SSDs that Think

Mats Öberg, Associate Vice President, Marvell

Compute where the data is

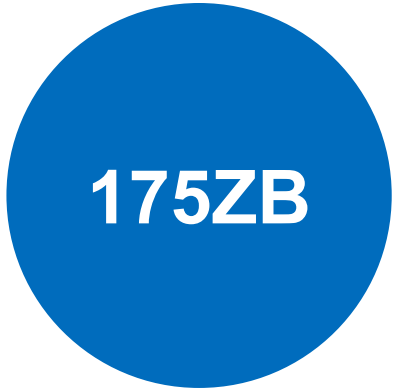
- Increase value of data
- Avoid major data movements
- Efficiency
  - Power
  - Performance
  - Cost

SSDs with compute engines for data processing

People have been mining forever

Data is the mine
of the 21st century
Information is the gold

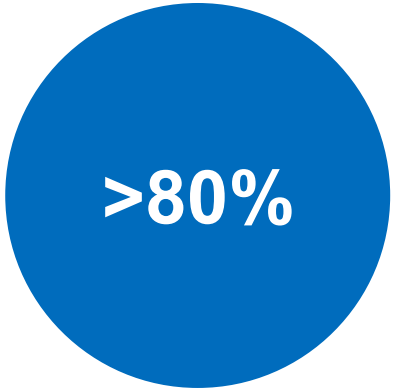
Data mining opportunity



DATA GENERATED

in 2025

26% CAGR



UNSTRUCTED DATA

of all stored data is unstructured “dark data”

BIG DATA ANALYTICS MARKET

In 2028

13.2% CAGR

Sources:
https://www.seagate.com/our-story/rethink-data/
https://www.cio.com/article/220347/ai-unleashes-the-power-of-unstructured-data.html
https://www.fortunebusinessinsights.com/big-data-analytics-market-106179

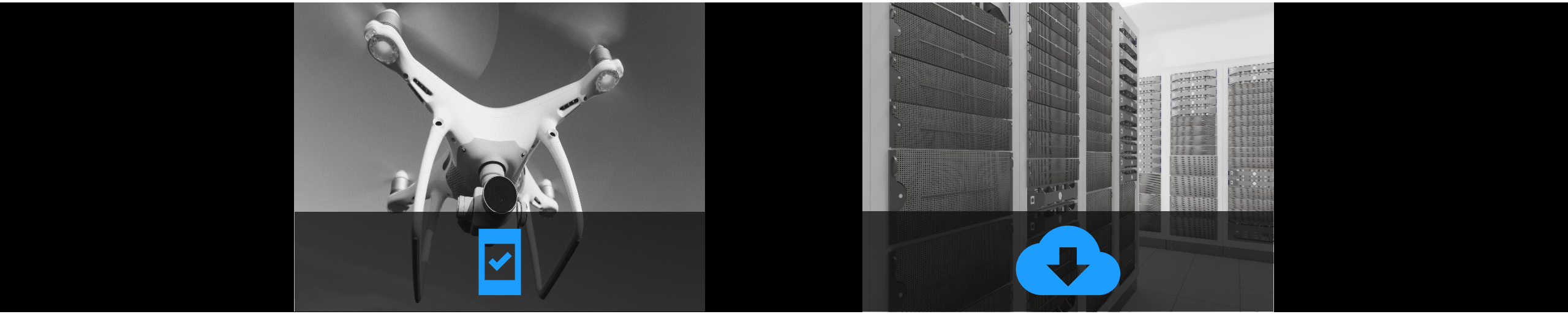
Dark data

Engineer's viewpoint: Dark data is unstructured data that is not analyzed

* Source: https://www.bmc.com/blogs/dark-data/

Computational storage

Moving data around is expensive



EDGE

Upload throughput

Latency

Power

CLOUD

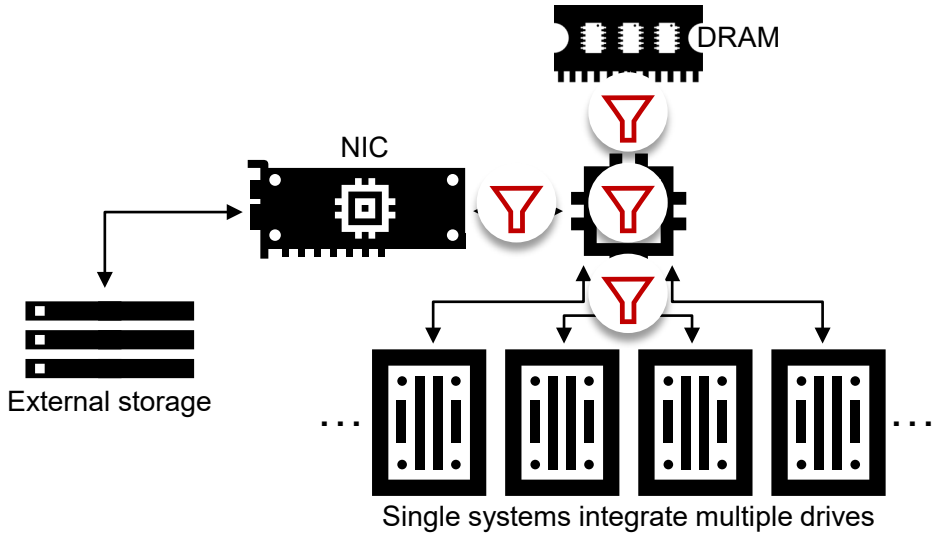
Efficiency

Heterogeneity

Network throughput

Traditional → computational storage

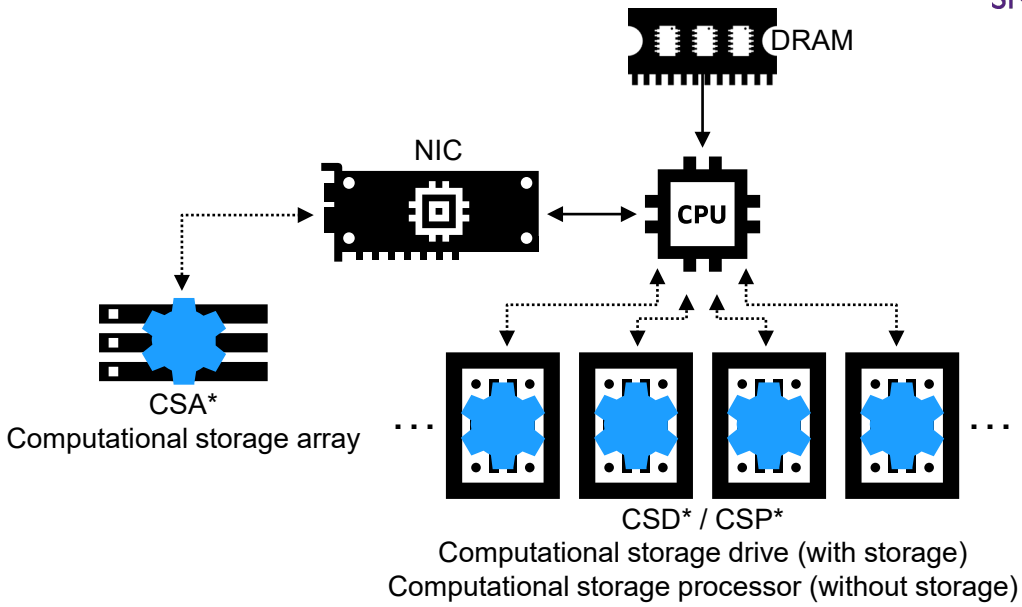
Traditional CPU-Centric architecture



- Centralized compute
- DRAM throughput and capacity challenges
- Massive data movement (in-server & network)
- Fixed compute as workload capacity grows

Computational storage architecture

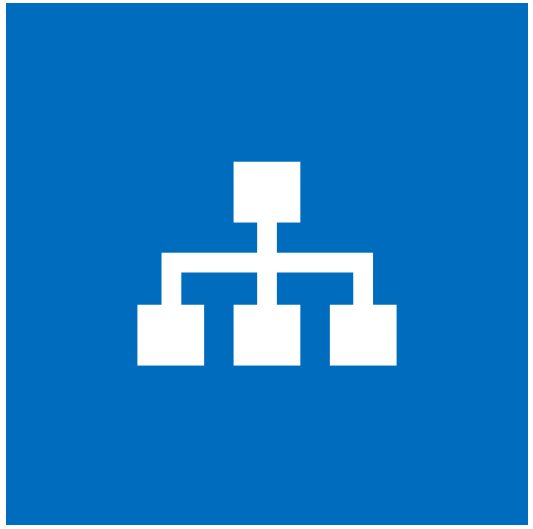
* SNIA Defined



- Parallelize compute (utilization ↑)
- Optimize DRAM throughput & utilization
- Minimize data movement (power/latency)
- Scale compute with capacity

Computational storage value

Analyze

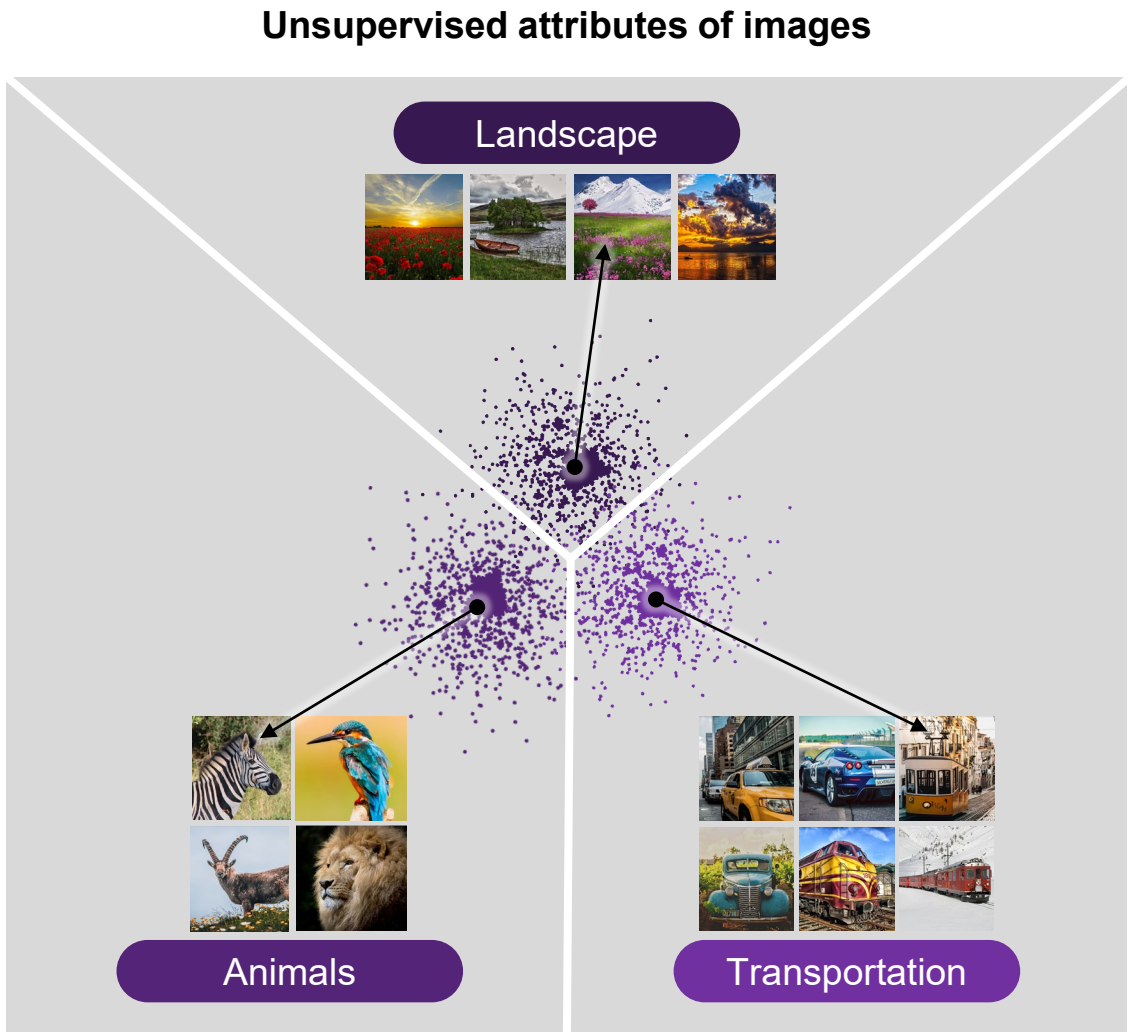


Structure

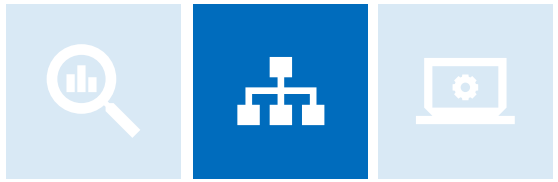
Compute

Analyze: Figure Out What the Data Is

- Transform data into knowledge
- “Metadata”: Compact, high-level representation of data
- Make data understandable and generate value
- Performed inside the storage device
  - Limited data movement
  - Minimized host CPU resource usage
  - Parallelizable over enormous sets of dark data

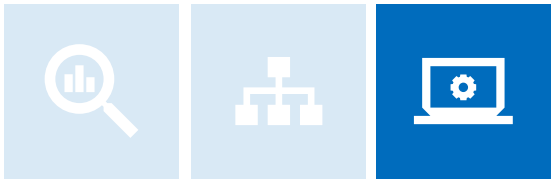


Structure: preprocess data in storage



- Prepare data for ingestion by ML application
  - Decode and resize jpeg images on drive for ML
  - Normalize and prepare database for ML, e.g. encode words into numbers France→1, Germany→2,Canada→3, …
- Filter and select only relevant data to send back to host for processing
  - Provide only images containing cars for processing by host CPU

Compute: process data in storage

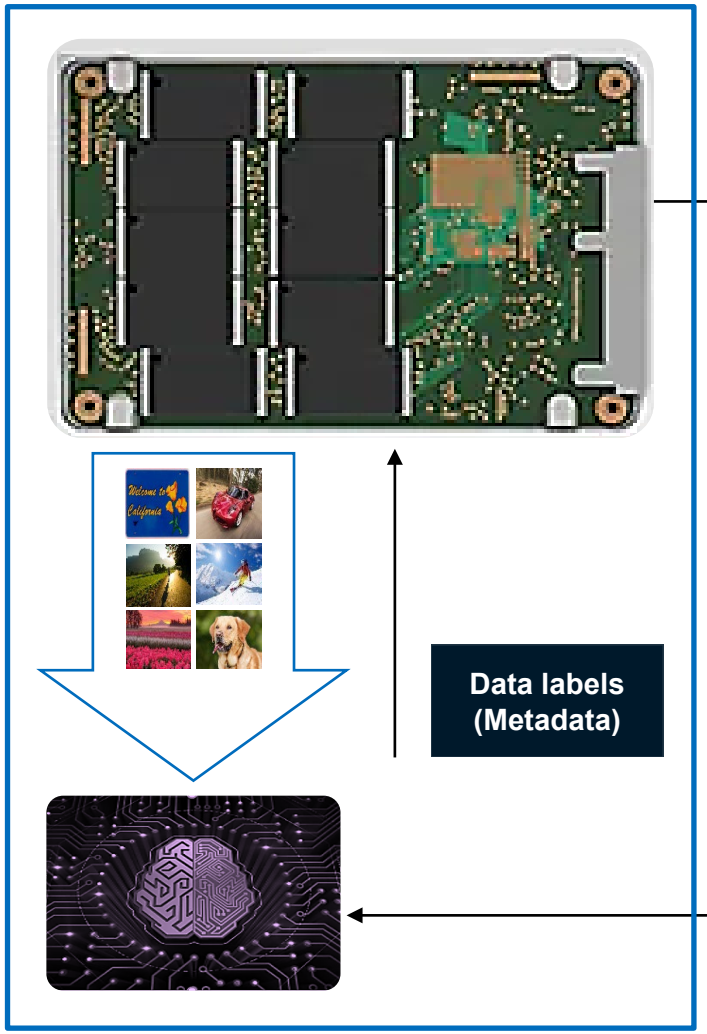


- Apply a neural network inferencing model on data in storage
  - Evaluate new detection feature for autonomous cars on data stored within vehicles prior to live release
- Offline video transcoding
- Evaluate stored data with new algorithm
  - Medical images

Big data analytics

Example

Storage edge compute



Data labels (Metadata)

Understanding content

Understanding relationship

Understanding importance

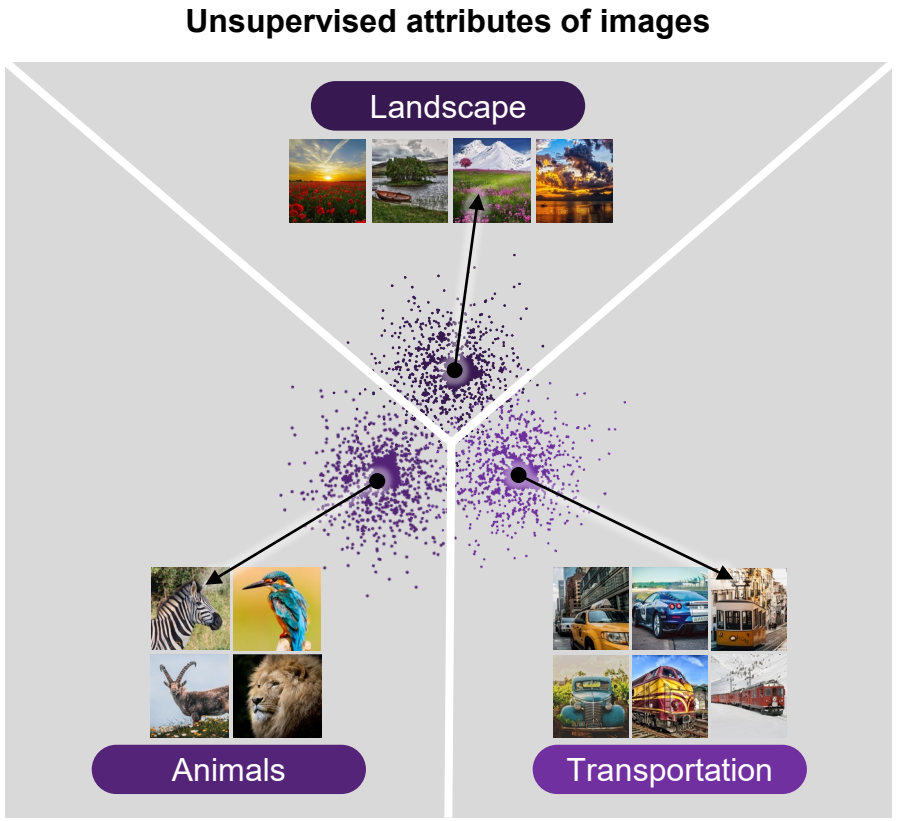
Storage market need for AI solution: Making sense of data

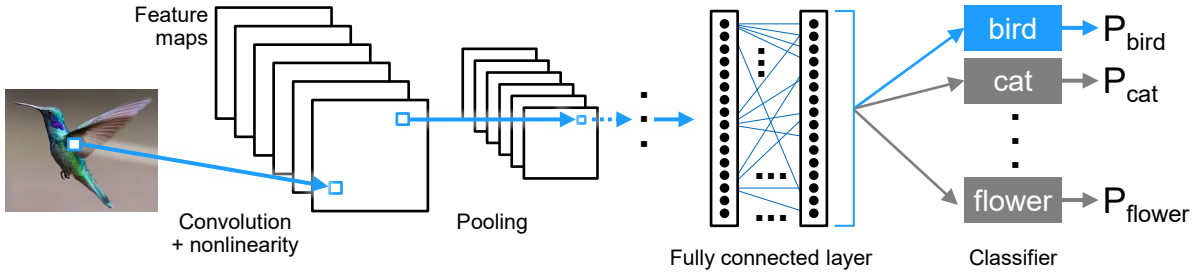
What is the data about?

How does it relate to other data?

Big data analytics: AI inferencing at the storage edge

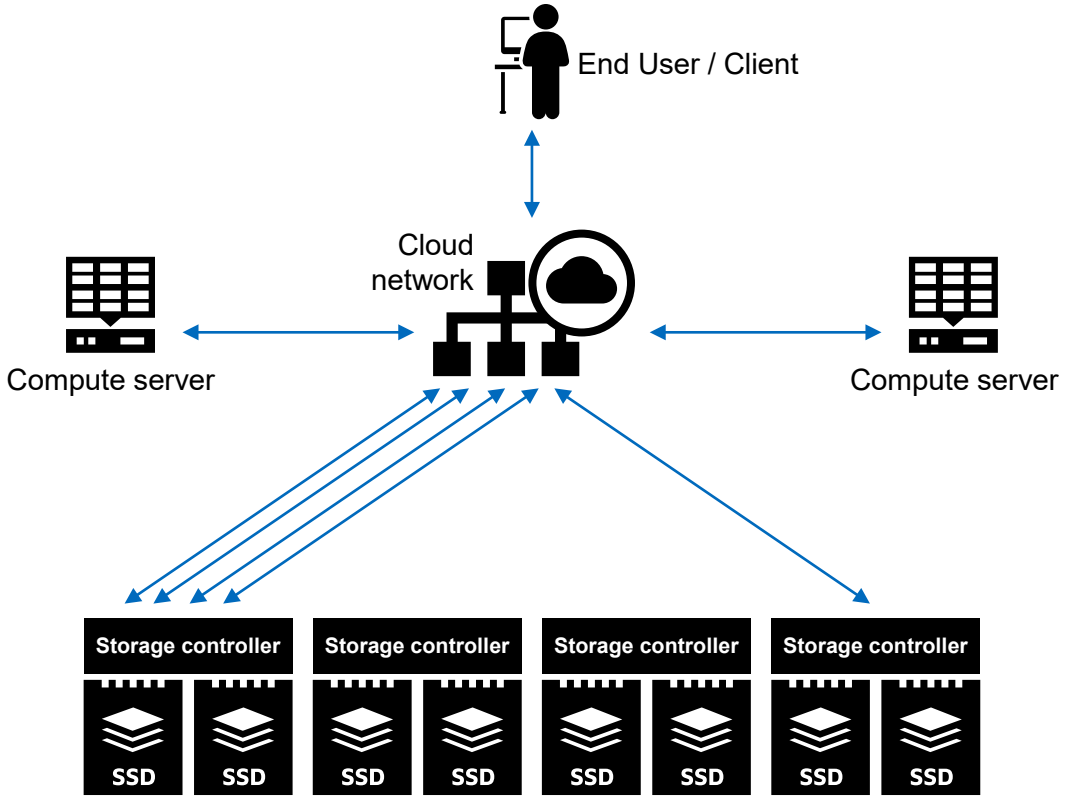
- AI Enabled SSD to generate metadata
- Emerging technologies rely on AI inferencing
- Load a trained AI model to AI engine in SSD





Why compute at the storage edge

- Produce metadata at host (CPU/GPU/FPGA)
  - Storage controller reads data from SSDs and returns to requester over the network
  - Remote processing/searching of data results in large amounts of data traffic
- Produce metadata at storage device
  - Significant reduction in network traffic
  - Storage device offloads host
    - Host CPU can be utilized for higher value computations
  - Scales with additional storage



Key takeaways

- Improve value of dark data with computational storage
  - Tag data for host that is relevant for the host's context
- Offline processing of data at the storage device
  - Analyze and tag data with meta data
  - Pre-process data for host processing
  - Process data
- Ubiquitous AI inferencing drives many opportunities for computational storage

Please take a moment to rate this session.

Your feedback is important to us.

