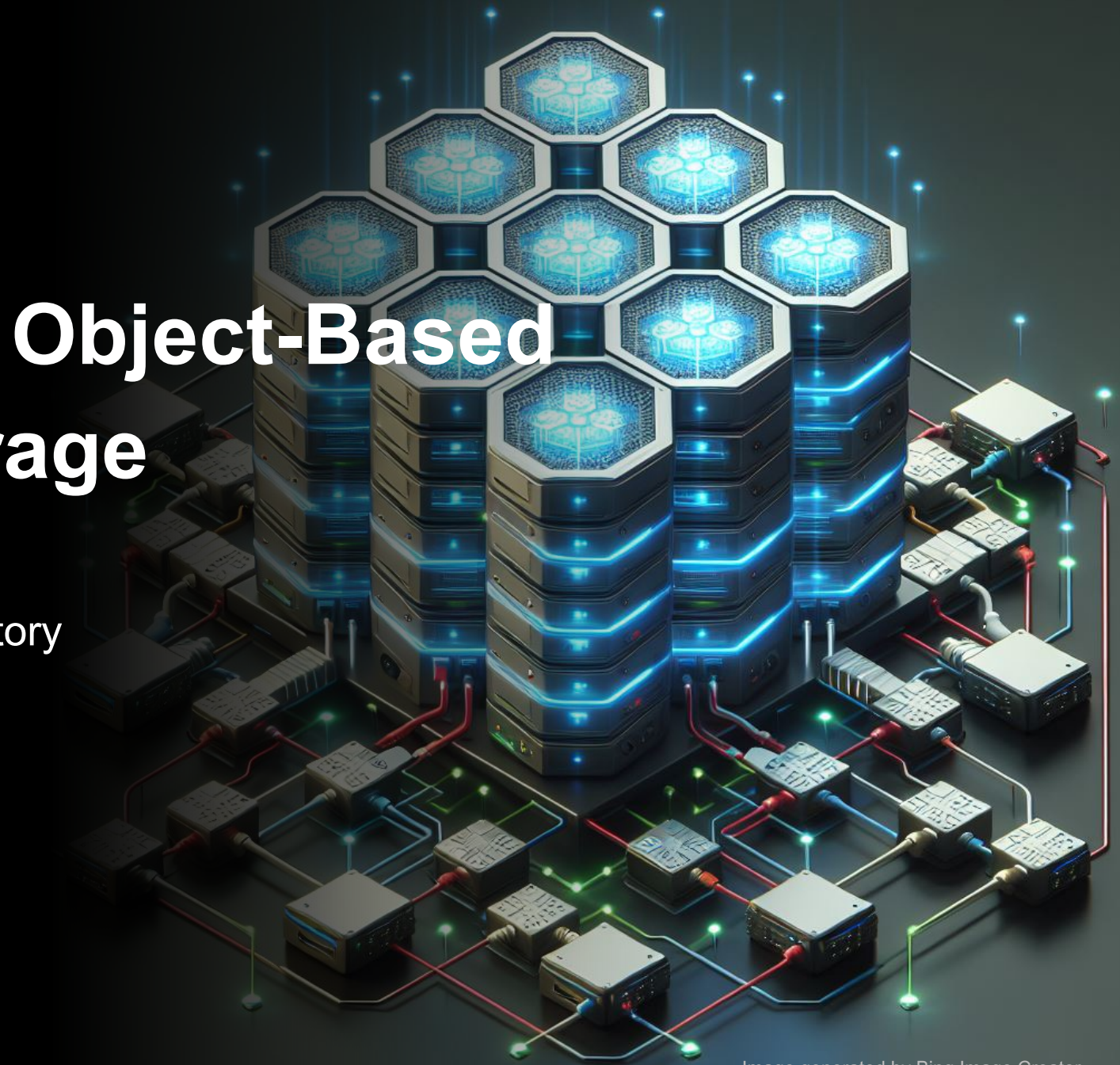
Los Alamos NATIONAL LABORATORY

OCS: Toward Open Object-Based Computational Storage

Qing Zheng, Los Alamos National Laboratory

3/5/24

LA-UR-24-21993



NNSA National Nuclear Security Administration Managed by Triad National Security, LLC, for the U.S. Department of Energy's NNSA.

Image generated by Bing Image Creator.

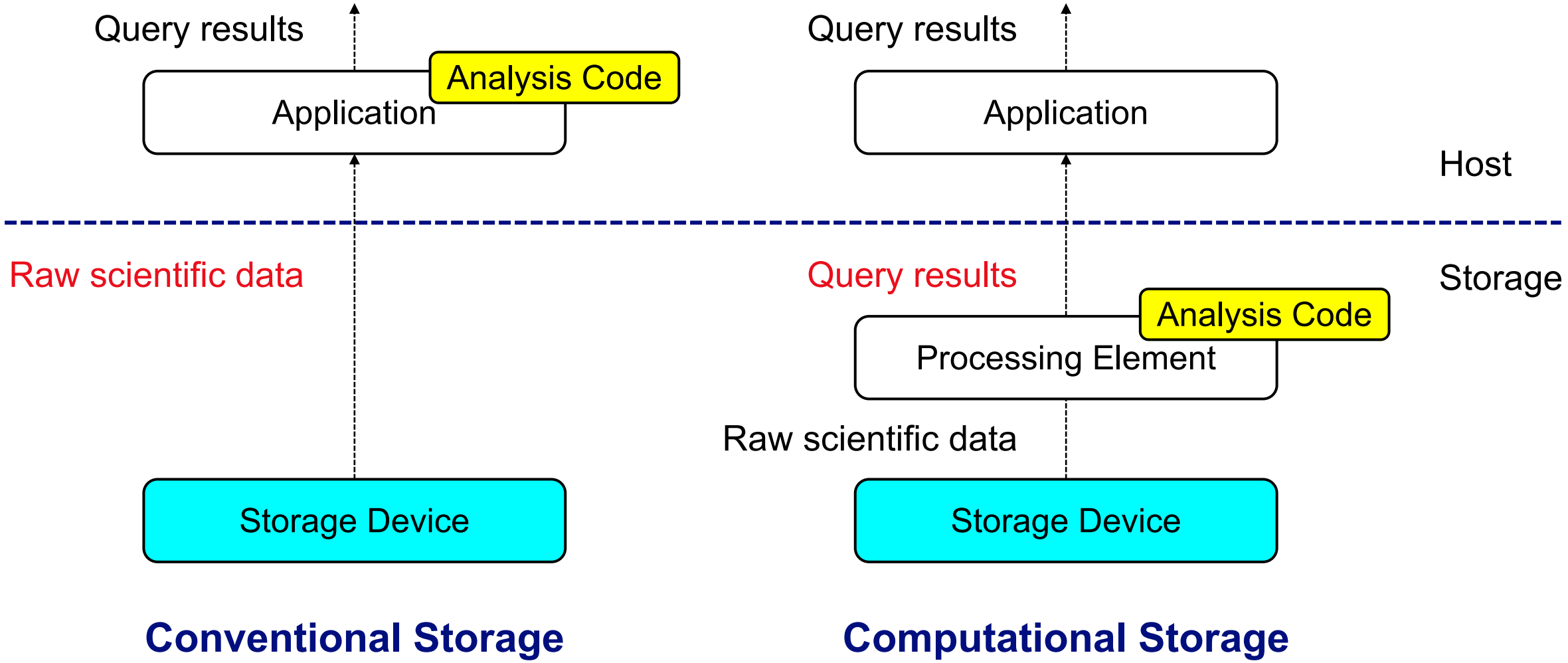
3 Things About Scientific Data Analytics

Data is big

Moving data is expensive

Queries often target a tiny portion of a large dataset

Pushing Queries to Storage

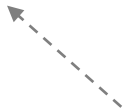


Target Applications

PIC (particle-in-cell) codes

- **Row-based apps**
- **A few** columns (~10)
- **Single-dimensional** queries
- **All** columns are retrieved during analysis

- **H/W-accelerated KV store (KV-CSD)**



We discussed KV-CSD yesterday

Grid codes

- **Columnar apps**
- **Many** columns (10 – 100)
- **Multi-dimensional** queries
- **A subset** of columns are retrieved during analysis

- **In-storage SQL processing**

This project

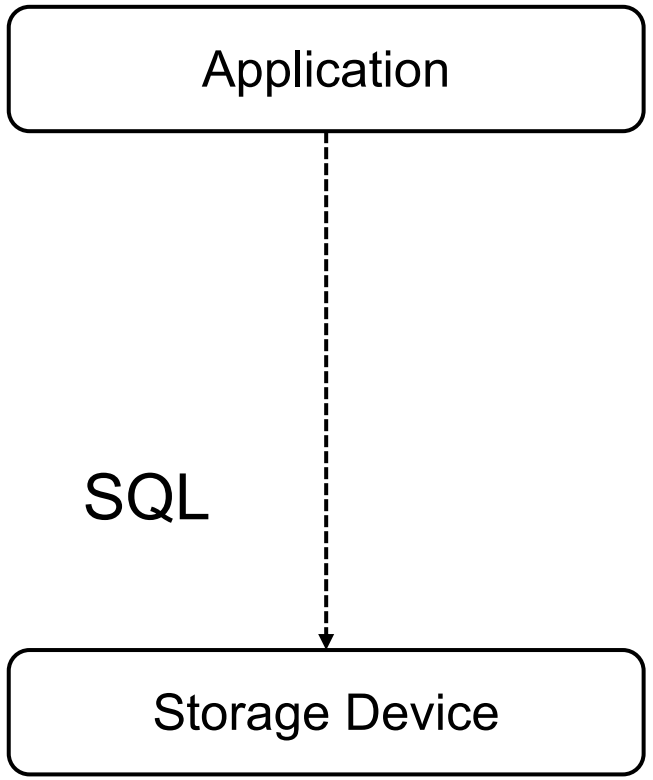
Data Agnostic vs Data Aware

Data Agnostic

- Storage does not know what's in the data (view data as byte streams)
  - Like what POSIX filesystems do today
- Ways to achieve SQL offloads: custom risc-v, eBPF functions

Data Aware (OCS is data aware)

- Storage and apps agree on a data format (e.g., Apache Parquet) and a query format (e.g., Substrait)



Storage Interface: Block? File? Object?

Block

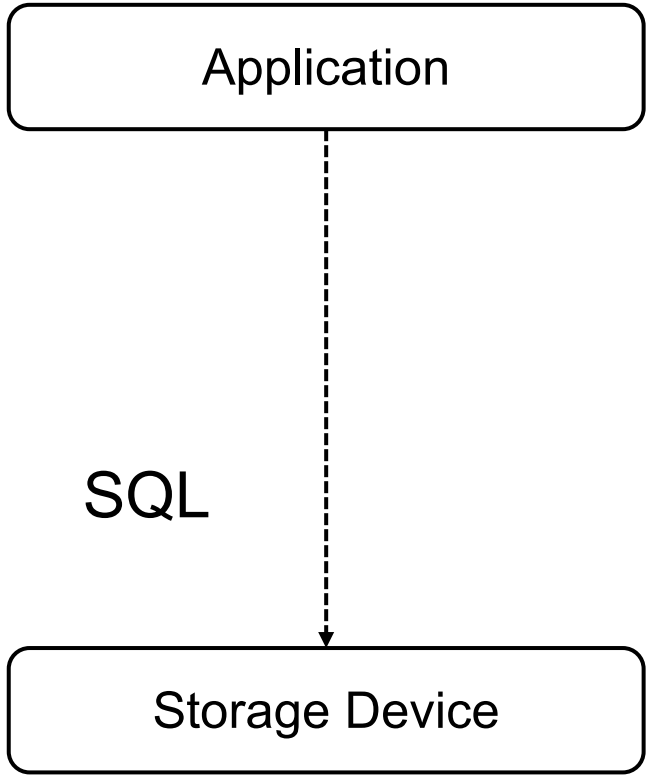
- Applications don't usually talk blocks
- Best for **data agnostic operations** (such as compression, encryption)

Object (OCS is object based)

- Increasingly popular
- Has a growing ecosystem (around S3)

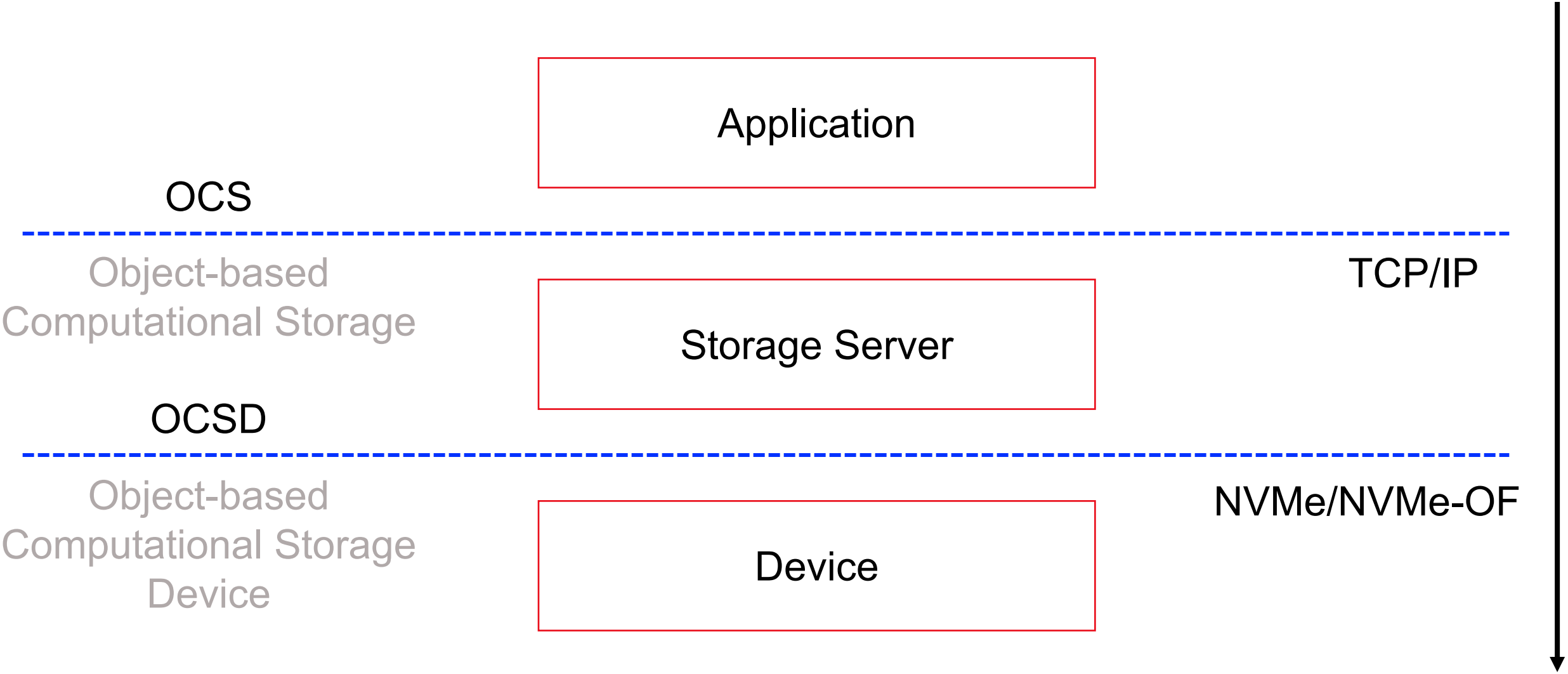
File

- This is another way to do it (e.g.: in the form of NFS-capable storage devices)



Standardization

Query Pushdown



Work Distribution

Device level

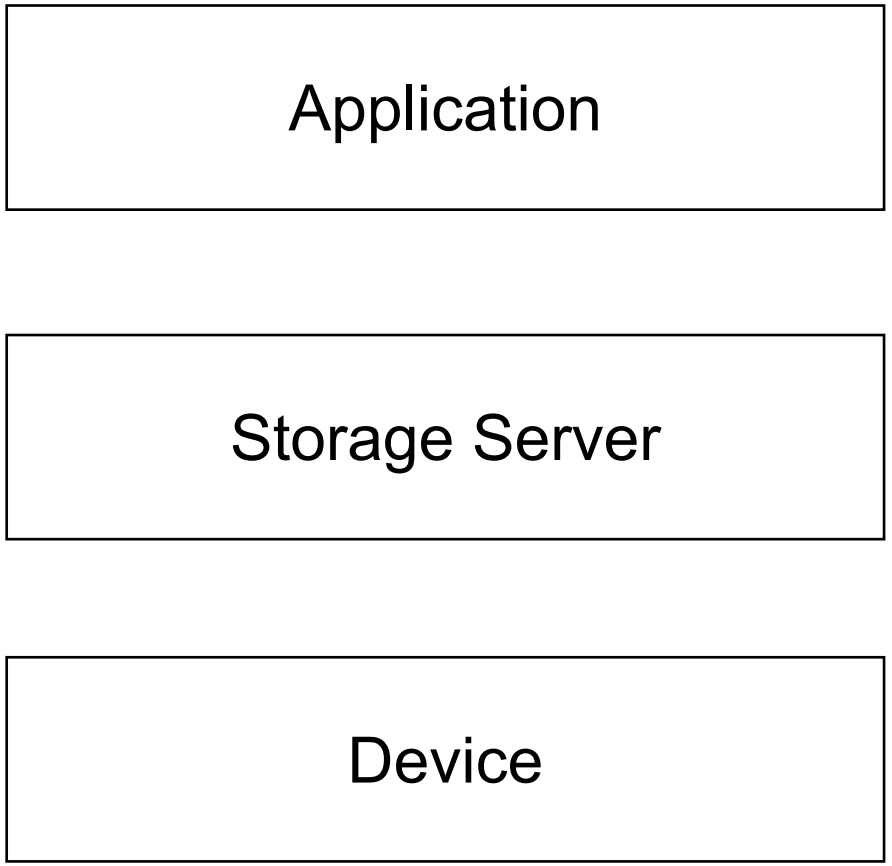
- Single parquet fragment aggregation/projection/filtering

Storage-server level

- Multi-fragment aggregations

Application-level

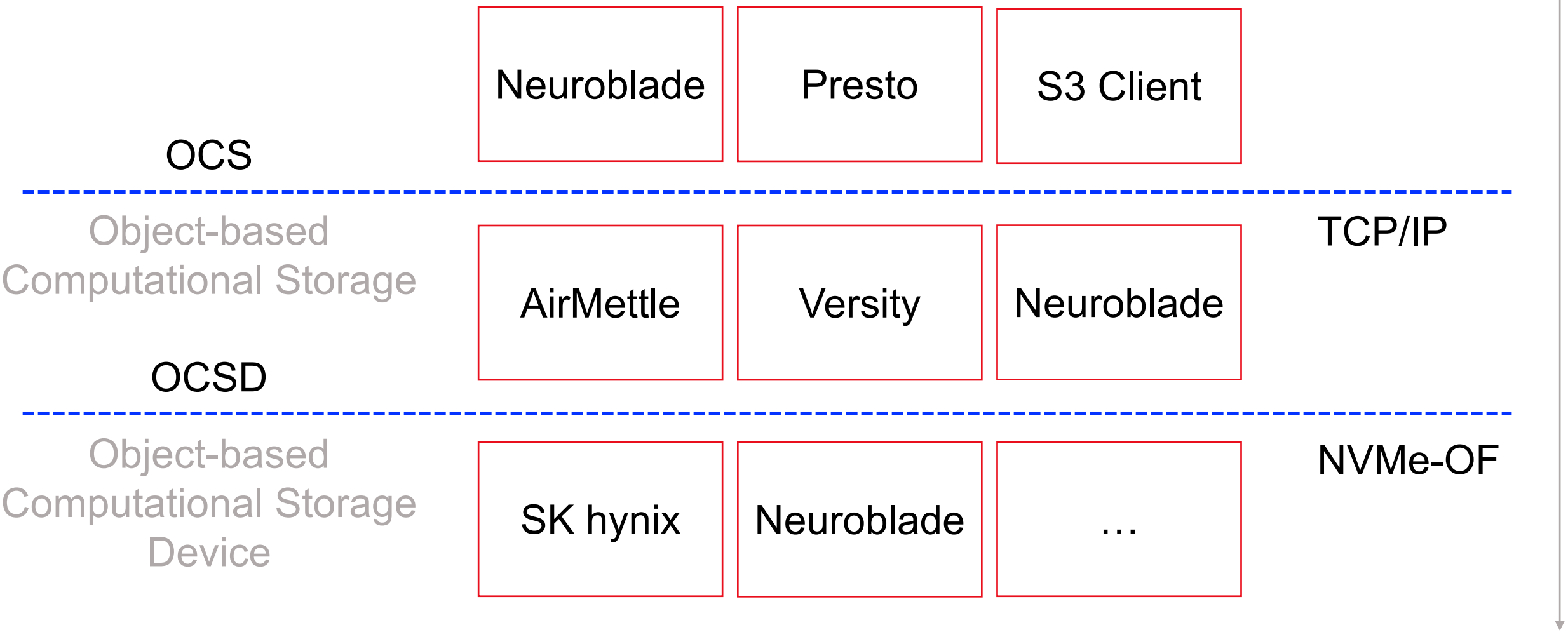
- Multi-table joins



Each layer tries to do as much as it can to process a query

Industry Ecosystem

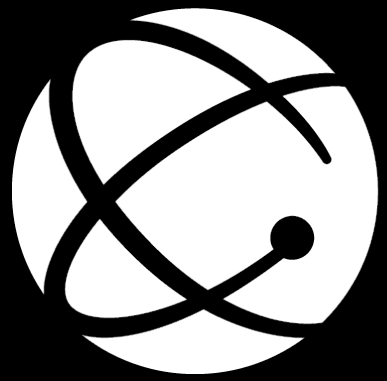
Query Pushdown



Versity Gateway

- Open-source software
- Implements S3 protocols & extendable
- Allows for different backends
- Stateless (scalable)
- Written in go (high performance)
- Friendly community

https://github.com/versity/versitygw



Los Alamos

NATIONAL LABORATORY