An Evaluation of NVMe-over-Fabrics for Disaggregated Databases over Fast Networks

BTW 2025 Short Paper

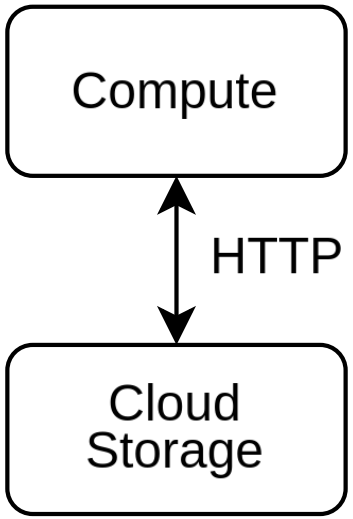
Jigao Luo[1], Nils Boeschen[1], Tobias Ziegler[2], Carsten Binnig[1, 3]

2025-03-05

[1]Technische Universität Darmstadt
[2]Technische Universität München
[3]DFKI

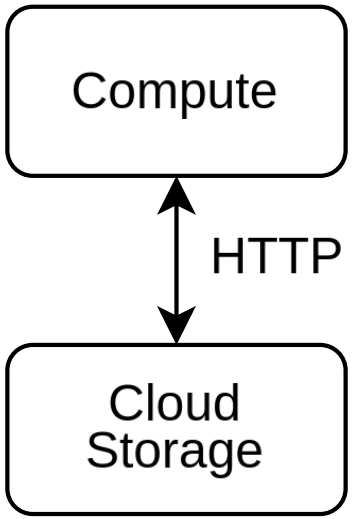
Motivation: Disaggregation

- Storage disaggregation in cloud DB:



Motivation: Disaggregation

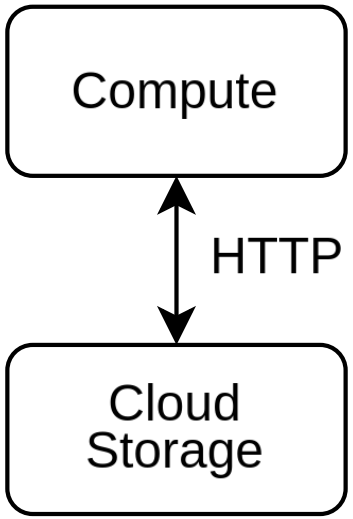
- Storage disaggregation in cloud DB:



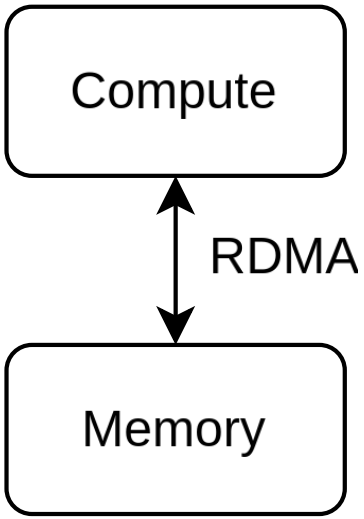
- Fast network on cloud: RDMA

Motivation: Disaggregation

- Storage disaggregation in cloud DB:

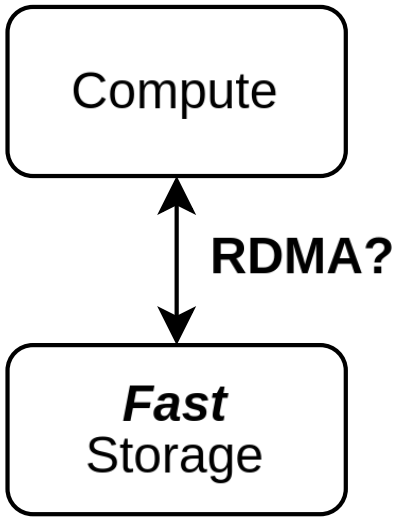


- Fast network on cloud: RDMA
- Memory disaggregation in DB:

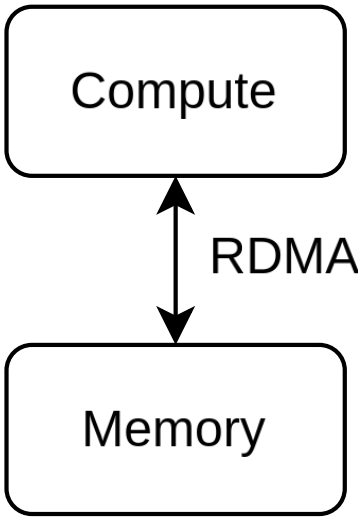


Motivation: Disaggregation

- Storage disaggregation with **RDMA**?

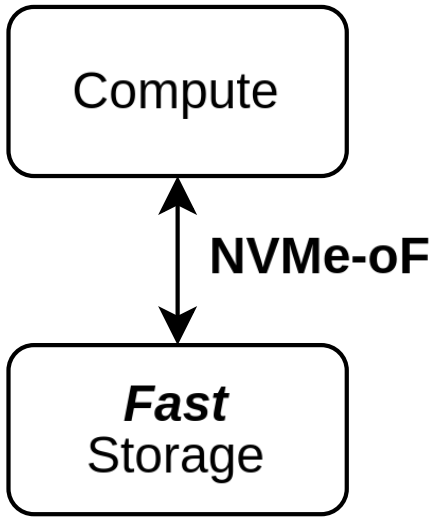


- Fast network on cloud: RDMA
- Memory disaggregation in DB:

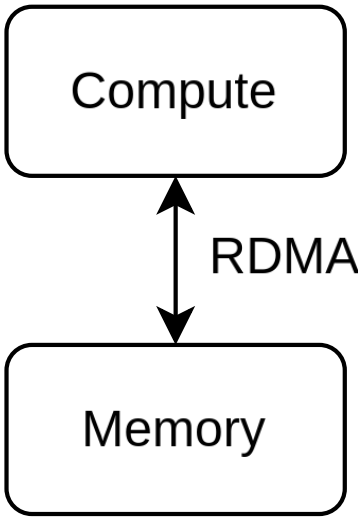


Motivation: Disaggregation

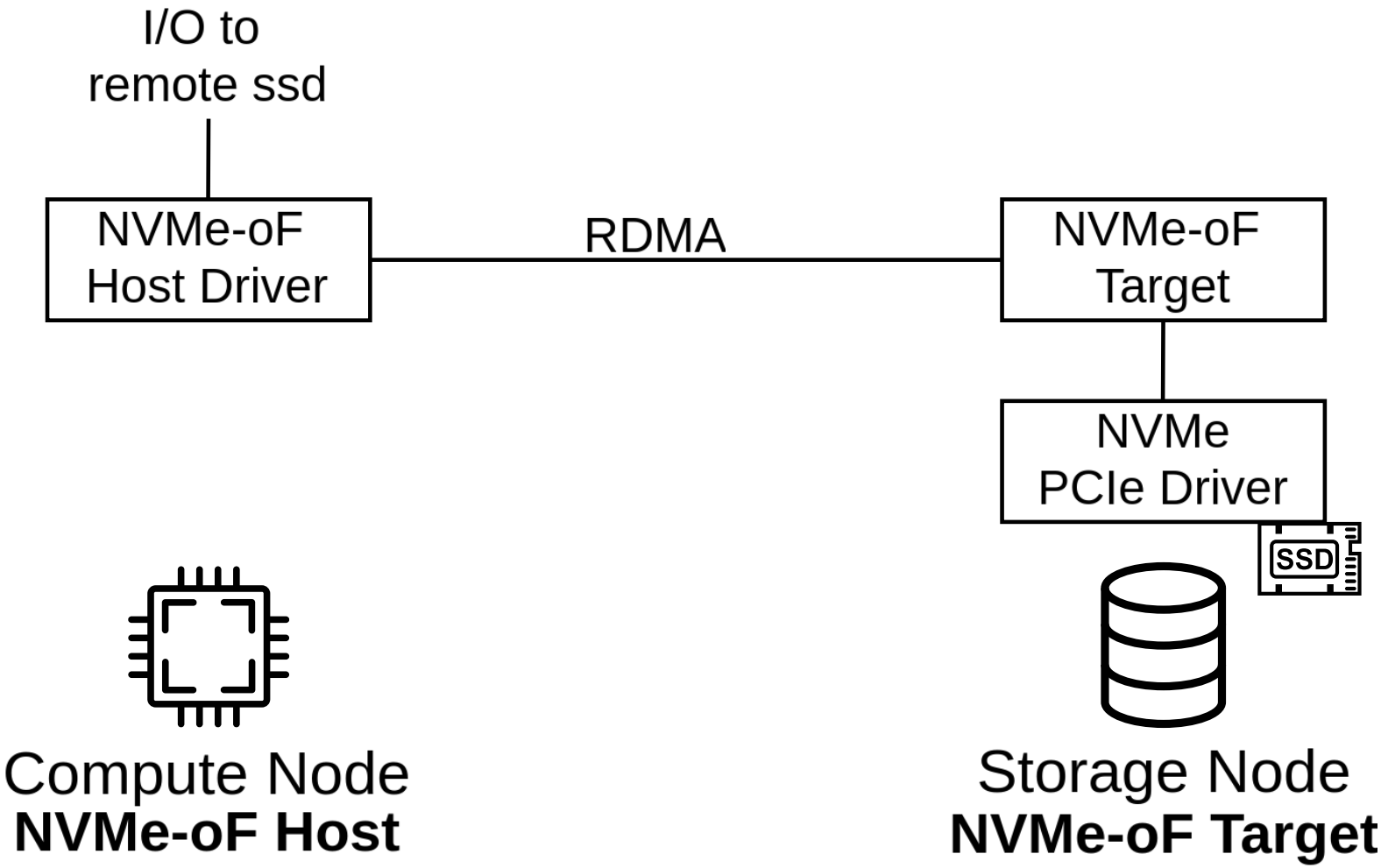
- Storage disaggregation with **NVMe-oF**!



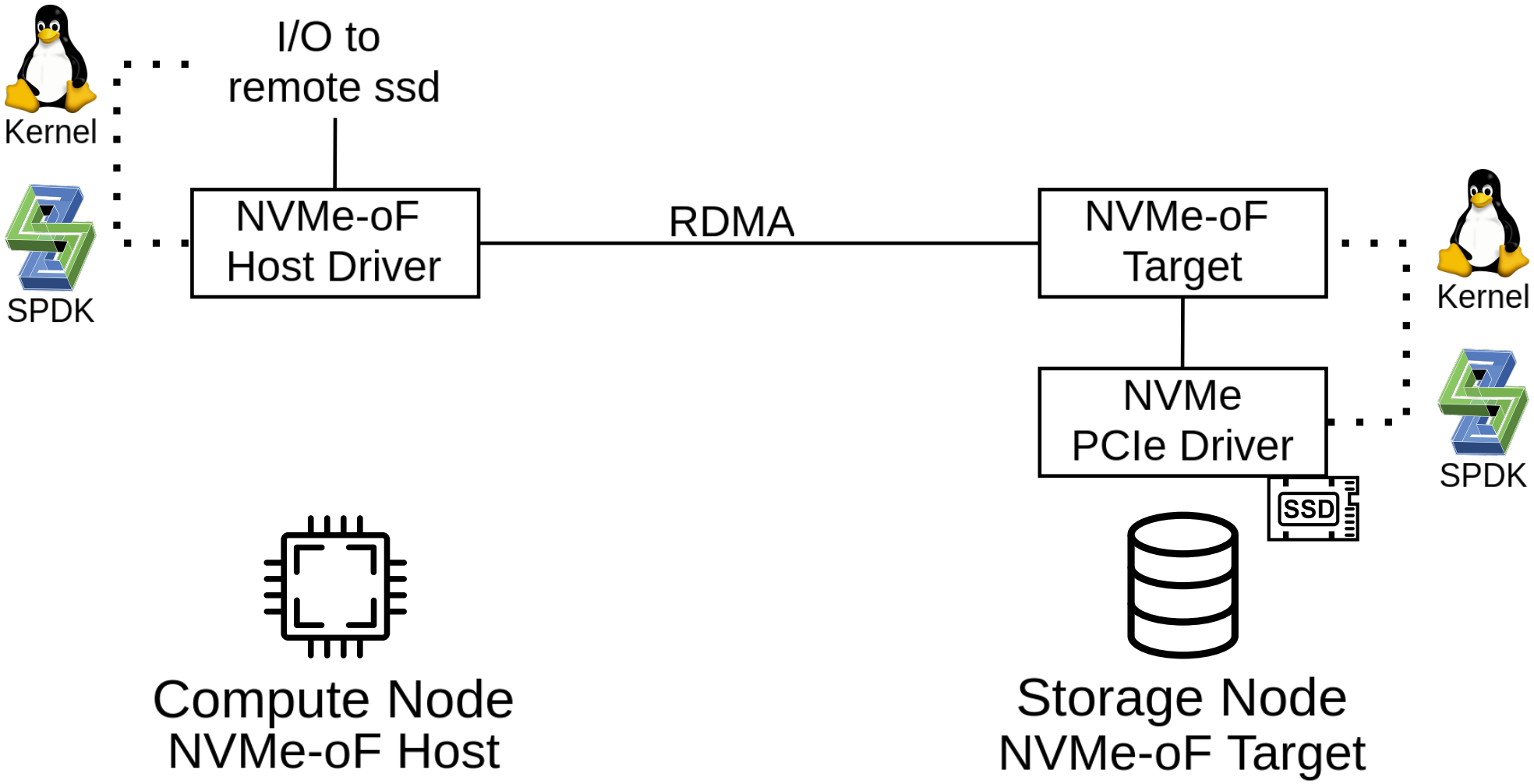
- Fast network on cloud: RDMA
- Memory disaggregation in DB:



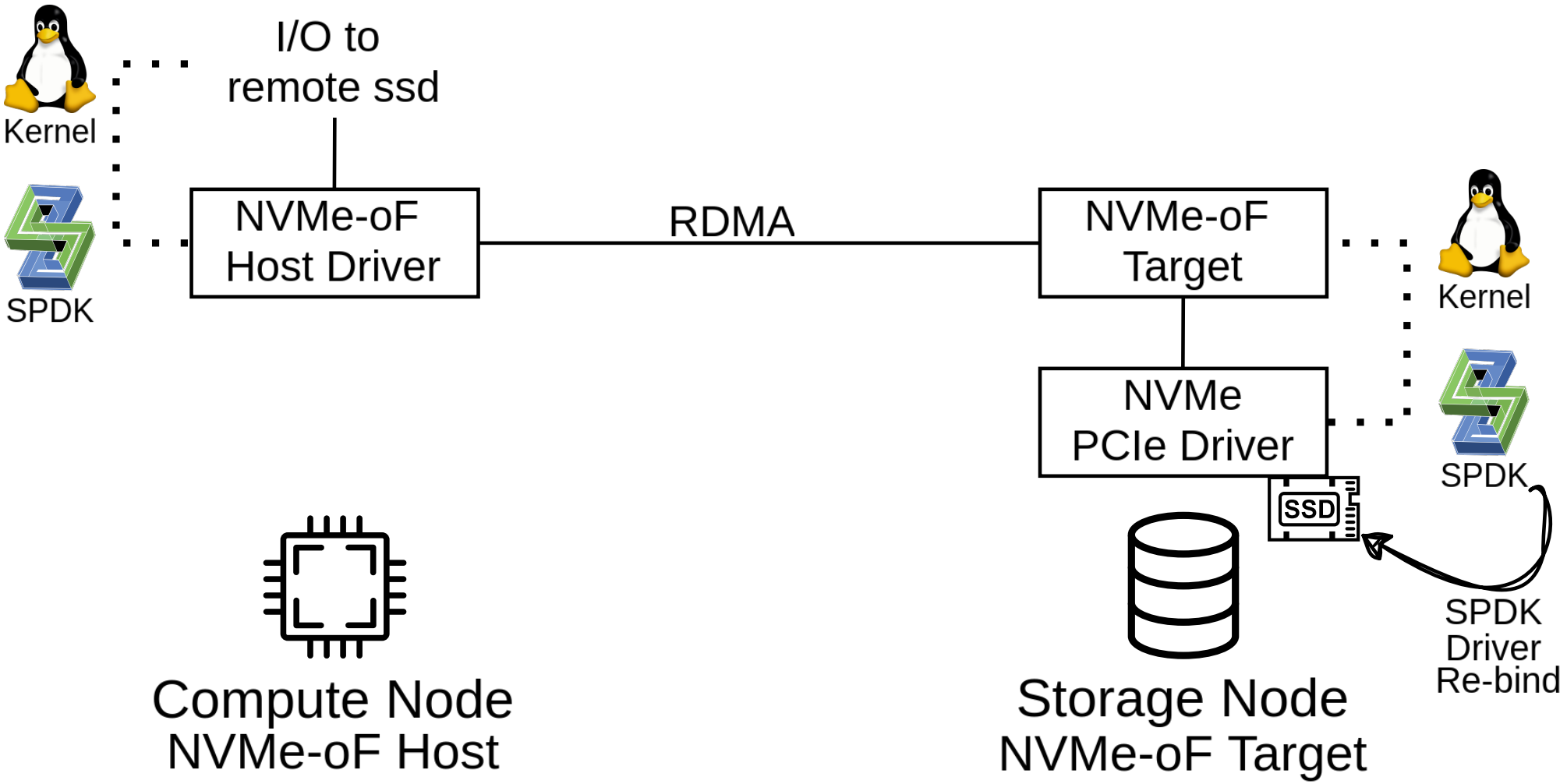
NVMe-oF Basics: Components



NVMe-oF Basics: Components



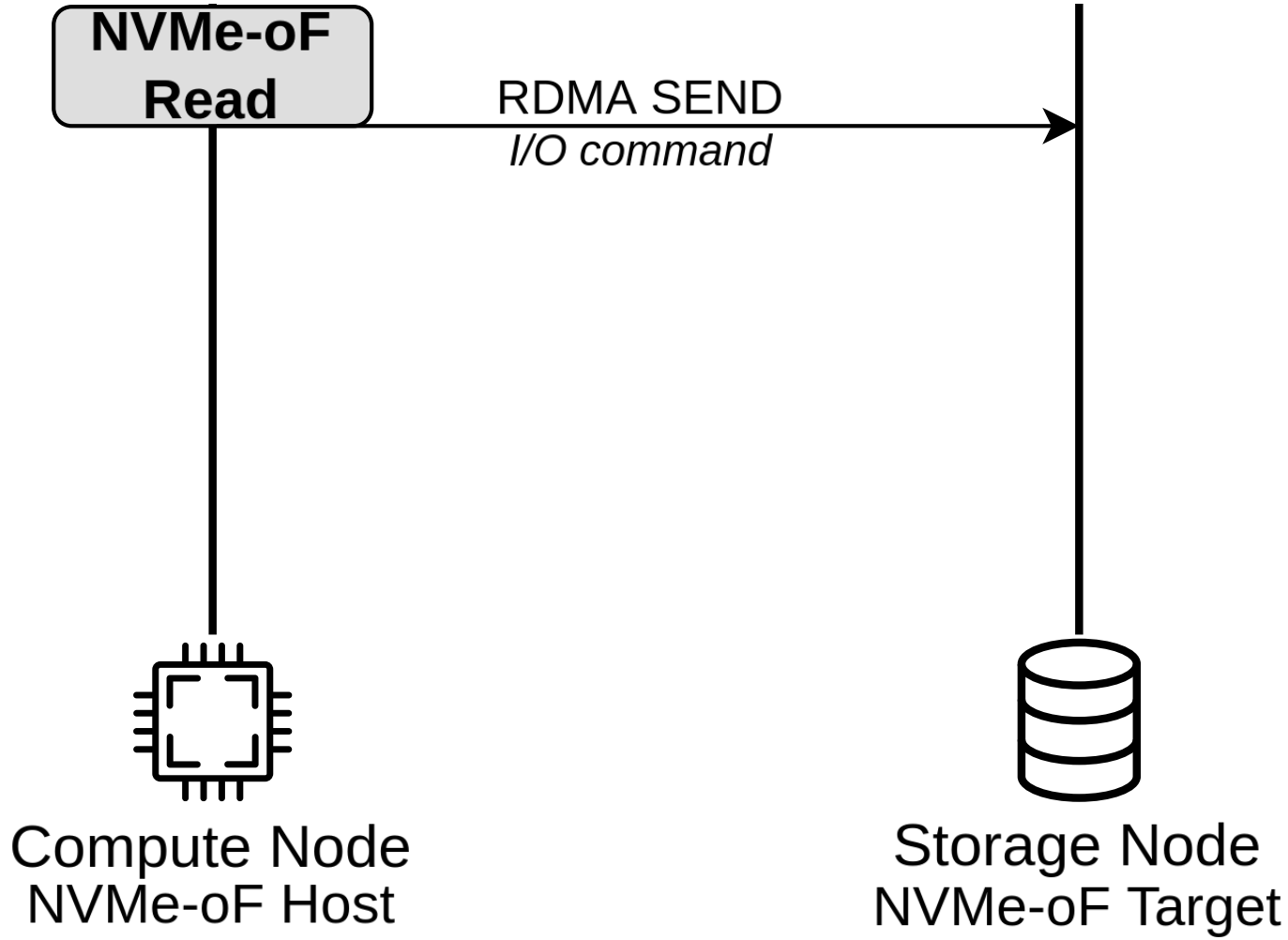
NVMe-oF Basics: Components



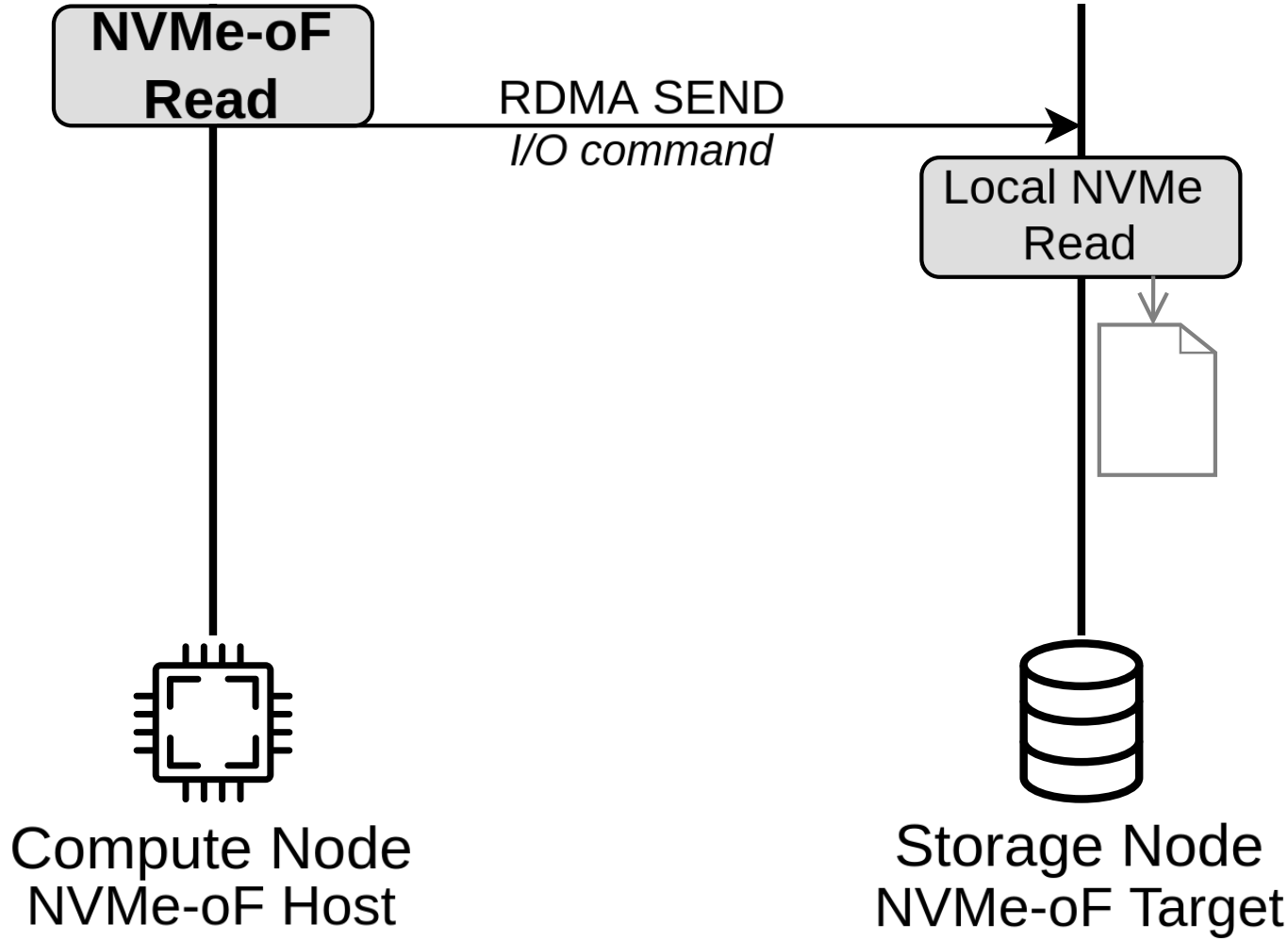
NVMe-oF Basics: Stack Comparison

	Kernel	SPDK user-space
I/O completion	interrupt / polling	polling
I/O Efficiency	medium	high
Setup	simple	complicated

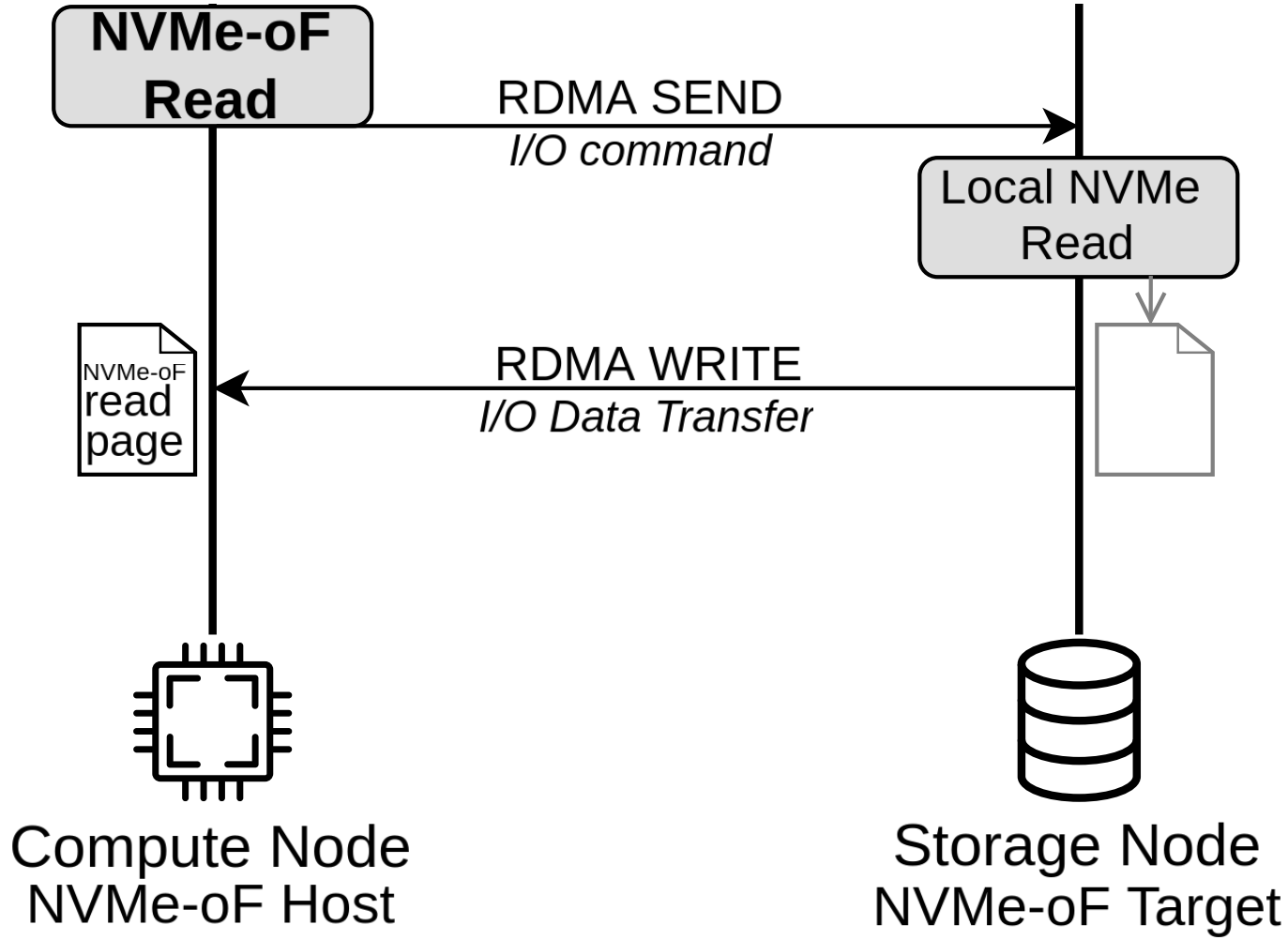
NVMe-oF Basics: Read



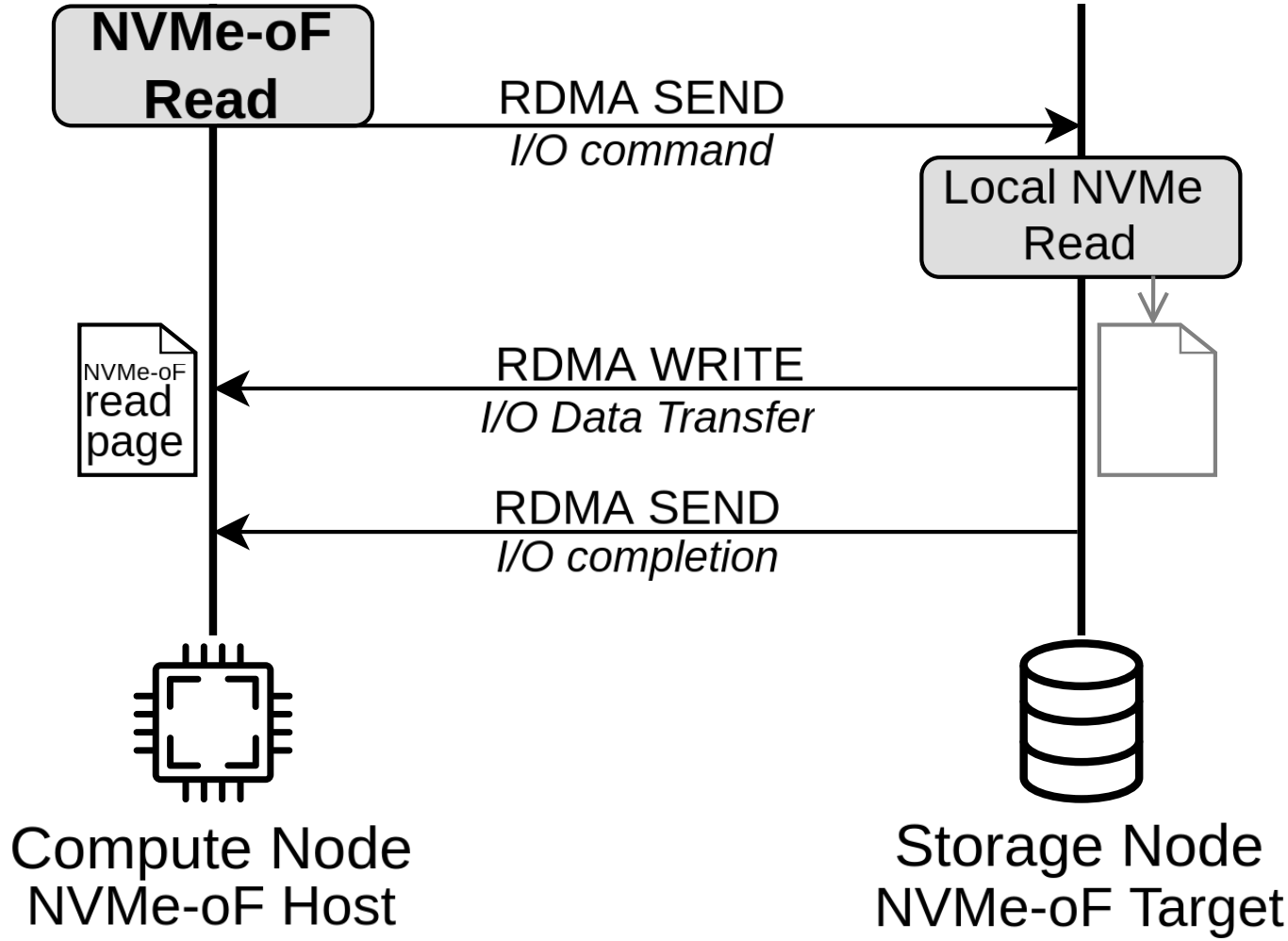
NVMe-oF Basics: Read



NVMe-oF Basics: Read

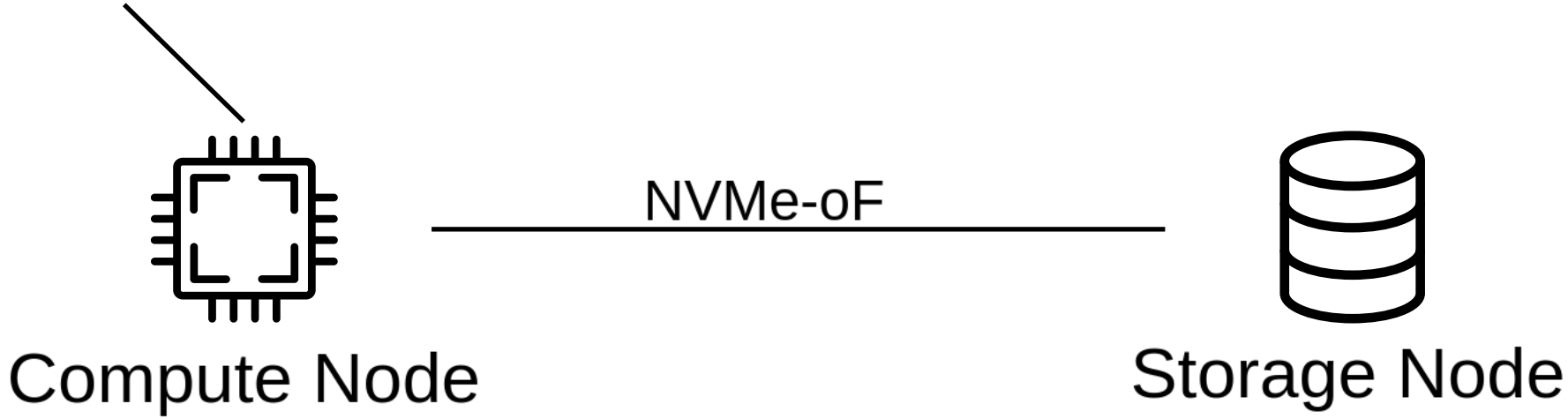


NVMe-oF Basics: Read



NVMe-oF in DB Design?

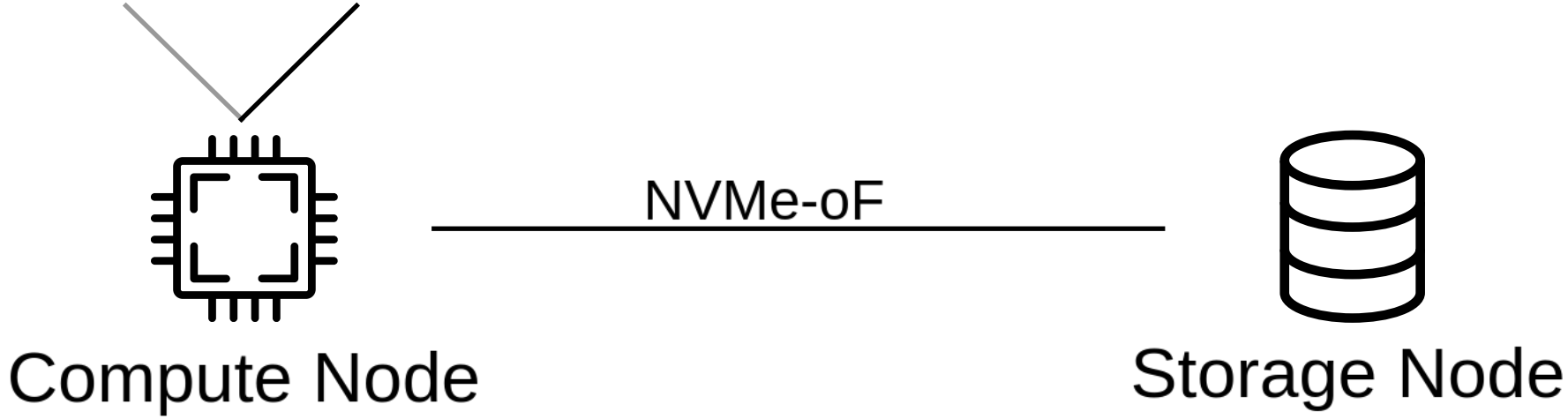
OLTP:
- Stack?
- Throughput?
- Latency?



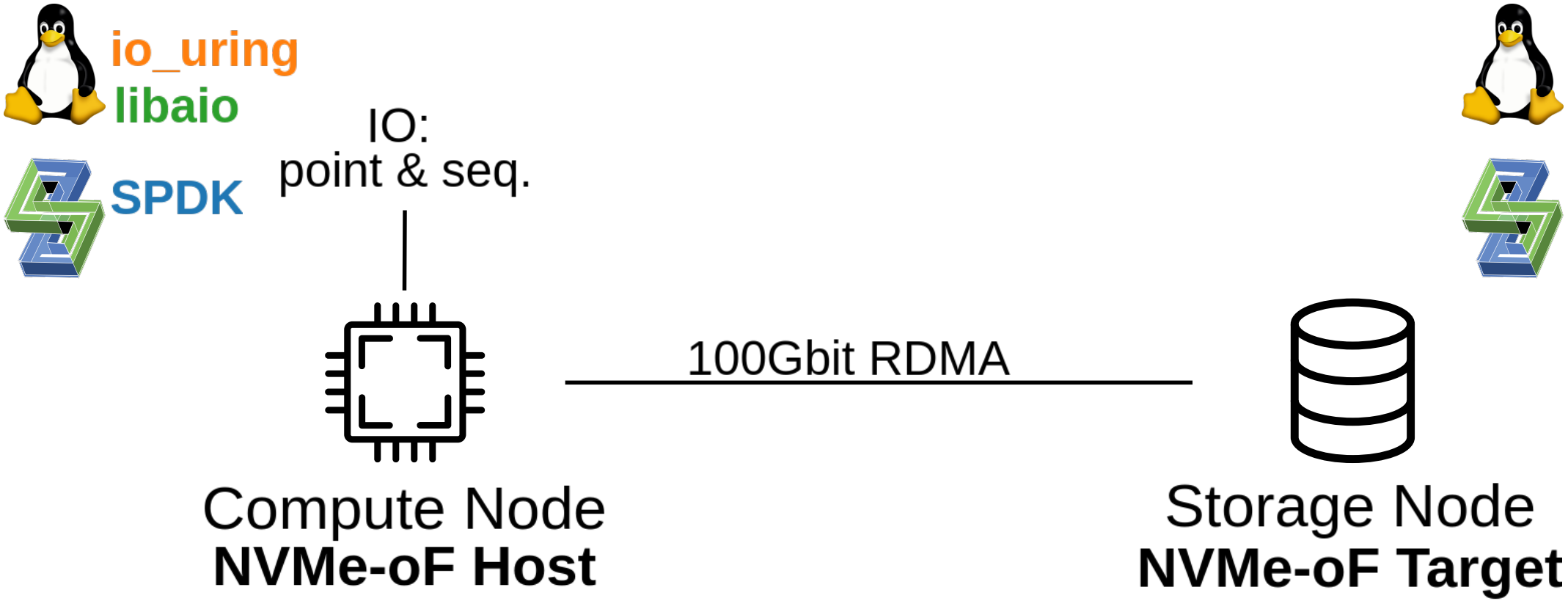
NVMe-oF in DB Design?

OLTP:
- Stack?
- Throughput?
- Latency?

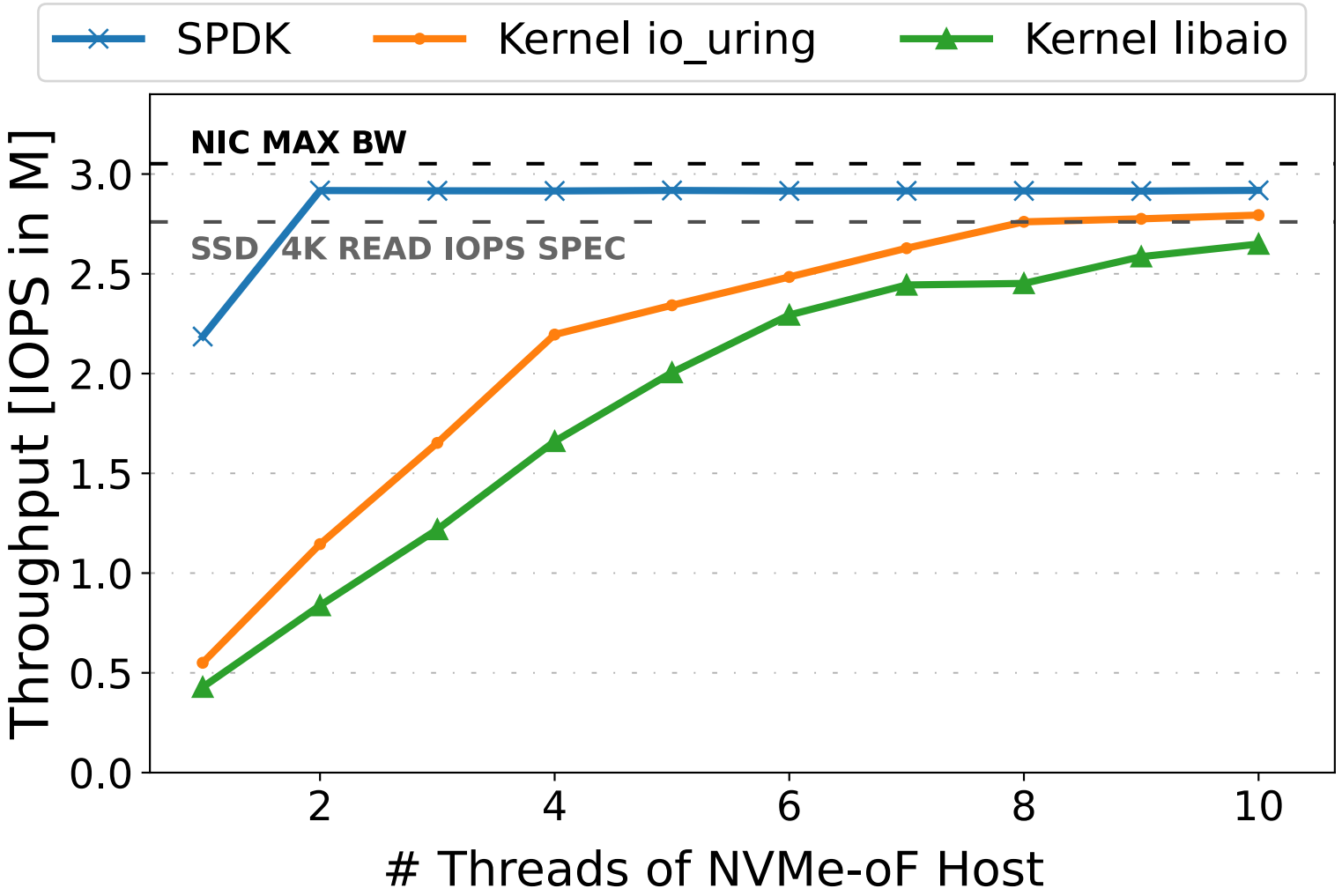
OLAP:
- Stack?
- Throughput?



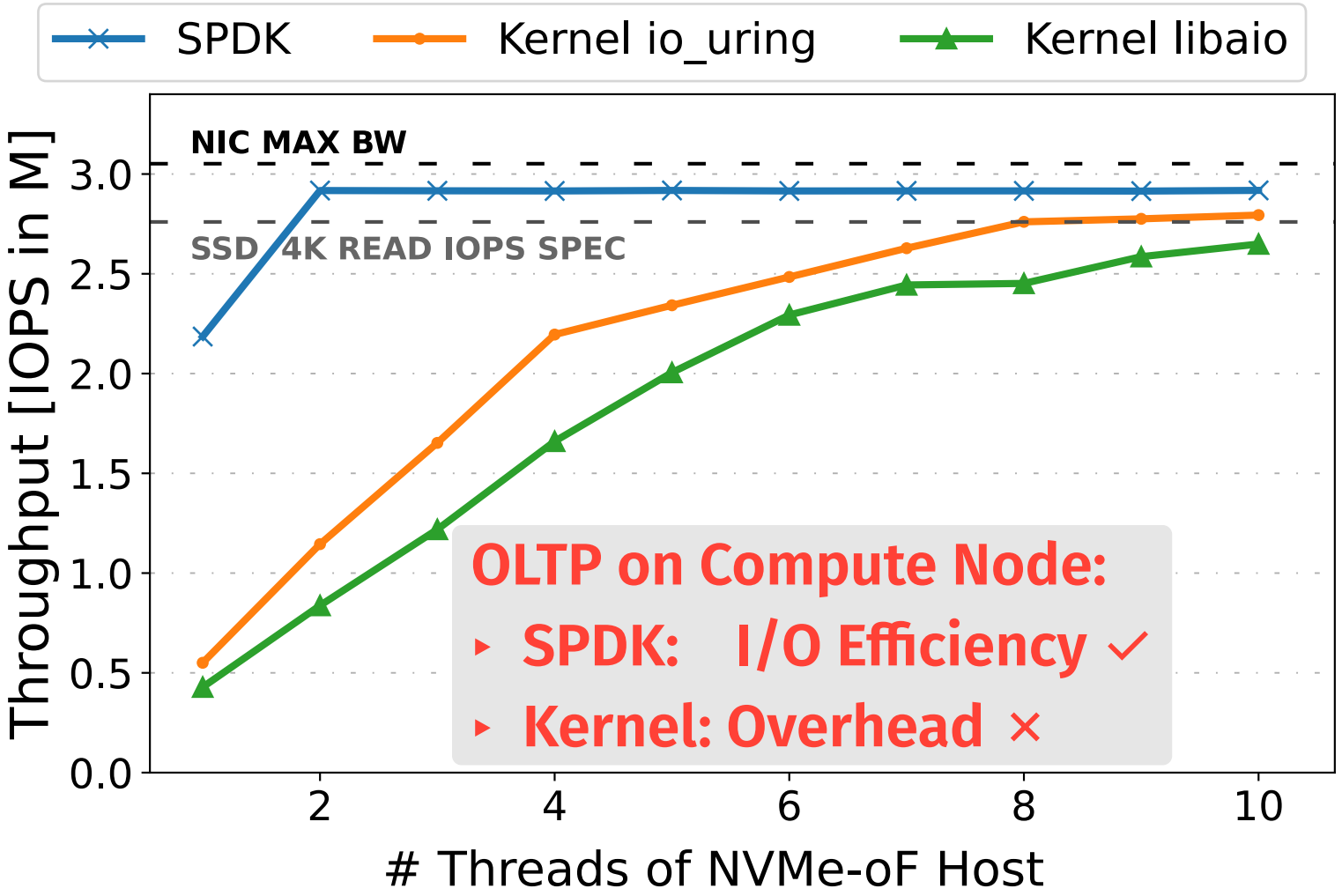
Evaluation Methodology



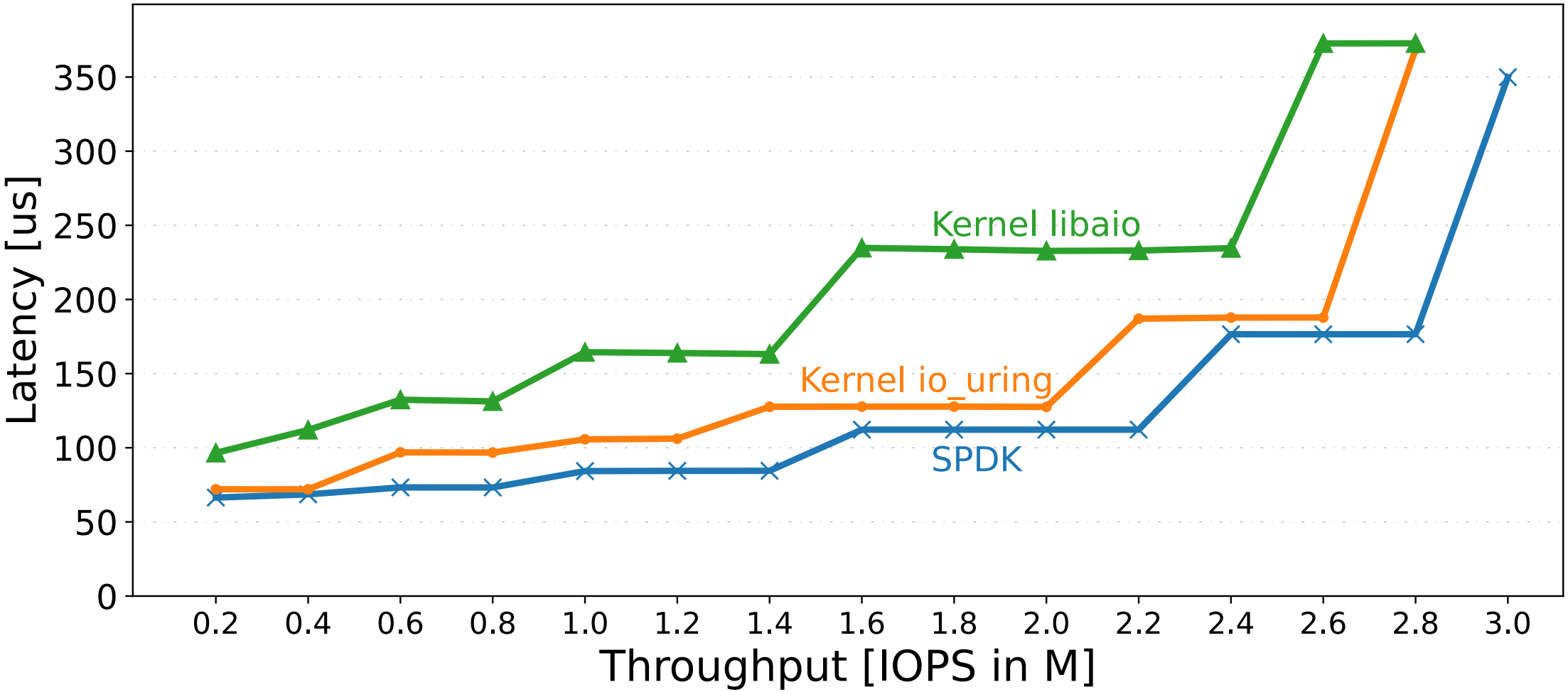
OLTP: 4K Random Read Throughput



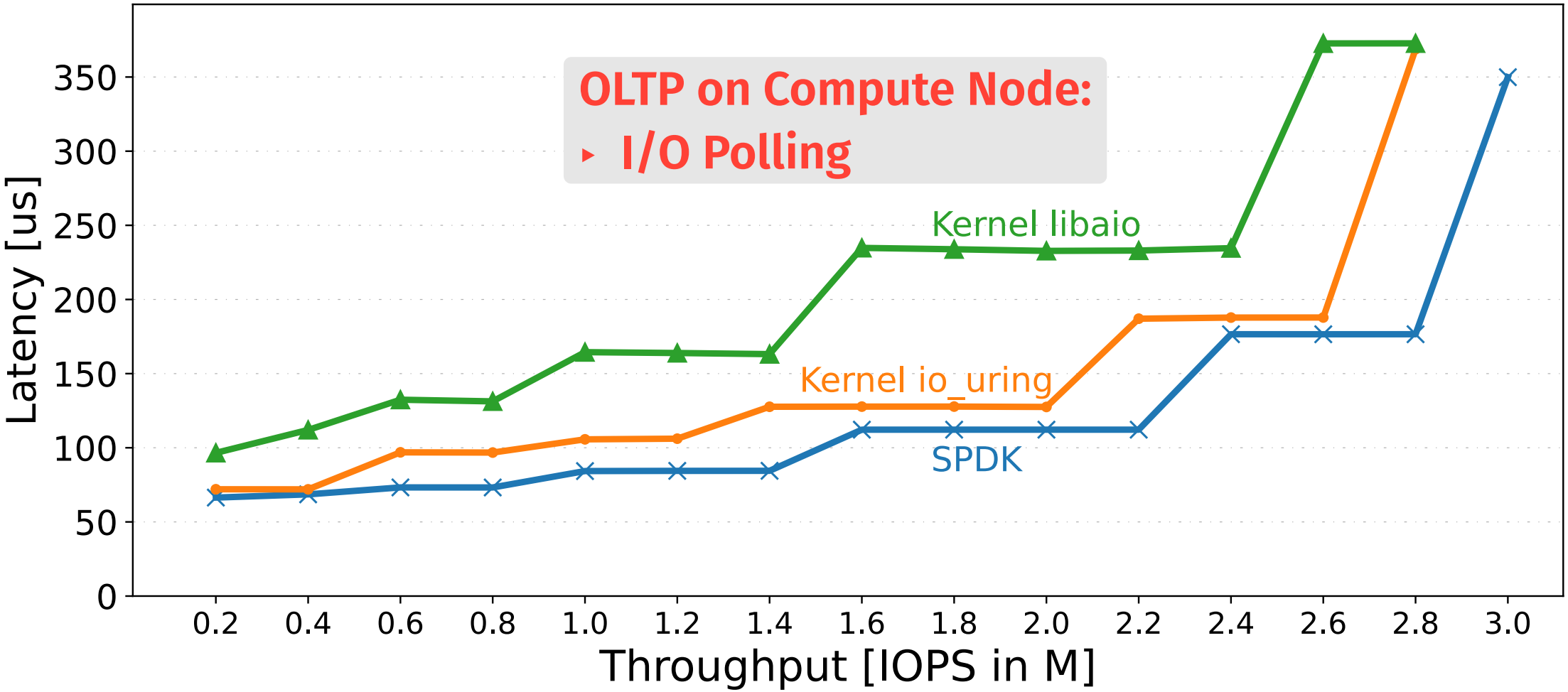
OLTP: 4K Random Read Throughput



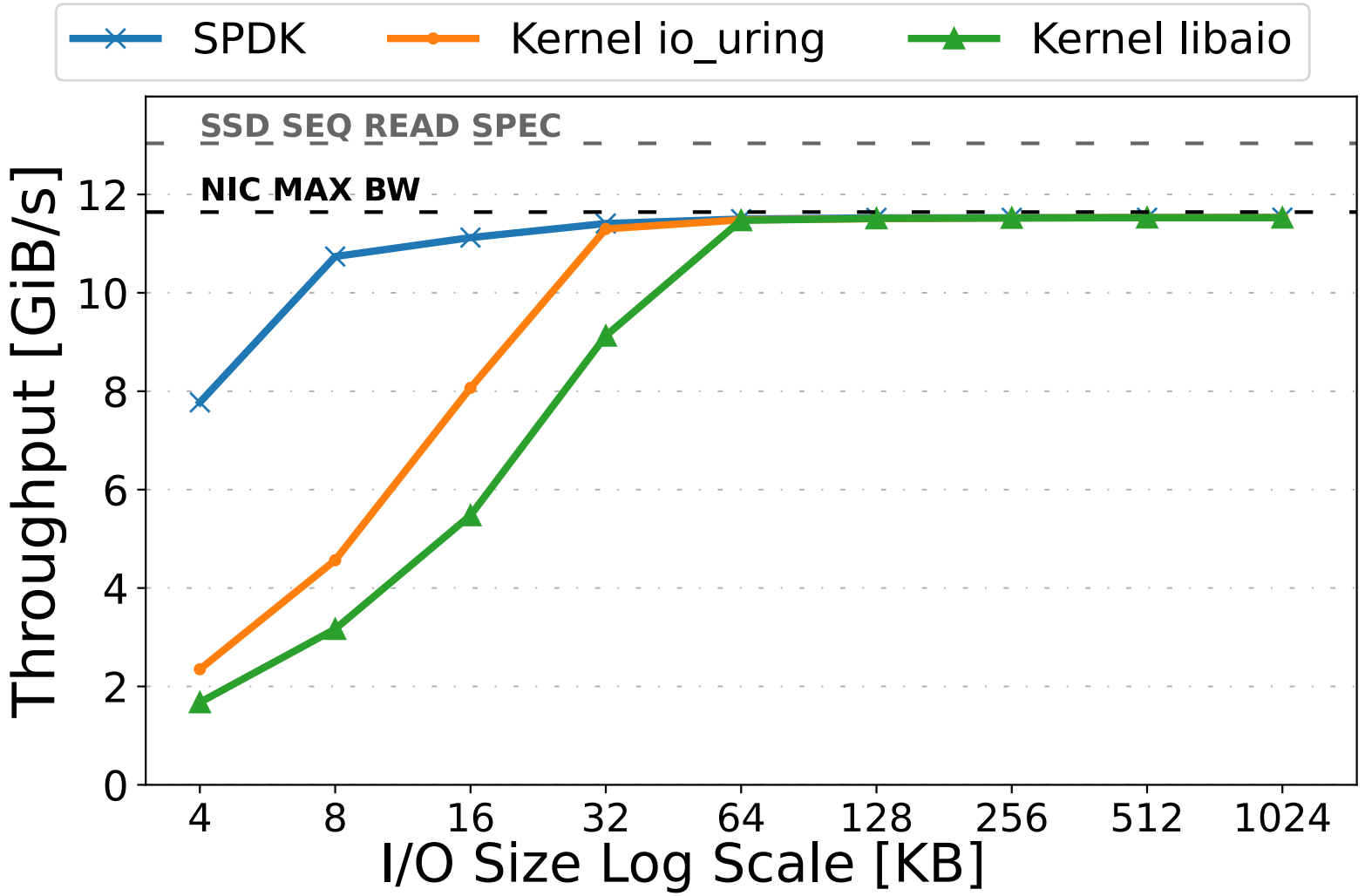
OLTP: 4K Random Read Latency



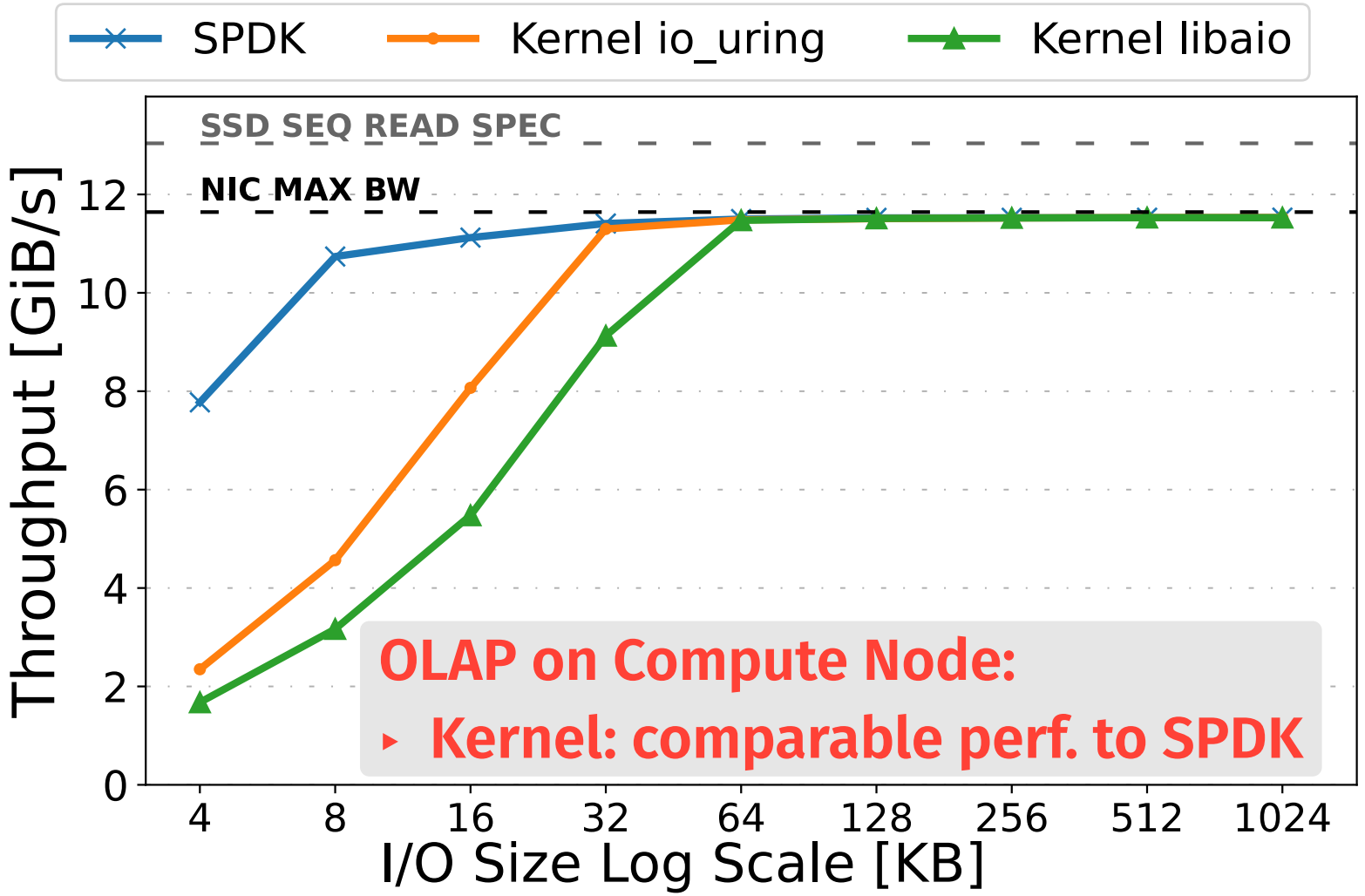
OLTP: 4K Random Read Latency



OLAP: Sequential Read Throughput



OLAP: Sequential Read Throughput



Summary

- NVMe-oF enables storage disaggregation
- Stack trade-off in NVMe-oF
- OLTP:
  - **SPDK**: I/O efficiency ✓
  - **Kernel io_uring**: general-purpose ✓
- OLAP: **Kernel io_uring**

Thanks for your attention