



Western Digital®

NVMe Computational Storage Processor for Edge Applications

Anand Kulkarni

Western Digital Research

Apr 19, 2022

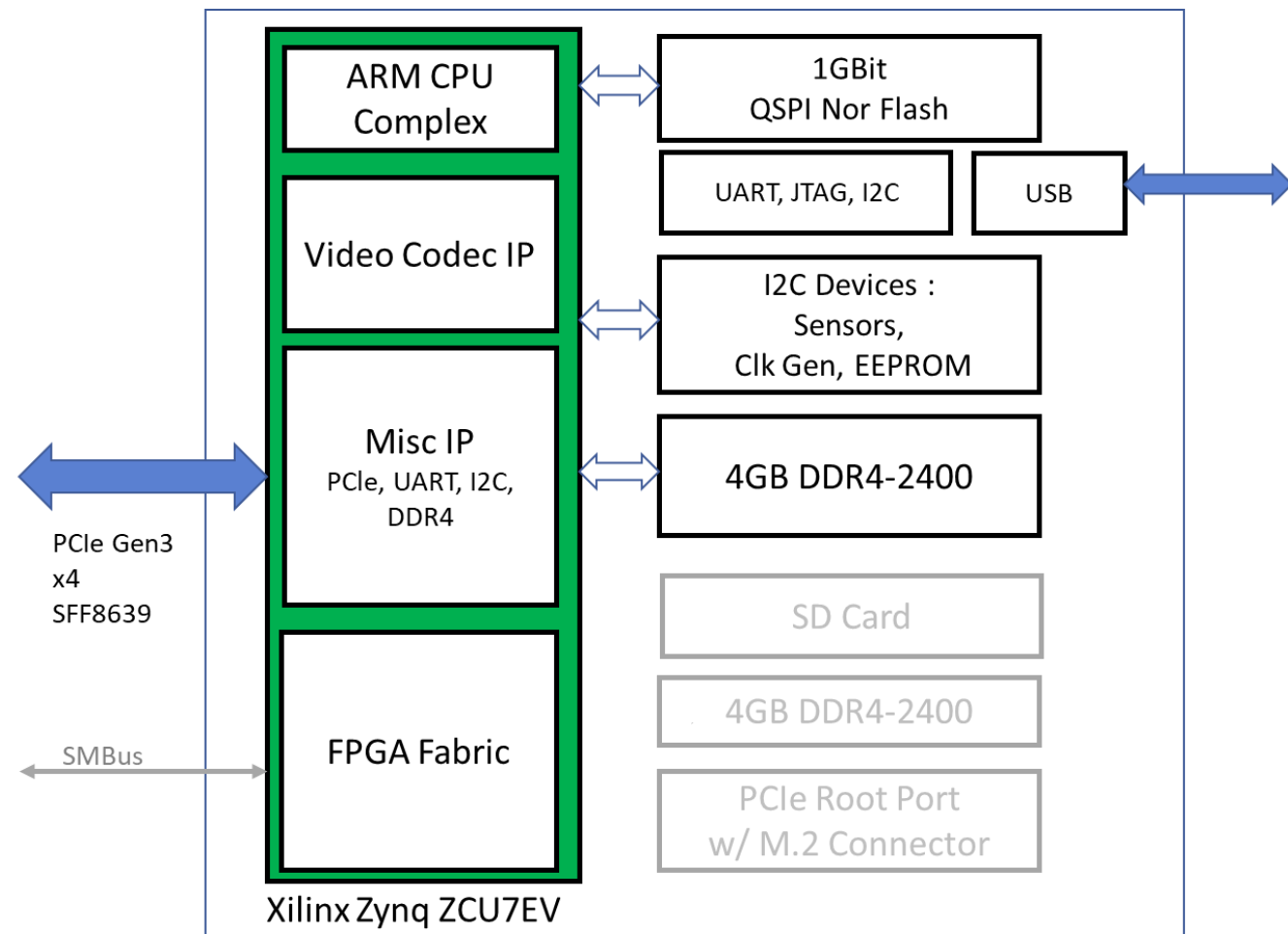
Outline

- Introduction to Western Digital Accelerator Platform
- Target Use cases
 - Video Transcoding
 - Machine Learning
 - Computational Storage
- Western Digital Computational Storage Development Platform
 - TFLite Offload and Acceleration
 - DB Search Offload and Acceleration
 - DB Search Acceleration with PCIe P2P DMA
- Conclusion

Western Digital Compute Accelerator Platform

FPGA Based PCIe ML/AI Accelerator Device in U.2 Formfactor



- Xilinx® UltraScale+™ MPSoC XCZU7EV
- 4GB DDR
- Gen3 x4 PCIe 2.5" SFF
- 25W Max Power




Western Digital Compute Accelerator Platform

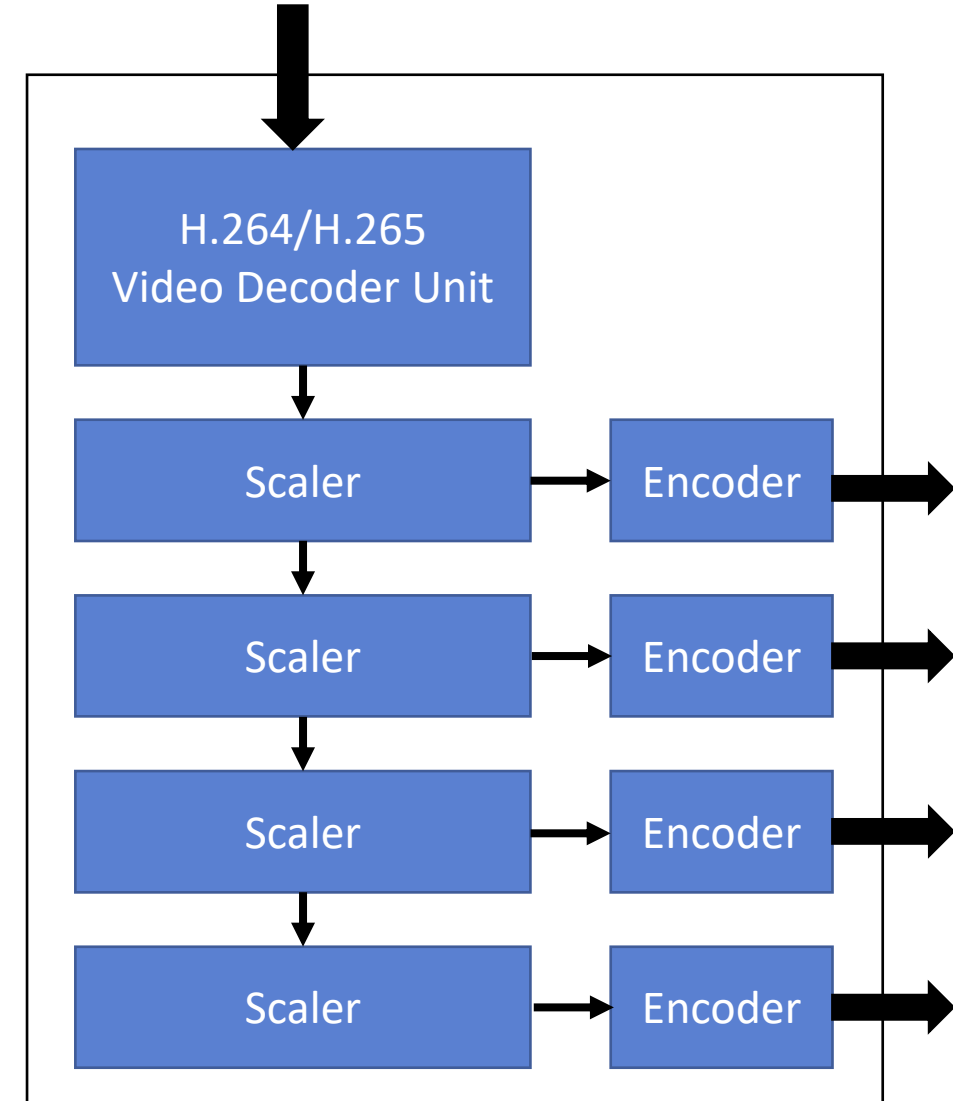
- Versatile & Scalable
- Data center ready



Use case	Market
Video transcoding (H264/H265)  XILINX.	HD/UHD video streaming VoD, Sports, Gaming
AI-Inference: image/video 	Image/Video: Classification, Segmentation, Super Res, Pose Est., etc. Video Surveillance Edge GW Smart City Medical Imaging
Computational Storage NVMe™ & eBPF Support	TP4091 Prototyping Analytics Acceleration Video Applications Database Applications

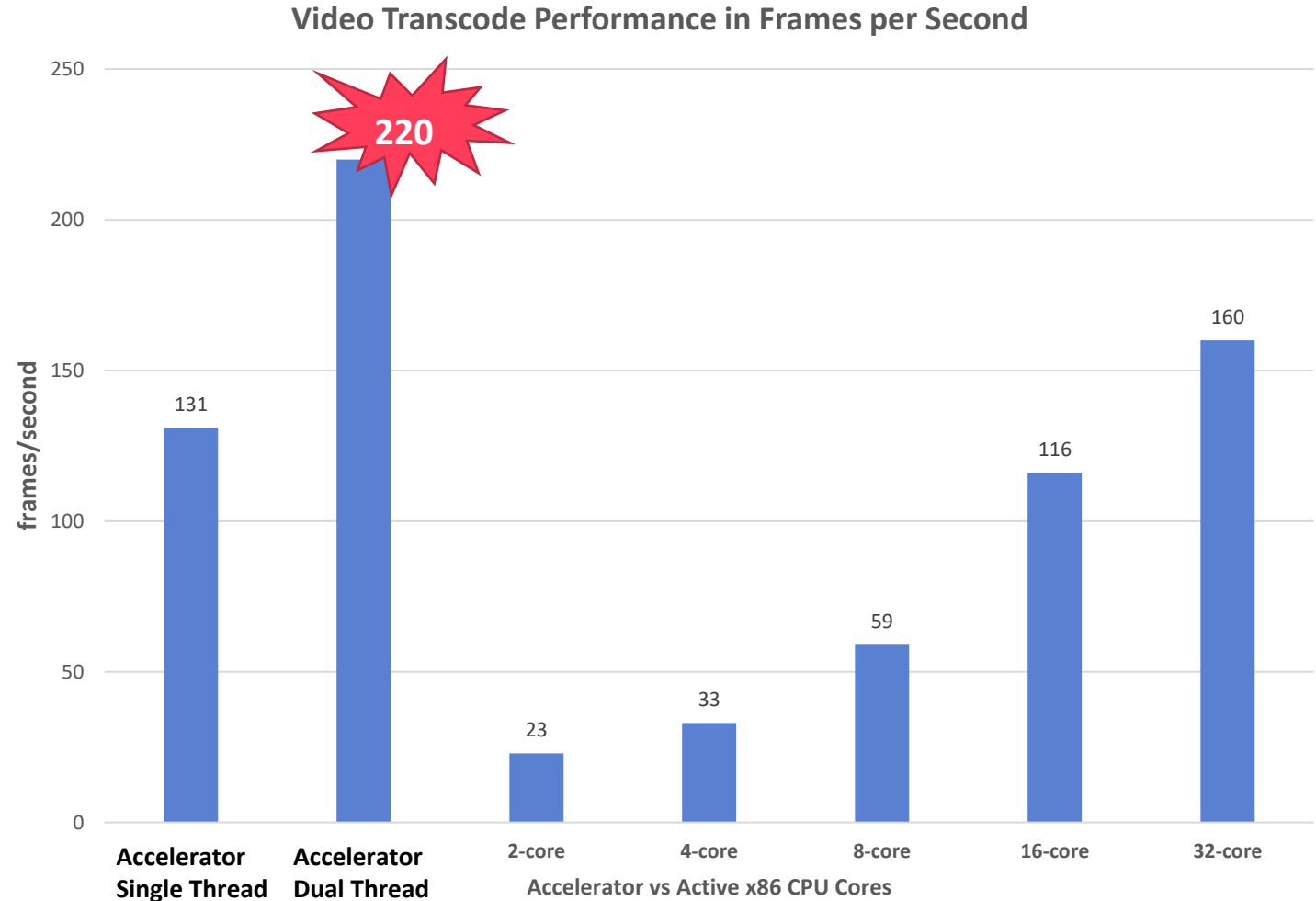
Accelerator for Video Transcoding Applications

- Developed in Partnership with  XILINX.
- Decode/Encode H.264/H.265 Streams
- Integrated with FFmpeg
- Supports inputs up to 4K@60fps
- Transcode multiple video streams
- High Performance
- Excellent Performance/Power
- Scalable, Supports Hot Plug
- Limited Availability now
 - Xilinx PN A-U2MA-P04G-PQG-021



Video Transcode Performance with Western Digital Accelerator

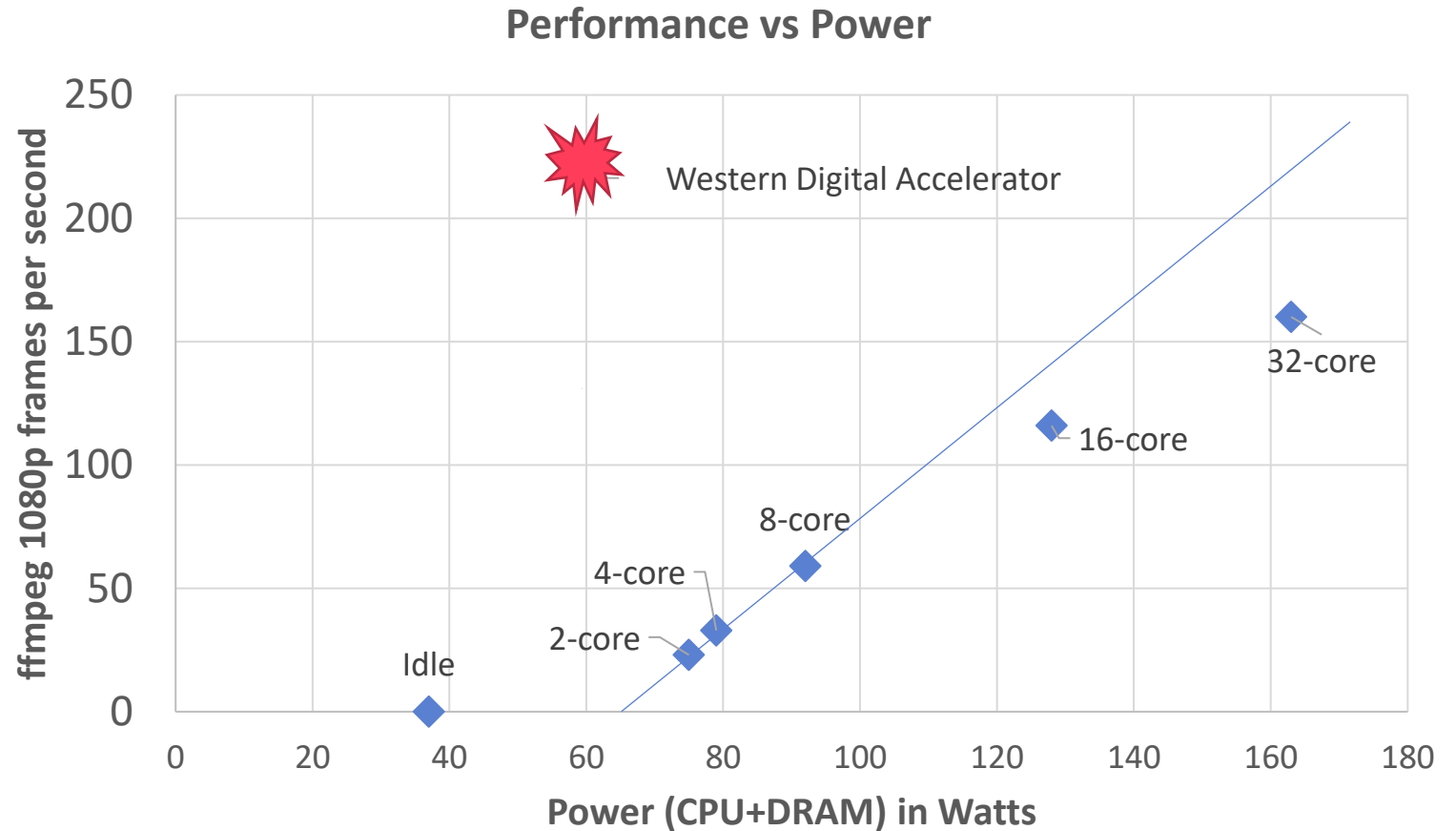
- Western Digital Accelerator beats performance of 32 x86 CPU cores
- Effective Compute Offload
 - 100x fewer instructions executed
 - 10x to 20x fewer Stall cycles on x86 CPU
 - 50x to 100x reduction in DRAM traffic
- Scalable performance
 - Multiple accelerators per Server



Multiple ffmpeg instances on sets of 4 x86 CPU Cores, each at utilization of ~85%

Video Transcode Performance to Power ratio

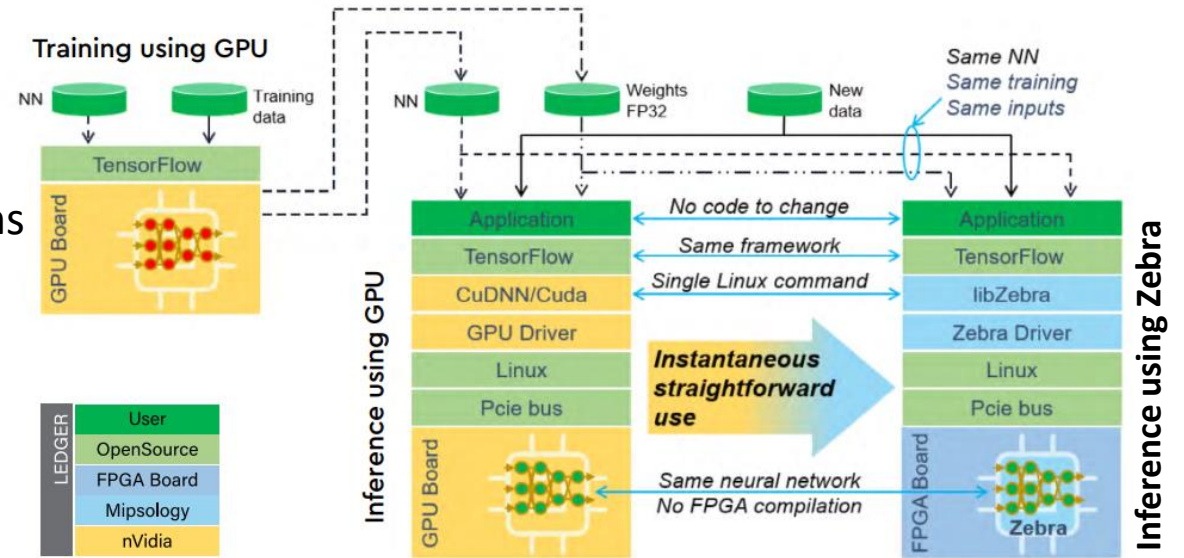
- Western Digital Accelerator
 - Less than 20W at peak performance
 - CPU close to Idle
- 32-core x86 CPU
 - Almost 100W higher for similar or lower performance
 - Each x86 CPU core adds ~3W of power consumption



- ffmpeg invoked to transcode 1080p 60fps input stream into multiple lower resolution streams
- Multiple ffmpeg instances on sets of 4 x86 CPU Cores, each at utilization of ~85%

ML Inference Accelerator

- Supported Frameworks
 - TensorFlow, PyTorch, MXNet, Caffe
- Run any pre-trained CNN models without modifications
 - Automatic quantization, No pruning required
 - Up to a Billion weights, Million Layers
 - Concurrently run two separate networks
- Supports wide variety of Networks
 - ResNet, Yolo, Inception...
 - SSD, EfficientDET, MaskRCNN...
 - SRGan, AlpaPose...
 - List continually updated
- Networks can be split
 - Compute Intensive Layers accelerated in FPGA
 - Unusual layers can be kept on CPU
 - Allow new networks and architectures
 - EfficientDet/Net, BERT, LSTM, etc.
- In partnership with Mipsology Inc.





Network	Performance (img/sec)
Inception V3	246
Inception V4	118
ResNet 50	479
ResNet 152	198
Yolo V1	66
Yolo V2	72
Yolo V3	22

Western Digital Compute Accelerator Platform

- Versatile & Scalable
- Data center ready



Use case	Market
Video transcoding (H264/H265) 	HD/UHD video streaming VoD, Sports, Gaming
AI-Inference: image/video 	Image/Video: Classification, Segmentation, Super Res, Pose Est., etc. Video Surveillance Edge GW Smart City Medical Imaging
Computational Storage NVMe™ & eBPF Support	TP4091 Prototyping Analytics Acceleration Video Applications Database Applications



Features Wishlist for Computational Storage

- Demonstrate Performance/Power/Cost benefits
- Additional criteria to be serious candidate for datacenter deployment
 - Demonstrate clear benefits in “real-world” user environments
 - Support variety of applications
 - Adapt to multiple workloads in each application
 - Support Industry Standards
 - Show Scalability
- This will be a long road
 - Stay adaptable to moving targets and changing targets
 - Allow for continuous improvements
 - Expect additional feature requests
 - Multi tenancy
 - Security
 - New protocols/interfaces

Western Digital NVMe Computational Storage Platform

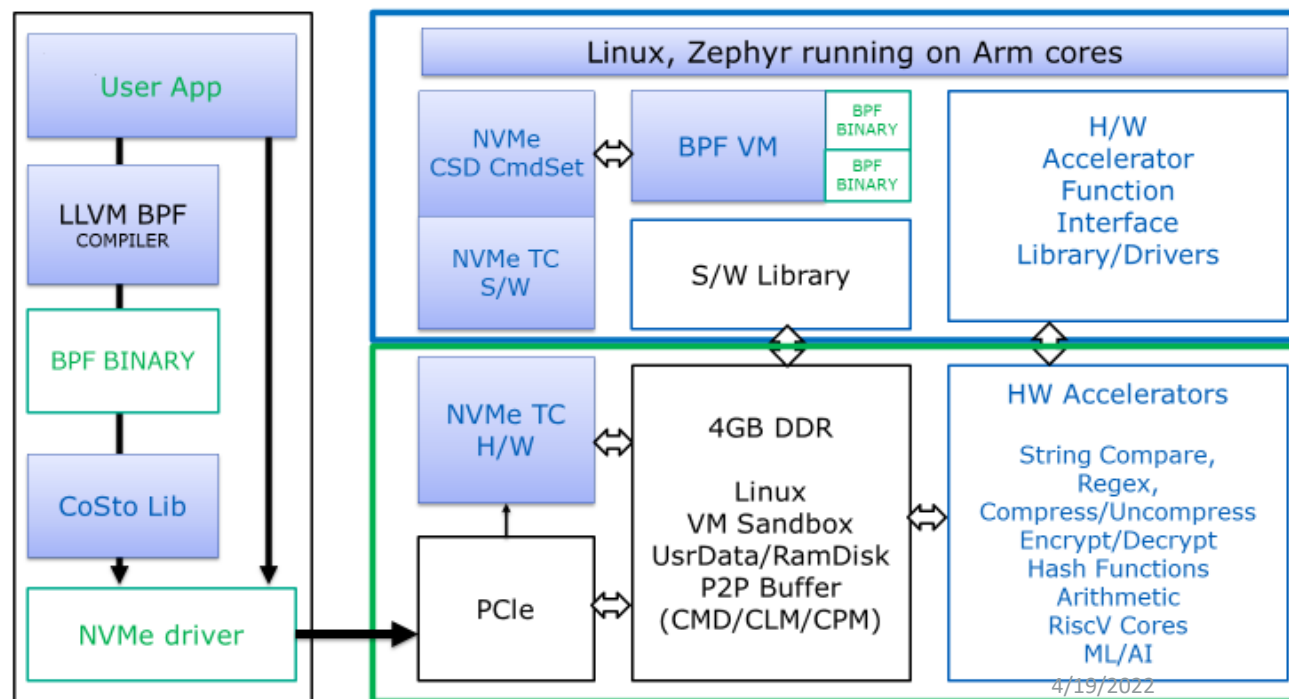
- ARM® for SW tasks
 - Linux® on A53 Cores
 - Zephyr RTK for latency sensitive tasks
- FPGA (>70%) available for HW tasks
 - AXI Interfaces for ease of integration
- NVMe Target Controller
 - Implementation split between FPGA and ARM Cores
 - Maximize FPGA resources for Accelerator functions
- NVMe TPAR 4091 CSD Command Extensions
 - Download and Execute SW Kernels built as eBPF code binaries running in a VM
 - Offload compute to SW/HW Kernels on the Compute Storage Processor (CSP) or Device (CSD)
- uBPF VM for running eBPF kernels
 - Link application specific libraries and export corresponding APIs
 - Integrate custom libraries and drivers for HW Accelerators implemented in the FPGA



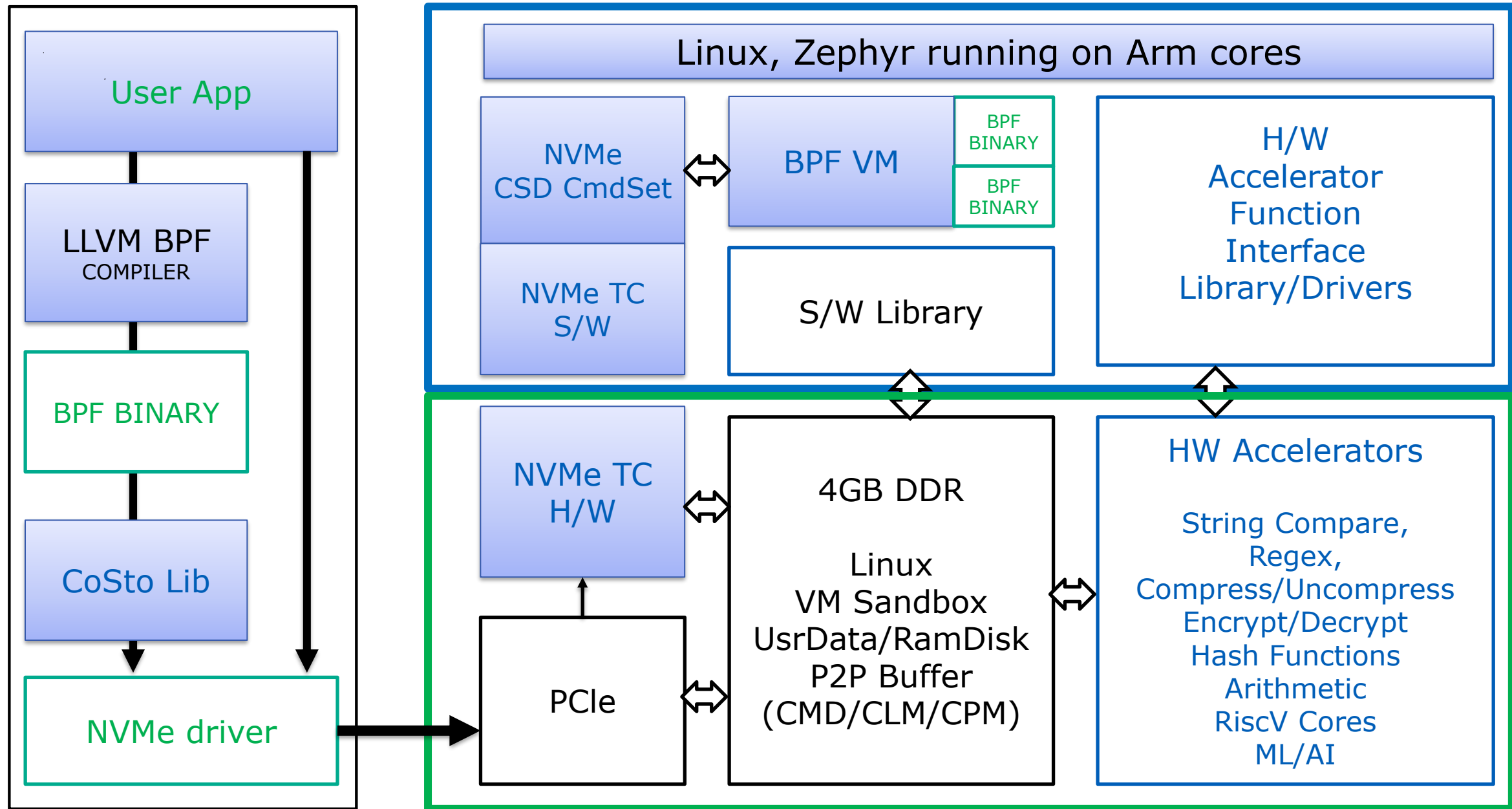
Western Digital NVMe Computational Storage Platform

- Demonstrate Custom Accelerators on NVMe CSD
 - Complete development environment will be available
 - Standard AXI interfaces to simplify integrating HW IP into the design
 - Allows for an elastic boundary between HW, RPU Firmware (Zephyr) and APU Firmware (Linux)
 - Implement features in Firmware and migrate features to HW over time
- Wide variety of Potential Applications
 - Image/Video Analytics, ML/AI
 - Database Search Acceleration
 - Genomics
- Actively under development
 - Developed In partnership with Antmicro
 - Expecting to publish in the near future

Acceleration Platform for Remote BPF code

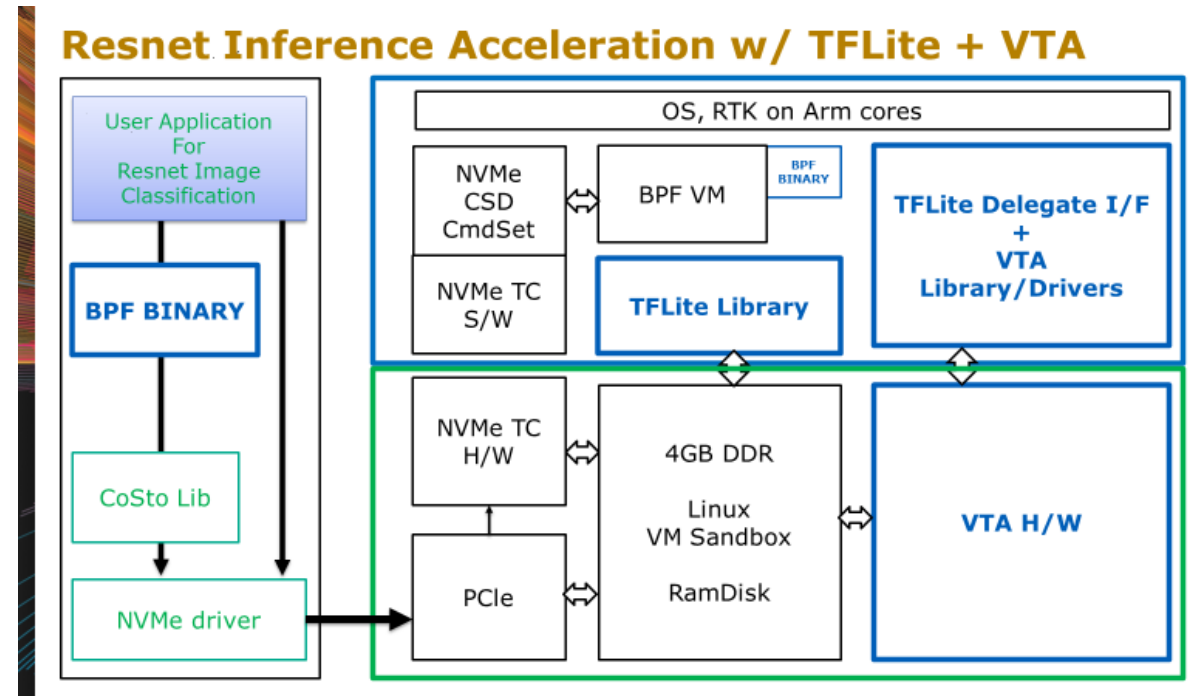


Acceleration Platform for Remote BPF code

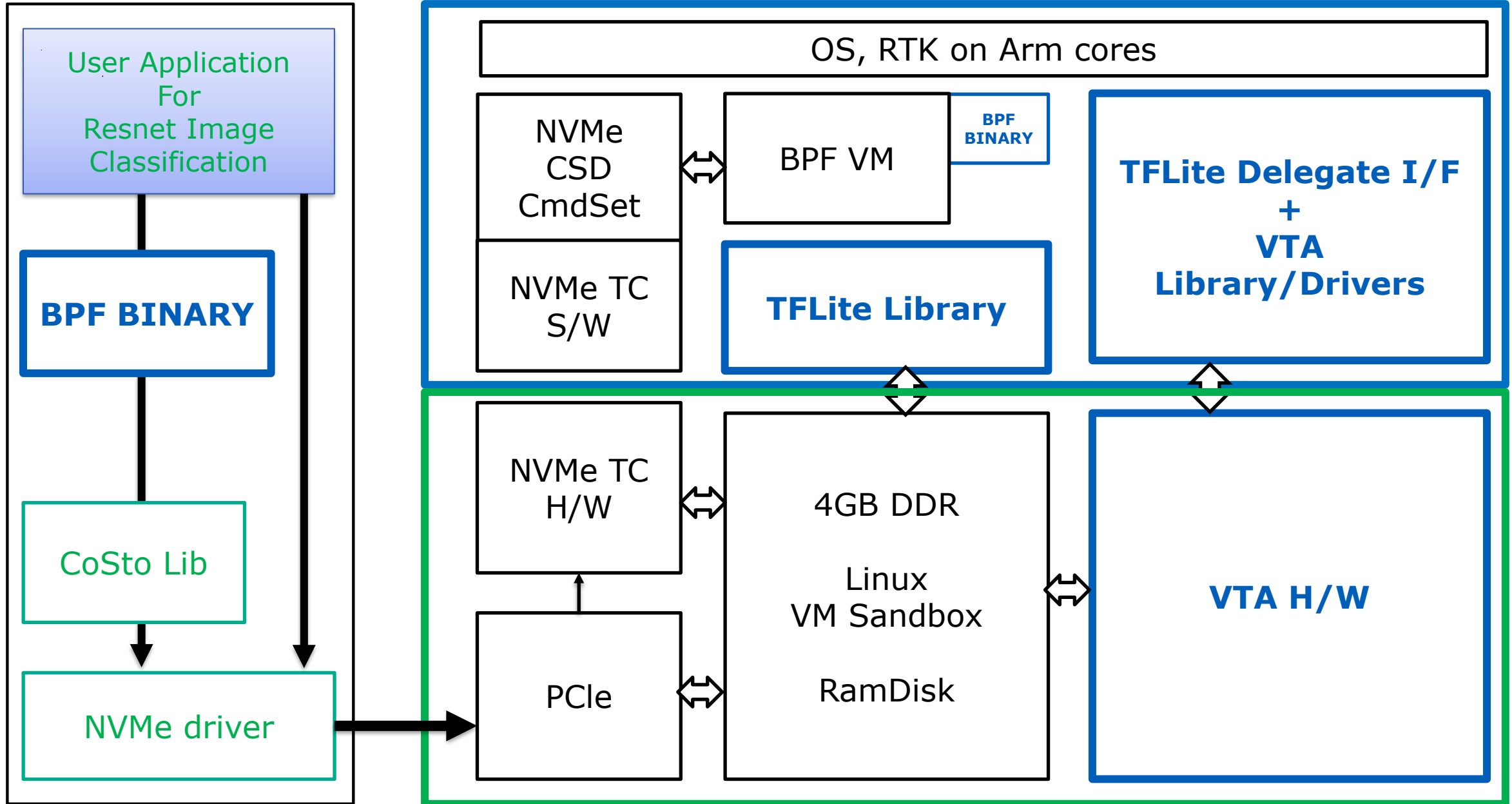


Resnet Inference Acceleration w/ TFLite + VTA

- Reference design for Western Digital NVMe Computational Storage Platform
 - Accelerator can run trained TFLite models on data in Ramdisk
 - Demonstrate scope and value of Computational Storage
- Integrates TFLite and TVM-VTA into the framework
 - TFLite linked as a library into BPF VM
 - TFLite invokes VTA drivers via Delegate Interface for specific operators
 - VTA HW integrated into the FPGA design
- Compute operations Accelerator focused
 - Lightweight Host Application offloads to TFLite
 - TFLite delegates to VTA H/W when possible
 - Delegating 2DConv layers for most acceleration
- Actively under development
 - In partnership with Antmicro, Poland
 - Expecting to publish repository soon
 - Potential for variety of enhancements



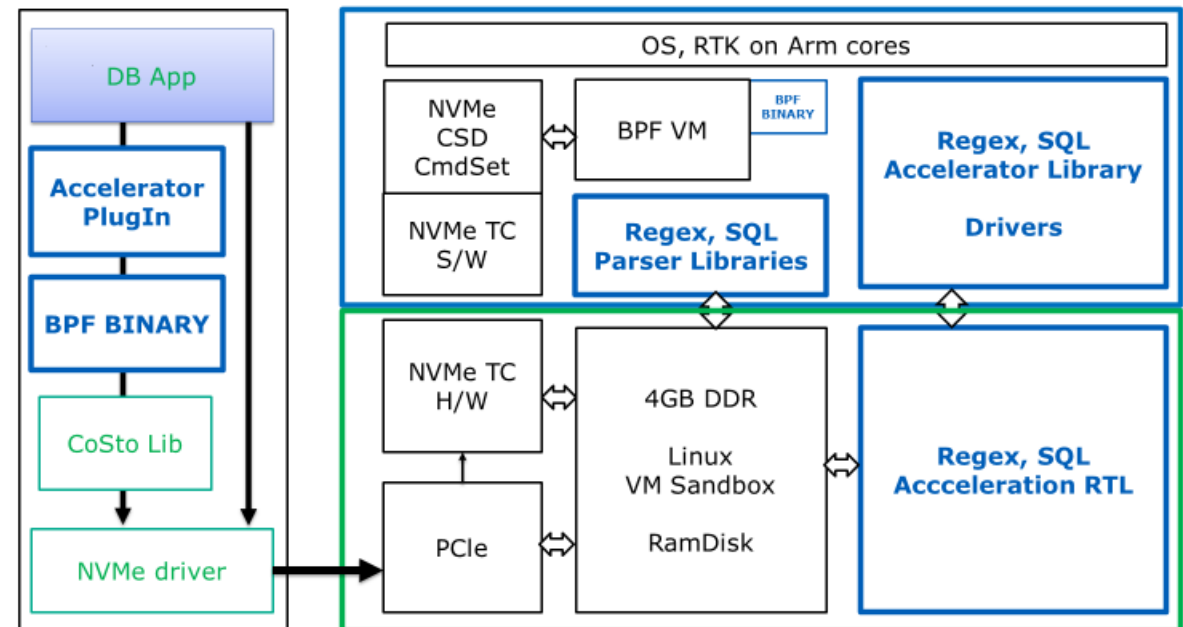
Resnet Inference Acceleration w/ TFLite + VTA



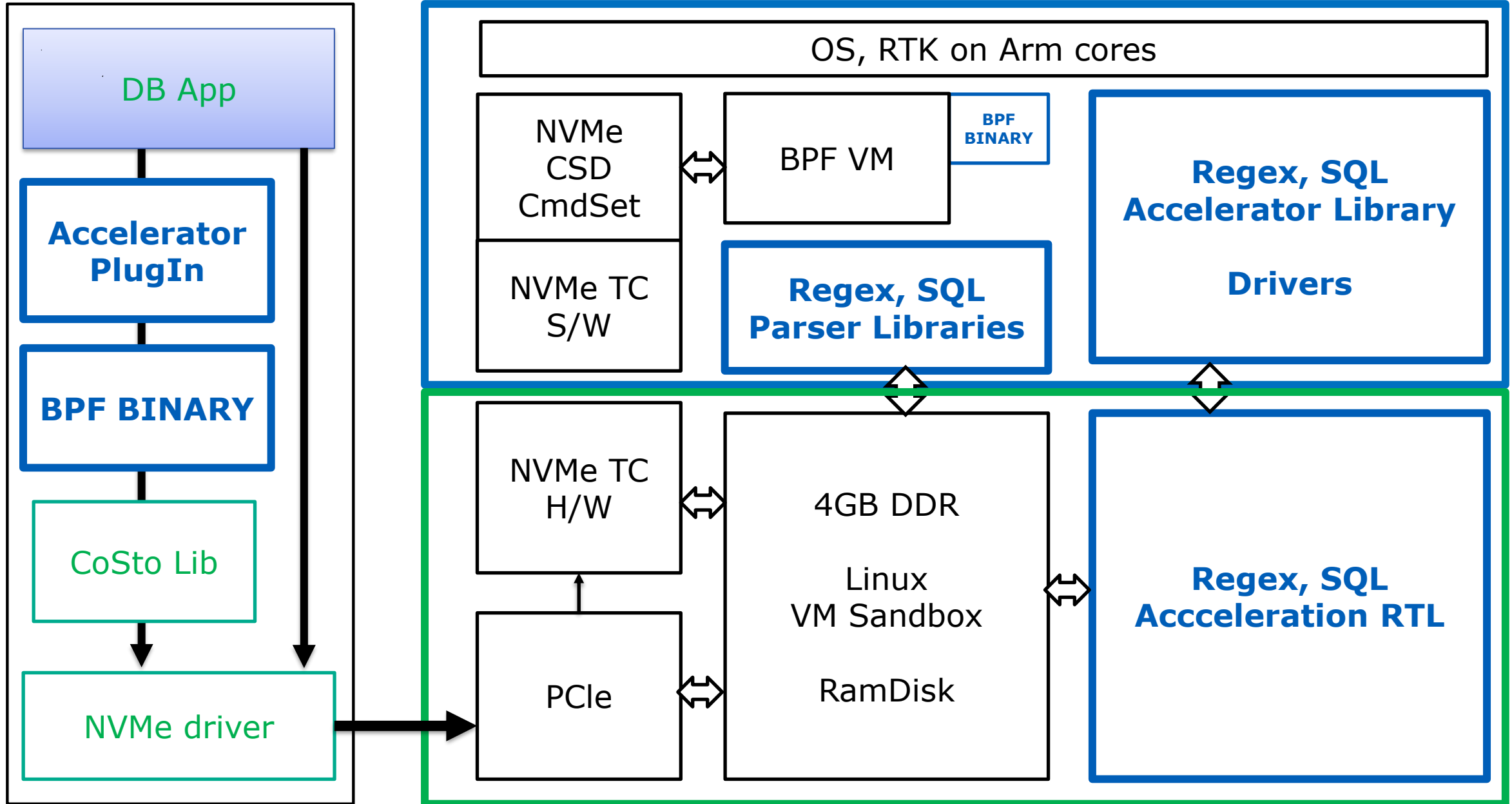
DB Search Offload and Acceleration

- Western Digital Computational Storage Platform enables study of variety of targets and approaches
 - Target DB Application: PGSQL, SPARK, MySQL, ...
 - CSD or CSP : DRAM as Ramdisk or CPM/CLM buffers
 - Complete DB Search Offload to Device or only partial Assist in the Device, rest in Host S/W
 - Full featured HW accelerator or assists in HW with full featured firmware for offload
 - Partition work between HW, FW and Host Application SW
 - Features to target specific DB App or multiple DB applications
- Proposal 1: CSP Device
 - Host copies DB Files into RAMDisk
 - Search Assist in H/W
 - Full Search in F/W
 - Plugin Validates results in Host S/W
- Proposal 2: CSD Emulation
 - RAMDisk mounted by Host as DataStore
 - Full SQL implementation in H/W
 - Plugin Validates results in Host S/W
- Architecture and Design in progress

DB Search Offload and Acceleration



DB Search Offload and Acceleration



DB Search Offload and Acceleration w/ P2P DMA

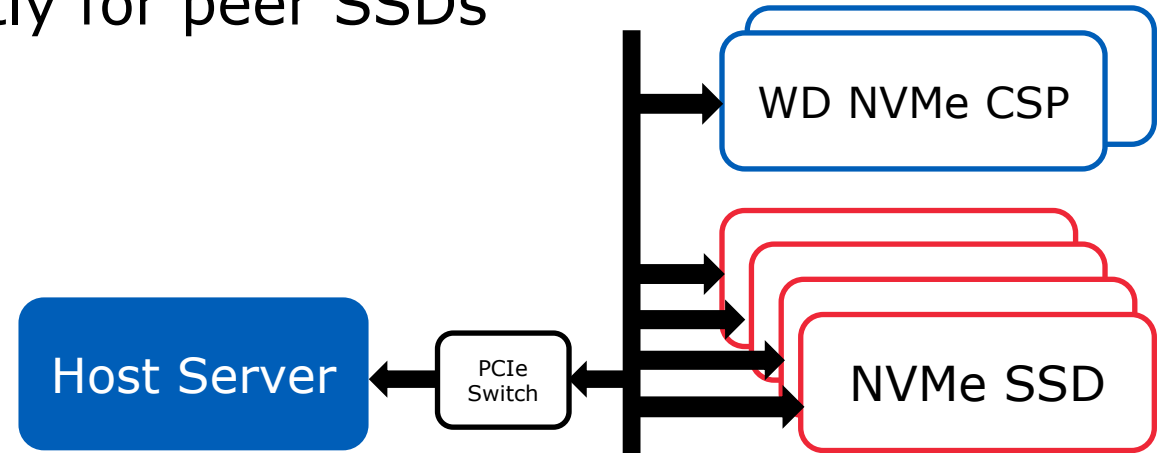
- Use PCIe P2P DMA to serve data directly for peer SSDs

- CSPs can service multiple SSDs
- Avoids need for bounce buffers in Host DRAM
- Reduces DRAM traffic at the Host
- Reduces PCIe traffic to the Host
- No special features required in the SSDs

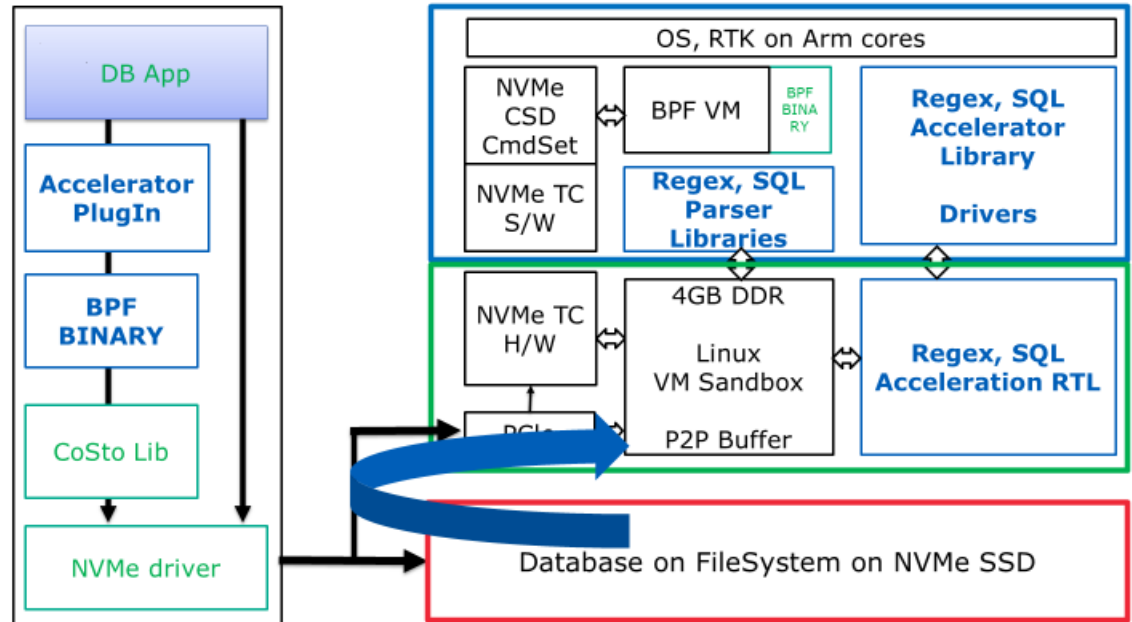
- Host CPU orchestrates data exchanges

- Host manages the file system on SSDs
- Host issues Read Cmds to the SSDs
- SSDs transfer directly into P2P Buffers on CSP

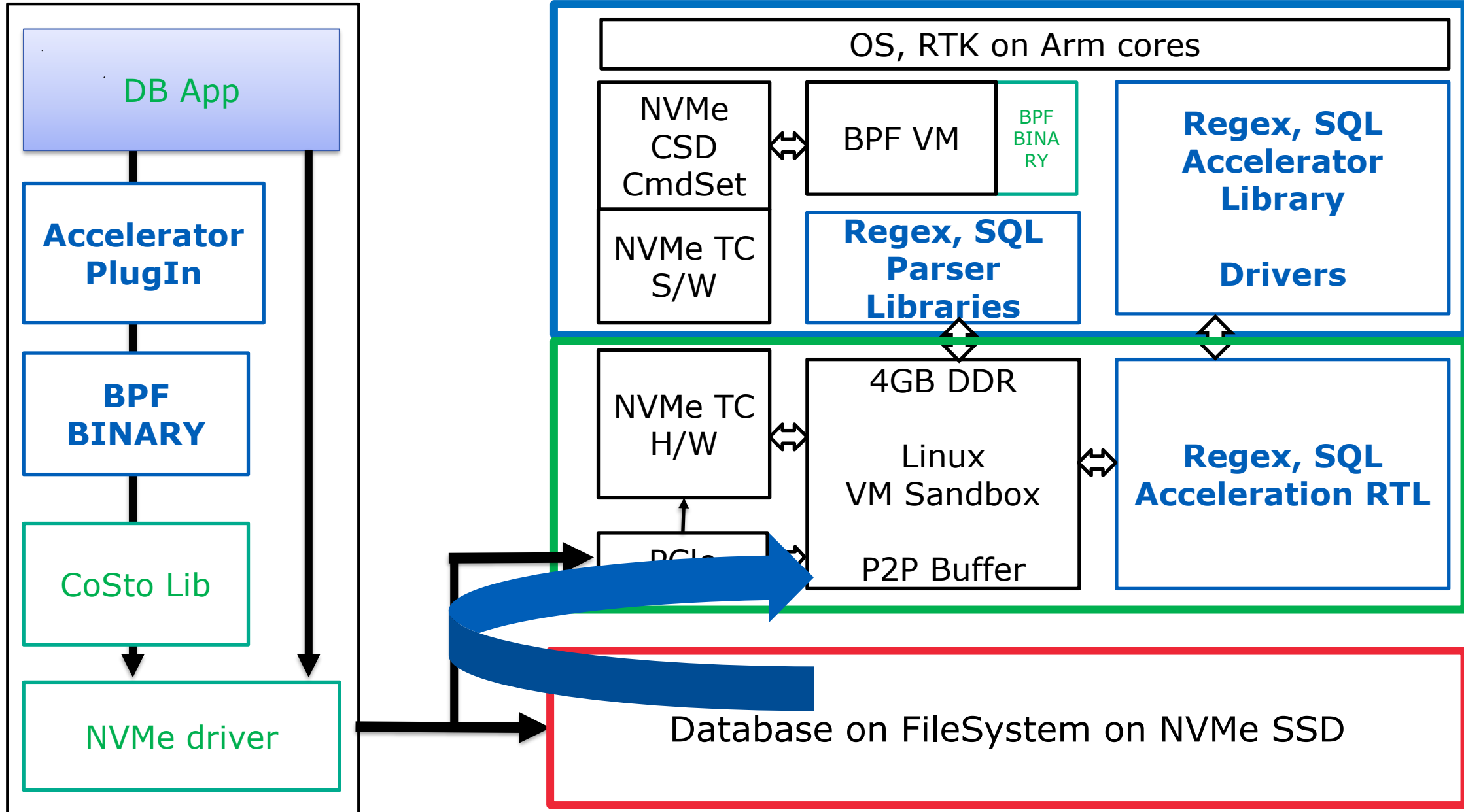
- Architecture and Design in progress



DB Search Offload and Acceleration w/ P2P DMA



DB Search Offload and Acceleration w/ P2P DMA



Conclusion

- Western Digital Compute Accelerator is a versatile device in a useful formfactor
 - Solutions available for Video Transcoding and AI/ML Inference Acceleration
 - FPGA based design allows for multitude of potential applications.
- Introduced Western Digital Computational Storage Accelerator Platform
 - NVMe + eBPF to offload CPU workloads, disaggregate from Compute Servers
 - P2P DMA to Storage devices to further reduce traffic on the host
 - FPGA and CPU resources available to implement custom accelerators
- Example designs underway
 - AI/ML Inference offload using TFLite with VTA for acceleration
 - DB offload and acceleration for Search operations
- Contact for more information
 - Anand.Kulkarni@wdc.com



Western Digital®

Western Digital and the Western Digital logo are registered trademarks or trademarks of Western Digital Corporation or its affiliates in the US and/or other countries. Linux® is the registered trademark of Linus Torvalds in the U.S. and other countries. The NVMe word mark is a trademark of NVM Express, Inc. All other marks are the property of their respective owners.