FlexSDS All Flash Storage

FlexSDS Inc.

https://www.flexsds.com/

What is FlexSDS?

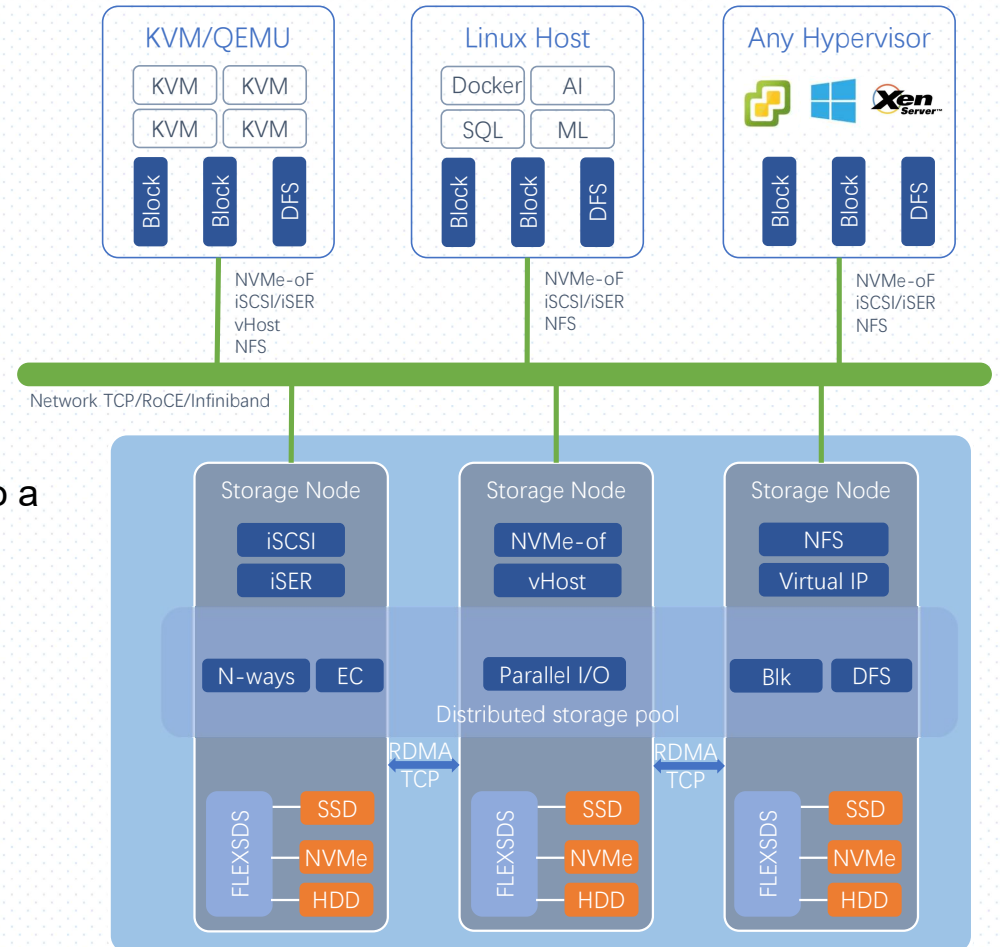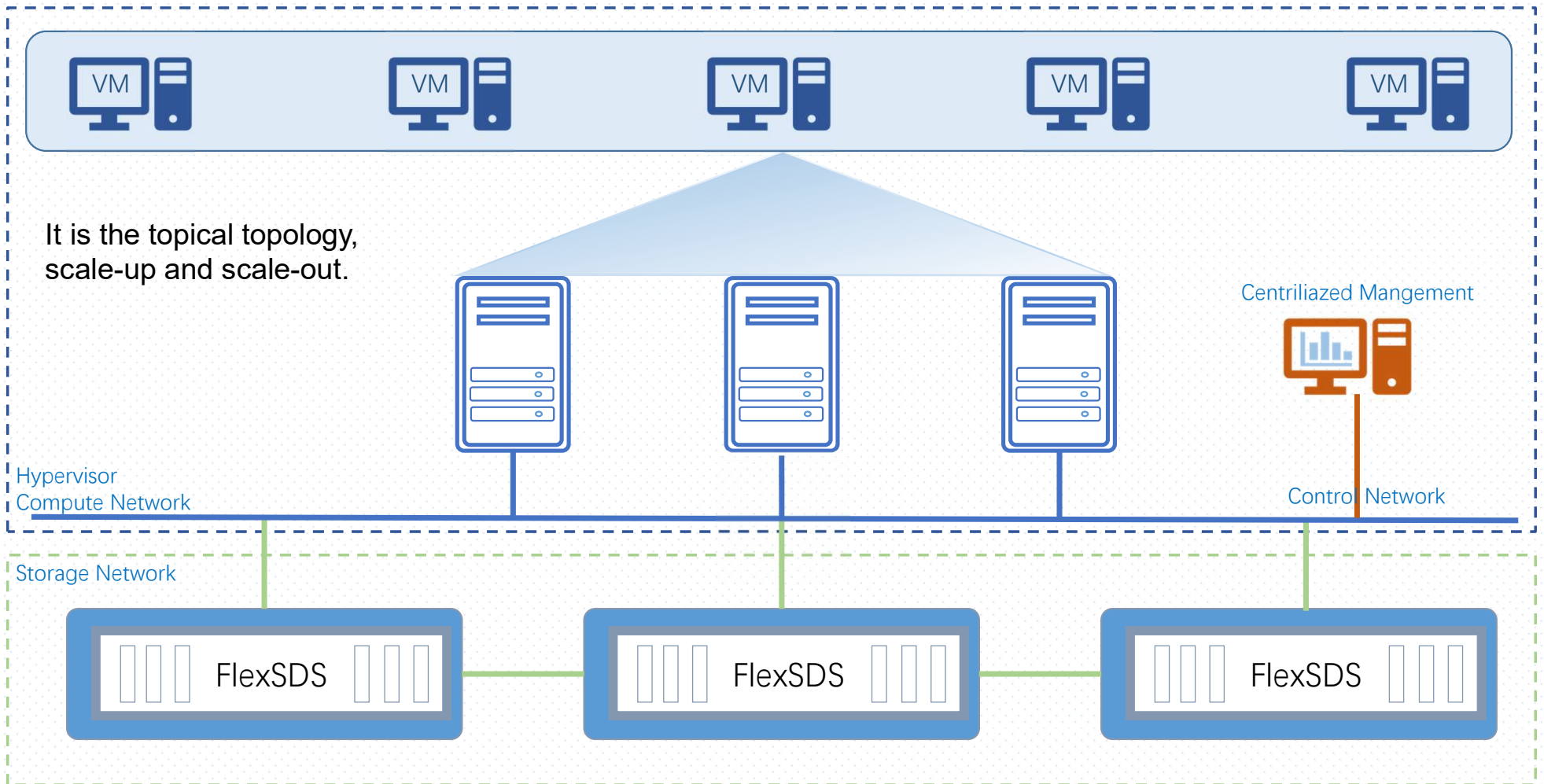X86 Standard Servers + FlexSDS = Distributed Storage Cluster
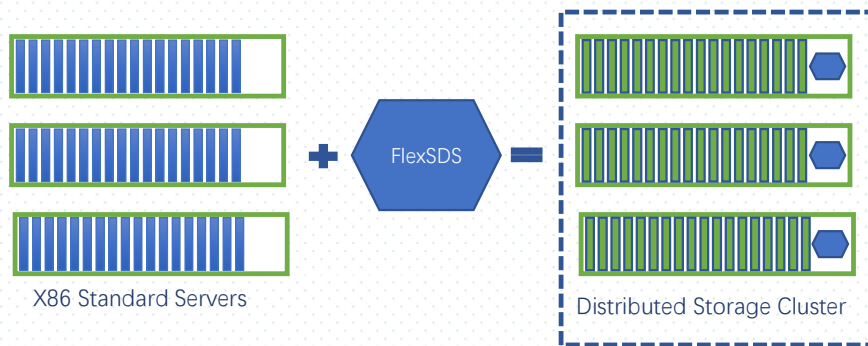
FlexSDS is software only high-performance scale-out storage software that convert any x86 based machine into a scale-up, scale-out, software defined storage cluster.
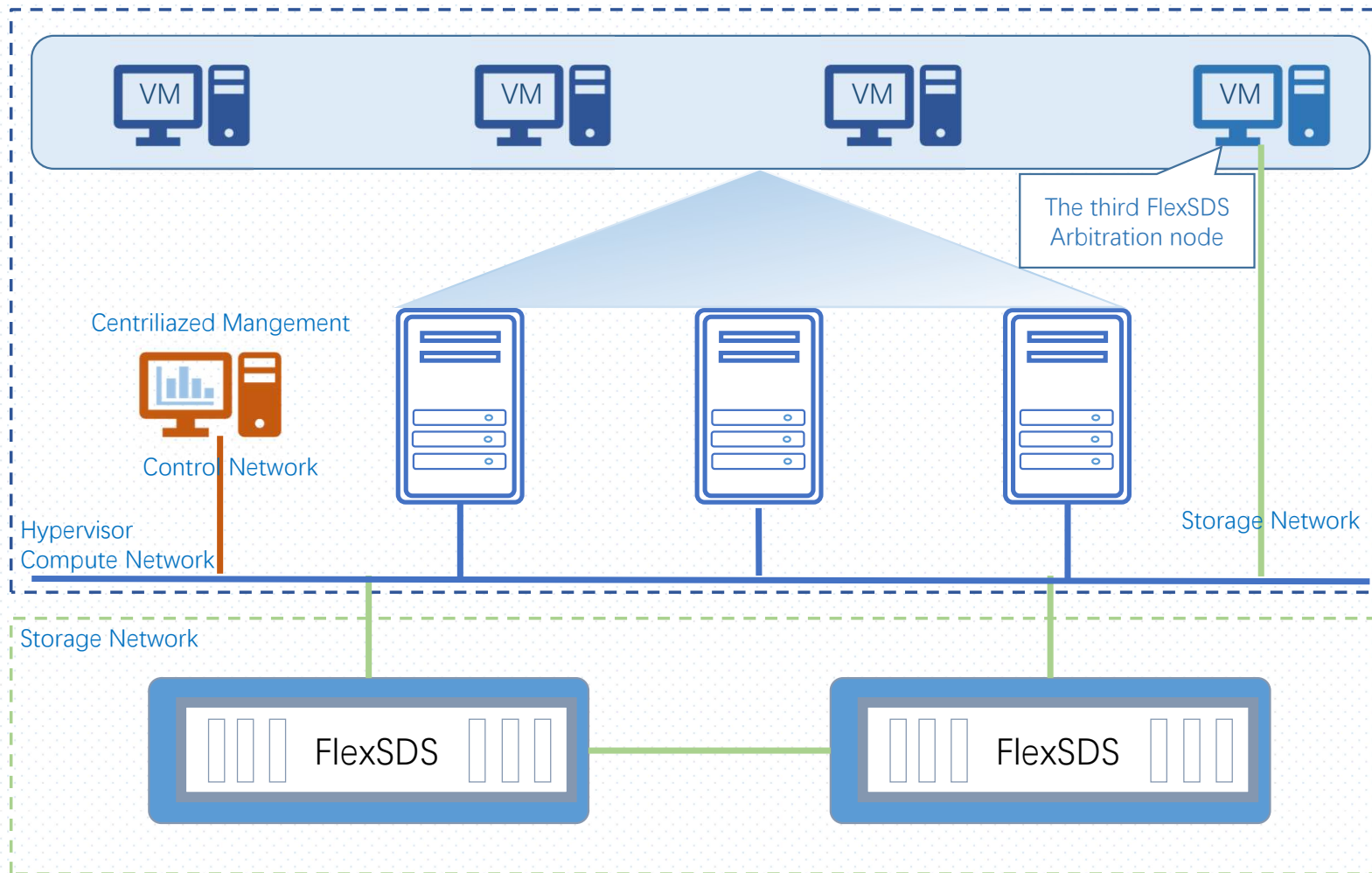
- Dynamic Scalability: 1-1024 scale out
- Block and Distributed File System (DFS)
- Erasure Coding (EC)
- Multi-Copy Redundancy
- All async and parallel I/O
- RAW & Thin Provisioning
- Elastic Storage Pools
- Unified Storage Interface
- Any type of backend storage
- Web-Based Management Console

KVM/QEMU
KVM | KVM
KVM | KVM
Block | Block | DFS

NVMe-oF
iSCSI/iSER
vHost
NFS

Linux Host
Docker | AI
SQL | ML
Block | Block | DFS

NVMe-oF
iSCSI/iSER
NFS

Any Hypervisor
Block | Block | DFS

NVMe-oF
iSCSI/iSER
NFS

Network TCP/RoCE/Infiniband

Storage Node
iSCSI
iSER
N-ways | EC
FLEXSDS — SSD | NVMe | HDD

Storage Node
NVMe-of
vHost
Parallel I/O
FLEXSDS — SSD | NVMe | HDD

Storage Node
NFS
Virtual IP
Blk | DFS
FLEXSDS — SSD | NVMe | HDD

Distributed storage pool

RDMA TCP

RDMA TCP

Topology: 3(+)nodes scale-out

VM VM VM VM VM

It is the topical topology, scale-up and scale-out.

Centriliazed Mangement

Hypervisor
Compute Network

Control Network

Storage Network
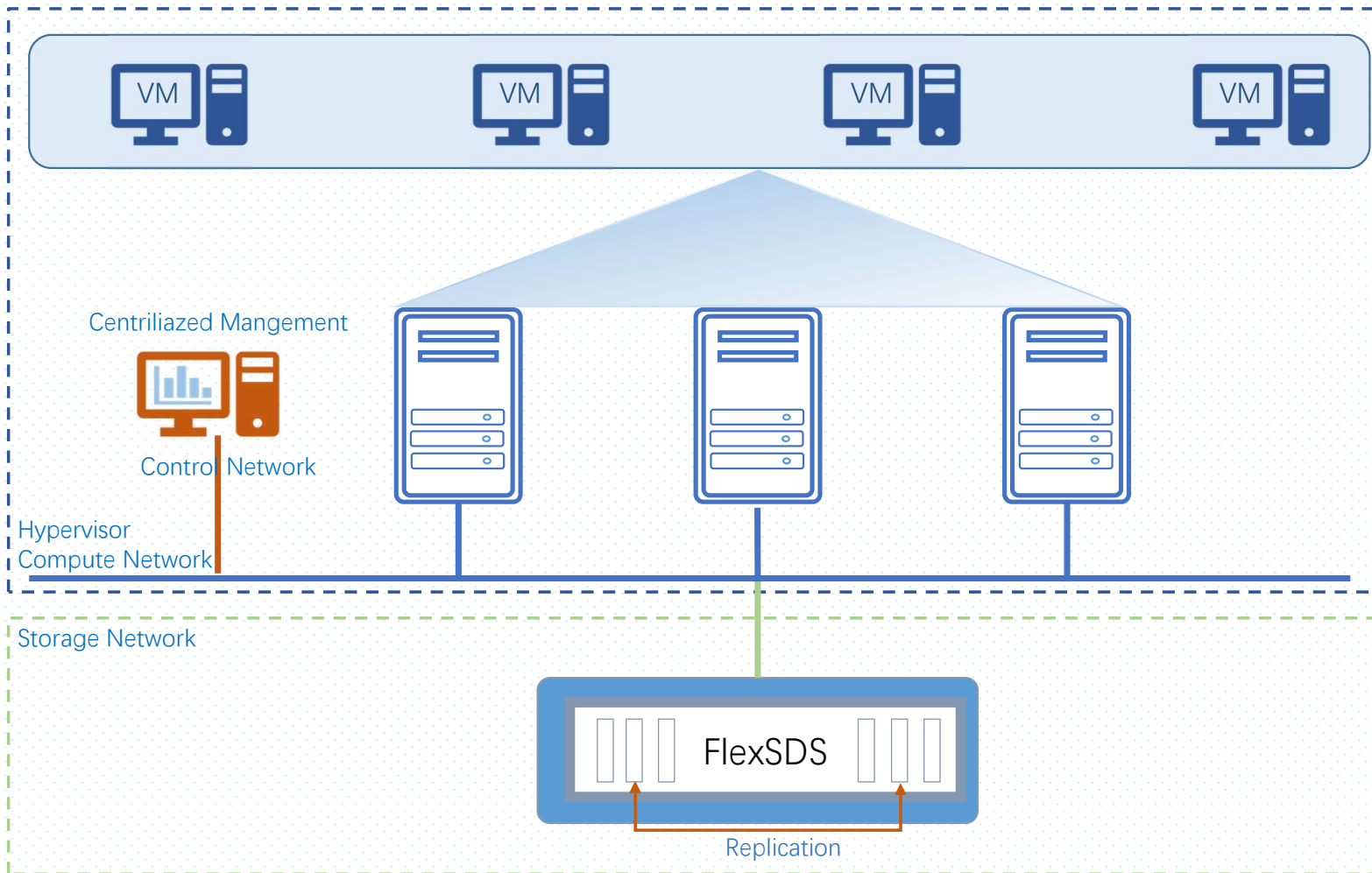
FlexSDS   FlexSDS   FlexSDS

Topology: Two nodes HA

- Using only two storage nodes.
- Lowest TCO.
- Using thin machine as the thrid node for arbitration.
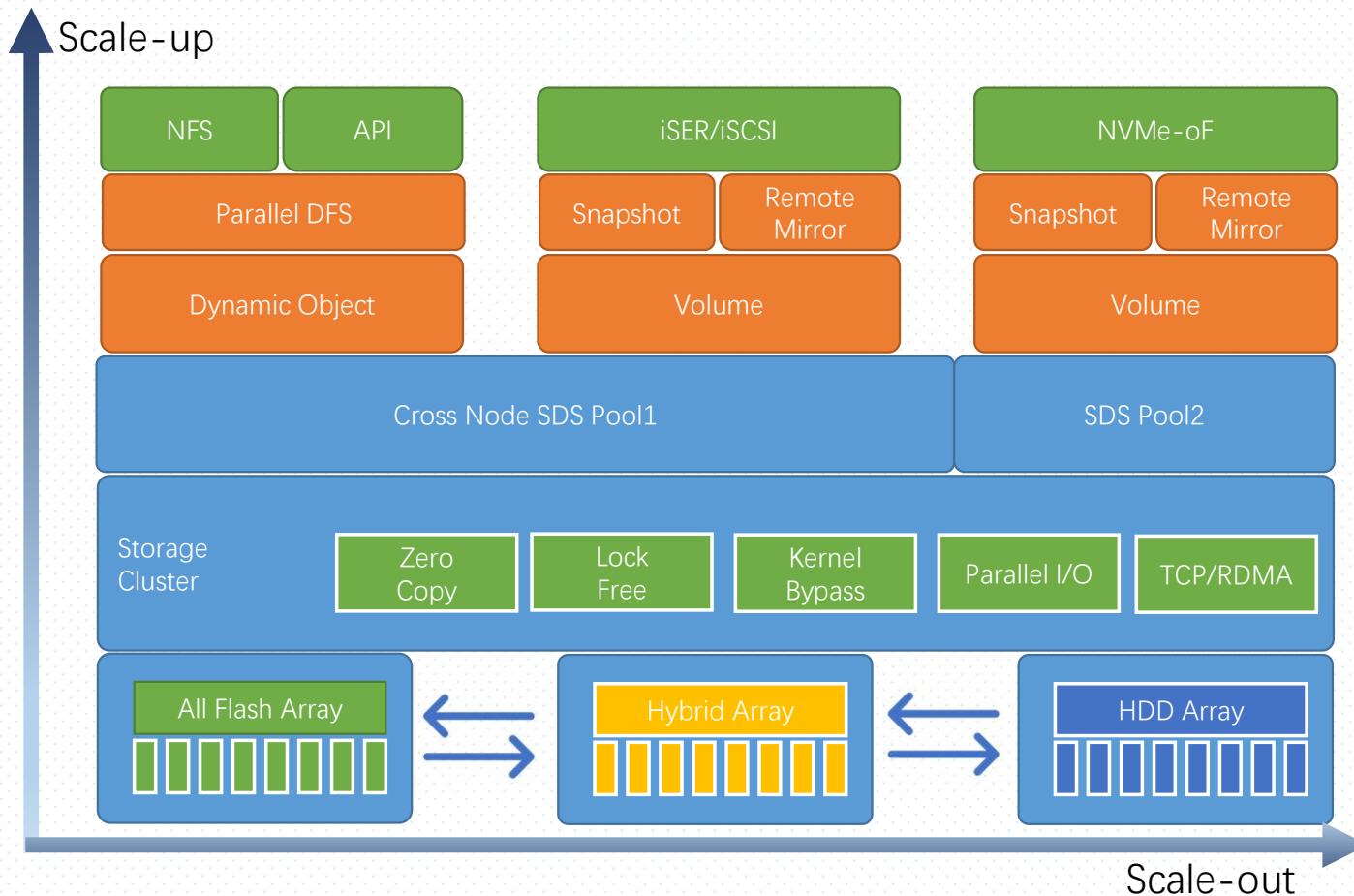- No storage and fast network requires in arbitration node.

The third FlexSDS Arbitration node

Centriliazed Mangement

Control Network

Hypervisor Compute Network

Storage Network

Storage Network

FlexSDS

FlexSDS

Topology: Single node SDS with replication

Single node only, with:
- data redundancy
- Auto-recovery

without High Availiability.

Centriliazed Mangement
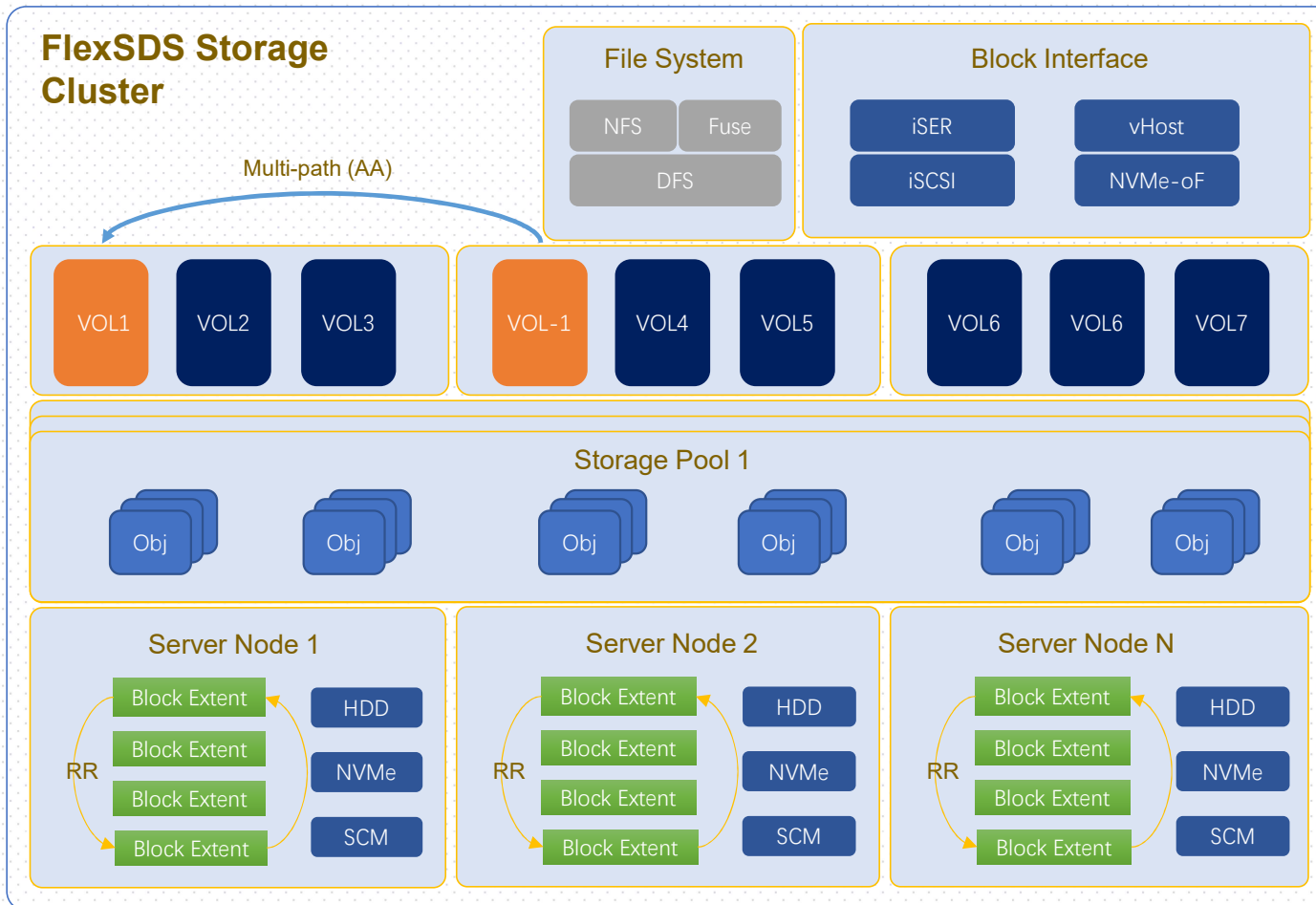
Control Network

Hypervisor
Compute Network

Storage Network

FlexSDS

Replication

The FlexSDS Software Stack

Scale-up

NFS	API

Parallel DFS

Dynamic Object

iSER/iSCSI	
Snapshot	Remote Mirror

Volume

NVMe-oF	
Snapshot	Remote Mirror

Volume

Cross Node SDS Pool1 | SDS Pool2

Storage Cluster

Zero Copy	Lock Free	Kernel Bypass	Parallel I/O	TCP/RDMA

All Flash Array ⟷ Hybrid Array ⟷ HDD Array

Scale-out

Completely desinged, optimized for All-Flash systems, also support for hyrbrid and HDD systems, the stack performance featue:

- File system and block
- Zero-copy I/O stack
- Kernel (OS) Bypass
- Lock Free (not using system lock)
- Fully support RDMA, TCP
- Fully concurrent, parallel I/O
- 100% SDS Pool, Dynamic objects
- RAW, Thin, Logged Volumes
- Zero-data-copy snapshot
- Sync, async remote mirror
- iSCSI, iSER and NVMe-oF
- Single Node SDS and Multiple Nodes Scale-out
- High availability by VIP and Multipath

FlexSDS Cluster Architecture

FlexSDS Storage Cluster

Multi-path (AA)

File System
- NFS
- Fuse
- DFS

Block Interface
- iSER
- vHost
- iSCSI
- NVMe-oF

VOL1 · VOL2 · VOL3 · VOL-1 · VOL4 · VOL5 · VOL6 · VOL6 · VOL7

Storage Pool 1
Obj · Obj · Obj · Obj · Obj · Obj

Server Node 1
- Block Extent
- Block Extent
- Block Extent
- Block Extent
- RR
- HDD
- NVMe
- SCM

Server Node 2
- Block Extent
- Block Extent
- Block Extent
- Block Extent
- RR
- HDD
- NVMe
- SCM

Server Node N
- Block Extent
- Block Extent
- Block Extent
- Block Extent
- RR
- HDD
- NVMe
- SCM

High Availability
- Active-Active/Active-Passive Multipath.
- Floating IP: Simplifies DFS access.

Redundancy:
- Replication & erasure coding (EC) for data durability.
- Failover: Auto rerouting to healthy nodes.
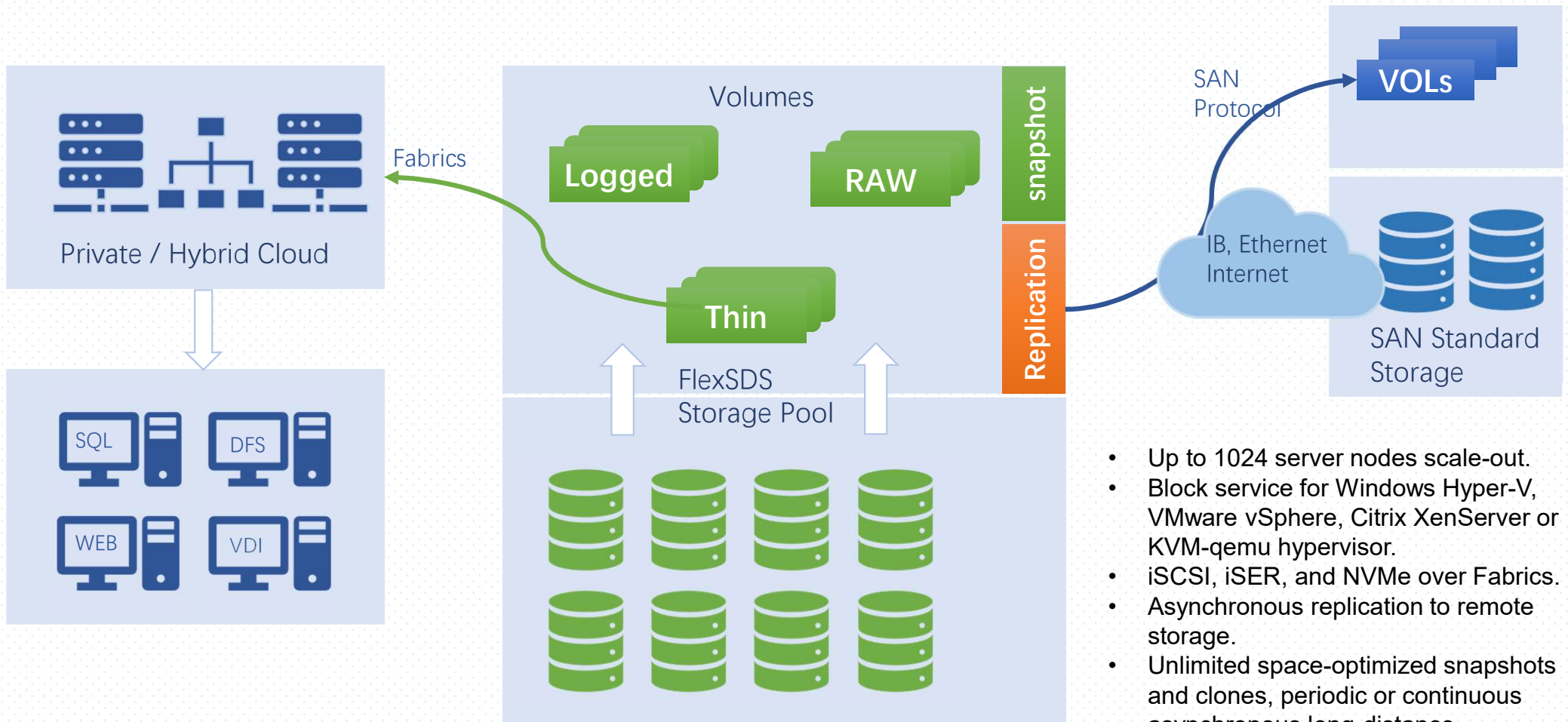- Uptime: High availability with minimal downtime.

Efficient Resource Utilization
- Dynamic Allocation: Thin provisioning to prevent over-provisioning.
- Snapshots: Zero-data-copy for instant backups.
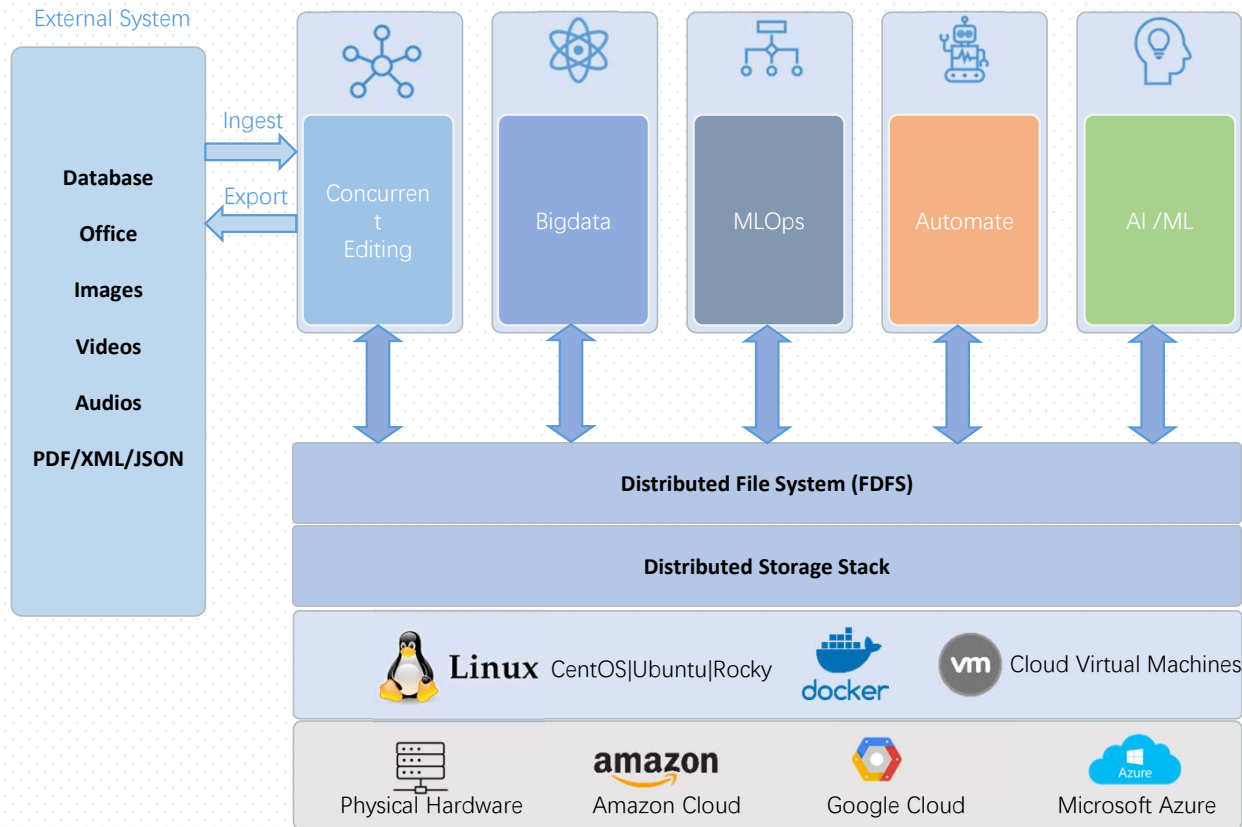
Data Protection & Recovery
- Volume-based asynchronous replication to secondary sites/cloud.
- Self-Healing: Auto-detection and repair of inconsistencies.

FelxSDS Block Storage for Cloud Infrastructure

Private / Hybrid Cloud

SQL DFS

WEB VDI

Fabrics

Volumes

Logged

RAW

Thin

snapshot

Replication

FlexSDS Storage Pool

SAN Protocol

VOLs

IB, Ethernet Internet

SAN Standard Storage

- Up to 1024 server nodes scale-out.
- Block service for Windows Hyper-V, VMware vSphere, Citrix XenServer or KVM-qemu hypervisor.
- iSCSI, iSER, and NVMe over Fabrics.
- Asynchronous replication to remote storage.
- Unlimited space-optimized snapshots and clones, periodic or continuous asynchronous long-distance replication.

FelxSDS Distributed File System (DFS)

External System

Database

Office

Images

Videos

Audios

PDF/XML/JSON

Ingest

Export

Concurrent Editing

Bigdata

MLOps

Automate

AI /ML

Distributed File System (FDFS)

Distributed Storage Stack

Linux CentOS|Ubuntu|Rocky

docker

vm Cloud Virtual Machines

Physical Hardware

Amazon Cloud

Google Cloud

Microsoft Azure

Typical use cases:
- AI/ML Data Processing.
- Big Data Analytics.
- Media and Entertainment Workflows.
- Cloud Storage Integration.
- Scientific Research and Simulations.
- Database Backends.
- Enterprise File Sharing and Collaboration.
- Virtualization and Container Workloads.
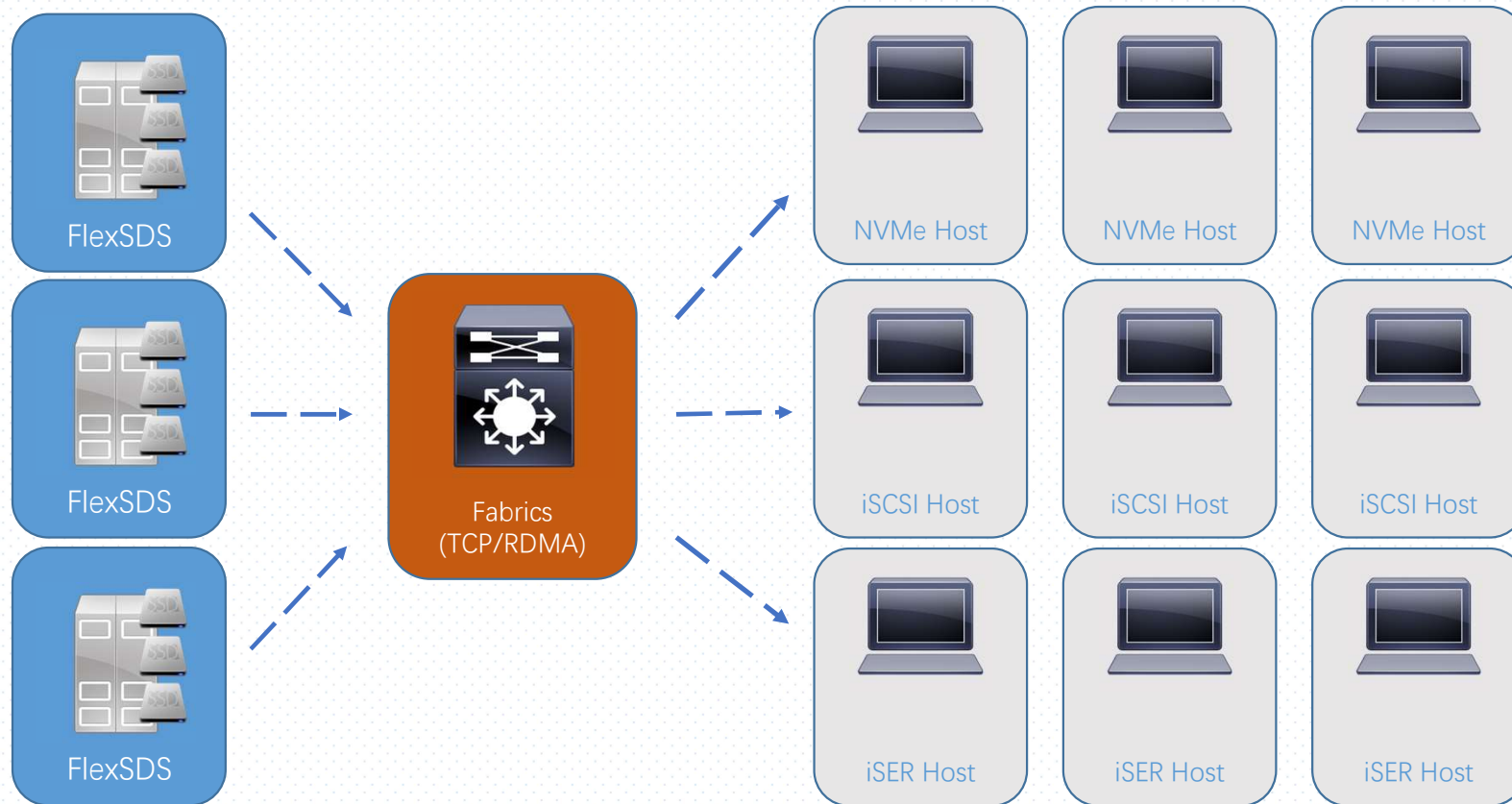- Backup and Archival Storage.

Backend storage:
- Any local attached block devices.
- JBoF, JBOD, ESSD etc.
- Network shared block devices (IP SAN).
- Block devices or files in virtual environment.
- Storage devices in public cloud.

FelxSDS Storage for HCI
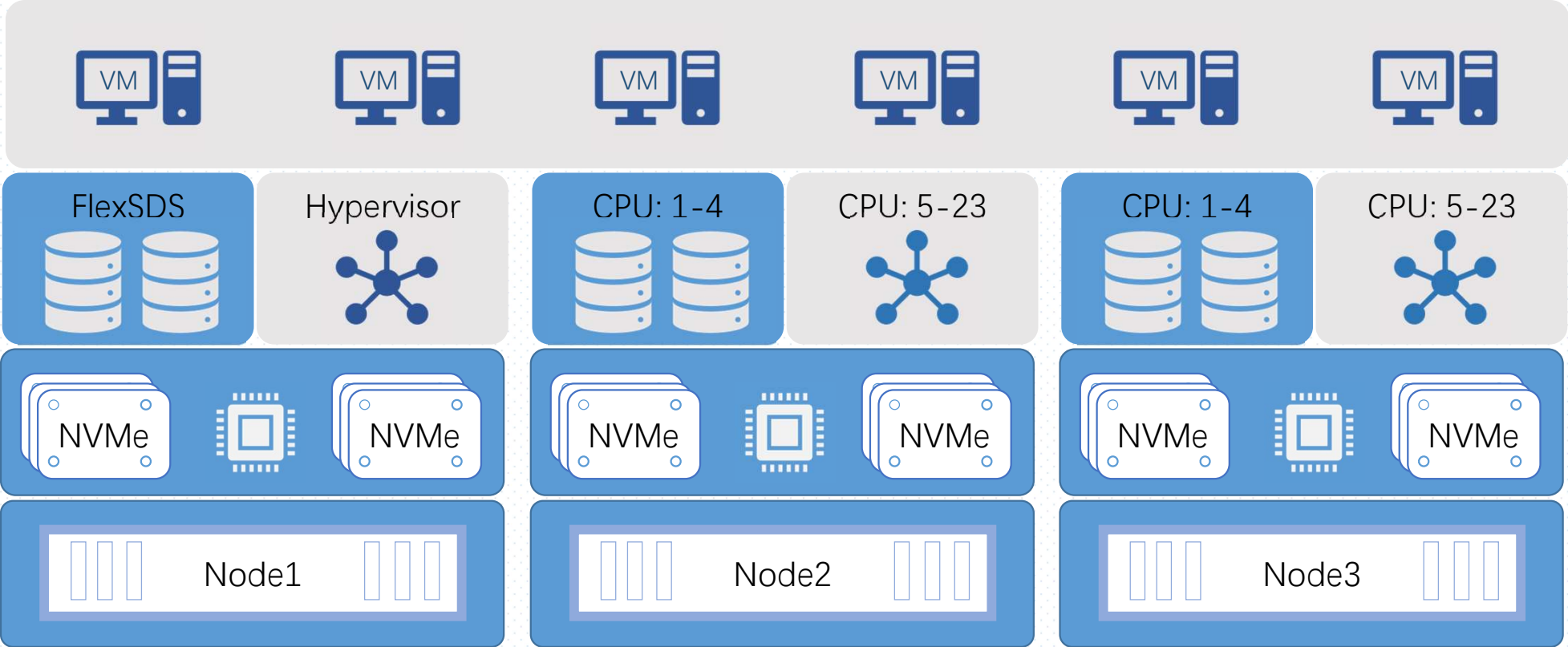
Block service for Windows Hyper-V, VMware vSphere, Citrix XenServer or KVM-qemu hypervisor.
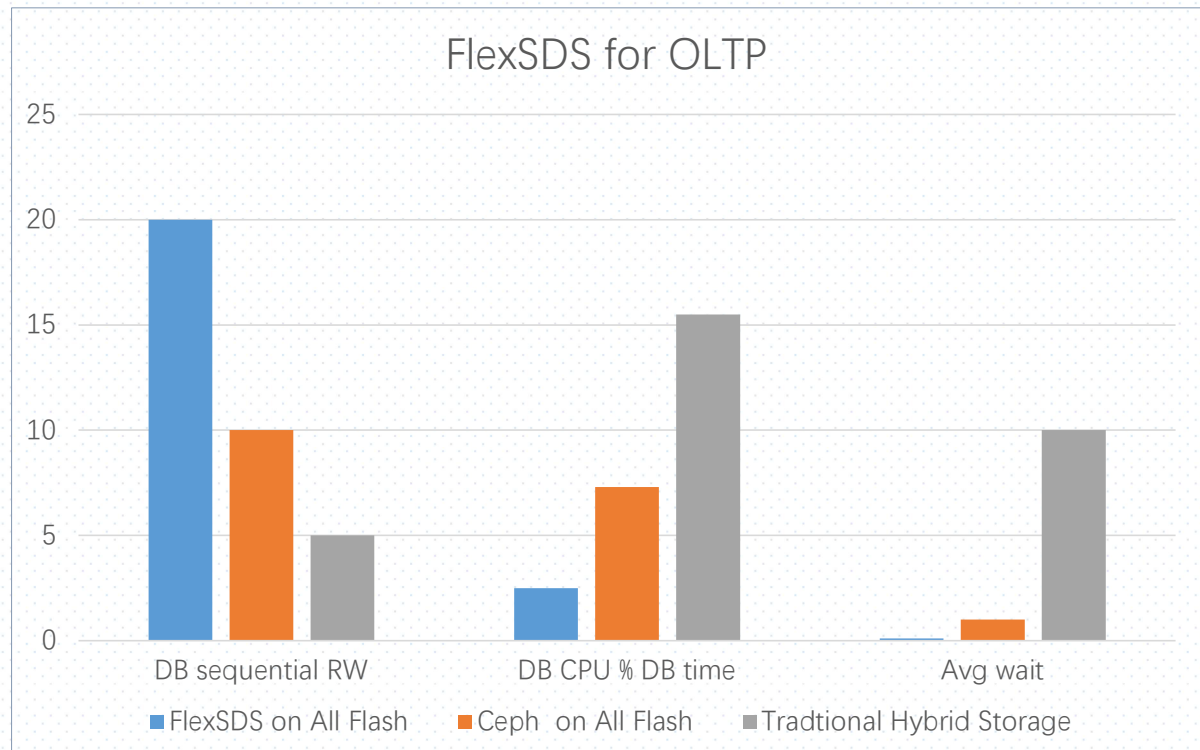
FelxSDS ServerSAN

Exporting snapshot enabled volume over iSCSI, iSER, and NVMe over Fabrics (NVMe-oF).
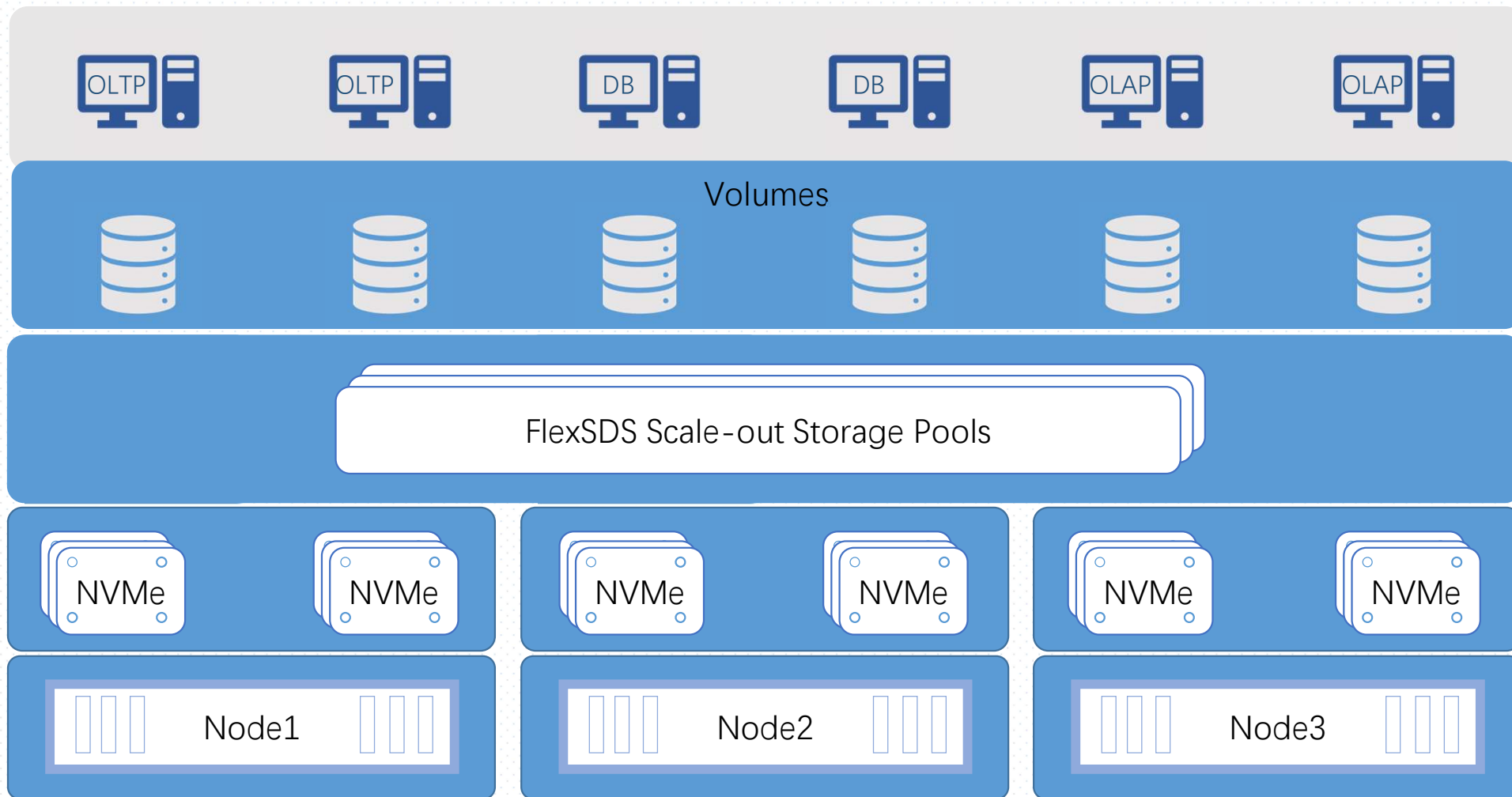
FelxSDS Storage for OLTP

In OLTP services, it is unlikely for a database to cache all data in the buffer. Therefore, the calculations and data access operations are performed in alternating orders, and the CPU have to frequently wait for I/Os to be completed before it can compute data, therefor OLTP much care about latency.

- Run any application at any scale, fast time-to-value, and non-disruptive scale-up or scale-out.
- Lock-free model, zero-mem-copy, high CPU cache optimize to provide extreme low latency.
- Designed for next generation hardware, kernel-bypass, Parallel and high concurrent I/O technology and high utilization of storage and network resources.
- Multiple node and data copies, high availability and auto-recovery.
- Low total cost of ownership (TCO)

FlexSDS for OLTP

(Chart categories: DB sequential RW, DB CPU % DB time, Avg wait)
Legend: FlexSDS on All Flash, Ceph on All Flash, Tradtional Hybrid Storage

FelxSDS Storage for OLTP & DB

OLTP OLTP DB DB OLAP OLAP

Volumes

FlexSDS Scale-out Storage Pools

NVMe NVMe NVMe NVMe NVMe NVMe

Node1 Node2 Node3

FlexSDS Storage for AI & BigData

Accelerating a data analytics project is critical for storage vendors that sell storage systems designed for big data analytics, artificial intelligence (AI), and machine learning (ML). Speed is hugely important when it comes to AI.

- Support NVMe SSD, RDMA (Infiniband, RoCE, and iWarp).
- End-to-End NVMe.
- Fast deployment and scaling of storage resources, and dynamic workload scaling and balancing.
- Designed to provide extreme low latency.
- Parallel and high concurrent I/O.
- High utilization of storage and network resources.
- Multiple node and data copies, high availability and auto-recovery.
- Efficient asynchronous replication.
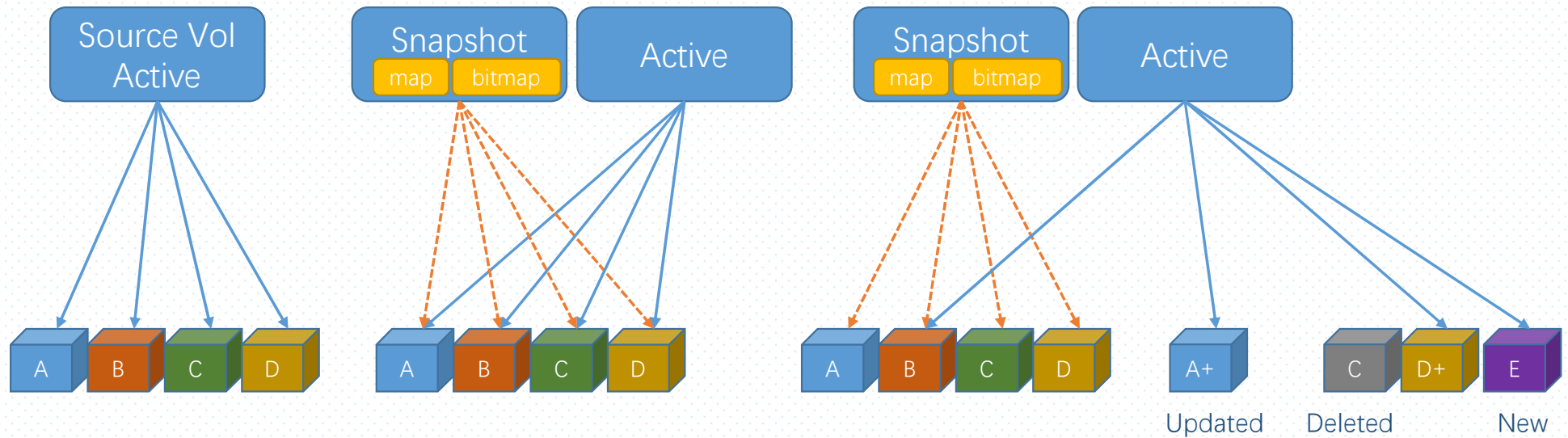- Supports up to 1024 server nodes.
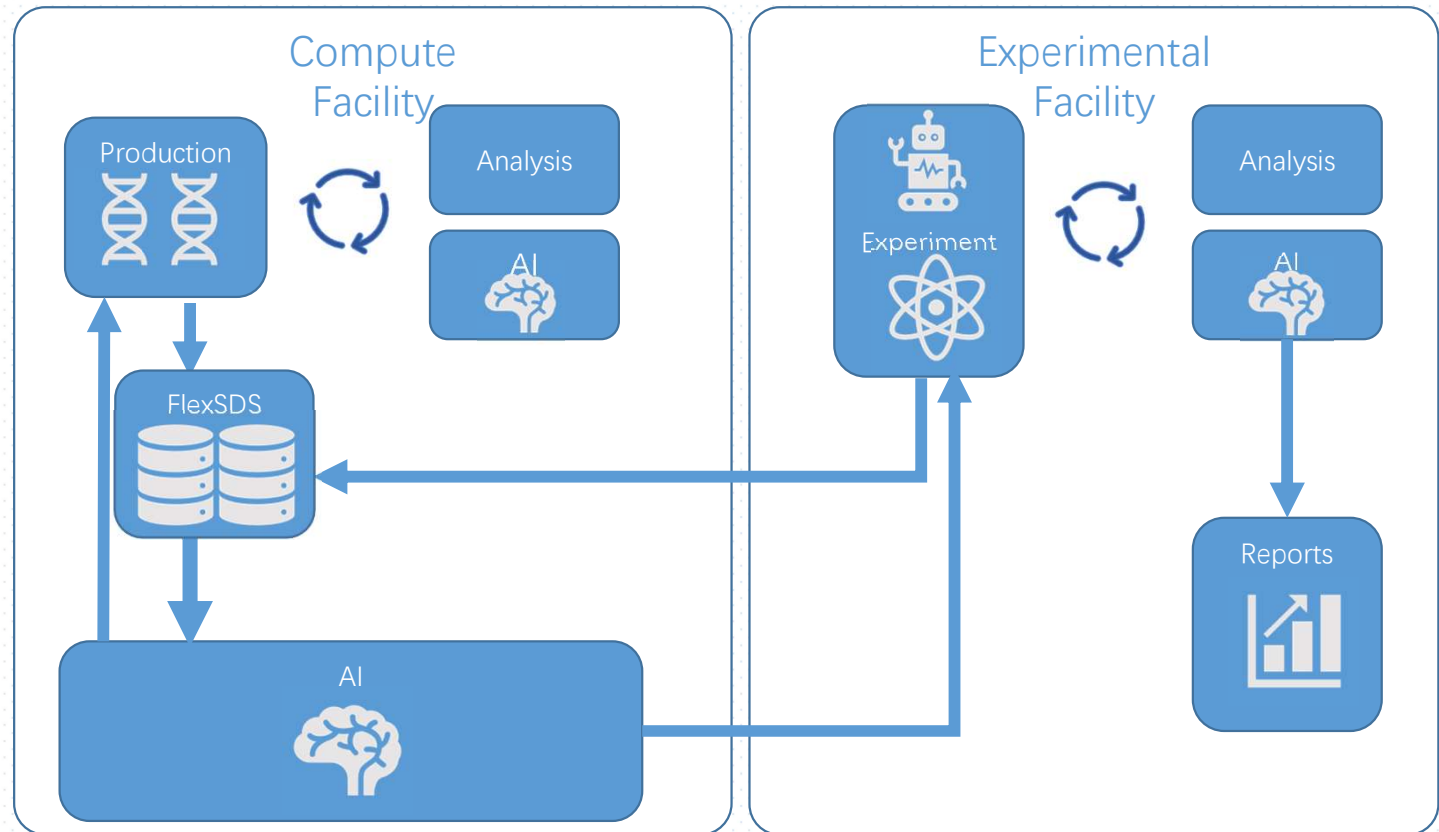
I/O Processing
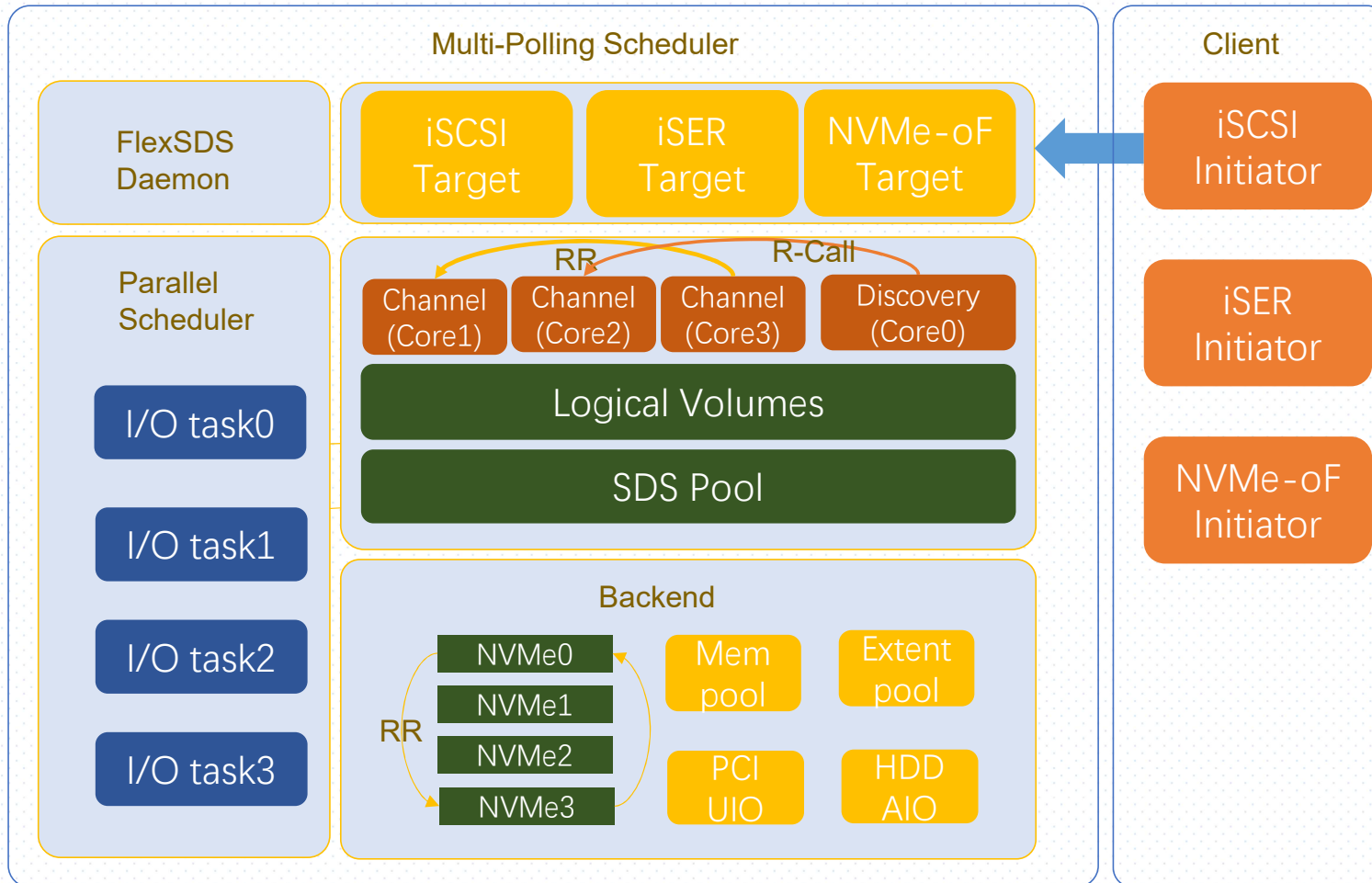
- Main core for listening.
- Multi-cores round-robin.
- Redirect to a core when connection come.
- Parallel, async I/O tasks.
- Disks are served as RR mode.
- Full stack user mode, zero-copy, lock-free, os-bypass, zero-context-switch.
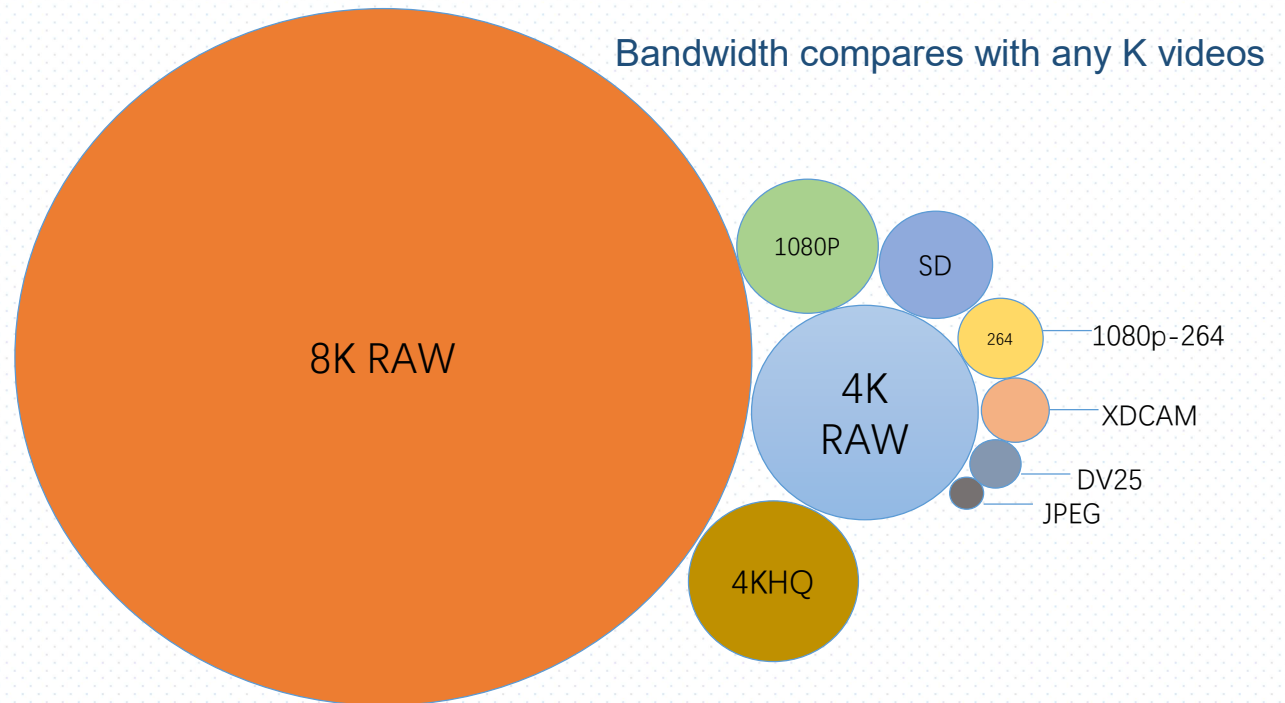
FlexSDS Snapshot

- Unlimited snapshots, read/write performance no related to the snapshot count.
- Zero data copy when creating snapshot, there are only small data of mapping table need to be copies, no user data need to be copied.
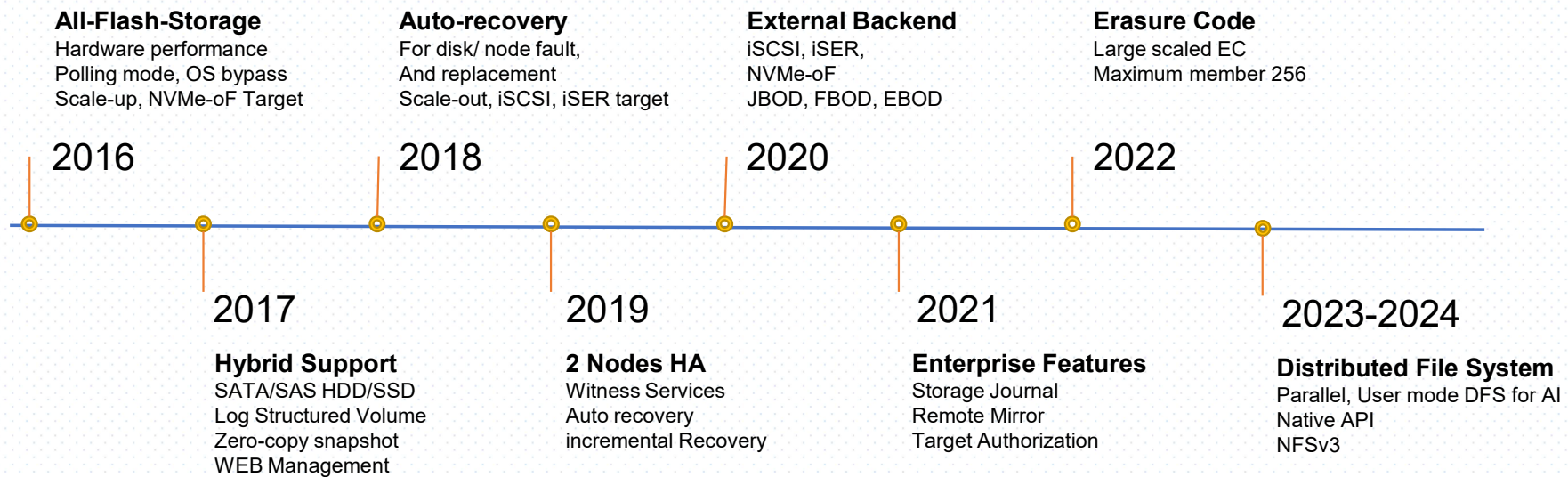- Metadata usage rate is less than 1/200

FlexSDS Storage for 8K Video

Provide editors workstations with access to high bandwidth and low latency performance to accommodate streaming editing and reel presentation to any workstation at any time.

- Delivers full performance of NVMes over the network.
- Linearly scale performance and capacity
- High IOPS and high bandwidth.
- Extreme low latency.
- Supports current and future media workflows
- Low total cost of ownership (TCO)
- Provider block interface.

Bandwidth compares with any K videos

8K RAW

1080P

SD

4K RAW

264 — 1080p-264

XDCAM

DV25

JPEG

4KHQ

FlexSDS Storage Software Roadmap

All-Flash-Storage
Hardware performance
Polling mode, OS bypass
Scale-up, NVMe-oF Target

Auto-recovery
For disk/ node fault,
And replacement
Scale-out, iSCSI, iSER target

External Backend
iSCSI, iSER,
NVMe-oF
JBOD, FBOD, EBOD

Erasure Code
Large scaled EC
Maximum member 256

2016

2018

2020

2022

2017

2019

2021

2023-2024

Hybrid Support
SATA/SAS HDD/SSD
Log Structured Volume
Zero-copy snapshot
WEB Management

2 Nodes HA
Witness Services
Auto recovery
incremental Recovery

Enterprise Features
Storage Journal
Remote Mirror
Target Authorization

Distributed File System
Parallel, User mode DFS for AI
Native API
NFSv3

Q&A

THANKS

FlexSDS Inc.

https://www.flexsds.com/