

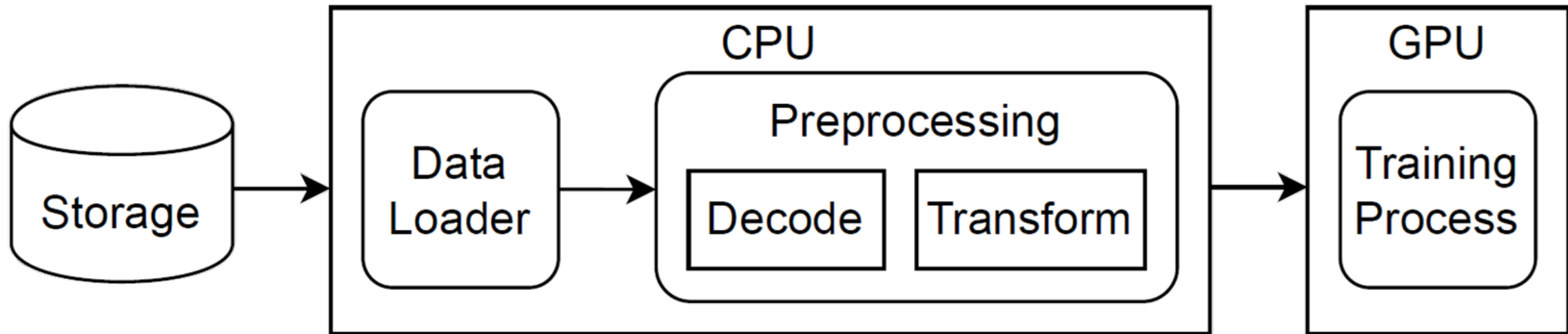
data path in deep learning

Pinar Tözün

Associate Professor, IT University of Copenhagen

pito@itu.dk, pinartozun.com, [@pinartozun](https://twitter.com/pinartozun)

journey of data in deep learning training



CPU feeds the accelerators

- 16-64 cores per GPU
- 96 cores per TPU*

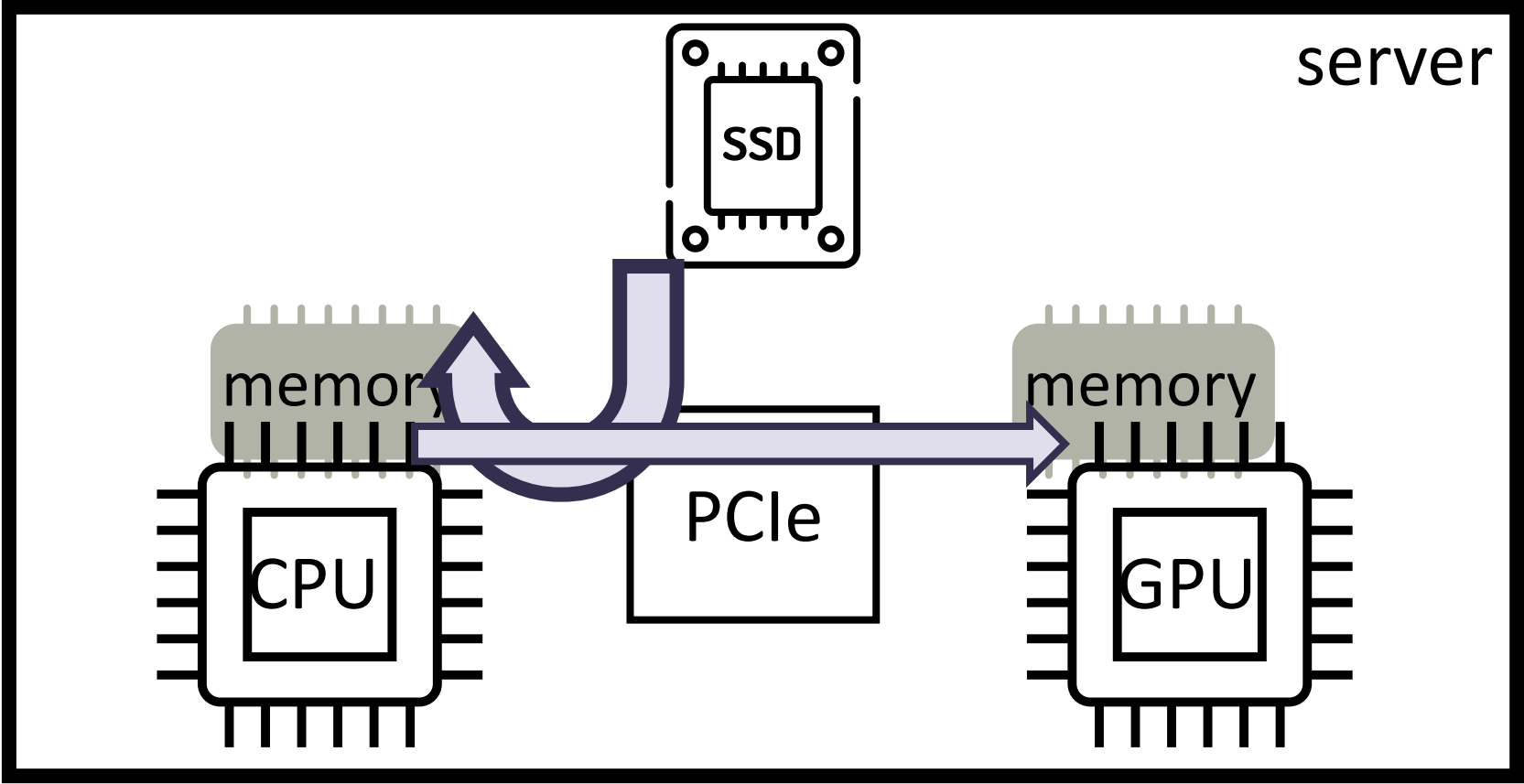
➔ otherwise, accelerator may be underutilized!

*Audibert et al., “tf.data service: A Case for Disaggregating ML Input Data Processing.” ACM SoCC 2023

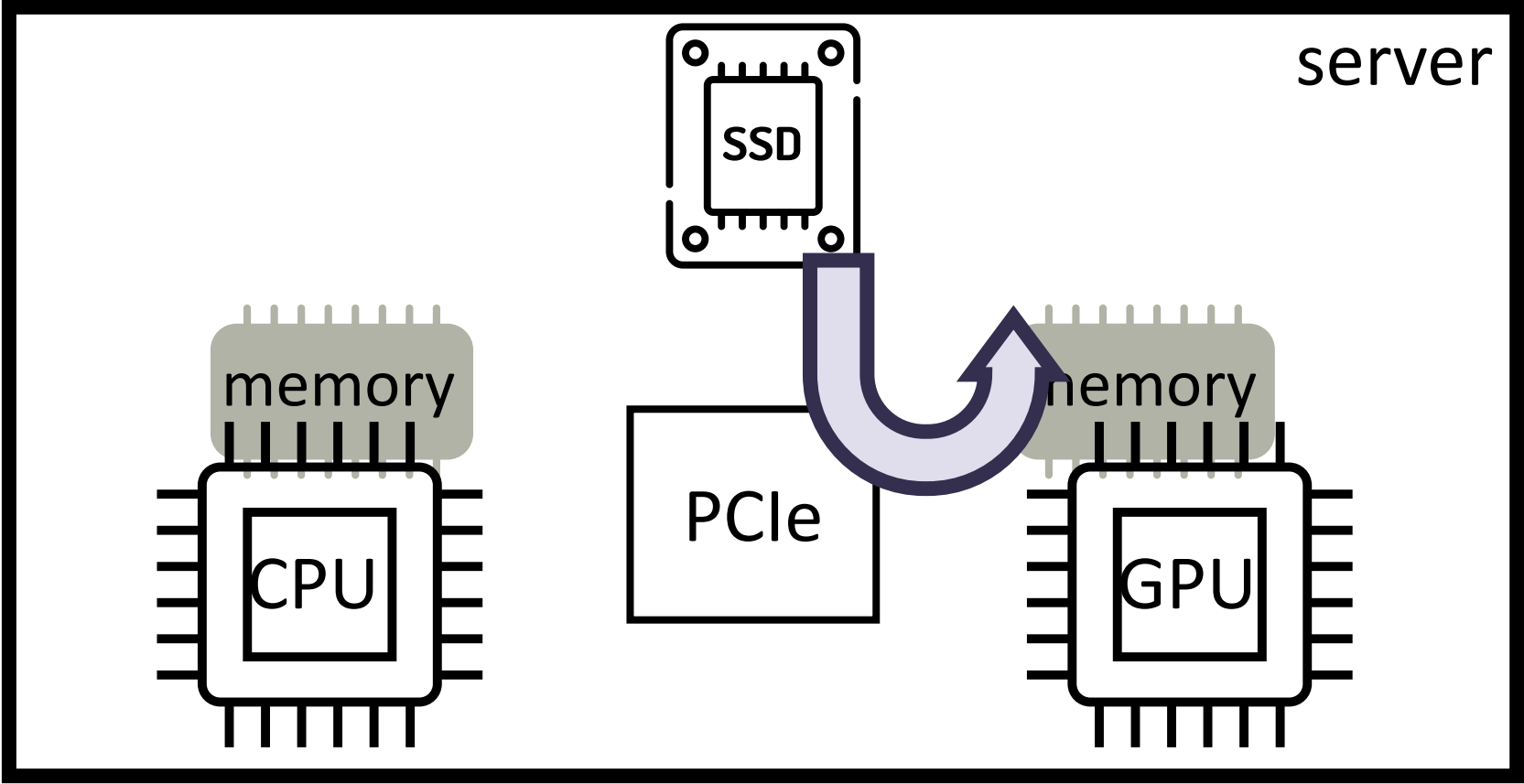
reducing the CPU needs for deep learning

- *GPU-initiated I/O*

conventional data movement



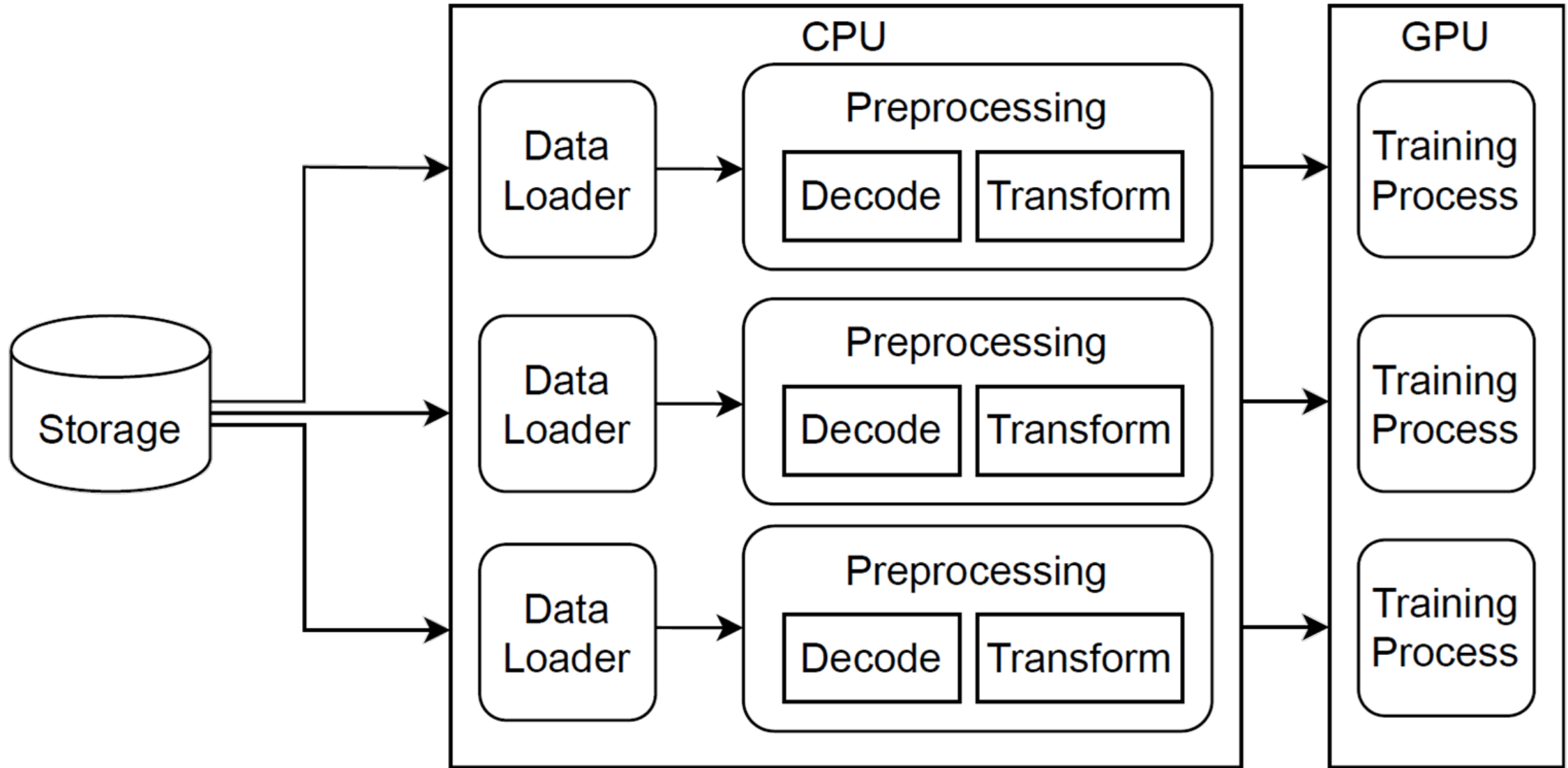
GPU-initiated I/O



reducing the CPU needs for deep learning

- ***GPU-initiated I/O***
 - ✓ reads data from storage to GPU directly, bypassing CPU
 - ✗ software needs maturing & has to be accessible/easy-to-use
 - ✗ still need to preprocess / transform data on the GPU
- ***data & work sharing while training***

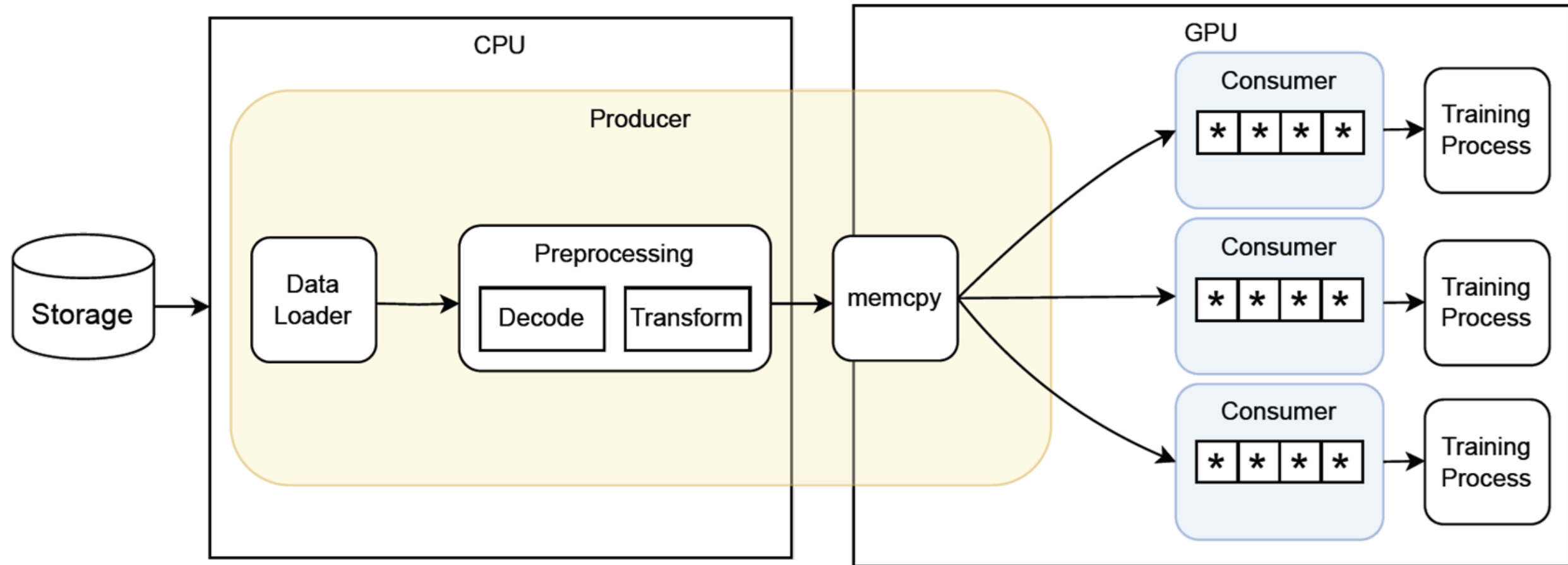
repeated training processes



redundant work & hardware use!

data sharing for collocated training

TensorSocket



eliminates redundant work on CPUs
can achieve 50% reduction in cloud costs

reducing the CPU needs for deep learning

thank you!

- ***GPU-initiated I/O***
 - ✓ reads data from storage to GPU directly, bypassing CPU
 - ✗ software needs maturing & has to be accessible/easy-to-
 - ✗ still need to preprocess / transform data on the GPU

- ***data & work sharing while training***
 - ✓ eliminates redundant work & helps with cost savings
 - ✗ needs enough similar training processes to share effectively

