Western Digital®

Video Transcoding Acceleration in the Datacenter

Anand Kulkarni

Western Digital Research

Aug 27, 2021

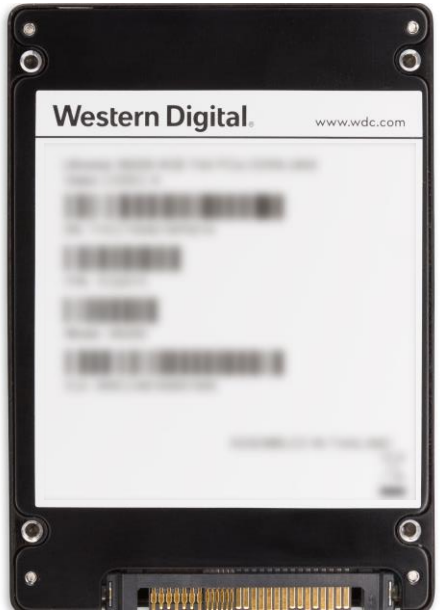
Outline

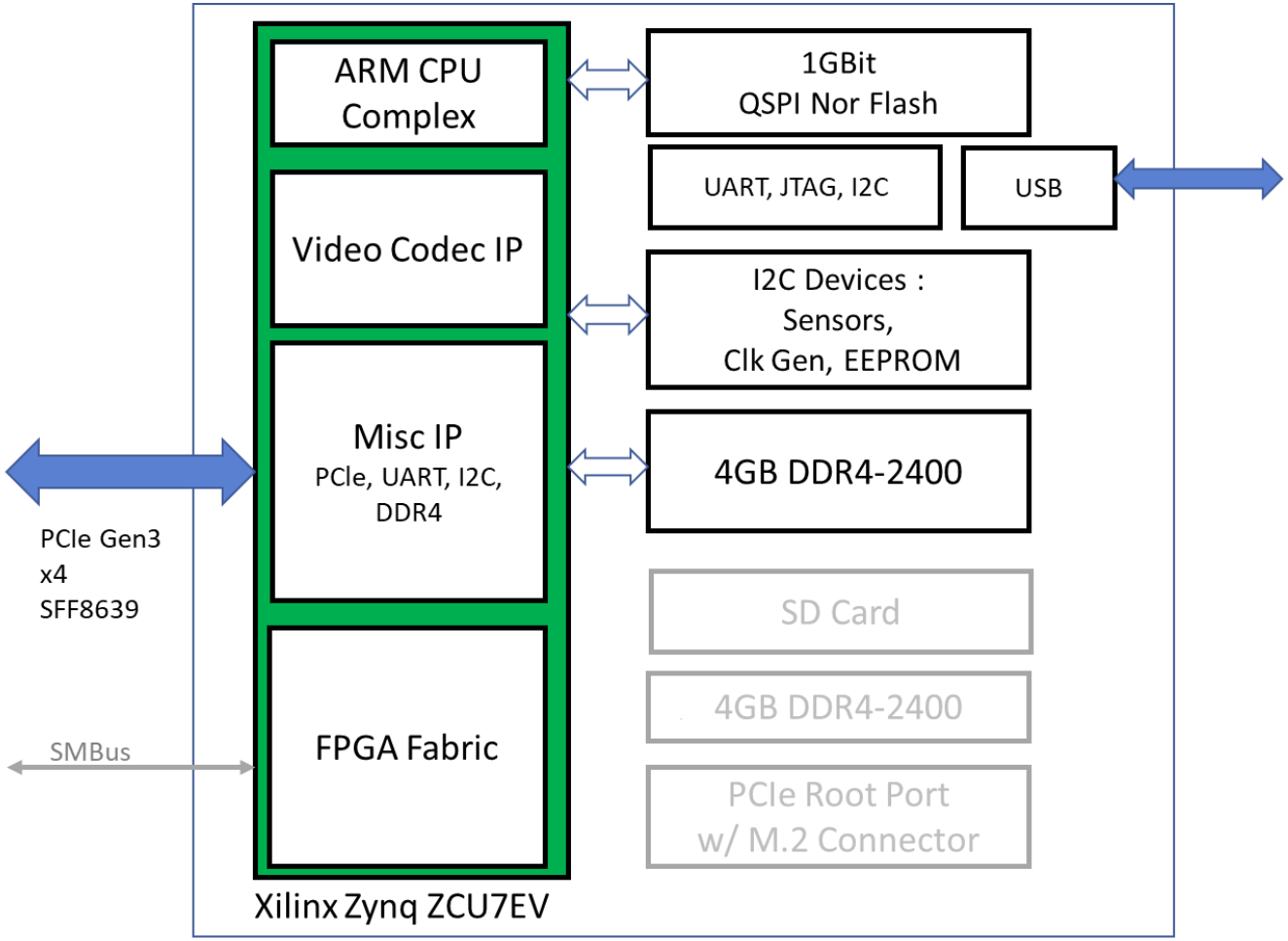
- Introduction to Western Digital Accelerator Platform
- Target Use cases
  - Video Transcoding
  - Machine Learning
  - Computational Storage
- Conclusion

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





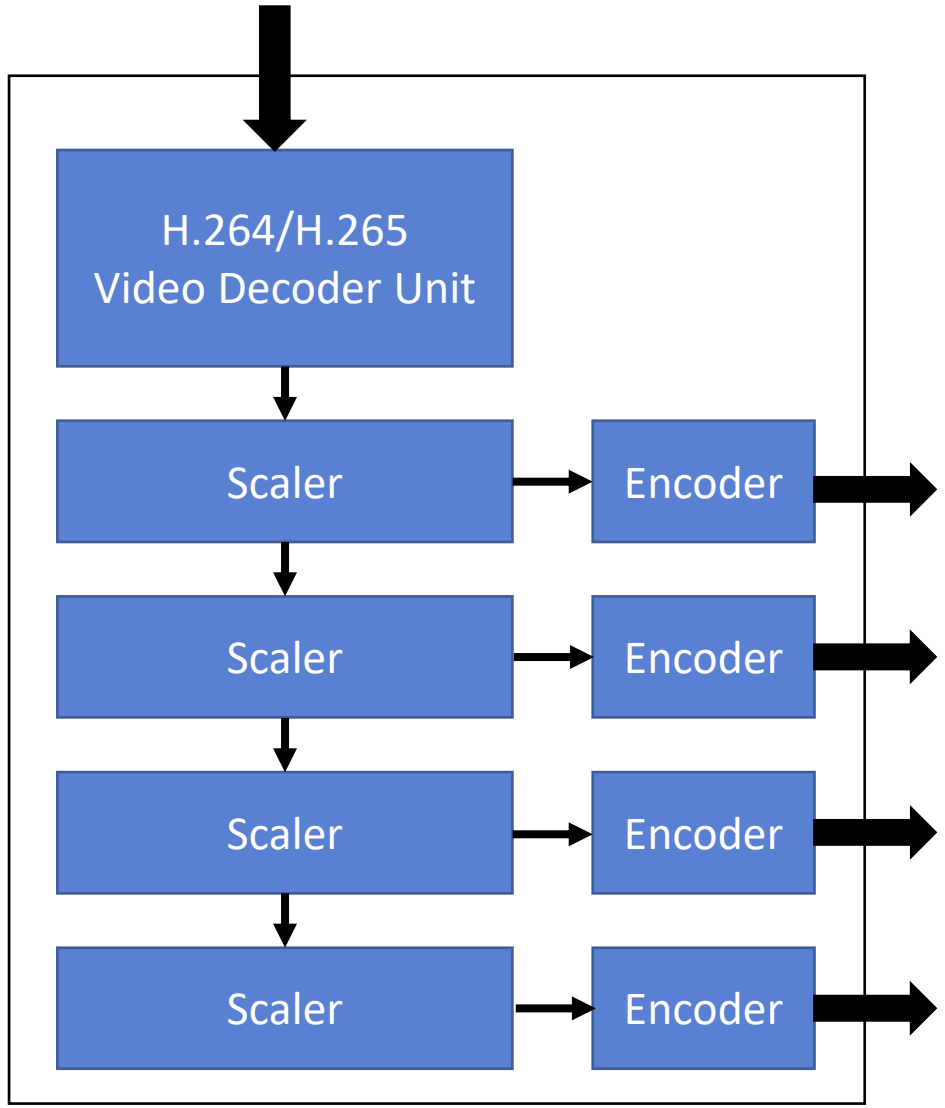
Western Digital Compute Accelerator Platform

- Versatile & Scalable
- Data center ready

Use case	Market
Video transcoding (H264/H265) XILINX	HD/UHD video streaming VoD, Sports, Gaming
AI-Inference: image/video Mipsology	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

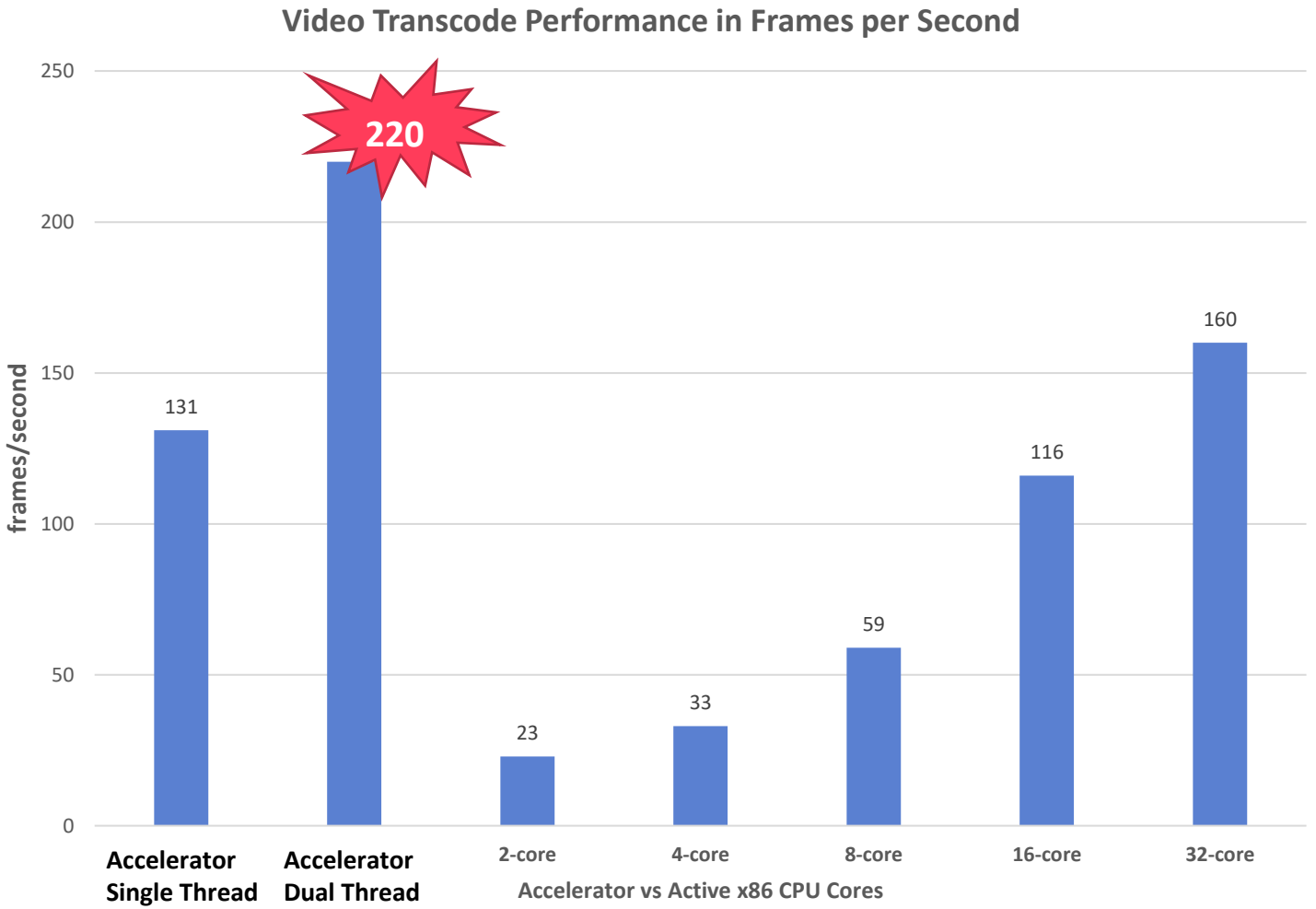
Use case: XILINX Video Transcode Accelerator

- 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

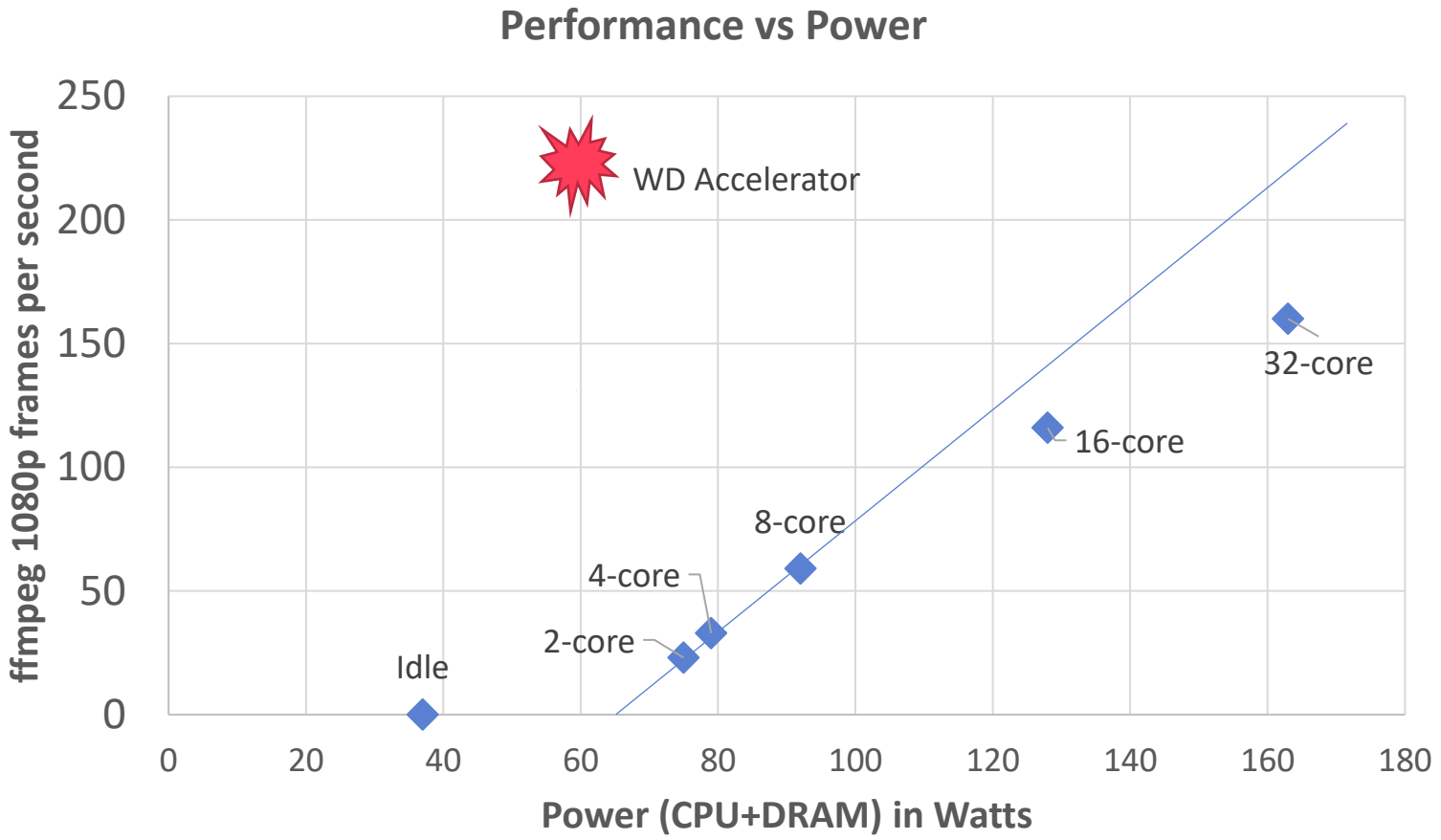
- Western Digital Accelerator beats performance of 32 x86 CPU cores
- Effective Compute Offload
  - 100x fewer instructions executed
  - 10x to 20x fewer Stall cycles on x86 CPU
  - 50x to 100x reduction in DRAM traffic
- Scalable performance
  - Multiple accelerators per Server



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

Performance Power ratio

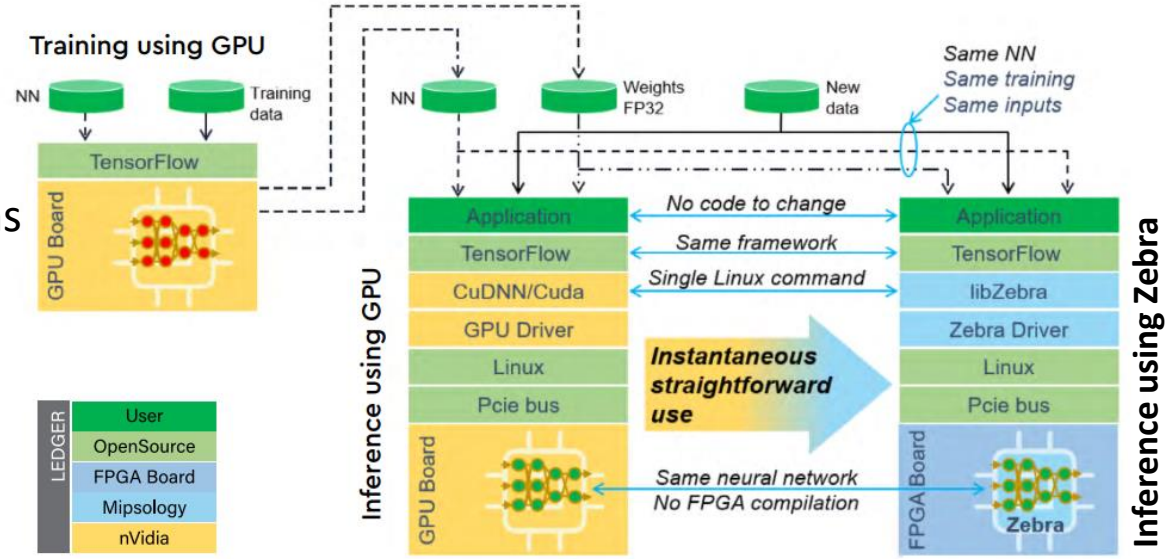
- Western Digital Accelerator
  - Less than 20W at peak performance
  - CPU close to Idle
- 32-core x86 CPU
  - Almost 100W higher for similar or lower performance
  - Each x86 CPU core adds ~3W of power consumption



- 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%

Target Use: ML Inference Accelerator

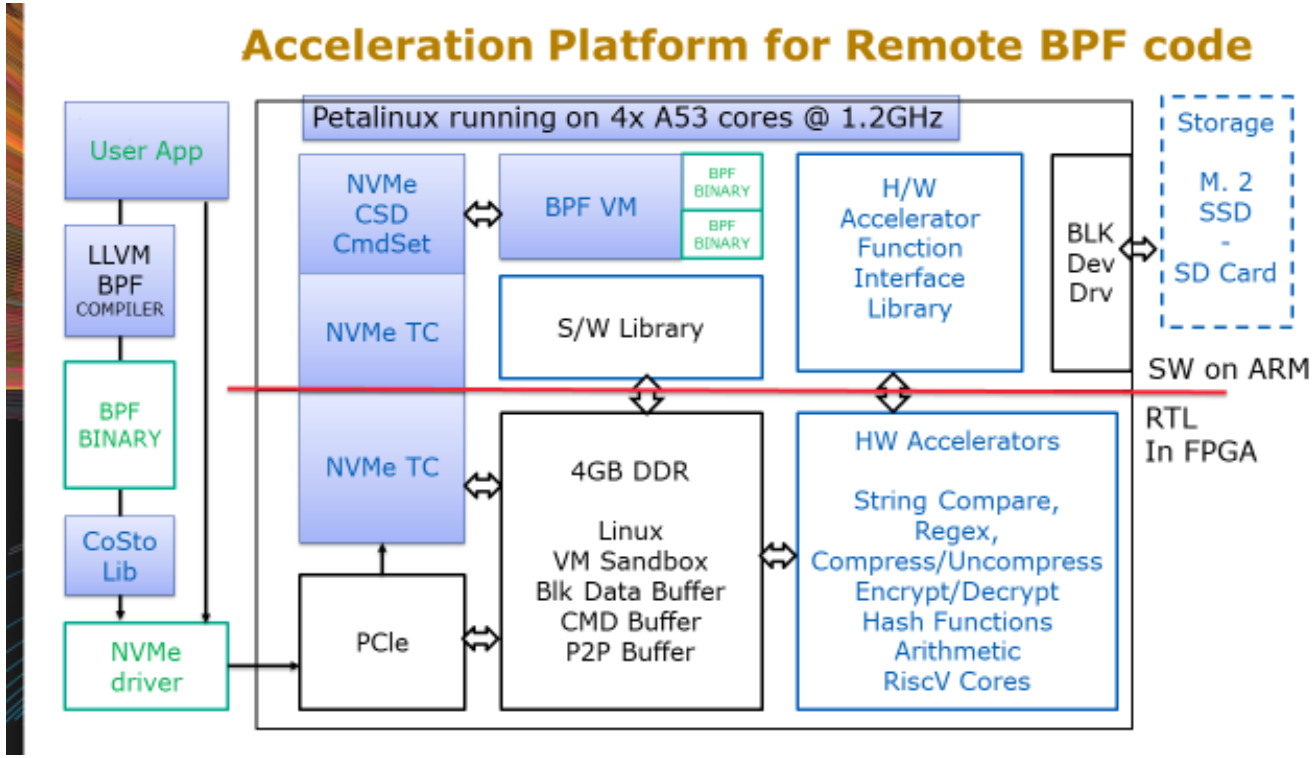
- Supported Frameworks
  - TensorFlow, PyTorch, MXNet, Caffe
- Run any pre-trained CNN models without modifications
  - Automatic quantization, No pruning required
  - Up to a Billion weights, Million Layers
  - Concurrently run two separate networks
- Supports wide variety of Networks
  - ResNet, Yolo, 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

Target Use: Computational Storage Platform

- Computational Storage over NVMe using eBPF
  - NVMe TPAR 4091 is a proposed standard to enable Compute Offload to Storage devices
    - 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)
    - HW Kernels via Custom RTL in FPGA
  - NVMe Target Controller split between FPGA and ARM Cores
    - Maximize resources for Accelerator functions
  - Support P2P DMA over PCIe
    - Reduction in PCIe/Network/DRAM traffic
  - Flexible boundary between HW & SW
    - Implement features in SW
    - Migrate features to HW over time
  - Applications development underway
    - Image/Video Analytics
    - Database Acceleration
    - Genomics
  - Actively under development



Conclusion

- Western Digital Compute Accelerator is a versatile device in a useful formfactor

- Video Transcoding is a compelling application for the Compute Accelerator Device
  - Significantly better performance
  - Lower total power and cost
  - Reduction in DRAM traffic
  - Free up CPU for other work
- Additional Compute Acceleration features/applications are within reach
  - ML Inference solution available now
  - ML Training solution in the works
  - NVMe + eBPF to offload CPU workloads, disaggregate from Compute Servers
  - P2P DMA to Storage devices to further reduce traffic on the host
- Contact for more information

Ted.Marena@wdc.com
Anand.Kulkarni@wdc.com

Western Digital®