Speeding-up Large-Scale Storage with Non-Volatile Memory

CERN openlab Open Day
10 June 2015

KL Yong
Sergio Ruocco
Data Center Technologies Division

about DSI

vision

Founded in 1992, DSI' vision is to be a vital node in a global community of knowledge generation and innovation, nurturing research talents and capabilities for world class R&D in next generation technologies.

mission

To establish Singapore as an R&D center of excellence in data storage technologies.

DSI
A*STAR

Core Competencies

HARD DISK DRIVE TECHNOLOGIES

- 10Tb/in^2 areal density technologies
- Thin Hybrid HDD (0.5TB 2.5”, 5mm, hybrid HDD)

NON-VOLATILE MEMORIES

- STT-MRAM
- ReRAM
- Signal Processing & Error Correction
- IC Design

DATA CENTER TECHNOLOGIES

- NVM System
- Active Hybrid Storage System
- Big Data Analytics Platform
- Data & Storage Security

ADVANCED CONCEPT & NANOFABRICATION TECHNOLOGIES

- Nanofabrication
- Spintronics
- Plasmonics
- Photo-Electronics
- Metamaterials and Small Particle Physics Research

Massive Data Key Challenge for Data Center

Performance

Scalability

Security

Energy Consumption

Space

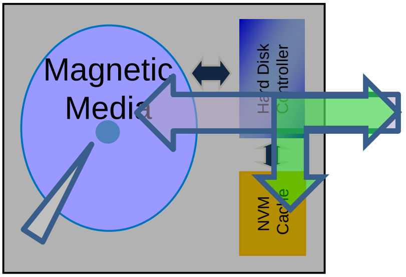
Manageability

- CAPEX cost for additional IT equipment - servers, networks and storage
- Driving the energy costs
- Larger footprint and space required
- Increasingly challenging and costly to scale and deliver performance
- Increasing complexity in operating and managing the data center
- Providing data protection and security for massive amount of data

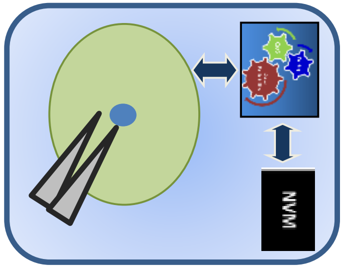
Future Data Center Architecture with Emerging Technologies

Integration of

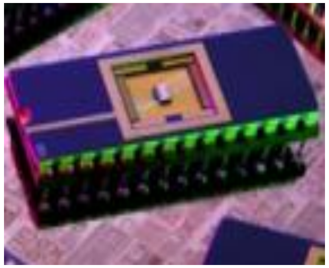
Hybrid Drive



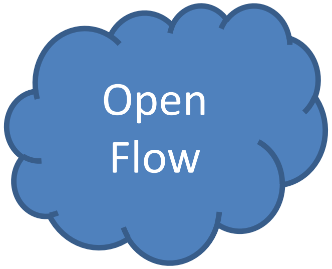
Active Drive



NVM



Software Managed



Homomorphic Security

Performance, scalable, secured, energy and cost efficient

Next Generation Non Volatile Memory (NVM)

Characteristics of next generation NVM:

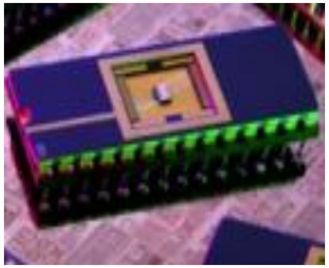
+ **high speed** ~ DRAM like
+ **data persistent** against power loss
+ **byte-addressable** (vs 4KB- 512KB blocks)
+ **endurance** (~DRAM like) >>> Flash
+ **no refresh** cycles/energy

Technology	Read	Write	Endurance Cycle	Read (V)	Write (V)	Maturity
HDD (15KRPM)	6000µs	6000µs	NA	5V, 12V	5V,12V	Product
SLC Flash	25µs	200µs/1.5ms (Program/Erase)	10^5 (1000x for MLC)	2	15	Product
DRAM	<10ns	<10ns	10^{16}	1.8	2.5	Product
STT-MRAM	**2-20ns**	**2-20ns**	**10^{15}**	**0.7**	**+ 1**	**Advanced Development**

DSI
A*STAR

NVM Research in DSI: Device to System

NVM Device

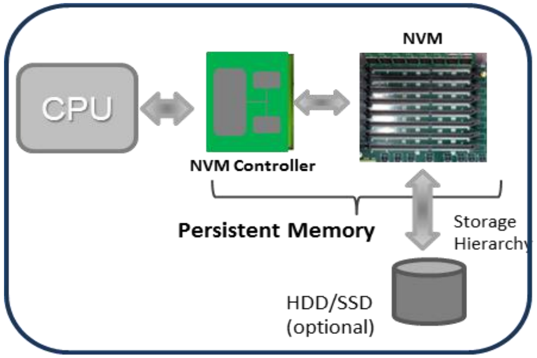


- Materials & Structure
- Process
- Signal Processing
- IC Integration

NVM Controller

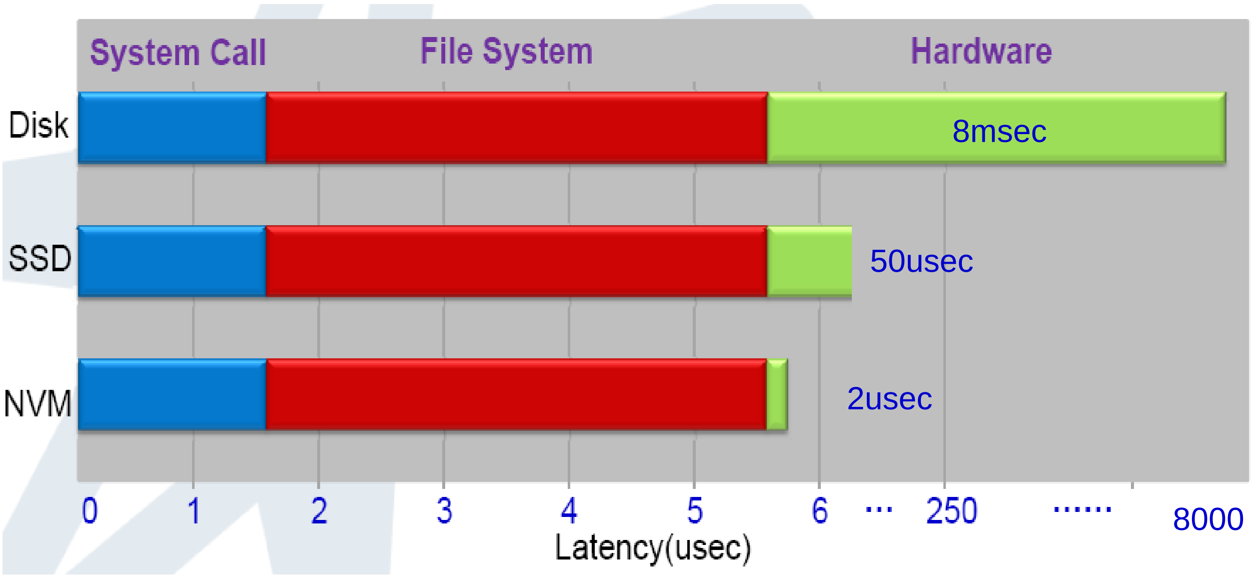
- I/O Flow, Scheduling, Buffering
- Wear Leveling
- Erasure Coding
- FPGA and firmware

NVM-based Systems



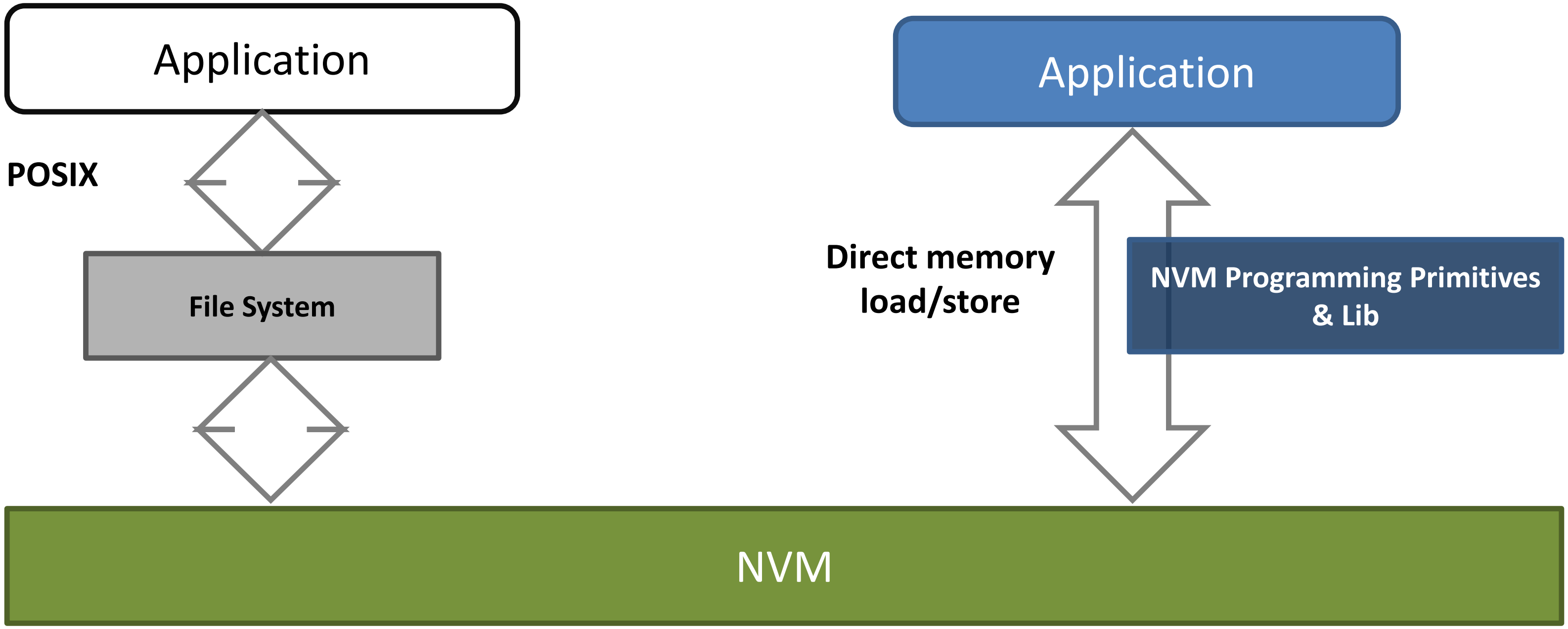
- NVM File System
- System, Memory & Storage Stacks
- Programming Model, Language & Toolchains
- NVM Systems Cluster

Next Generation Non-Volatile Memory



To fully exploit its performance, **the hardware architecture and OS stacks including programming model –** applications, languages, compilers/VMs, run-time libraries, middleware,... – **must change**

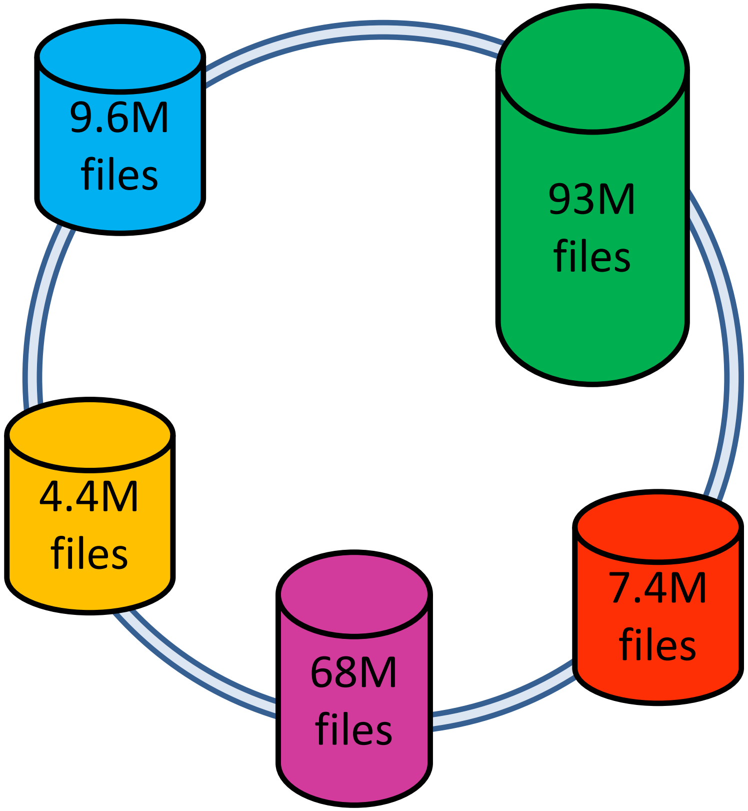
NVM Software Programming Model



New programming model for NVM provides data persistence integrated into the application programs:

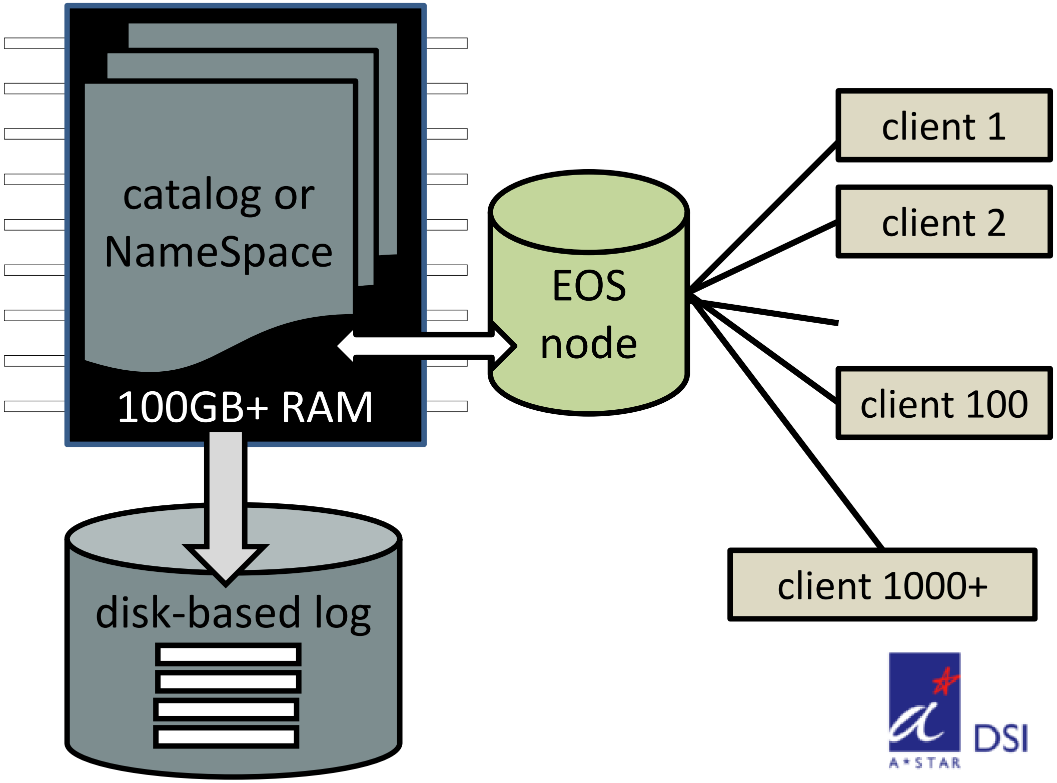
- Byte-addressable
- Load/Storage access without demand paging
- Memory performance

CERN EOS NameSpace



Metadata operations (create, rename, move, delete etc.) are sped-up by in-memory NameSpace, with a growing RAM footprint of 100+ GBs

Disk-based logs enable consistent reconstruction of NameSpace to recover after any hw & sw faults



50+ PB experimental data in 150M+ files across 5 experiments (nodes): ATLAS, CMS, LHCB, ALICE…

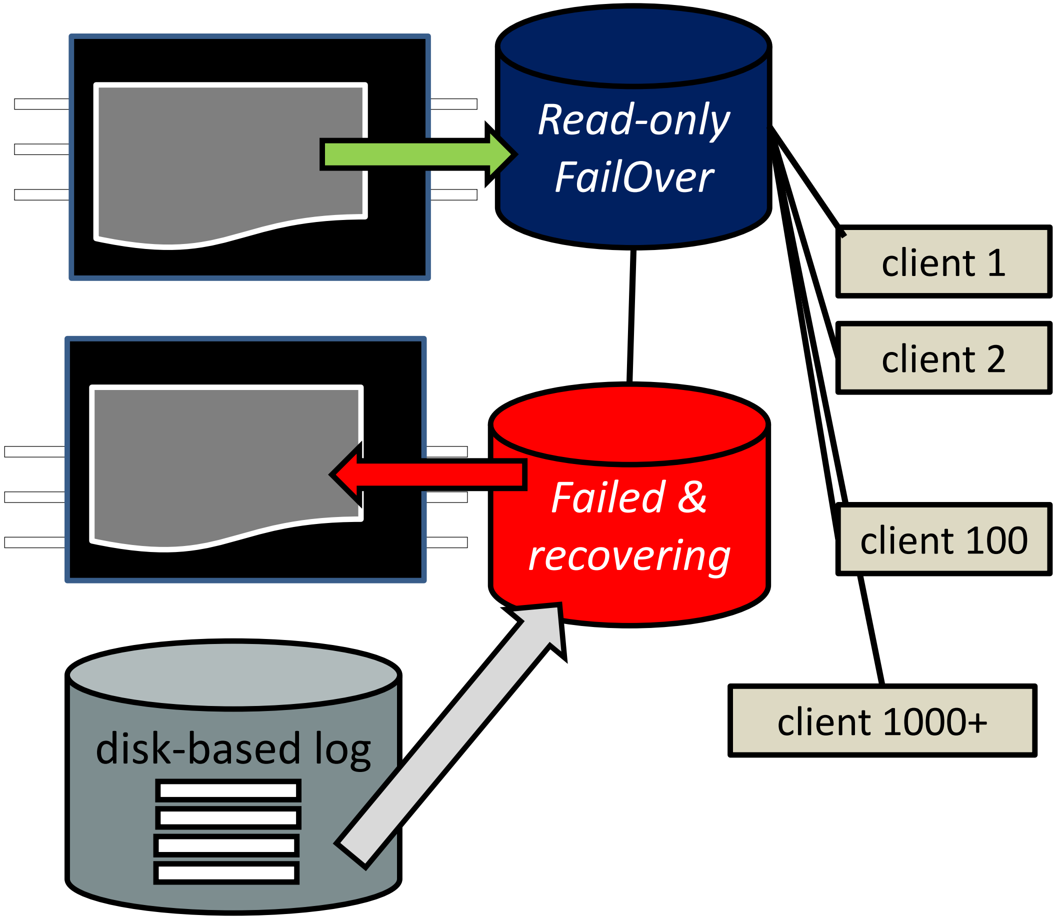
Node availability critical for the continued operation of thousands of clients

Challenges: Availability and Consistency

One of the challenges is the consistent journaling of metadata updates between memory and disk logs; but also across failures of the NS service, the hardware or power.

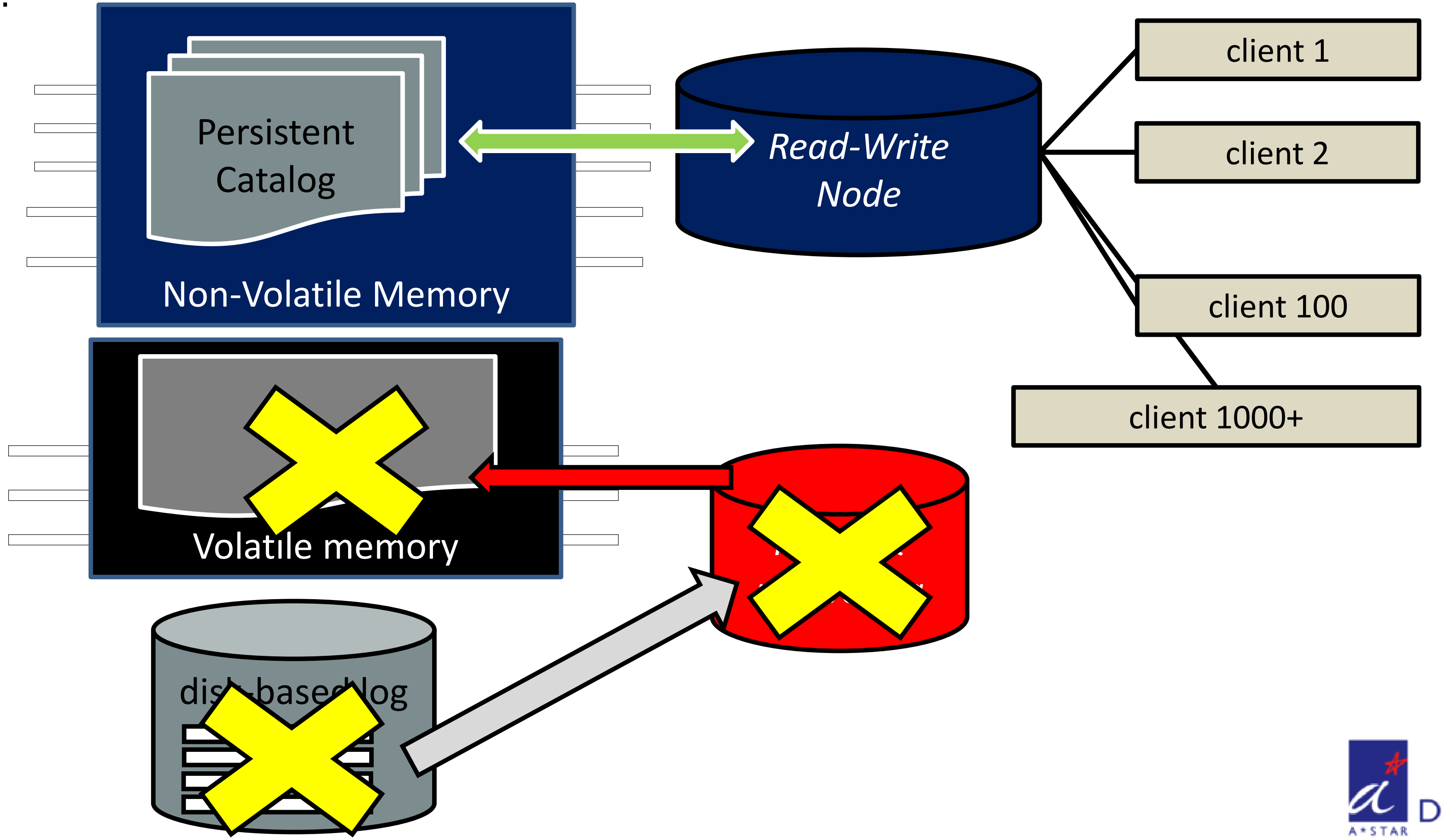
Reconstructing a 100GB+ Catalog can take even 10 minutes, disrupting client's work.

Reconstruction is not IO-bound but CPU-bound because data structures trade-off *lookup* speed against *insert* speed.



Proposed Solution: EOS Catalog in Non-Volatile Memory

Store the instance of the EOS Catalog in Non-Volatile Memory. NVM-based Catalog is persistent, fault-tolerant, and always consistent. No more slow reconstructions from logs
.



of Enabling
Storage
Technologies