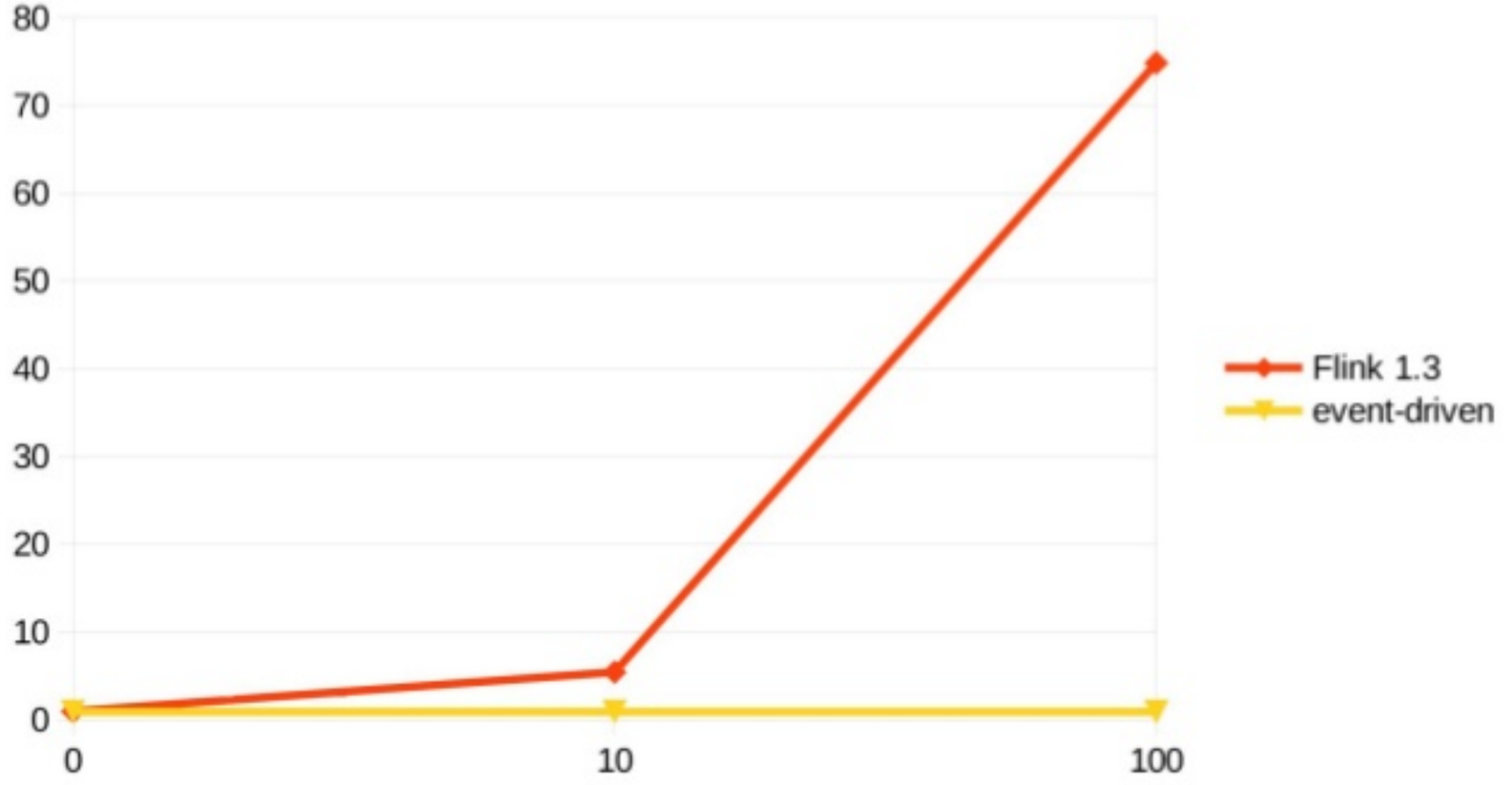
- Variant 4:
  add to subpartition queue but continue writing
  - Transmits whenever the network is ready (low latency)
  - Allows buffers to fill when not (high throughput)

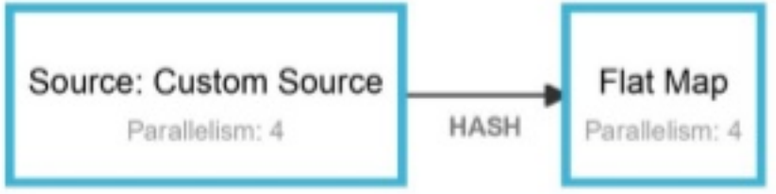


Latency with/out Output Flusher

Slow source:

- 1 event every 100ms



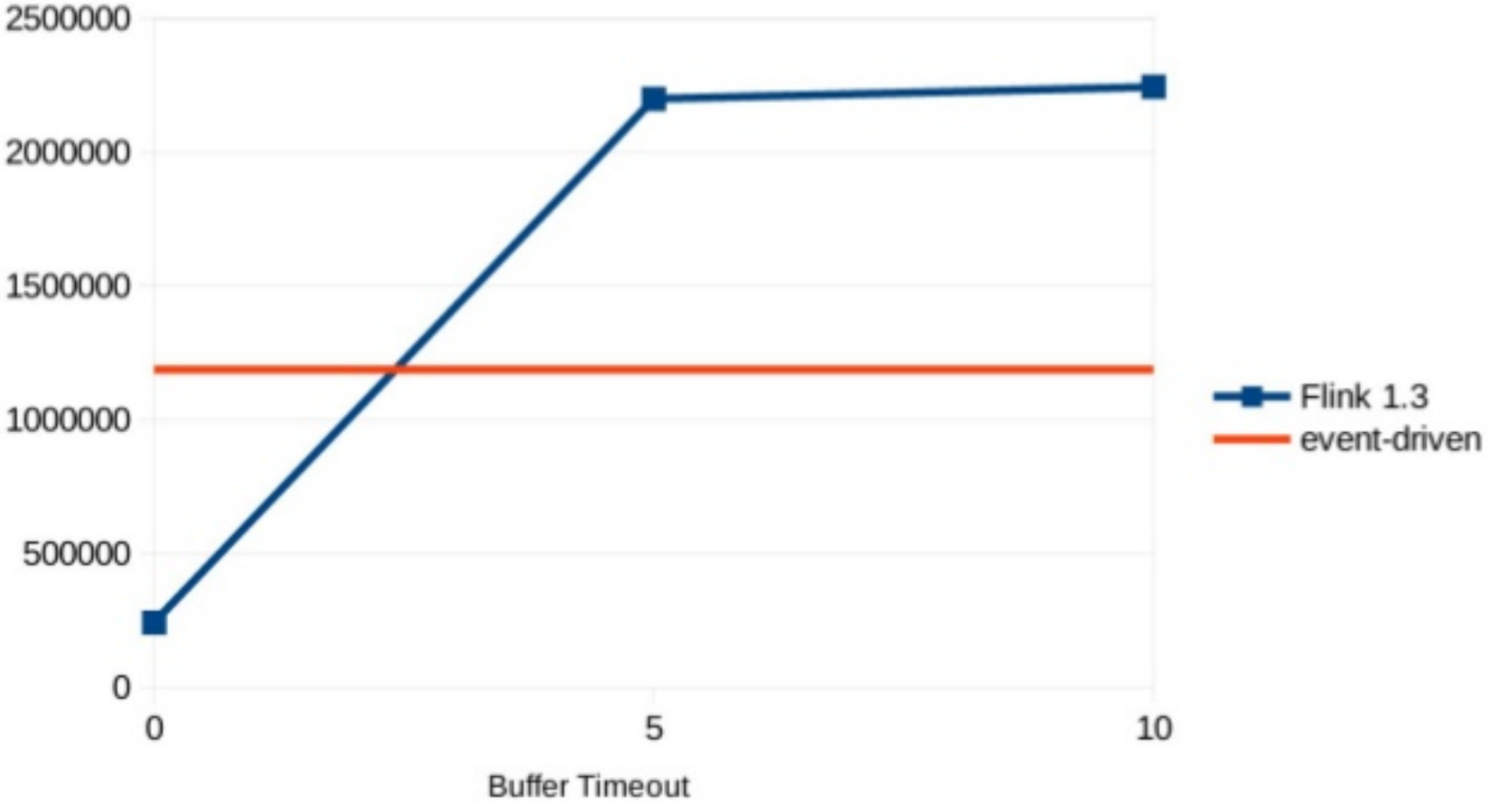


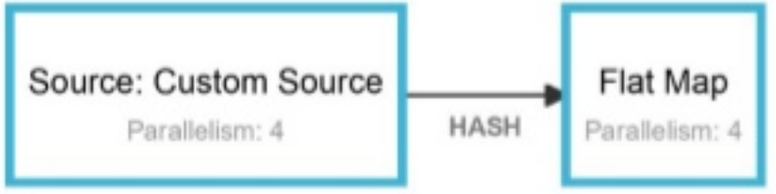
*on Amazon EC2, 5x m4.xlarge (1 JM, 4 TMs)

Throughput with/out Output Flusher

Fast source:

- generates events as fast as possible



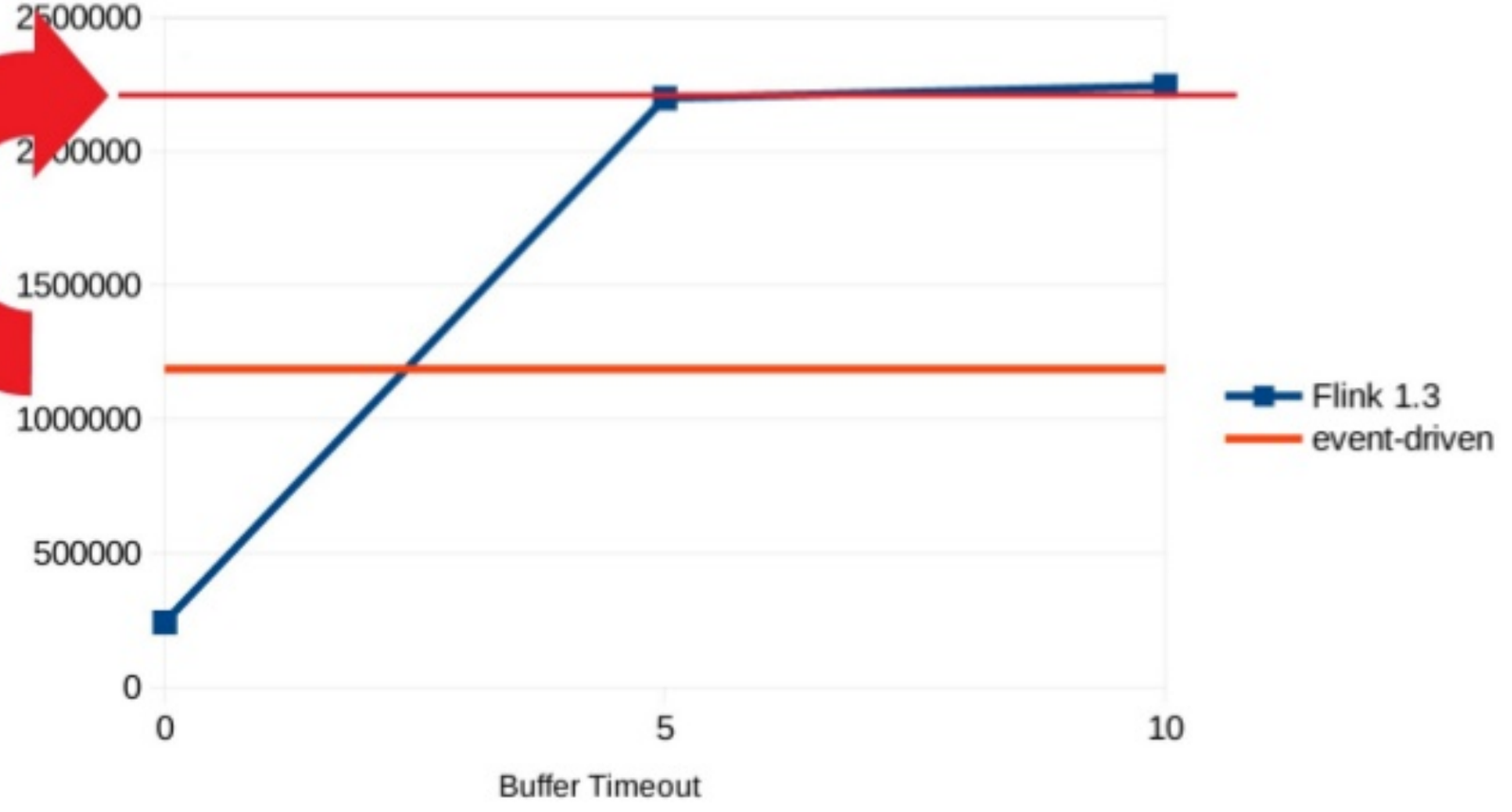


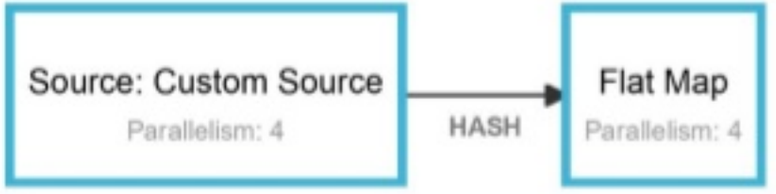
*on Amazon EC2, 5x m4.xlarge (1 JM, 4 TMs)

Throughput with/out Output Flusher

Target this!

Work in progress!
(FLINK-7612)

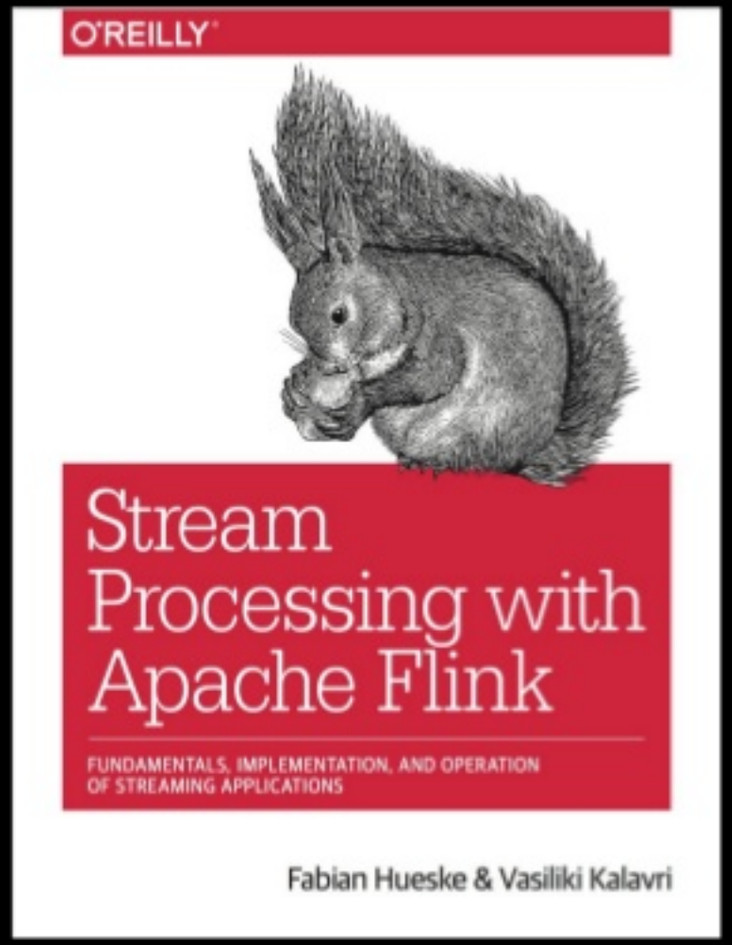




*on Amazon EC2, 5x m4.xlarge (1 JM, 4 TMs)

Event-Driven Flushes (FLINK-7612)

- flush based on network channel availability
- near perfect latency behaviour
- uses the all of the available capacity (expect only minor effects on the throughput)



Thank you!

@ApacheFlink
@dataArtisans

We are hiring!

data-artisans.com/careers