Next-generation NVMe storage for HPC systems with specialized hardware

Philippe Bonnet

phbo@itu.dk

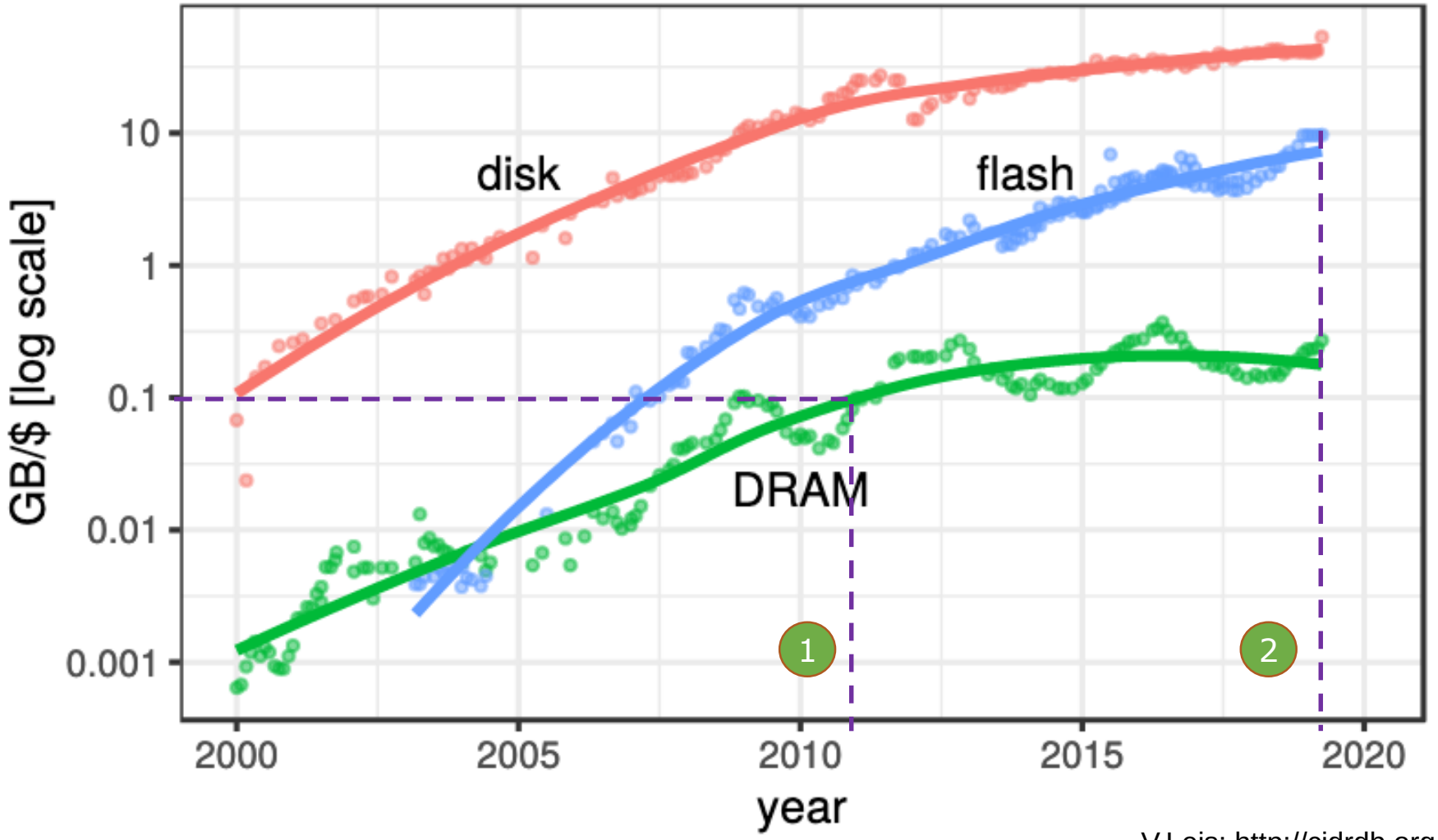
DASYA, Computer Science, IT University of Copenhagen

Storage Devices

- **Storage Drives**
  - Embedded storage media
  - Examples: Solid State Drives (SSDs), Hard-Disk Drives (HDDs), tape.
  - Composed of host-controller interface, **storage controller** and storage media.
  - Connected to a single host via interconnect or fabric.
- **Storage Hubs**
  - Expose a uniform interface to one or several underlying storage drives.
  - Examples: Disk arrays, Functional Accelerator Cards.
  - Composed of host-controller interface, **storage processor** and backplane interface to storage drives.
  - Connected to multiple hosts via interconnect or fabric.

Fundamental Trends:

1. Memory is what disks used to be
2. Flash is 20X cheaper than RAM



V.Leis: http://cidrdb.org/cidr2020/papers/p16-haas-cidr20.pdf

From HDD to SSD (*)


TUTORIAL FREE ACCESS

Not your Grandpa's SSD: The Era of Co-Designed Storage Devices

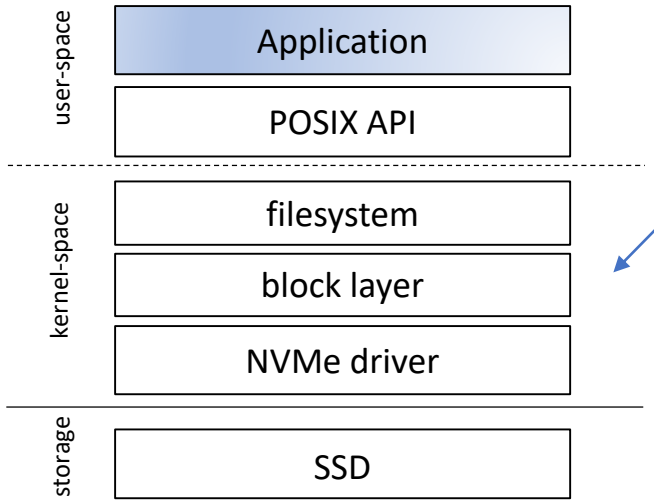
Authors: Alberto Lerner, Philippe Bonnet Authors Info & Claims

SIGMOD/PODS '21: Proceedings of the 2021 International Conference on Management of Data • June 2021 • Pages 2852–2858 • https://doi.org/10.1145/3448016.3457540

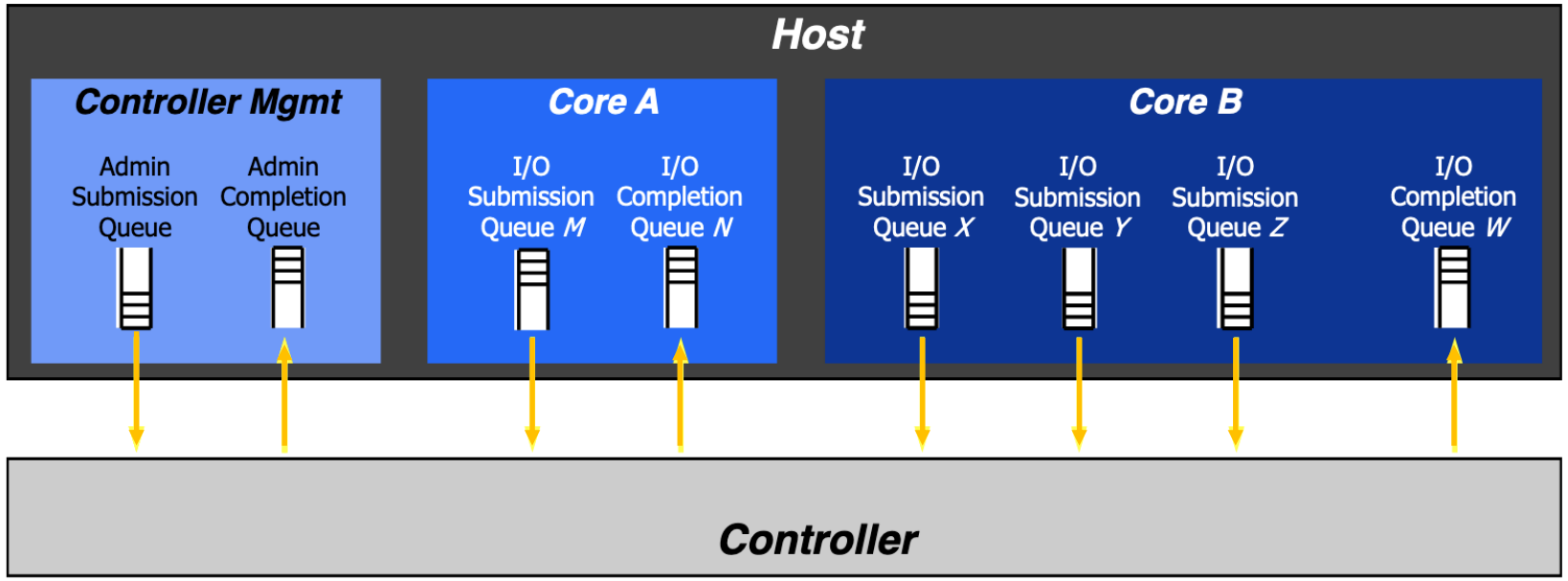
I/O Performance (2020)	I/O performance does not matter	I/O performance is crucial
HDD disk seek: 2msec 1 MiB seq. read: 718 usec	POSIX file with buffered I/O on top of Block-based HDD	Custom buffer management with direct POSIX I/Os on top of Block-based HDD
SSD (*) rand. read: 16 usec 1 MiB seq. read: 39 usec	POSIX file with buffered I/O on top of Block-based SSD	Beyond POSIX and Blocks with NVMe SSDs

(*) SSDs are not a uniform class of devices. Continuum from latency-optimized (Z-NAND) to archival.

Deeper Dive into NVMe



- NVMe is a host-controller interface specification
  - Designed to attach SSDs directly to the PCIe fabric.
  - First specification in 2011. Consortium led by Intel.
  - NVMe 2.0 released on 3/6/2021

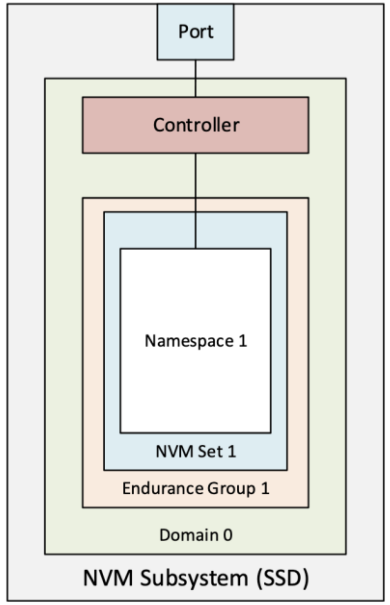


Hosts and controllers communicate through pairs of submission/ completion queues.

The queues are located in memory-mapped address space either on the host or on the device.

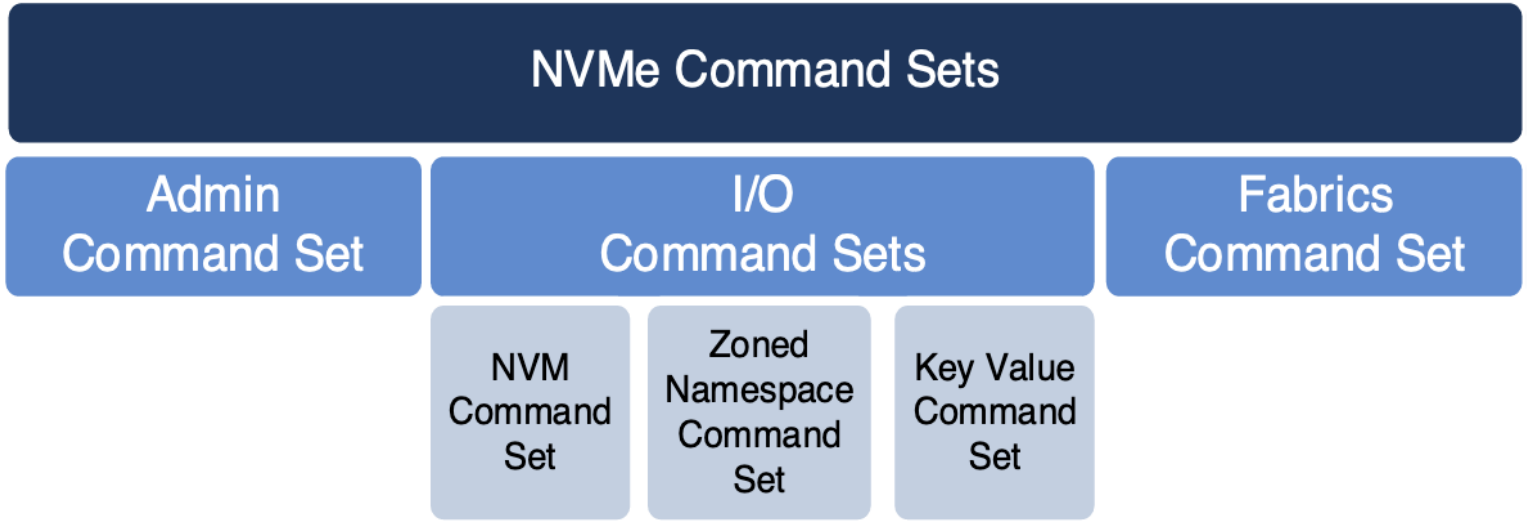
NVMe Abstractions

Namespaces

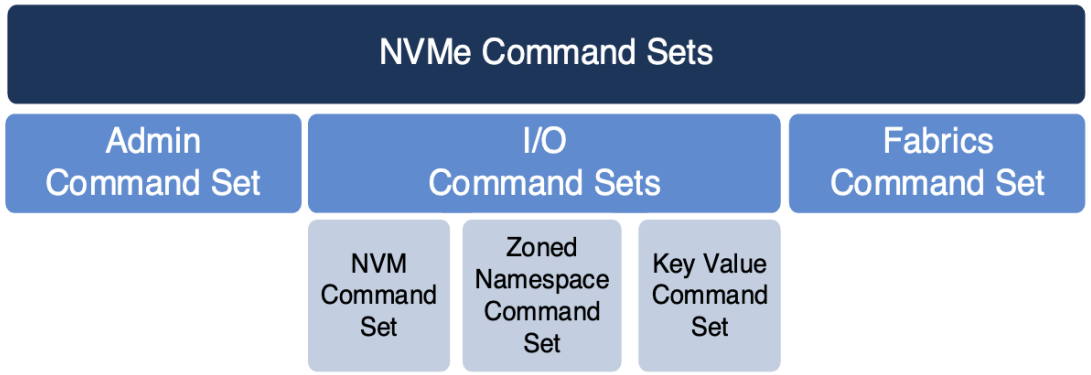


- A namespace is a formatted quantity of non-volatile memory that may be accessed by a host.
- Each namespace has an ID, a size, a capacity (max number of LBAs used), and a utilization

Command Sets



Command sets are the operations associated to a namespace



NVMe Command Sets

- Admin command set include the creation and deletion of submission/completion queues, as well as primitives for device identification, or getting log-pages, device capabilities, and features.
- Three types of I/O command sets:
  - NVM: The namespace is a collection of logical blocks, with read, write, write-zeroes commands. This is the block device abstraction.
  - Zoned: The namespace is partitioned into zones. Each zone is a collection of logical block addresses. The command set establishes that logical blocks must be written sequentially within a zone and that zones must be reset before they are written. It also defines the append command.
  - Key-Value: The namespace is organized as a collection of key-value pairs. The maximum key size is 16B. The commands supported are store/retrieve, list, exist, delete.

Copenhagen is where next-generation NVMe storage systems are developed

Computational Storage

Los Alamos announces details of new computational storage deployment

November 16, 2020

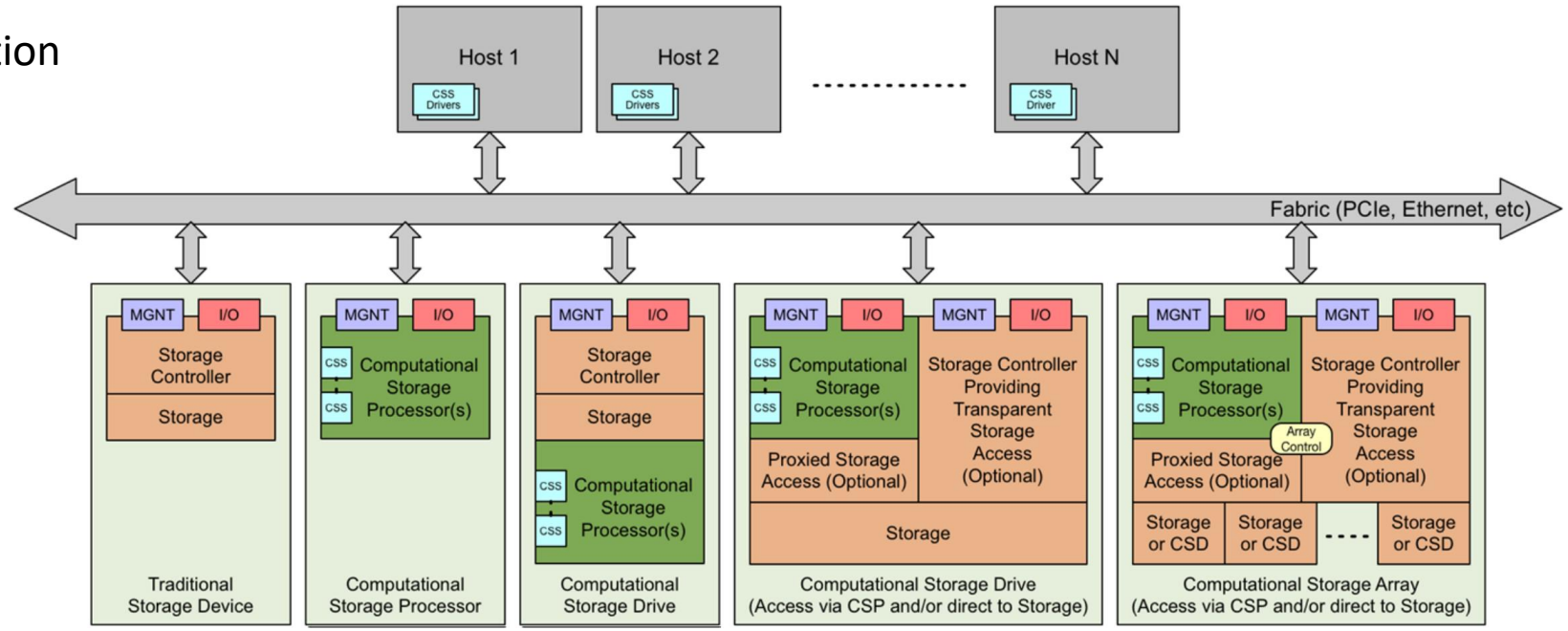
Computational Storage

- Programmable storage controller/processor:
  - Programmable storage drive: In-storage processing
    - Storage controller: MPSoC (Multi-Processor System on a Chip integrating ARM cores and FPGA)
  - Programmable storage hub: Near-data processing
    - Storage processor: CPU, MPSoC or Data Processing Unit (ARM core with hardware accelerated functions, e.g., NVMe controller).
- Computational storage is a means to
  - decrease data movement and thus improve the cost-performance of data-intensive applications
  - gracefully scale compute capacity with volume of stored data
  - (i) reduce power consumed – thanks to low power CPUs - and (ii) improve energy proportionality – thanks to low power modes when not processing.

Architecture View: SNIA Terminology

Computational Storage Processor (CSP)

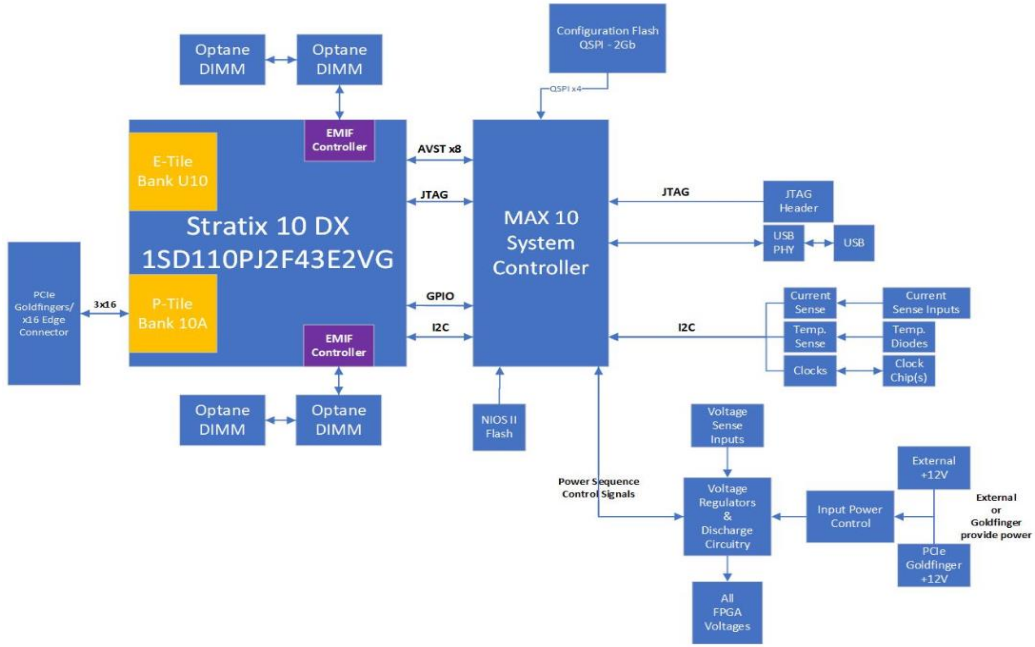
- Provides computational Storage Services (CSS)
  - Fixed: compression, encryption
  - **Programmable**: installed via code upload
    - OS image
    - Container application
    - FPGA bitstream
    - eBPF **bytecode**



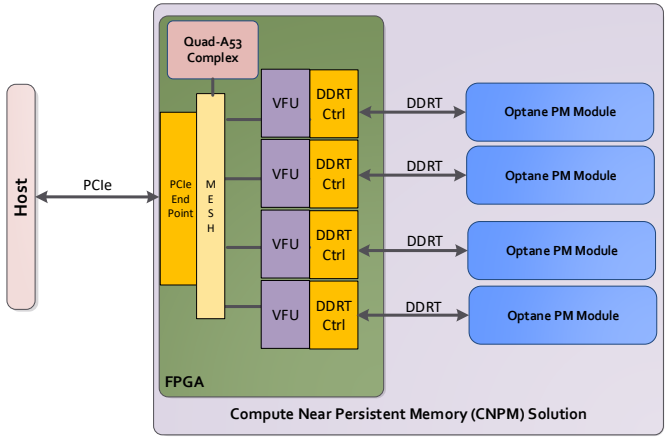
SNIA. Computational Storage Architecture and Programming Model. V0.5, Rev 1. Aug 2020.

Computational Storage Device - Intel Kestral

- Allows Optane Pmem connectivity over PCIe (without using Xeon DDR ports)
- Total capacity: 2TB, (4x512GB Pmem); PCIe Gen4 x16
- Key Components:
  - **Intel® Optane™ persistent memory 100 series**
  - **Intel Stratix 10 DX FPGA**
- Power: <150W
- Full Height; Half Length (FHHL) card (two slots)
- ARM core available
- Max 10 System Controller (Power Sequencing, Temp and Power Monitor, Remote System Update, debug...)
- Health monitoring for Optane DIMMs



Reference design:
~Q1 2022

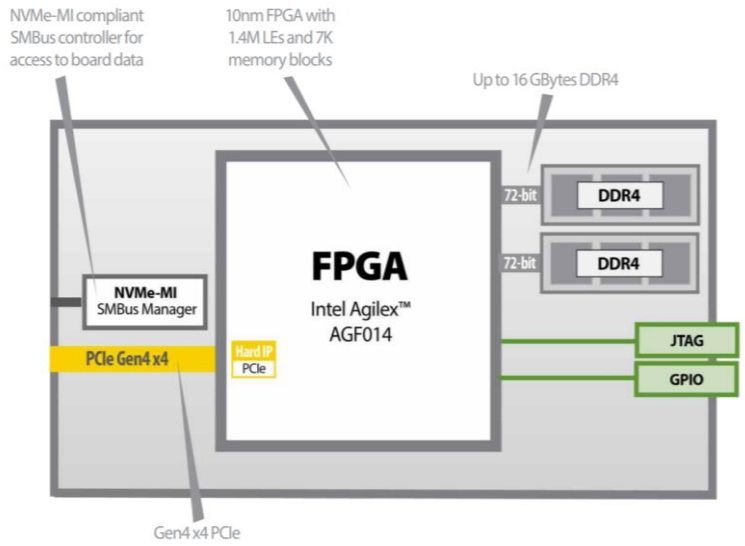


Potential usage scenario:

- Connected as a block device (Paging with DSA/NVMe)
  - Will likely gain a few us latency but could be prohibitively expensive for most usecases.
- Connected as a P2P compute device with large memory
  - For usecases benefiting from offloaded compute and large memory capacity
- Connected as a standalone computational memory accelerator
  - Independent acceleration without too much dependance or data movement from the host/peer devices. e.g. accelerating a training model with new data units (text, pictures)

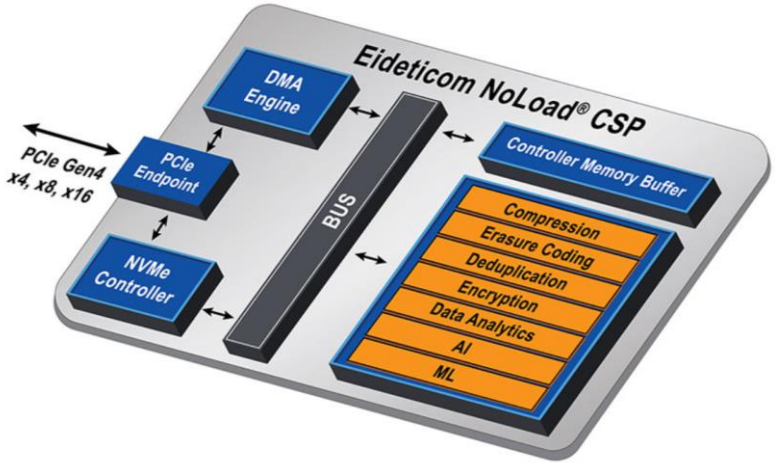
Computational Storage Device – Bittware IA-220-U2

Bittware PCIe FPGA Board:



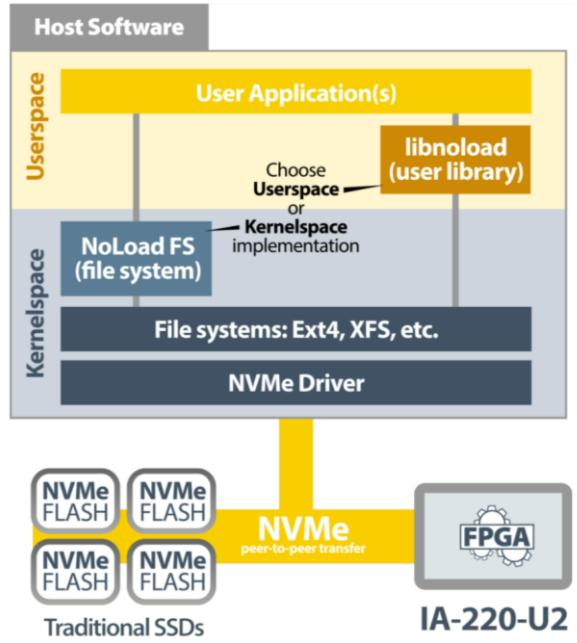
https://www.bittware.com/fpga/ia-220-u2/

Eideticom NoLoad IP:



HW components:

- NVMe font end : PCIe endpoint, NVMe controller, DMA engine...
- CSS HW –accelerated services (Compresion,...)



SW components (host SW):

- libnoload: modifing applications to use library (OS remain untouched)
- NoLoad FS: use filesystem as a shim (user appliactions needs zero changes)

„Intended to be deployed within conventional U.2 NVMe storage arrays (approximately 1:8 ratio) allowing FPGA-accelerated instances of:

- Erasure Coding and Deduplication
- Compression, Encryption & Hashing
- String/Image Search and Database Sort/ Join/Filter
- Machine Learning Inference”

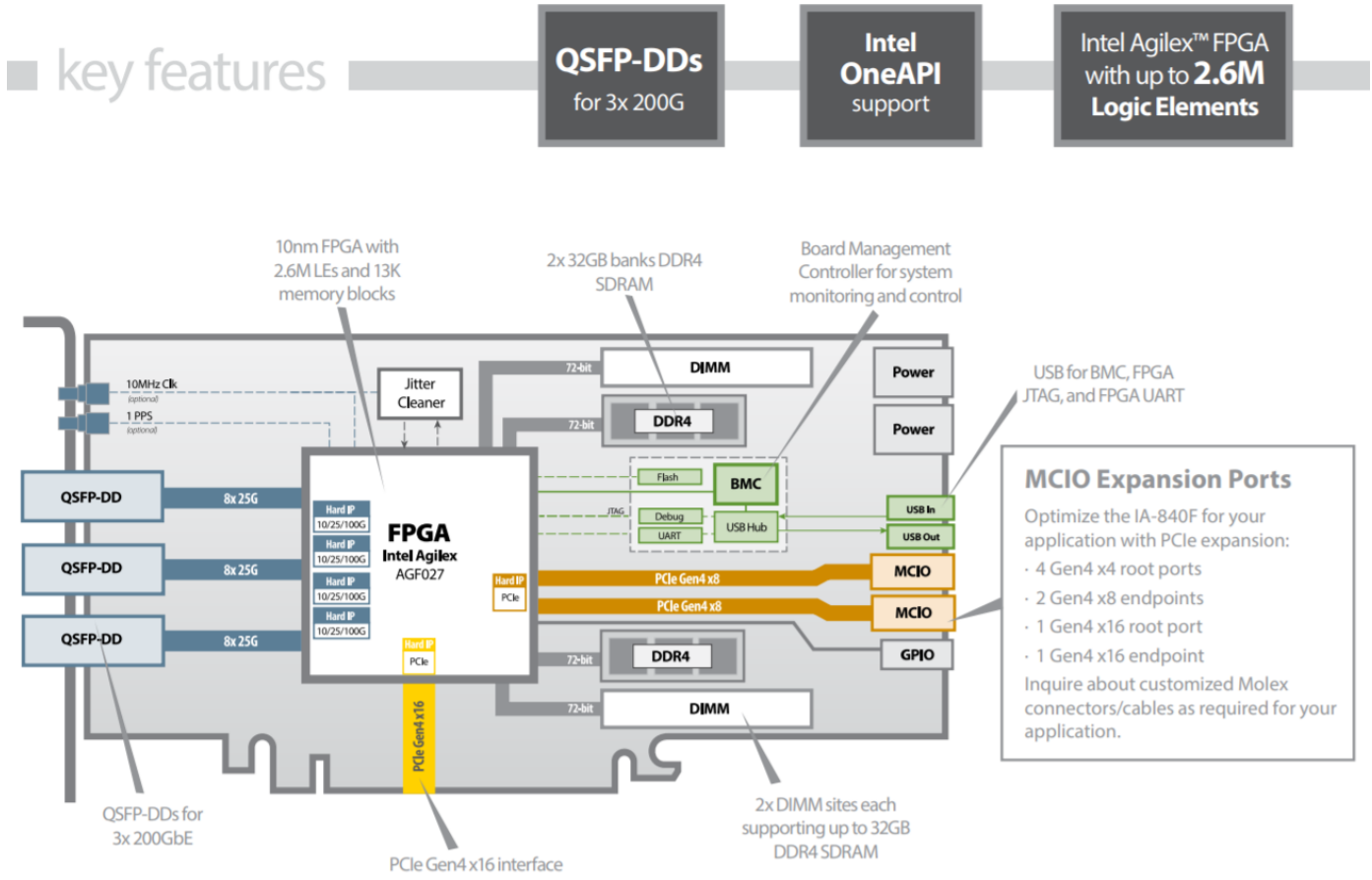
Computational Storage Device – Bittware IA-840F

https://www.bittware.com/fpga/ia-840f/

Enterprise-Class Intel Agilex Based
FPGA Accelerator

Intel OneAPI support!

Connectors for cable connection to NVMe SSDs.



Intel Agilex:

•2nd-Generation HyperFlex Architecture:

Up to 40 percent higher performance or up to 40 percent lower total power compared with Stratix 10 FPGAs.

•DSP Innovation:

•Supports hardened BFLOAT16 and up to 40 teraflops of digital signal processor (DSP) performance (FP16).

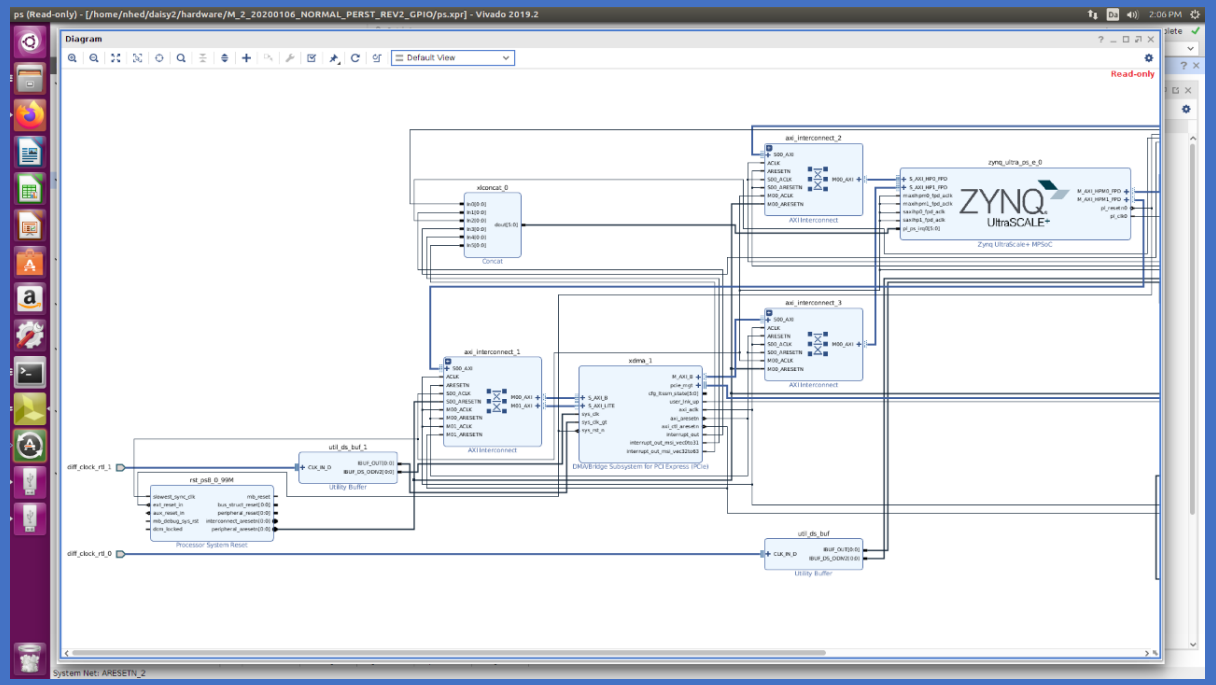
•Advanced memory support:

•DDR4, QDR-II+ (through custom BittWare DIMMs) and Intel Optane DC persistent memory.

OpenSSD Daisy

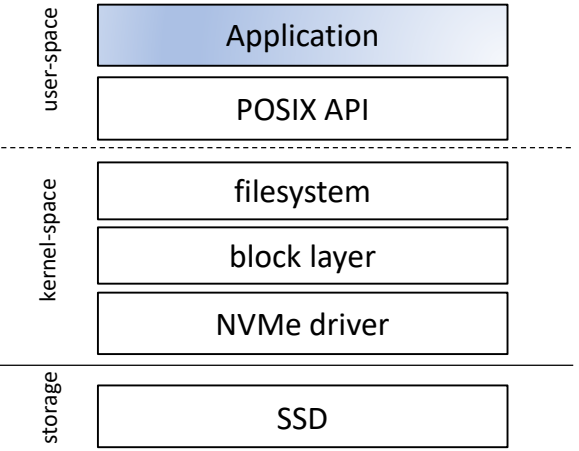
- PCIex16 Gen3 | 100GE
- MPSoC (Zynq Ultrascale+)
  - ARMv8 Cortex A-53 (4 cores)
    LPDDR4 (4GB)
    running Embedded Linux
  - FPGA
    connected to peripherals
    Xilinx IP
- 2xM.2 connectors

DAPHNE



Programming Computational Storage

- *Defining new storage interfaces:*
  - Computational storage enables a co-design of data-intensive applications and storage interfaces
  - *boot/config time via OS image, bitstream, container application*
- *Shipping code from host to storage*
  - Bytecode is generated on host (*at compile time*) and shipped to computational storage (*at run time*)
    - SNIA points out eBPF
      - Eid-Hermes: https://github.com/Eideticom/eid-hermes
    - Other alternatives should be explored

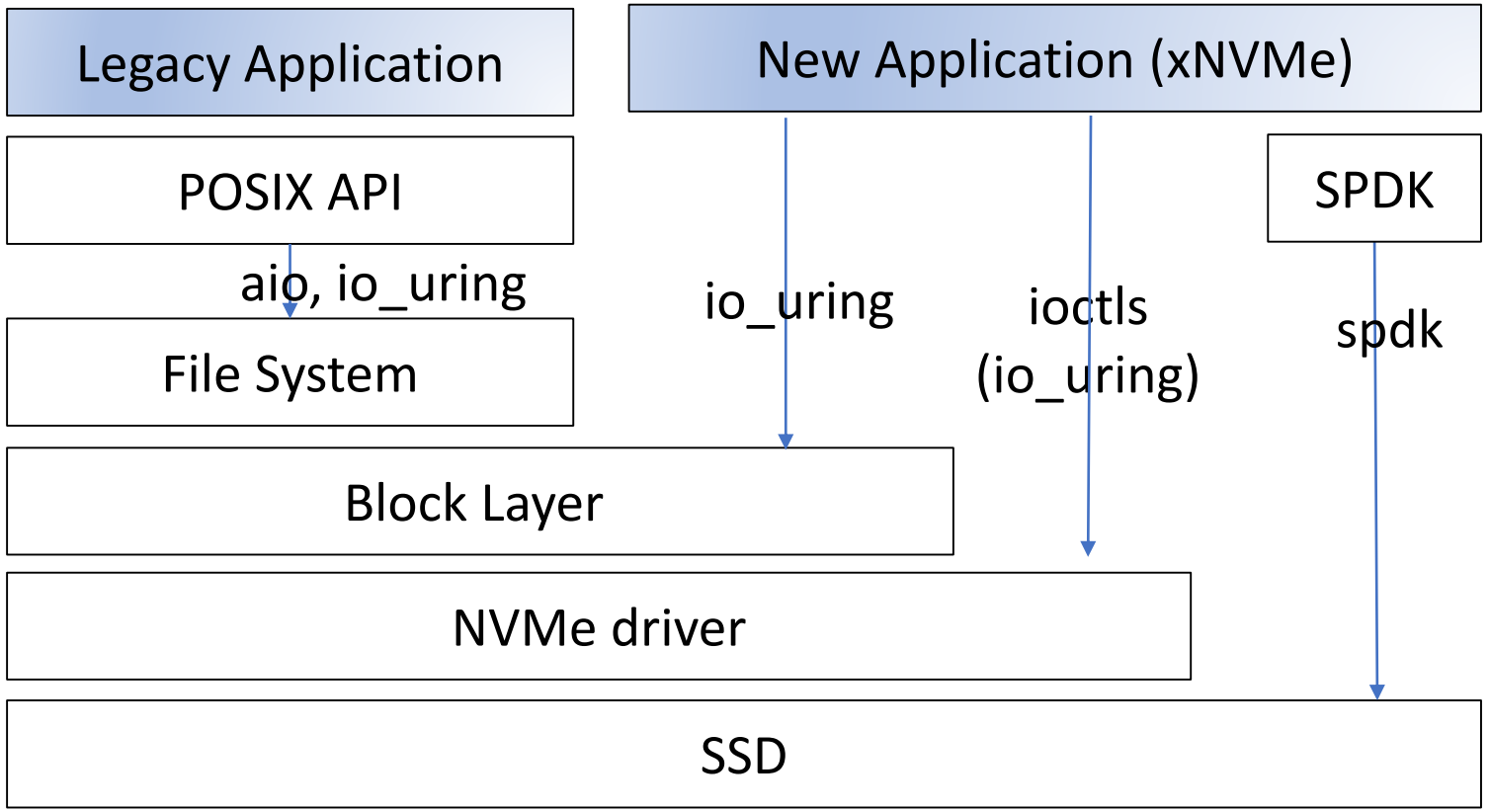


Linux I/O Frameworks

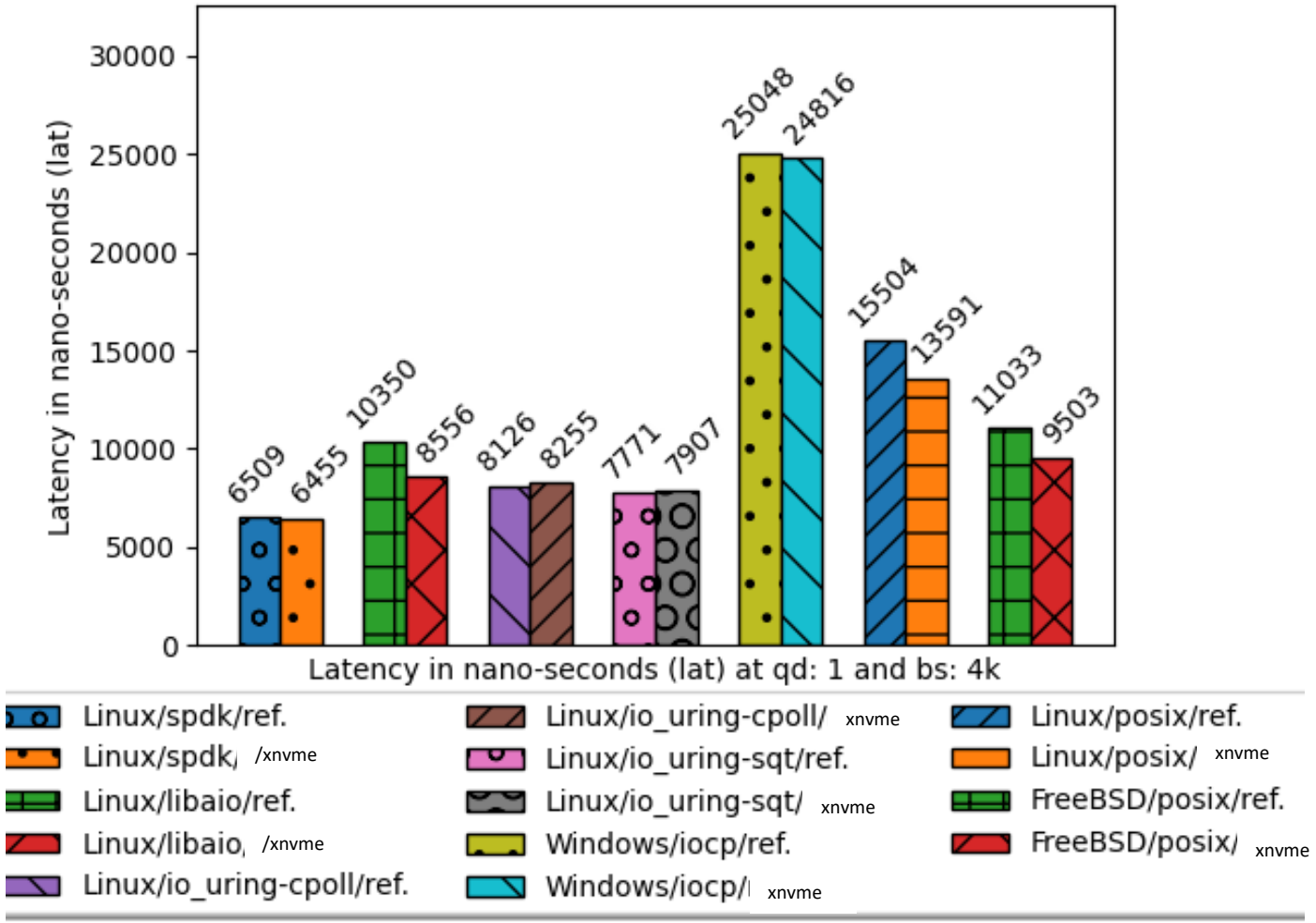
File system extensions for ZNS
In F2FS, XFS

mlq-blk for NVMe since 2013

No support for KV in NVMe driver yet



I/O Frameworks



Simon Lund et al. Under Submission.

Thesis

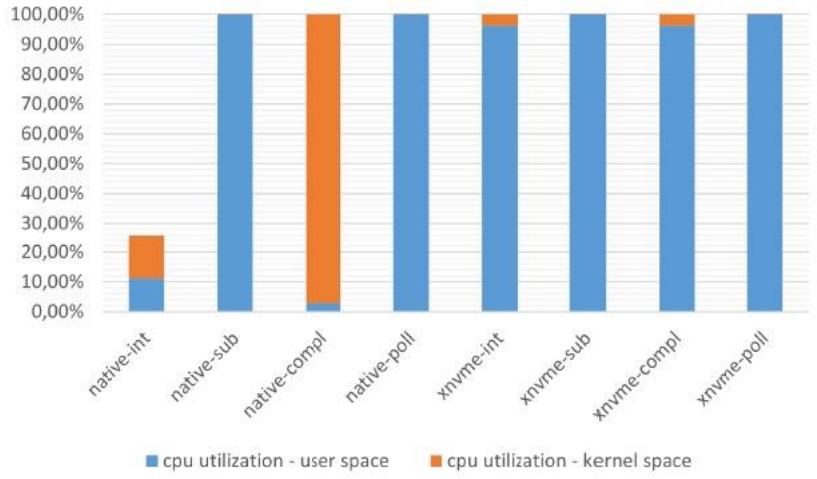
An Analysis Of New Generation SSDs and Storage Stack

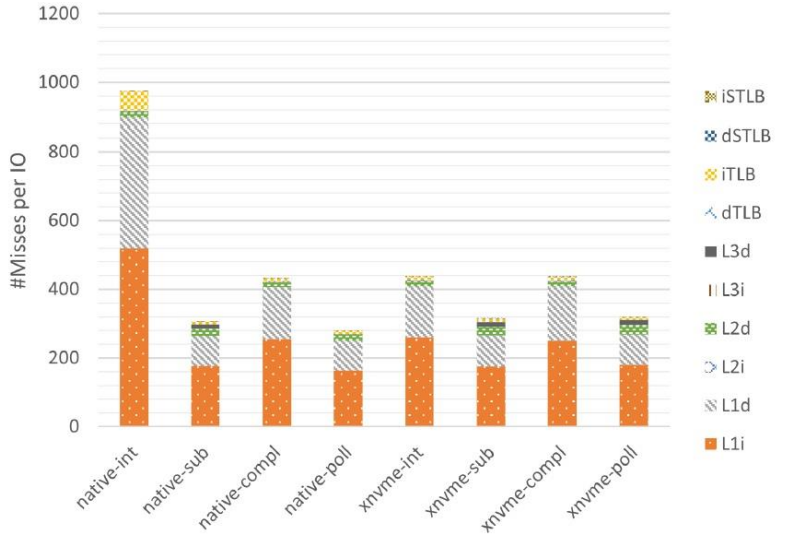
Andreas Blanke and Magnus Krøyer

IT-University of Copenhagen

Supervisors: Philippe Bonnet and Pinar Tözün

Spring 2021





Conclusion

1. The storage software stack must be adapted to leverage the capabilities of modern (NVMe) storage devices.

2. Computational storage as a means to improve cost-performance, scalability and energy-proportionality.

3. Standard expected in 2022 (NVMe Computational Storage command set). Products already on the market. Deployments in largest HPC clusters and hyperscalers.