



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's 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.

Core Competencies

HARD DISK DRIVE TECHNOLOGIES



NON-VOLATILE MEMORIES



DATA CENTER TECHNOLOGIES



ADVANCED CONCEPT & NANOFABRICATION TECHNOLOGIES



- 10Tb/in² areal density technologies
- Thin Hybrid HDD (0.5TB 2.5", 5mm, hybrid HDD)
- STT-MRAM
- ReRAM
- Signal Processing & Error Correction
- IC Design
- NVM System
- Active Hybrid Storage System
- Big Data Analytics Platform
- Data & Storage Security
- Nanofabrication
- Spintronics
- Plasmonics
- Photo-Electronics
- Metamaterials and Small Particle Physics Research

Massive Data Key Challenge for Data Center

Performance

Scalability

Security

Energy
Consumption

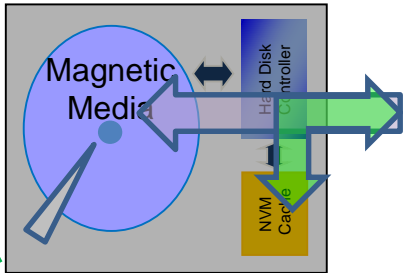
Space

Manageability

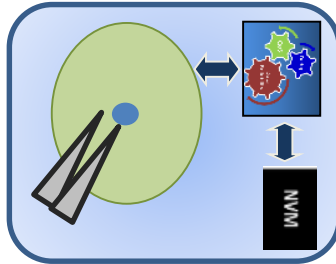
- CAPEX cost for additional IT equipment - servers, networks and storage
- Driving the energy costs
- Larger footprint and space required
- Increasingly challenging and costly to scale and deliver performance
- Increasing complexity in operating and managing the data center
- Providing data protection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

Integration of

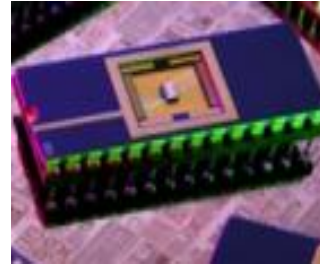
Hybrid Drive



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)	6000 μ s	6000 μ s	NA	5V, 12V	5V,12V	Product
SLC Flash	25 μ s	200 μ s/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

NVM Research in DSI: Device to System

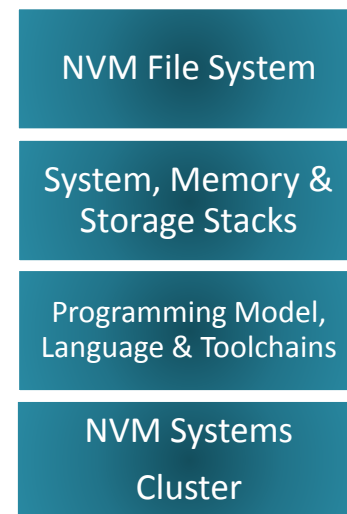
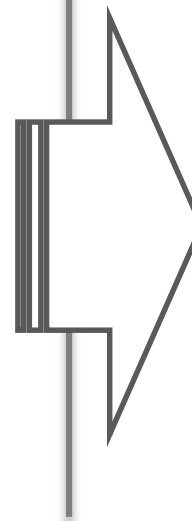
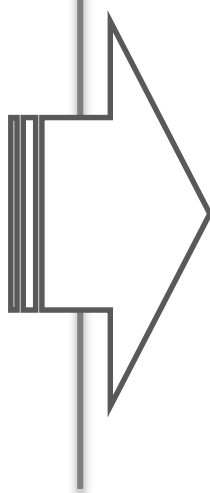
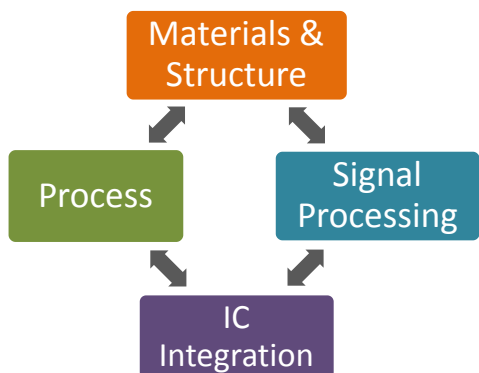
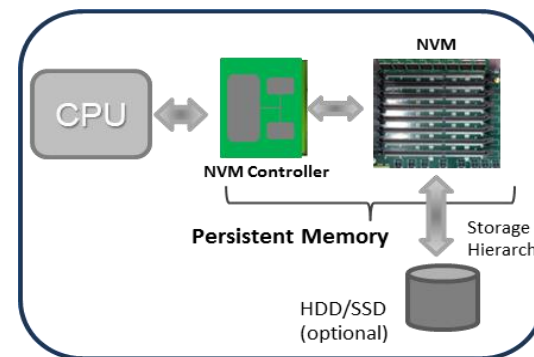
NVM Device



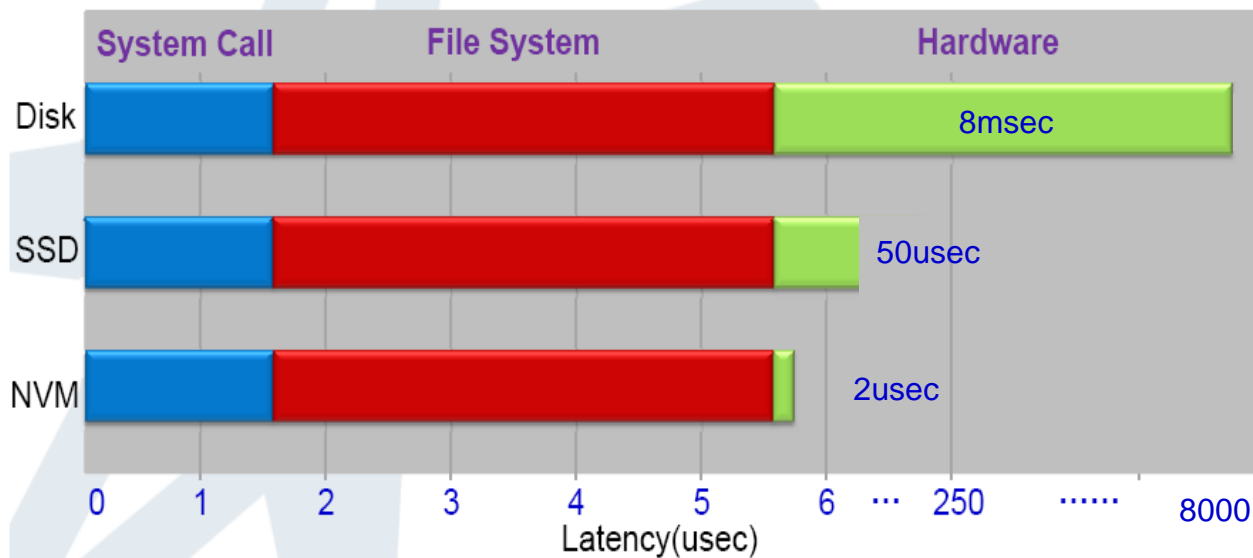
NVM Controller



NVM-based Systems

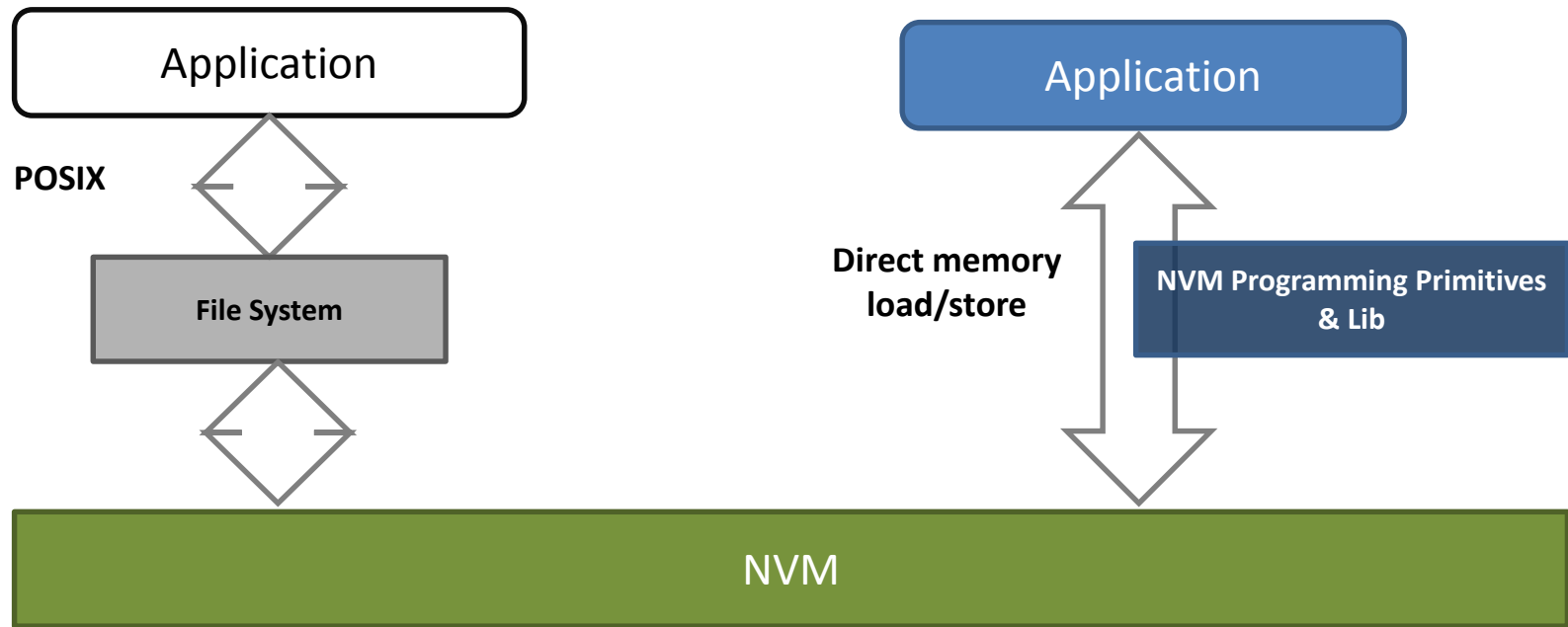


Next Generation Non-Volatile Memory



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 Catalog in NVM

OUTLINE

- 1) CERN EOS distributed, scalable, fault-tolerant filesystem serves data to 100s-1000s of clients.
- 2) To speed-up the service, it keeps and updates a large and *fast* data Catalog in memory (volatile RAM) that can reach 100+ GBs.
- 3) EOS server nodes are prone to software and hardware faults of any origin: bugs in libraries, operating system or hardware
- 4) When a server node crashes, even after a quick reboot, before resuming operations it must rebuild a new coherent Catalog in memory from small pieces stored in disk logs. Rebuilding the Catalog currently takes up to 10 minutes.
- 5) We propose to design a persistent, version of the EOS Catalog to be stored in the new Non-Volatile Memory (NV-DIMMs and future technologies)
- 6) *In principle* the persistent Catalog shall remain always-consistent and fault-tolerant

Therefore after faults the EOS server will be able to quickly resume serving the clients, as it can skip the slow reconstructions from disk-based logs
- 7) Challenges: all the consistency management previously left at 50+ years old disks and filesystem technologies shifts in the hands of the application and the programming model

CERN EOS Catalog in NVM

CERN EOS Namespace

“EOS” CERN storage system: 50+ PB experimental data in 150M files across 5 “experiments / nodes” (ATLAS, CMS, ALICE...)

The availability of each node is critical for the continued operation of thousands of local and remote clients that analyze continuously data for weeks or months.

For performance reasons each node maintains in **main memory** a “NameSpace” (NS) with the **updated filesystem metadata that currently has a RAM footprint of 100+ GB and growing.**

Disk-based logs support consistent reconstruction of NameSpace and recovery from faults

Challenges: Consistency and Availability

1) consistent journaling of metadata updates between memory and disk logs; but also across failures of the NS service, the hardware or power

2) After faults NS metadata reconstruction for 100 GB takes 10 minutes, disrupting client's work; CERN IT want to minimize it

(reconstruction is not IO-bound but CPU-bound because data structures trade-off *lookup* speed against *insert* speed)

CERN EOS Catalog in NVM

Goal

Store a persistent, always-consistent and fault-tolerant instance of NS metadata in NVM-based Persistent Virtual Memory

PVM-based data structure are persistent, fault-tolerant, and always consistent. **No more slow reconstructions from logs**



A * STAR

Data Storage
Institute

of Enabling
Storage
Technologies