Computational storage in a virtualized environment

Jinpyo Kim, Senior Staff Engineer, VMware
Michael Mesnier, Principal Engineer, Intel Labs

In collaboration with MinIO

Outline

- A computational storage research platform – Intel
- Implications/opportunities for virtualized environments – VMware

Computational storage research in Intel Labs

Application server

Server

- Application layer with support libraries (e.g., oneAPI)
- Scheduling layer for multiple compute engines
- Aggregation layer (e.g., erasure coding)
- Device layer with support for legacy block storage

Block storage

- Secure, multi-tenant programming layer (AI/ML, search, CRC, …)

CPUs & accelerators

Smart NICs (IPUs)

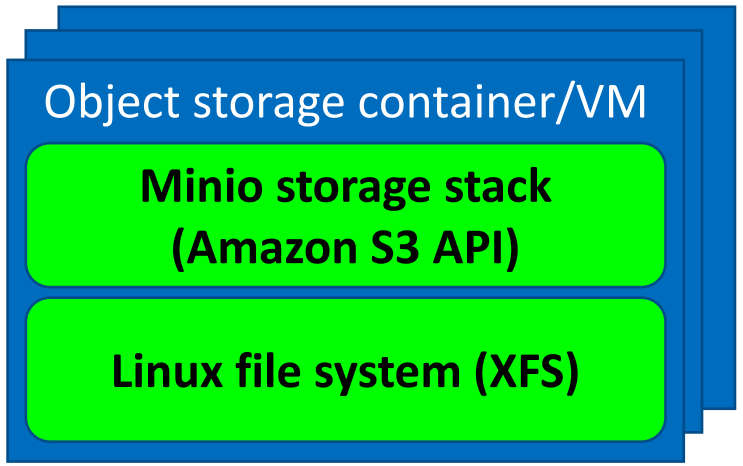
GPUs/FPGAs

NAND and Optane SSDs (may also be computational)

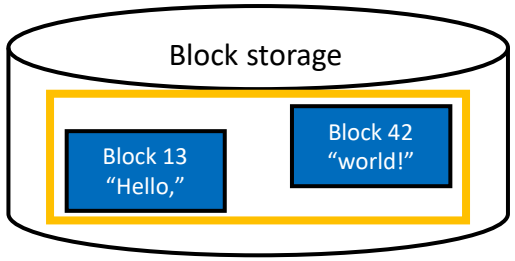
Software research vectors

Enabling silicon

Data scrubbing – a practical, microservices use case



READ/WRITE



MINIO ceph openstack cassandra ...

Multiple objects storage servers

Object servers regularly scrub all data –

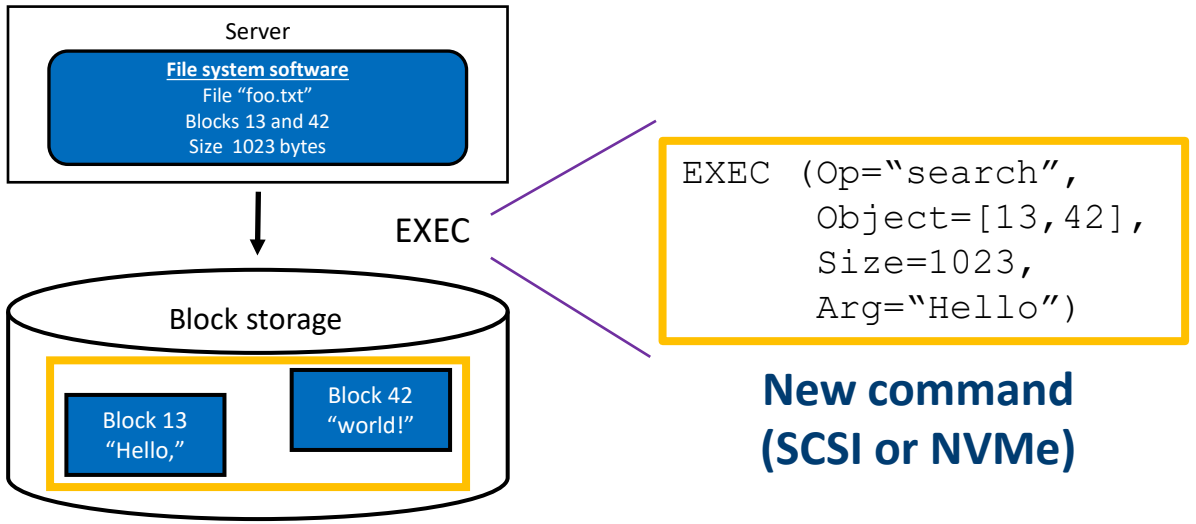
1. Read all objects from the FS
2. Calculate hashes (CRC-32C, MD5, Highway, …)
3. Compare with previously stored hashes

Significant READ traffic generated

Block storage (DAS or SAN)

Intel Labs research platform

1. Teach block storage about *data objects*
   - ❖ Files, directories, tables, records, …
2. Specify operations on objects
   - ❖ Search in text object A
   - ❖ Classify image object B
3. Execute on diverse HW
   - ❖ CPUs, GPUs, FPGAs and ASICs



Computational storage using virtual objects (Adams, Keys, Mesnier), HotStorage '19.

Under the hood

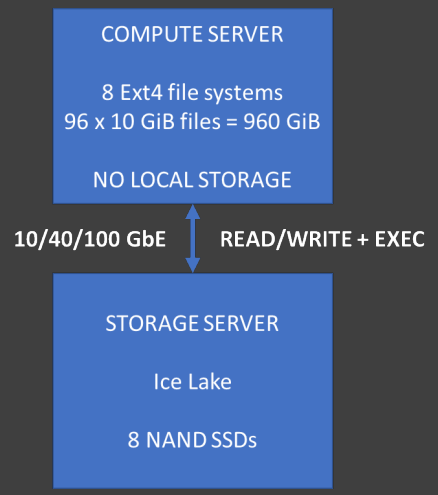
- ✓ **On target: modified stack to implement EXEC**
- ✓ **On host: FIEMAP (IOCTL) to get block mappings from FS**
- ✓ **On host: NVMe or SCSI pass-through to send EXEC via command line**

Example command line usage

```
/usr/bin/cs-hash "crc-32c" /mnt/foo /dev/nvme0n1
```

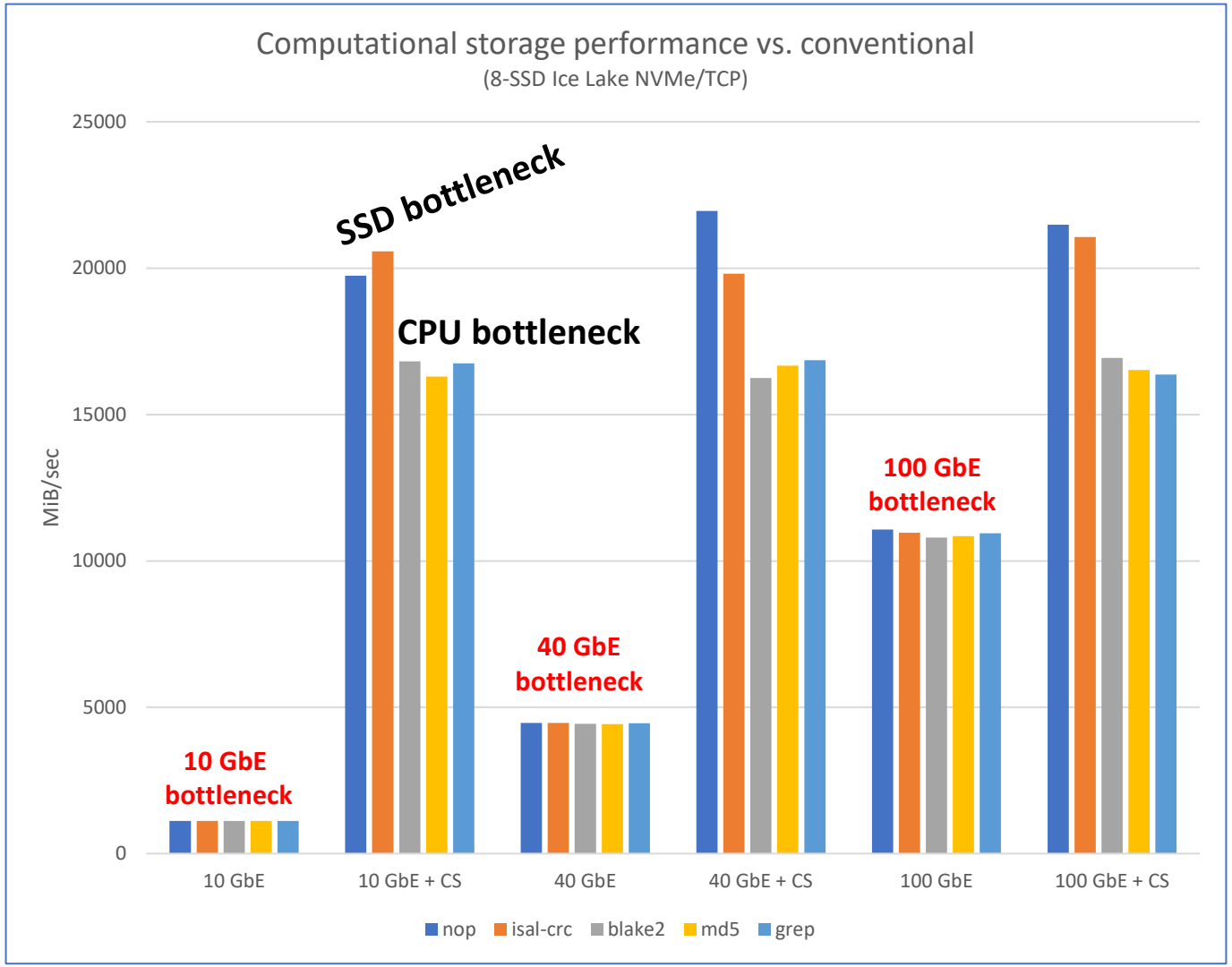

50% to 18x more scalable

(depending on link speed)



Storage primitives tested

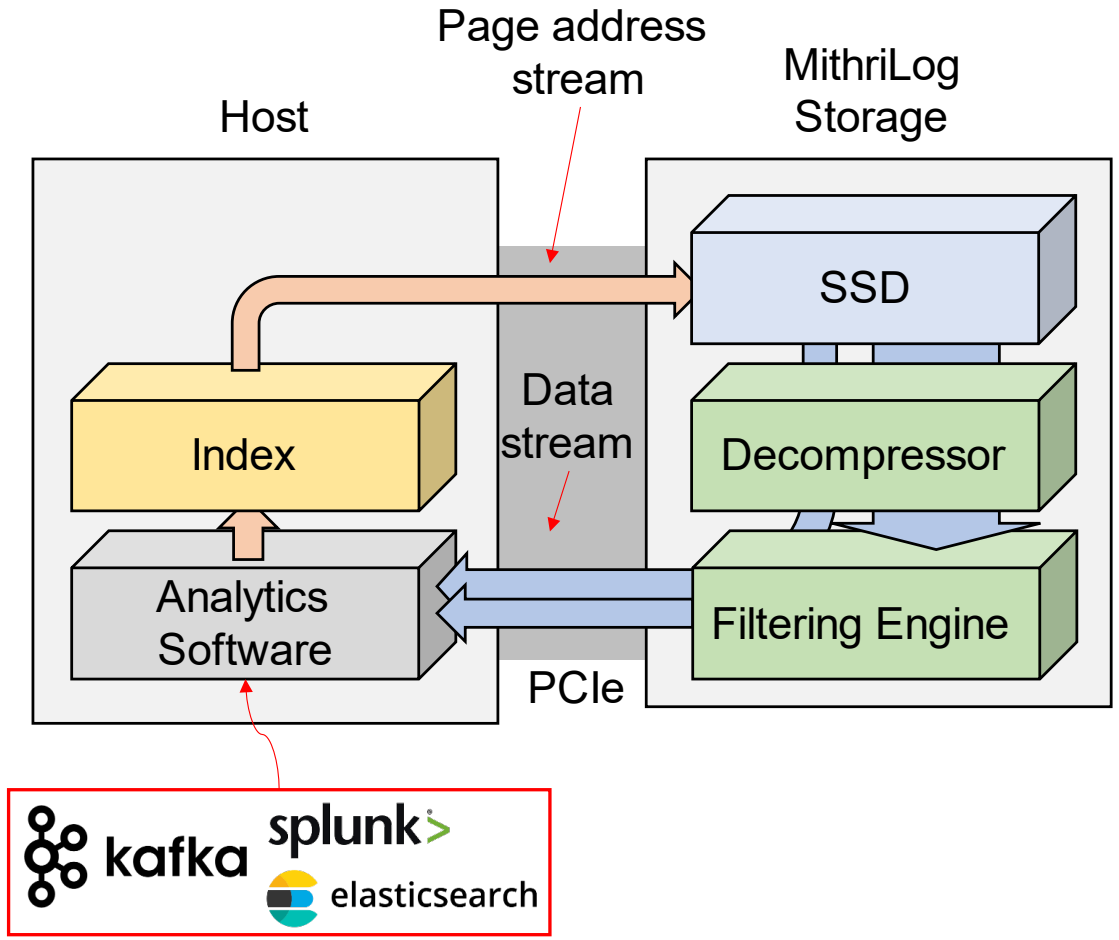
- Reading – NOP
- Scrubbing – CRC-32C
- Deduplication – MD5
- Search – GREP
- Intrusion detection – BLAKE2



Computational Storage at VMware

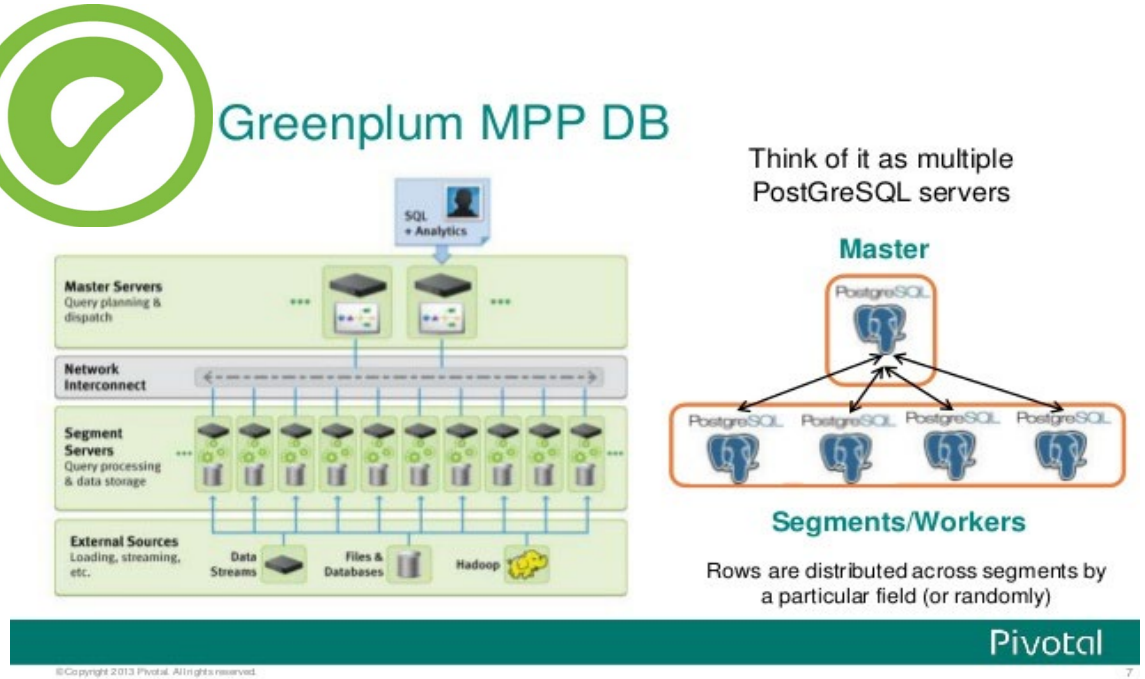
- Implications/opportunities for virtual environments

Computational Storage Device (CSD) at VMware (1)



- Near-Storage Log Analytics (Research Prototype with UC Irvine)
- One pipeline ~= 3.2 GB/sec
- 4x pipelines ~= 12.8 GB/sec
- Order of magnitude better query performance compared to software only (Splunk)
- Lower power consumption
- Implemented on MIT BlueDBM
- Currently porting to Samsung SmartSSD

Computational Storage Device (CSD) at VMware (2)



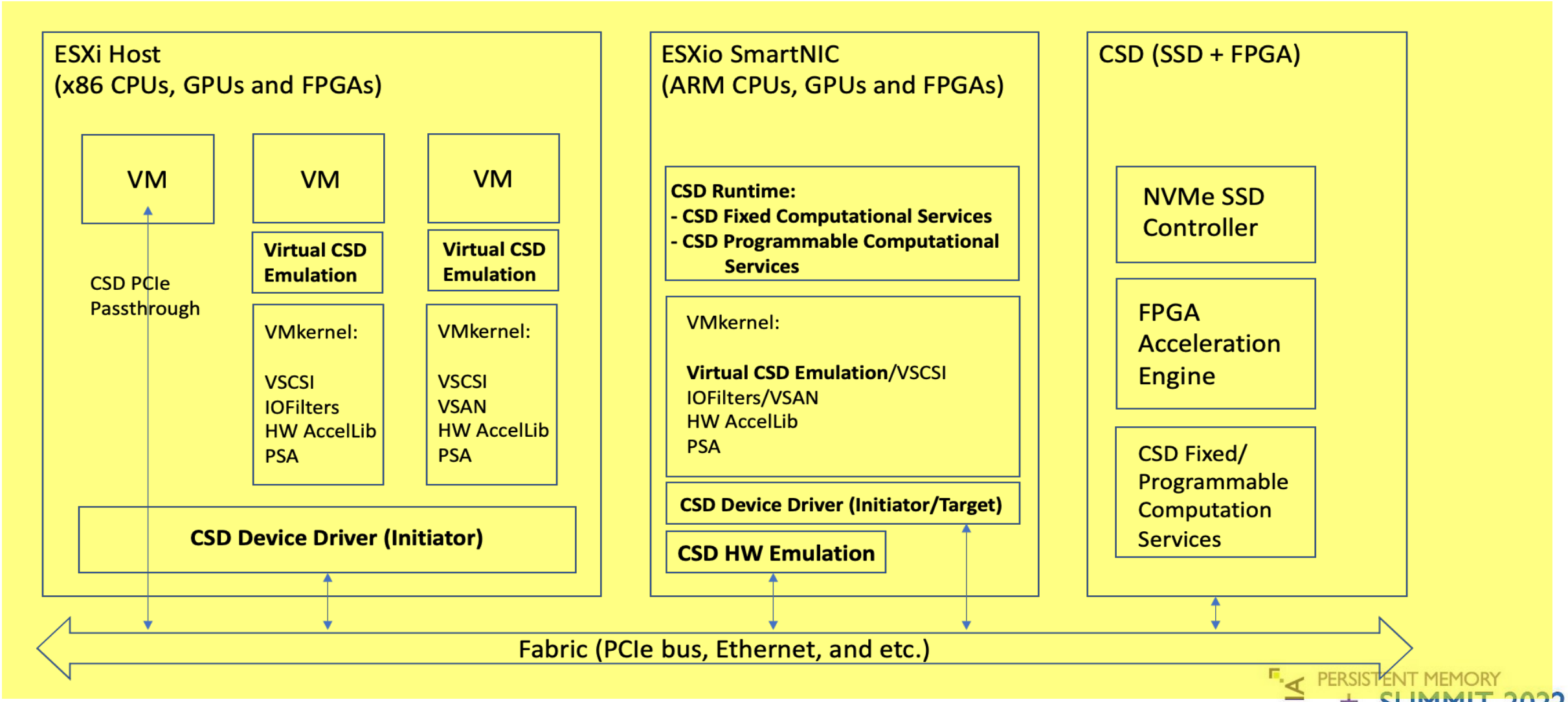
32 TB capacity and 1 GB/sec write speed per device

- Greenplum MPP DB with computational storage devices (Tech preview prototyping with NGD devices)
- Run real-time analytics workload at scale
- Embedded ML queries using Apache MADlib
- Opportunities to virtualize physical NVMe computational storge devices (CSD) into virtual NVMe computational devices (vCSD)

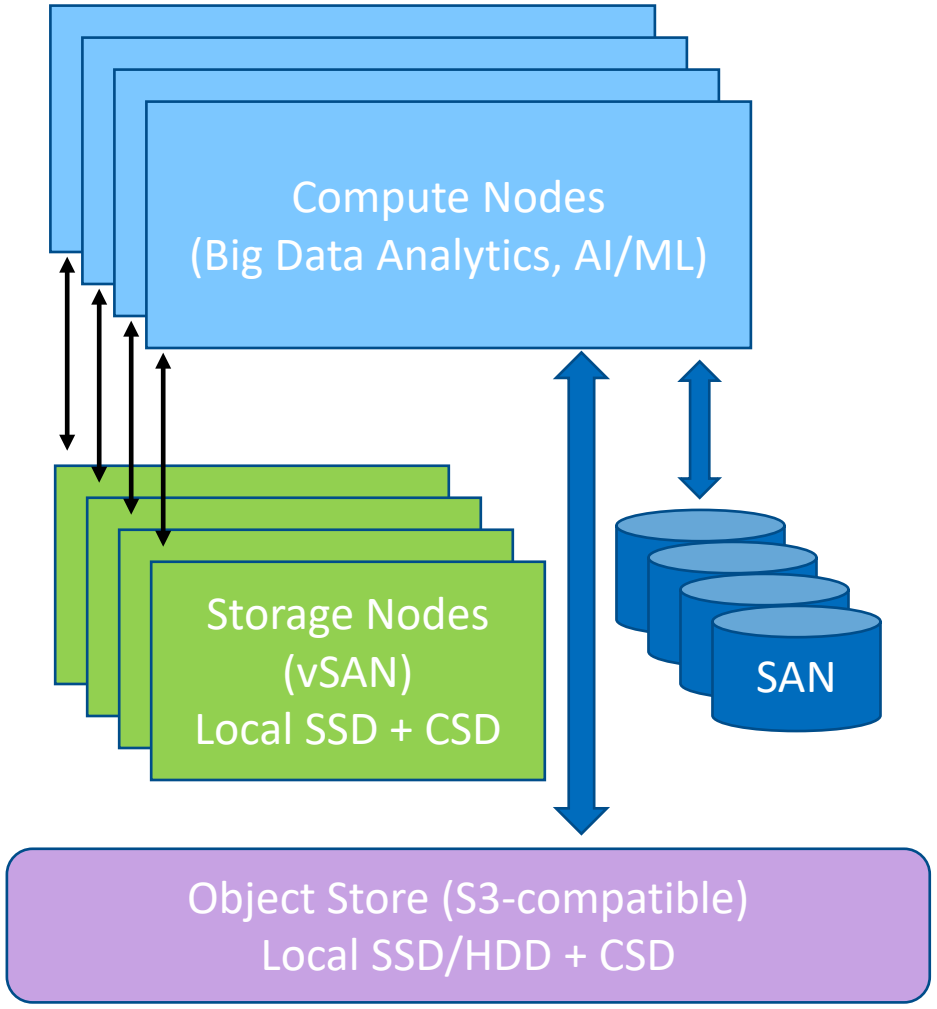
Benefits of virtualizing CSD (vCSD) on vSphere/vSAN

- vCSD = virtual NVMe (or hardware emulated) + CSD command set support
- Can share hardware accelerators more effectively
- Can migrate virtual CSD between compatible hosts
- Flexible composition of storage and near-storage computation engine (FPGA, SmartNIC cores and accelerators on host CPU chip sets)
- Fixed Computation offloading: Compression and Encryption.
- Programmable Computation Service offloading: Key-Value Engine, Database storage engine (containerized), Object Store (optimized network and CPU usage)

Virtual Computational Storage Architecture

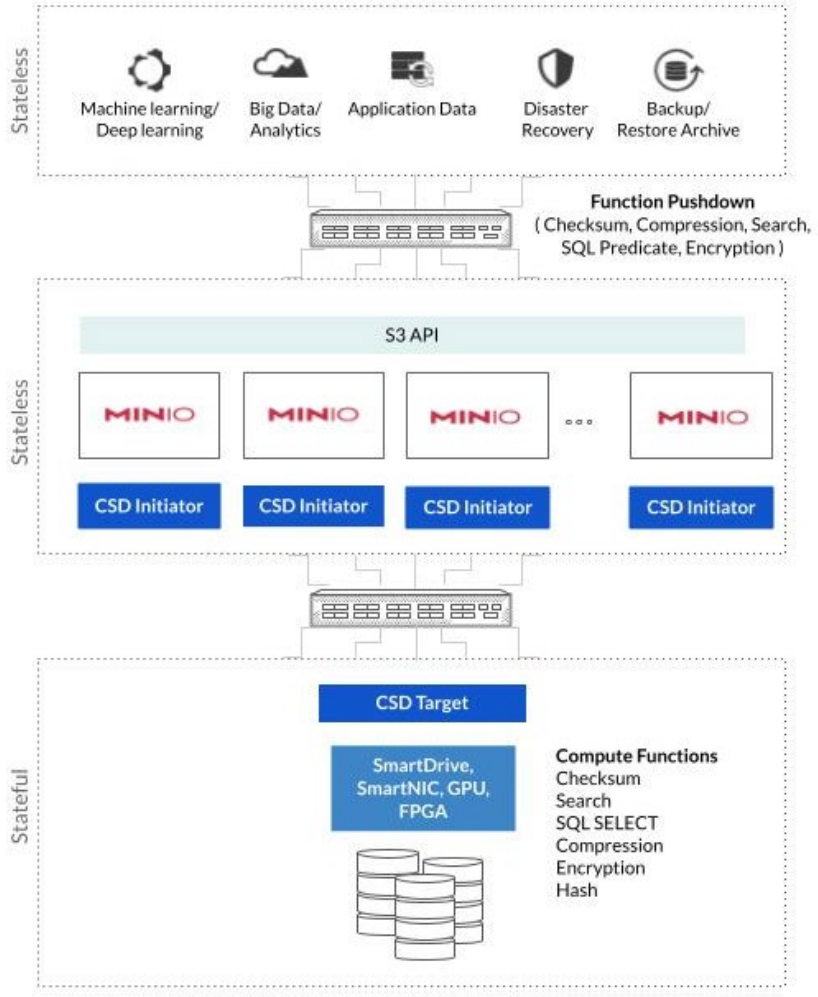


Computational Storage Device (CSD) at VMware Cloud



- Disaggregate cloud native apps (Big Data Analytics, DB, AI/ML) and offload storage intensive functions
- Example: AWS Redshift (compute) + AWS AQUA (storage) + AWS S3 (object store)
- CSD can be used in storage node and object store
- CSD helps minimize data move over the storage network and offload more software functions near storage (checksum, compression, encryption and database queries)

MinIO Object Storage with Computational Storage Device



- Cloud native apps push down computational functions (Checksum, Compression, Search, SQL Predicates and Encryption) through APIs to MinIO object store.
- MinIO object store can offload computational functions through CSD commands to CSD target.
- CSD initiator/target can be virtualized through vCSD.

We're collaborating on a complete end-to-end solution

- Data scrubbing plug-in
  - Streaming Highway Hash
  - Optimized for x86 (AVX-512)
  - Linux executable (reads from standard input)

MINIO

- Container execution environment
  - Tanzu K8s Cluster for running MinIO
  - Virtualized computational storage for vSphere/vSAN

vmware

- Computational storage backend
  - Linux NVMe/TCP extended with EXEC option
  - Pipes virtual object data into MinIO x86 plugin
  - Support for additional HW acceleration
    - ❖ FPGA vs. GPU via oneAPI; 3rd party computational SSDs

intel

Please take a moment to rate this session.

- Your feedback is important to us.

