DASYA

www.dasya.dk

RAD

rad.itu.dk

IT UNIVERSITY OF COPENHAGEN

www.itu.dk

quest to reduce dependency on CPUs in deep learning pipelines: *GPU-centric IO*

Karl B. Torp & Simon A. F. Lund

Samsung Semiconductor Denmark Research

Pınar Tözün

IT University of Copenhagen

pito@itu.dk

pinartozun.com

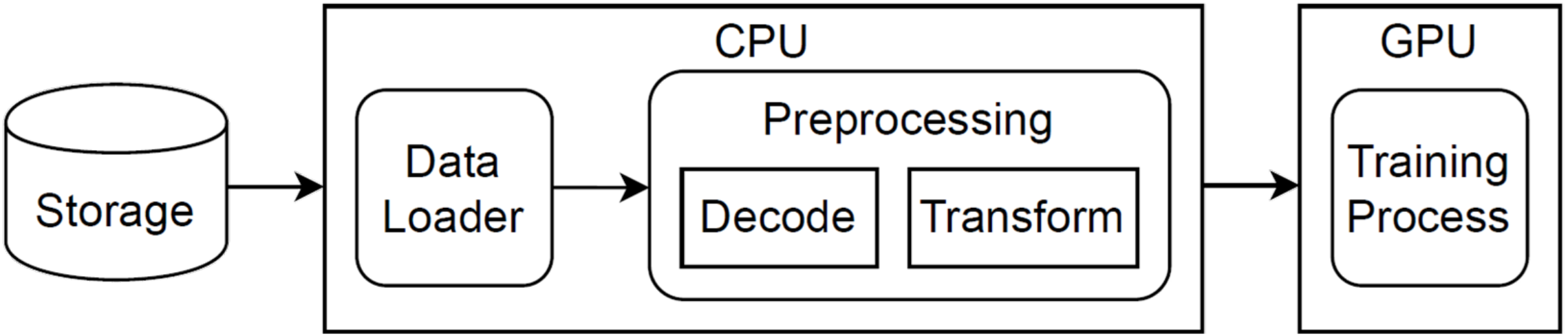
INDEPENDENT RESEARCH FUND DENMARK

novo nordisk foundation

CHEOPS

31/03/2025

journey of data in deep learning training



CPU feeds the accelerators

- 16-64 cores per GPU (recommended)
- 96 cores per TPU*

➔ **otherwise, accelerator may be underutilized**

➔ **can we do more with fewer CPUs?**

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

reducing the CPU needs for deep learning

- data & work sharing

  e.g., CoorDL [PVLDB'21], Joader [NeurIPS'22], tf.data service [SoCC'23]

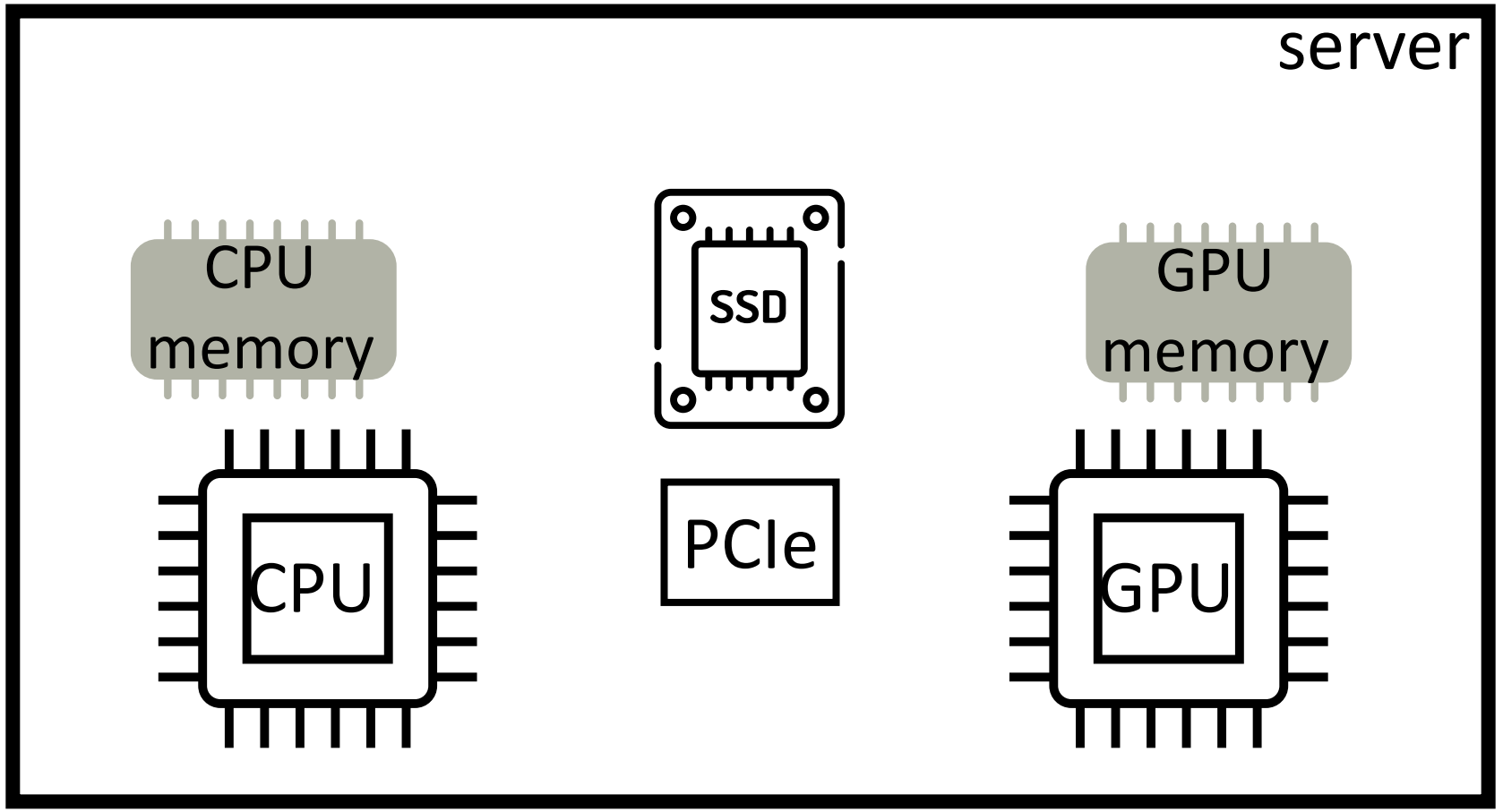
- data pre-processing on the accelerator

  e.g., DALI [NVIDIA], FusionFlow [PVLDB'24]

- GPU-centric I/O path
  - GPUDirect Storage (GDS)
  - Big Accelerator Memory (BaM)

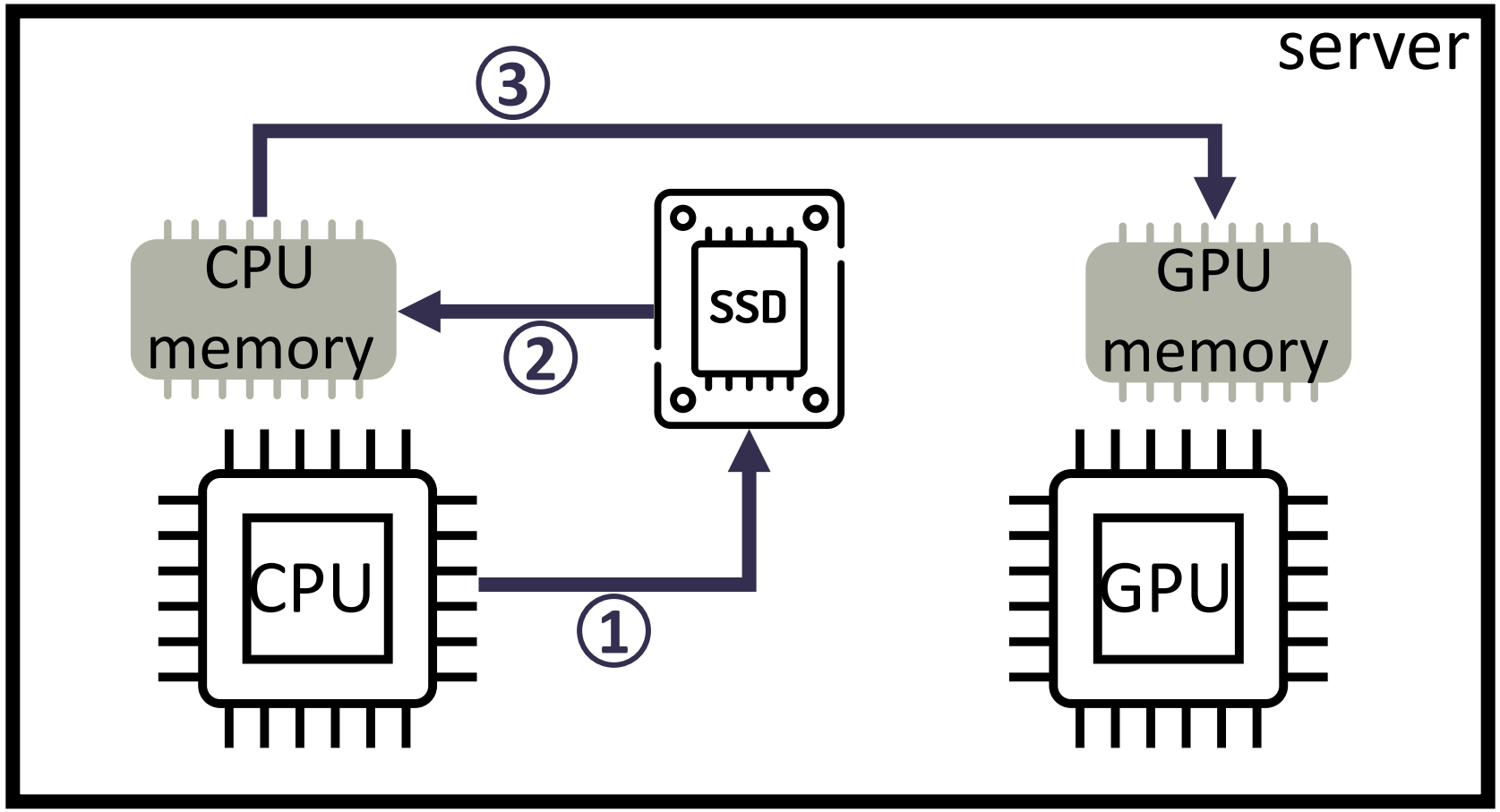
what are the trade-offs of different options?

target hardware setup

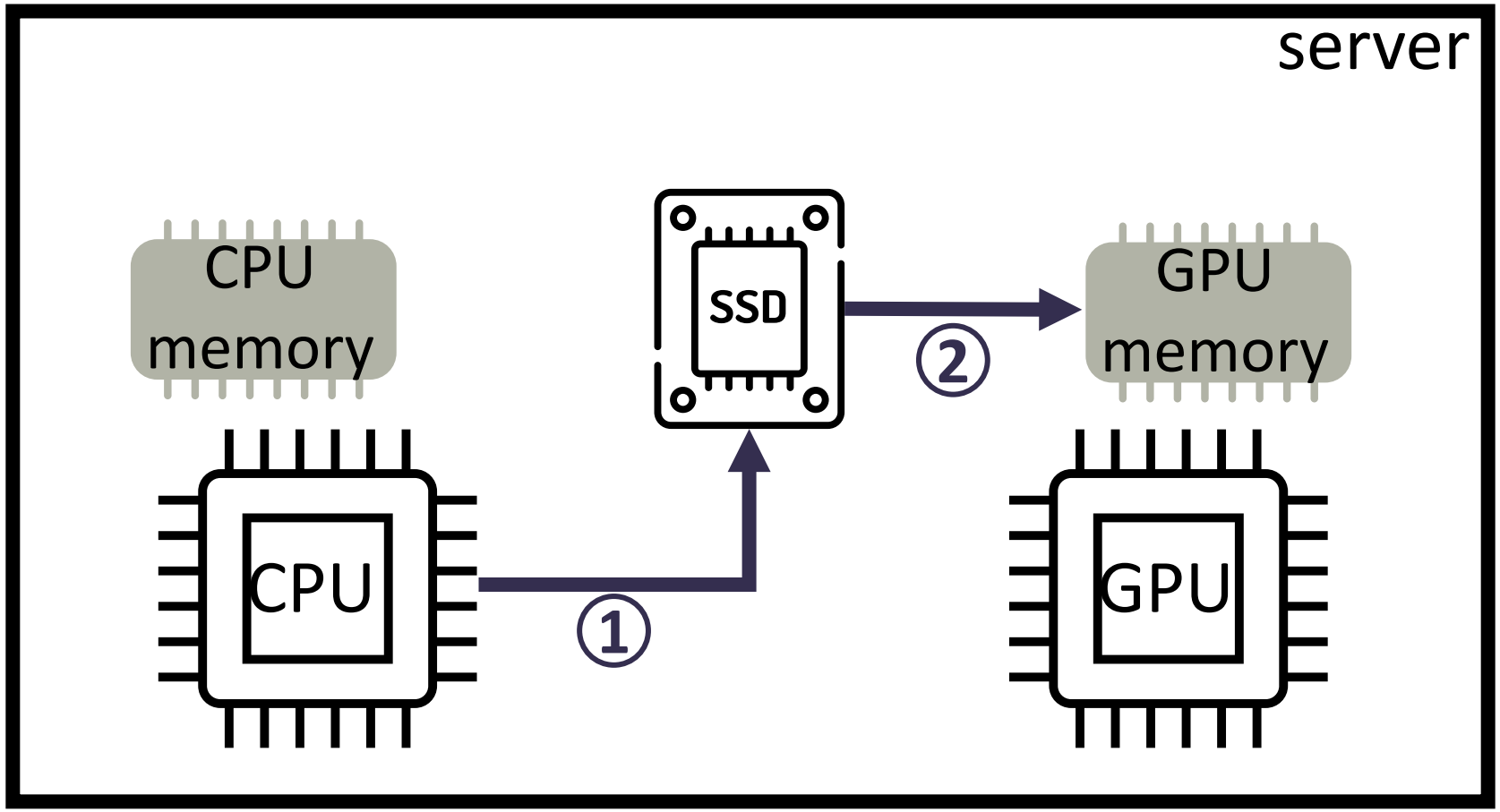


* PCIe is dropped in the remaining figures for the sake of simplicity in illustrations.

conventional: CPU-centric I/O

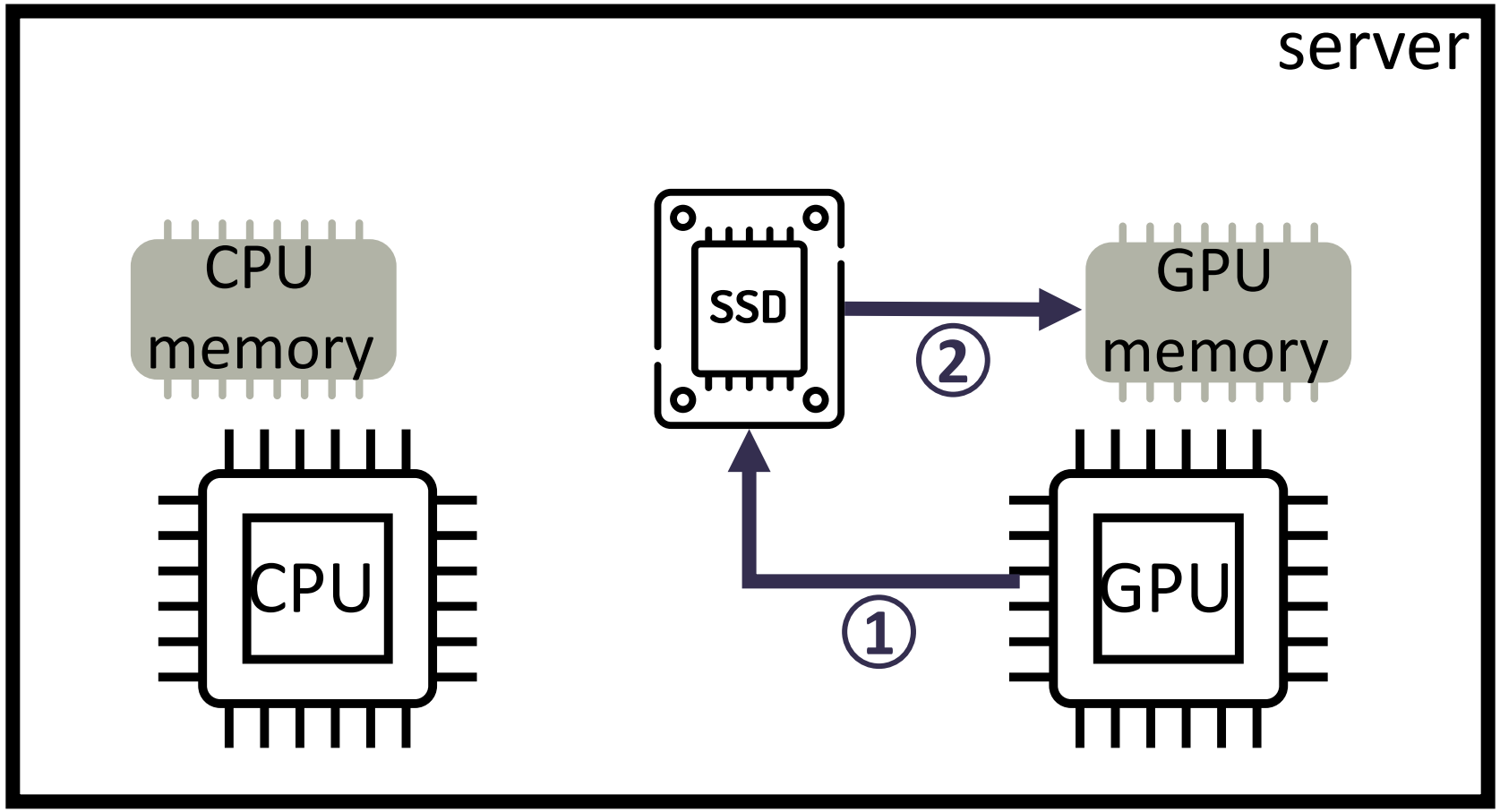


GPUDirect: GPU-centric & CPU-initiated



BaM: GPU-centric & GPU-initiated

Big Accelerator Memory [ASPLOS'23]



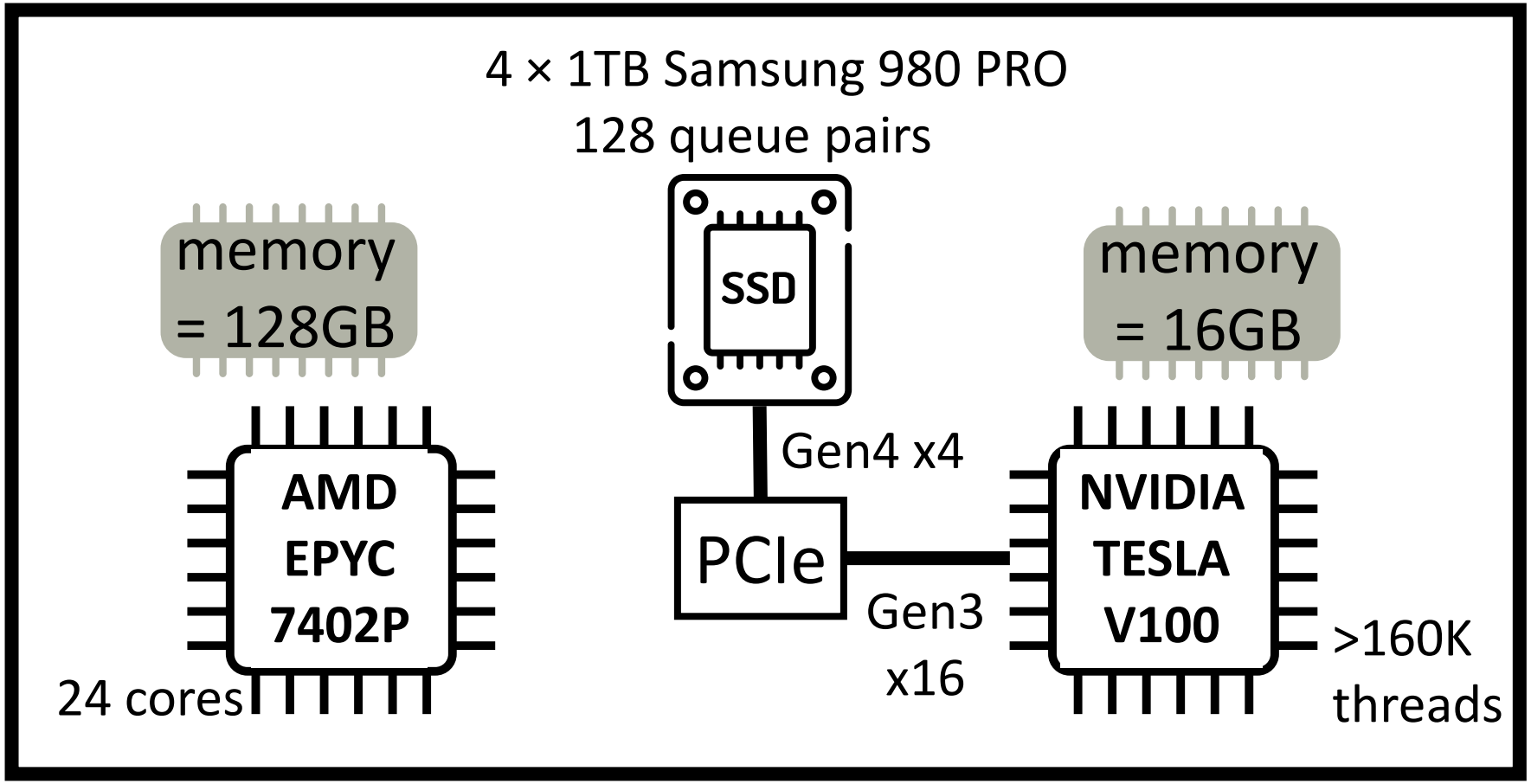
evaluation: CPU- vs GPU-centric I/O

mechanisms: CPU-centric: SPDK & GPU-centric: GDS, BaM

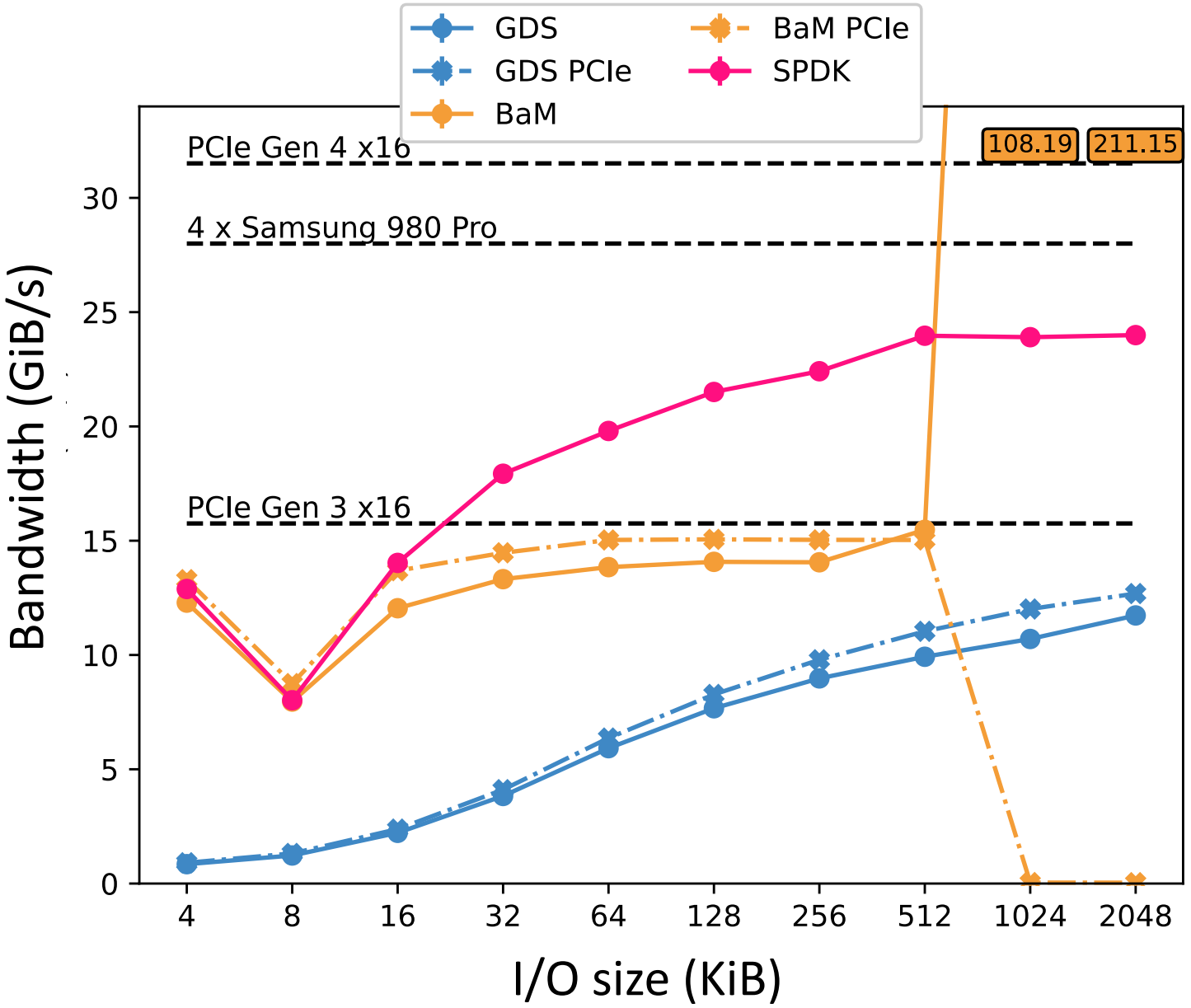
workload: random reads

- each mechanism has their own fio-like tool for benchmarking

hardware



bandwidth utilization – 4 SSDs & PCIe



GDS is CPU-compute heavy.

➔ 16 logical cores utilized

BaM is limited by the PCIe Gen3 link & heavy on the GPU resources.

➔ whole GPU utilized

SPDK is the most resource-efficient but has a longer path to the GPU.

➔ 2 logical cores utilized

path to GPU-centric I/O

➔ when to use which mechanism while being resource-aware?

➔ how to best integrate these mechanisms into popular deep learning frameworks to allow wider-scale use?

thank you!

pito@itu.dk