STORAGE DEVELOPER CONFERENCE
SDC 22
Fremont, CA
September 12-15, 2022
BY Developers FOR Developers
A SNIA Event

Next Generation Architecture for Scale-out Block Storage

Jaspal Kohli
VP, Storage Software
Fungible
jaspal.kohli@fungible.com

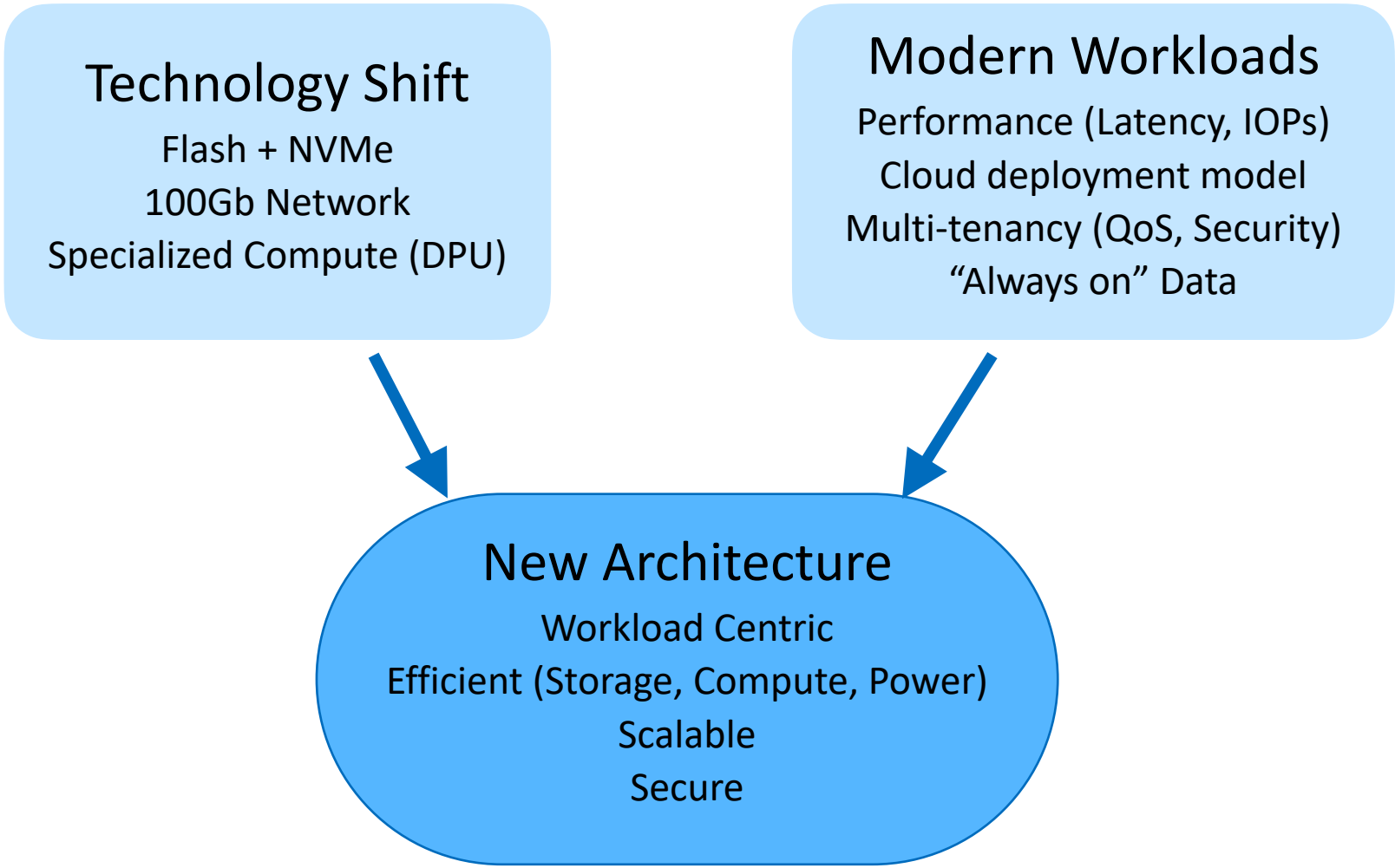
Agenda

- Motivation
- Architecture Principles
- Implementation
- Summary

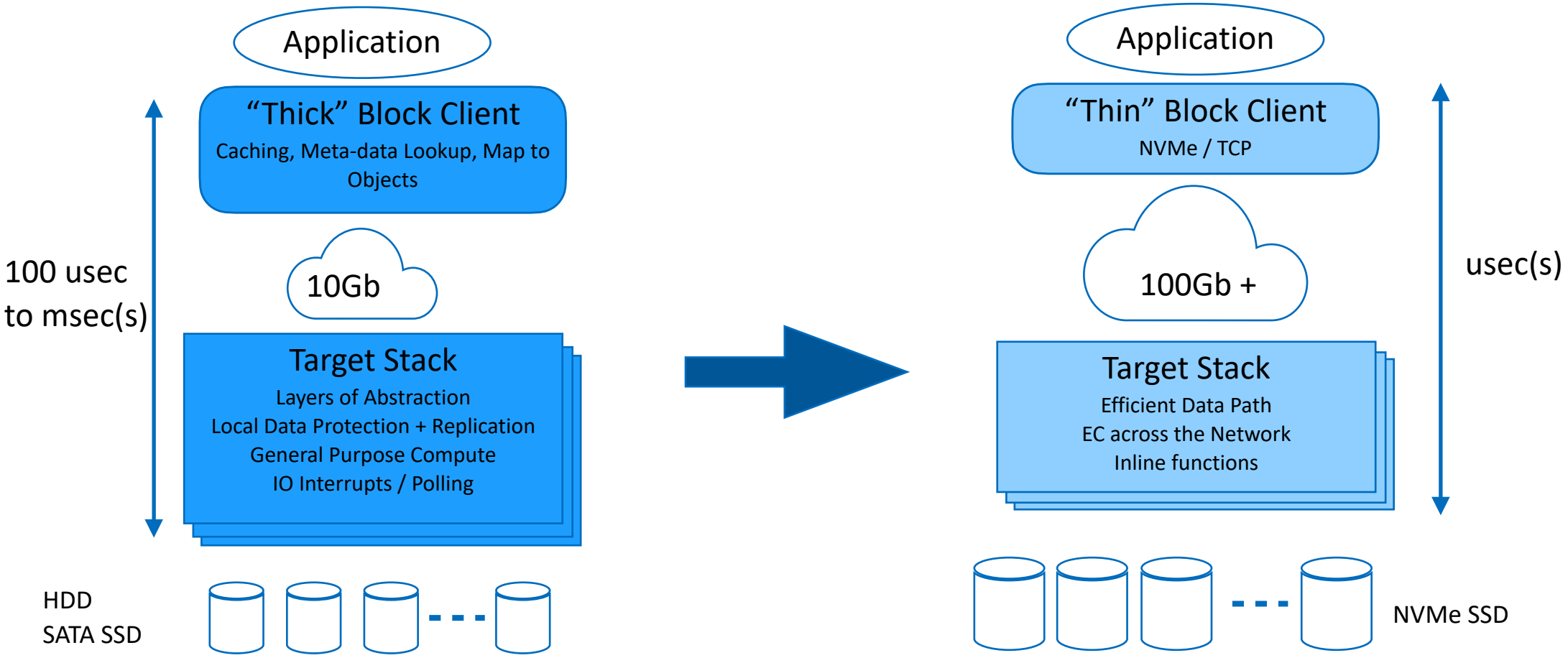
Motivation

"A Perfect Storm"

A Perfect Storm



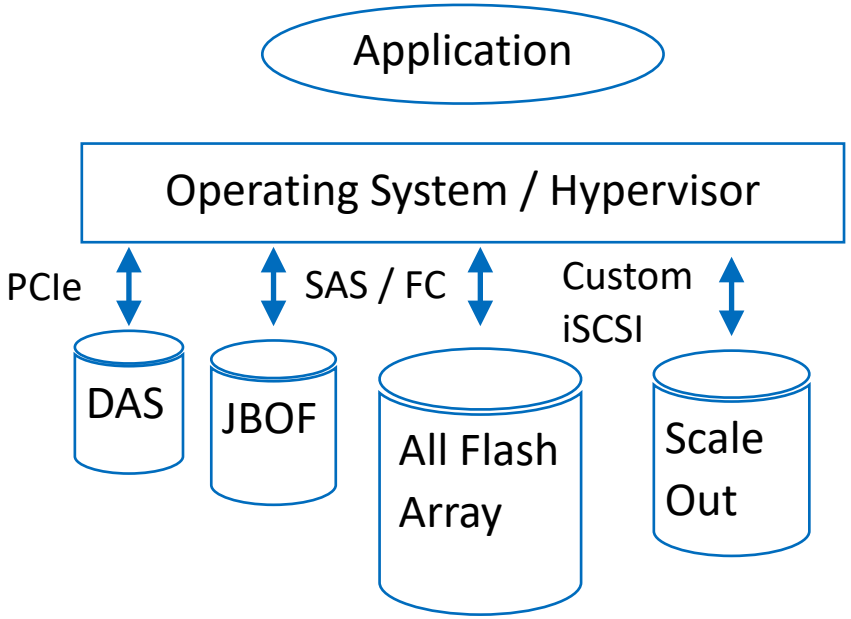
Technology Shift: IO leaps ahead of Compute



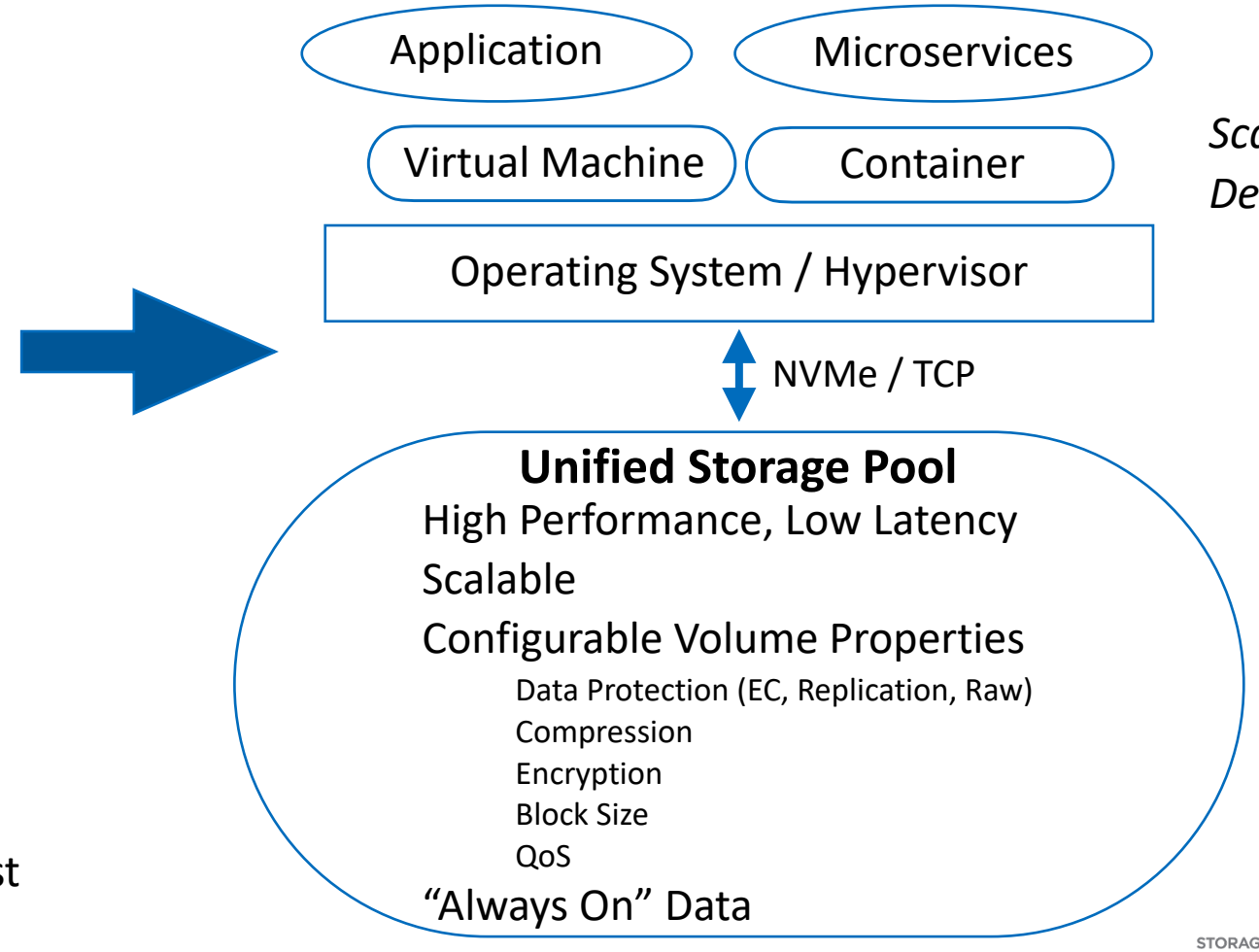
Striping/hashing required for performance

Each SSD can serve many applications

Modern Workloads: Need for Adaptive Storage



Storage silos => Poor utilization, Higher Cost



Scale-out Deployment

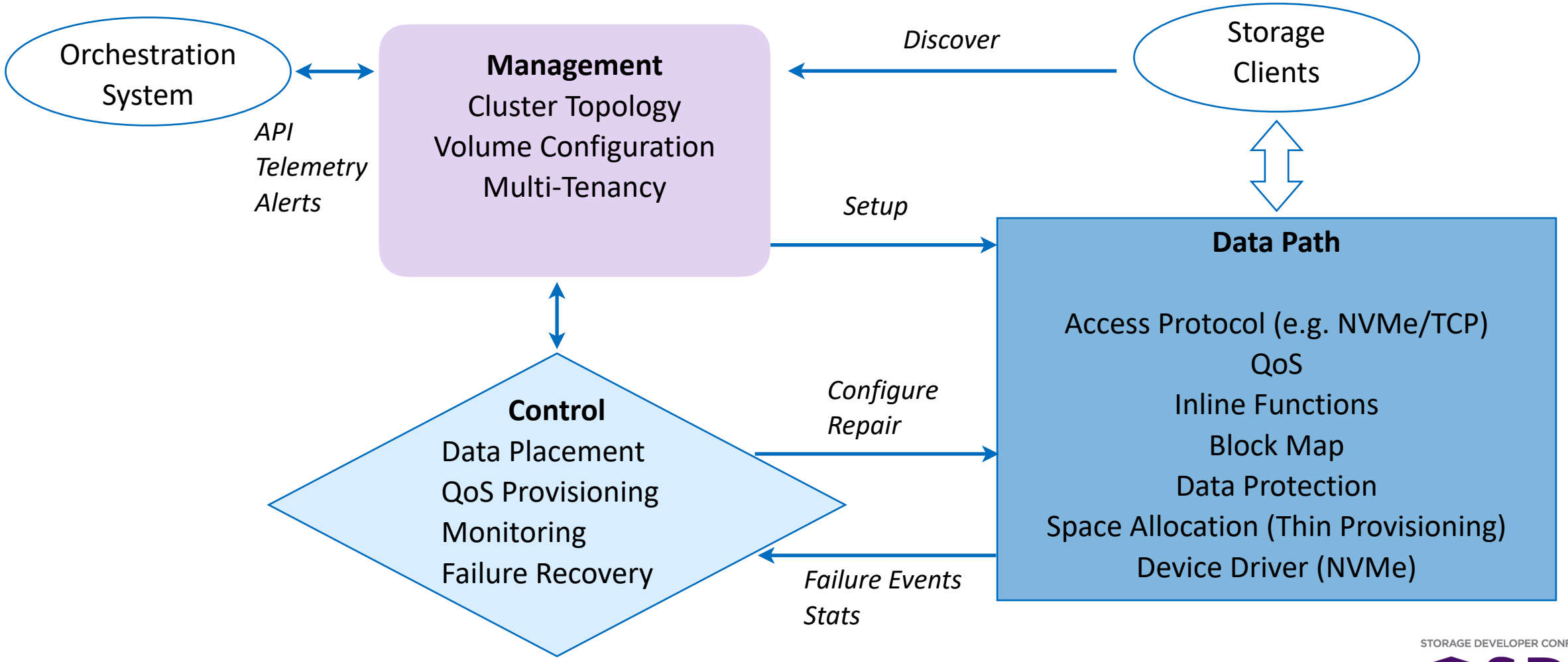
Architecture Principles

"The age of specialization"

Networking Analogy

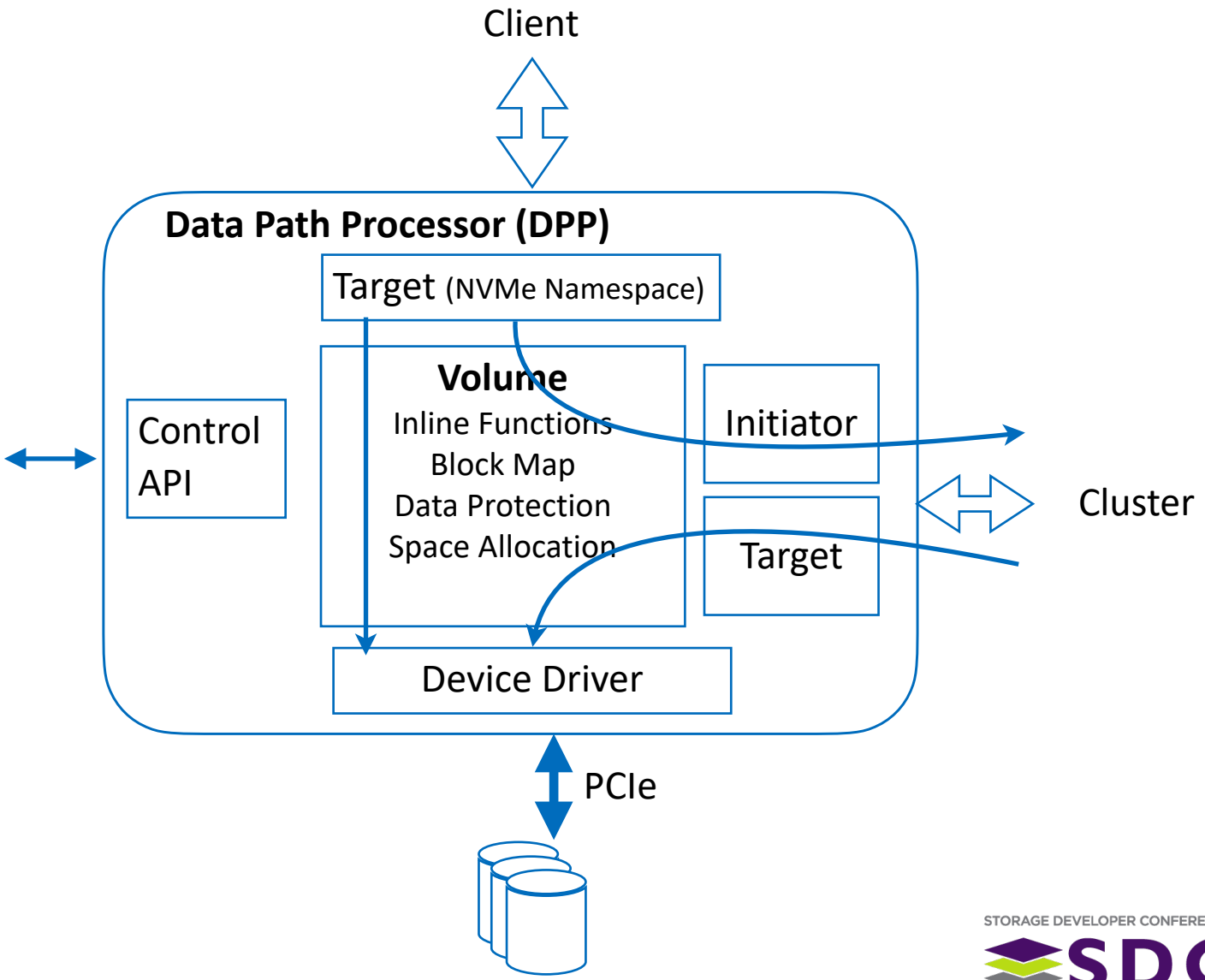
- The internet boom in the late 90s and early 2000s
  - Scale (number of hosts)
  - Network bandwidth
  - Latency
  - Security
- New Architecture for IP Packet Routing
  - Specialized data path processors
    - Latency and Performance (packets per second)
    - Inline functions (security, inspection, NAT, etc)
    - Telemetry
  - Clean separation of the Control Plane
    - Route tables
    - Monitoring
- Network Virtualization

Separation of Roles: Management, Control and Data

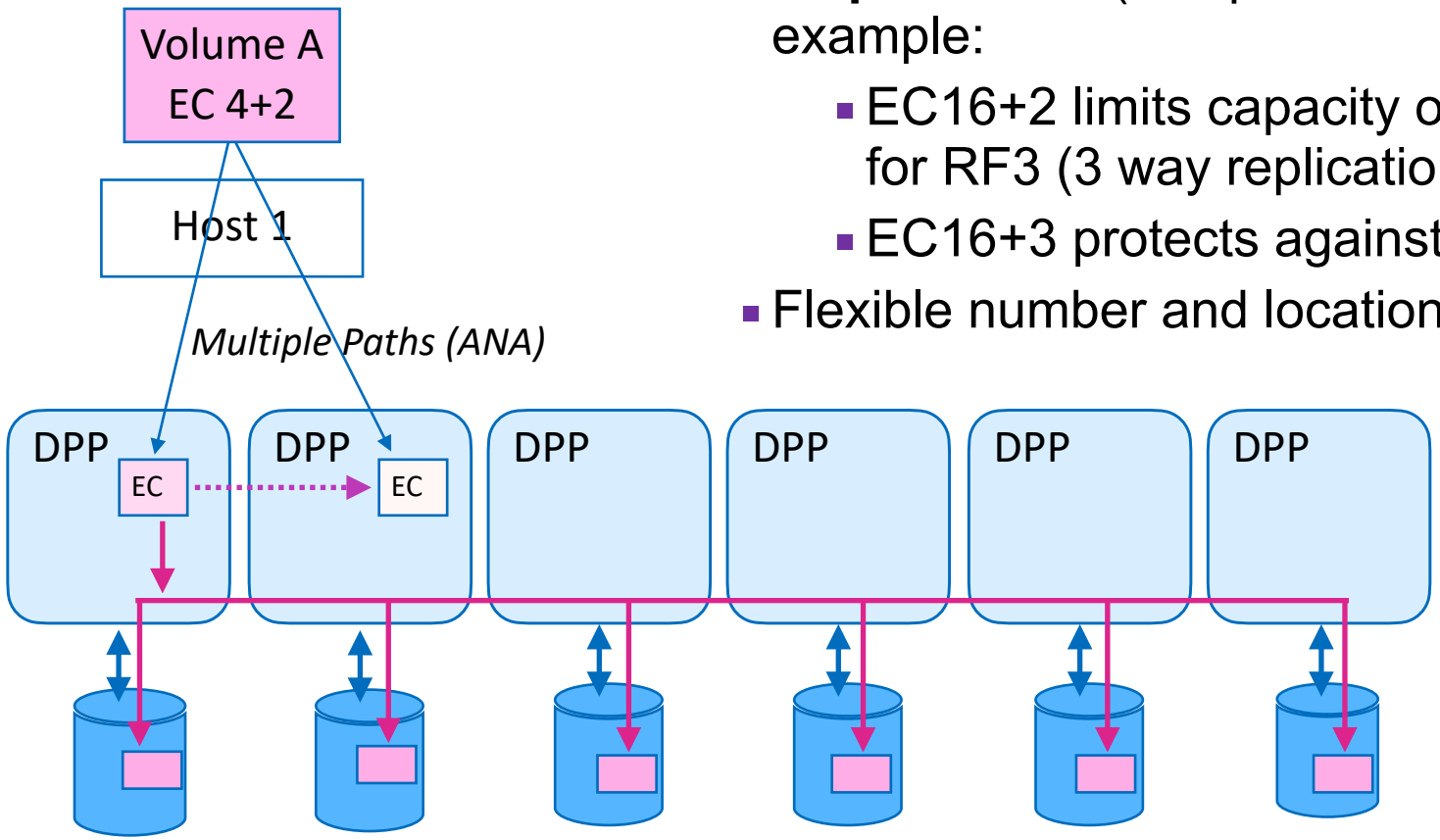


Storage Data Path: Performance and Flexibility

- Low Latency
  - Path length
  - Meta Data
- Composable pipeline per volume
  - Host map (NVMe Namespace)
  - Data protection scheme
  - Inline functions (e.g. compression, encryption)
  - Data placement
  - Computational Storage
- Scale
  - Fast context switch between pipelines
- Clustering and scale-out

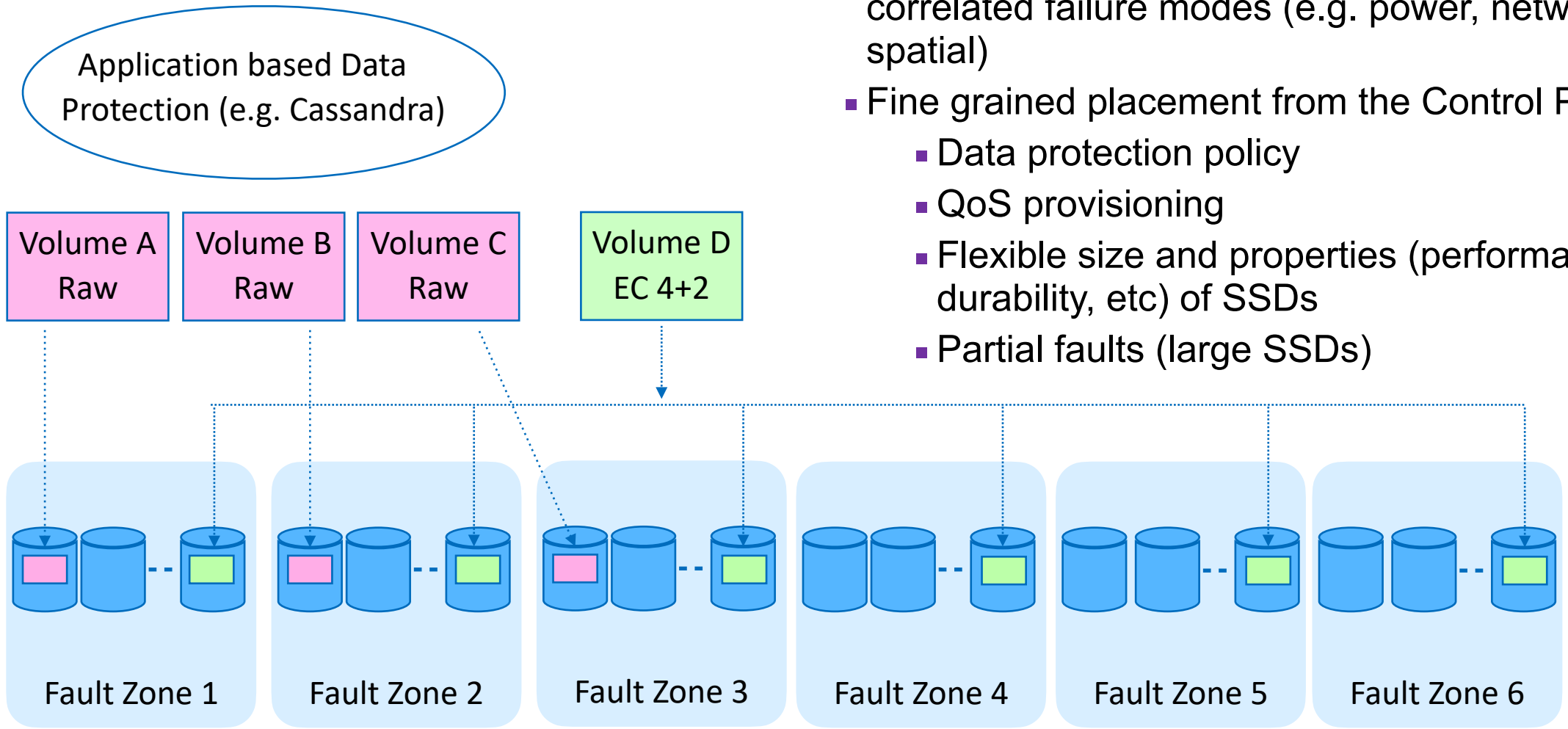


Efficient Data Protection across the Network



- Per volume configurable EC scheme across the network
- *Optimize* **data protection** as well **space overhead** and **write amplification** (compared to local EC and replication). For example:
  - EC16+2 limits capacity overhead to 12.5% relative to 200% for RF3 (3 way replication)
  - EC16+3 protects against 3 concurrent failures
- Flexible number and location of redundant paths

Controlled Data Placement



- Fault Zone: subset of the cluster with correlated failure modes (e.g. power, network, spatial)
- Fine grained placement from the Control Plane
  - Data protection policy
  - QoS provisioning
  - Flexible size and properties (performance, durability, etc) of SSDs
  - Partial faults (large SSDs)

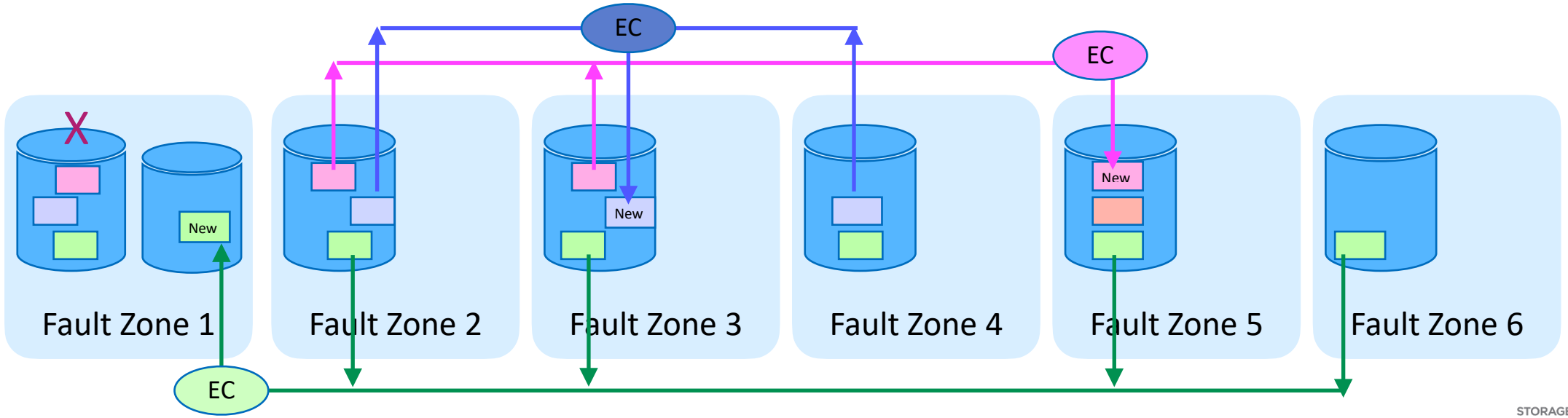
Efficient Rebuild and Cluster Resize

- Rebuild each volume impacted by a failure (e.g. SSD)
  - Scales with number of volumes
  - Only rebuild blocks in use
  - Executed by Primary DPP for the volume
- Cluster resize is policy driven from Control Plane
  - Minimal to no data relocation
- Both perform better with size of cluster

Volume A
EC 2+1

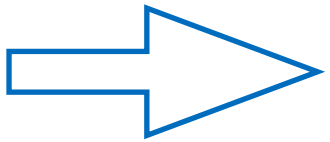
Volume B
EC 2+1

Volume C
EC 4+2



Linear Scaling

- Work done by a DPP remains independent of cluster size. For each volume, it can:
  - Host the Primary Path or Secondary Path
  - Execute the IO Pipeline (including inline functions)
  - Store data (or portion of it)
- A DPP does not have to keep track of other DPPs in the cluster
- Distribution of data, processing and inter-DPP interactions are determined by the Control Plane



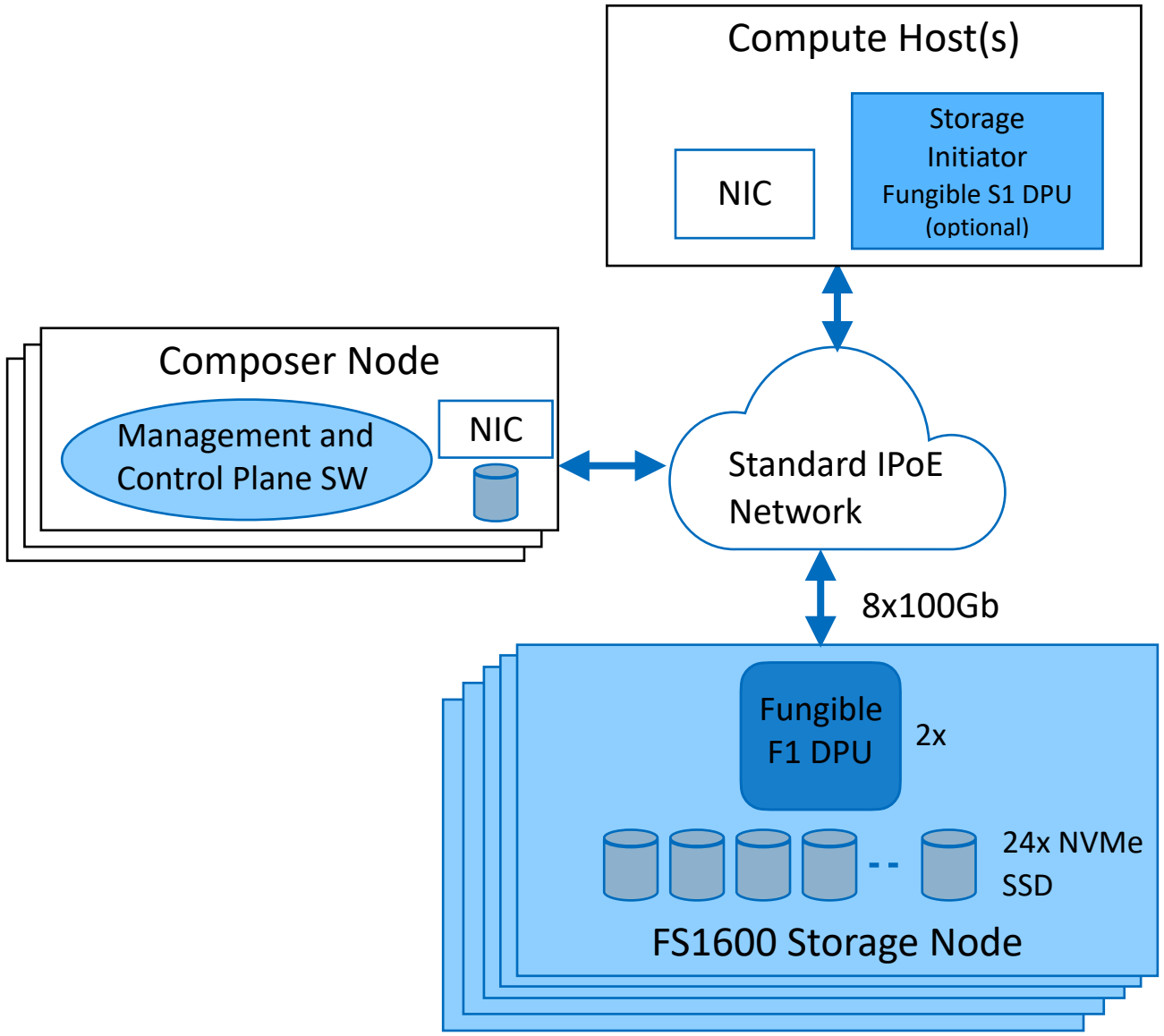
- Total IOPs that can be served by a cluster ***scales linearly with the number of DPPs***.
- Speed of recovery improves with cluster size
  - Rebuild on SSD failure
  - Path Failover on DPP failure

Implementation: Fungible Storage Cluster

"Architecture Specialization"

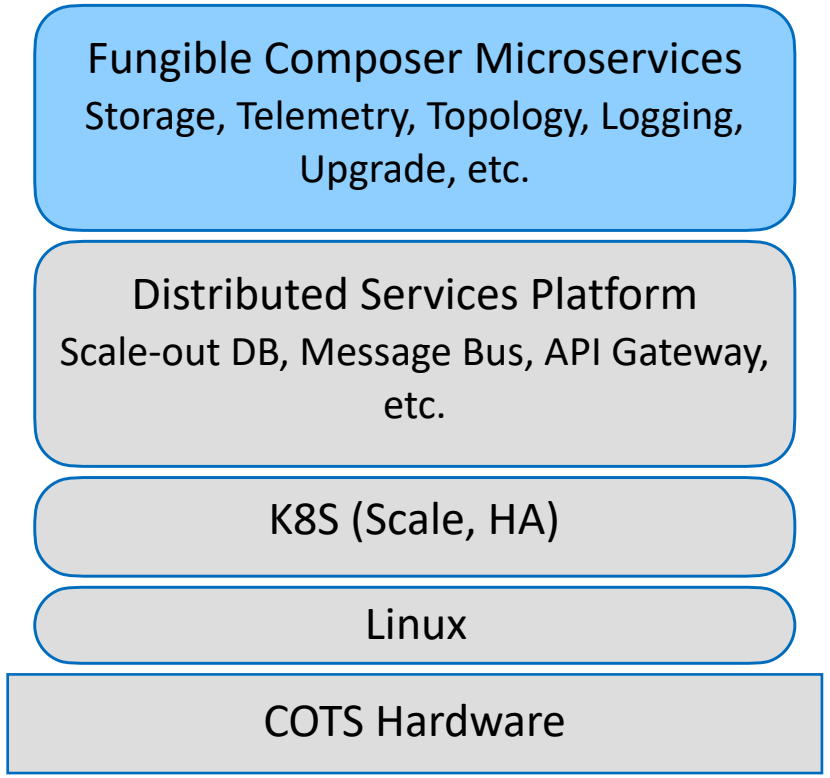
Overview

- Scale-out Disaggregated Block Storage
  - NVMe/TCP
- Clean separation of Management and Control Plane
- Fungible F1 DPU: High Performance Data Path
  - Linear Scaling
- Adaptive Storage: per volume policy
  - Data protection, Compression, Encryption, QoS, Block Size
- Data Integrity (Block CRC)
- Snapshots and Clones
- Intent API for integration with Orchestration Systems
  - Cinder Driver
  - CSI plugin



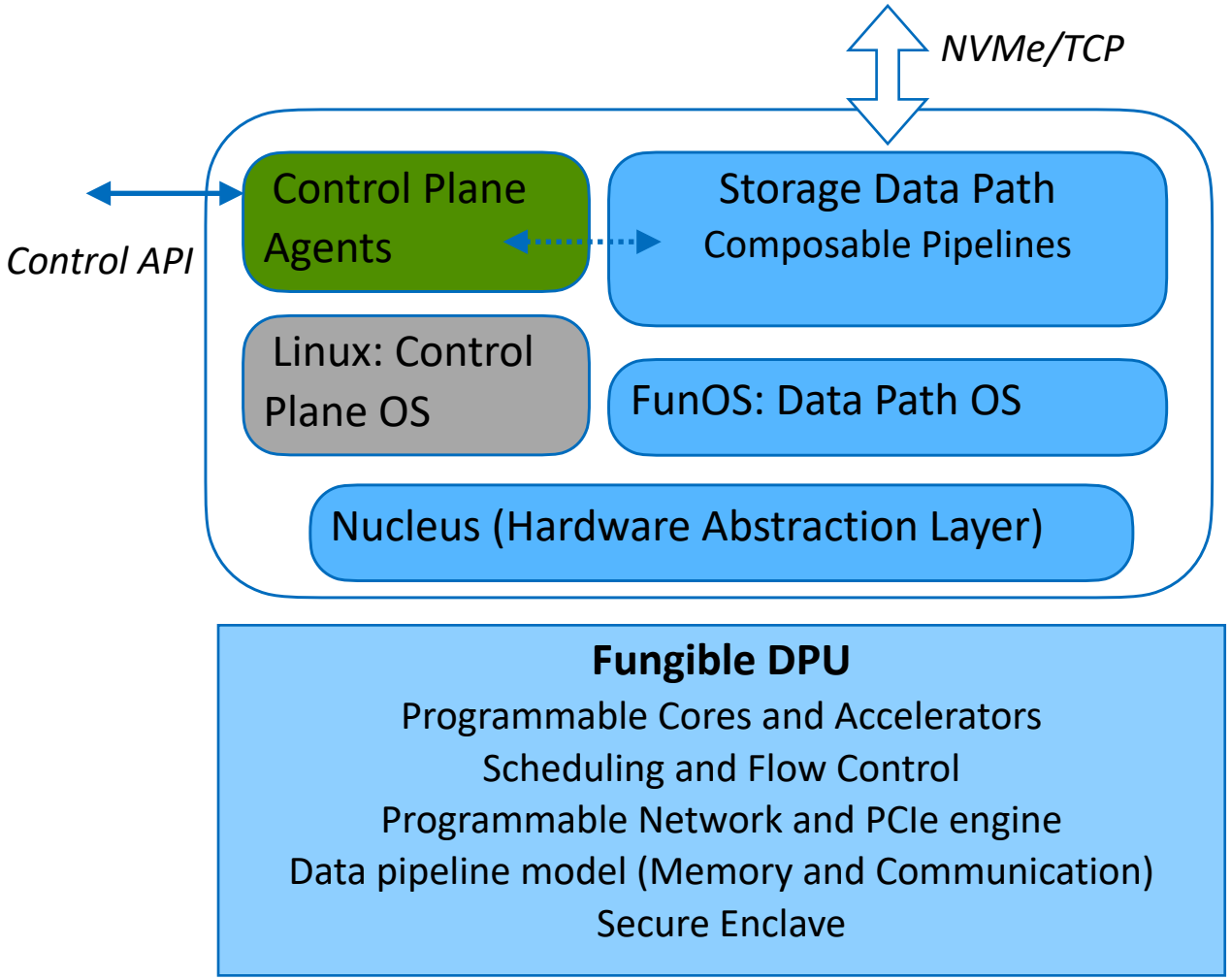
Fungible Composer Software

- Scale-out
  - HA, Volume count, API rate
- Modern application architecture
  - Microservices
  - Distributed Services Platform
  - High Level languages (Go, Python)
  - Agility



Storage Data Path on Fungible DPU

- Per DPU Performance for Random Read
  - 6.5M IOPs (at 4KB)
  - 37.5GB/s (at 16K)
- Path length through the DPU is <10us
  - Includes inline services (e.g. Compression, Encryption, Block CRC, QoS)
- Linear scaling with DPUs
  - Measured up to 32 DPUs
  - Plan to qualify 64/128 DPUs



Summary

Recap

- Technology shift + Workload evolution -> A perfect storm
- Core architecture principles
  - Clean separation of Planes: Management, Control and Data
  - High performance and composable data path
  - Efficient Data Protection over the network
  - Controlled Data Placement
  - Linear Scaling
- Fungible Storage Cluster
  - DPU as a Data Path Processor
  - Leverage distributed systems technology for Management and Control

Related Talks

- The Rise of DPU-based Storage System *by Jaishankar (Jai) Menon, Chief Scientist, Fungible*.
  - DPU Track, Wed at 4:35pm
- Data Processing Unit as a Storage Initiator *by Pratapa Reddy Vaka, Sr. Director, Storage Software, Fungible*
  - DPU Track, Wed at 3:35pm

Please take a moment to rate this session.

Your feedback is important to us.