Reconfigurable Vector Floating Point Accelerator on FPGAs

Himanshu Kumar Rai[1]
Sasi Snigdha Yadavalli[1]
Aishwarya Sridhar[2]
Nanditha Rao[3]

1: International Institute of Information Technology Bangalore, India (IIIT-B)
2: Infineon Technologies Semiconductors India Pvt. Ltd.
3: IBM India Pvt. Ltd - Research work done when Nanditha Rao was affiliated to IIIT-B

62 THE CHIPS TO SYSTEMS CONFERENCE

SPONSORED BY CEDA sigda

Motivation

Challenges in Traditional Floating Point Implementations: Requires multiple bit-widths. Fixed 32/64-bit designs lack precision for emerging AI/ML, signal processing and scientific simulations.

Vectorization and Parallel Processing: Existing architectures struggle with varying vector lengths, causing hardware underutilization and bottlenecks, while vectorization improves computational efficiency.

Reconfigurability for Diverse Applications (E.g, AI Inference favours low precision, training and simulations need higher precision, Signal Processing needs mixed precision and flexible vector length): Reconfigurable FPUs that support various precision formats and adjust exponent/mantissa widths, vector lengths can enhance performance.

Inspiration from Prior Research: Published research on VLIW instructions, Conventional SIMD, Packed SIMD, and transprecision vector units (all non reconfigurable) highlight the need for combining vectorisation and reconfigurability for efficient design.

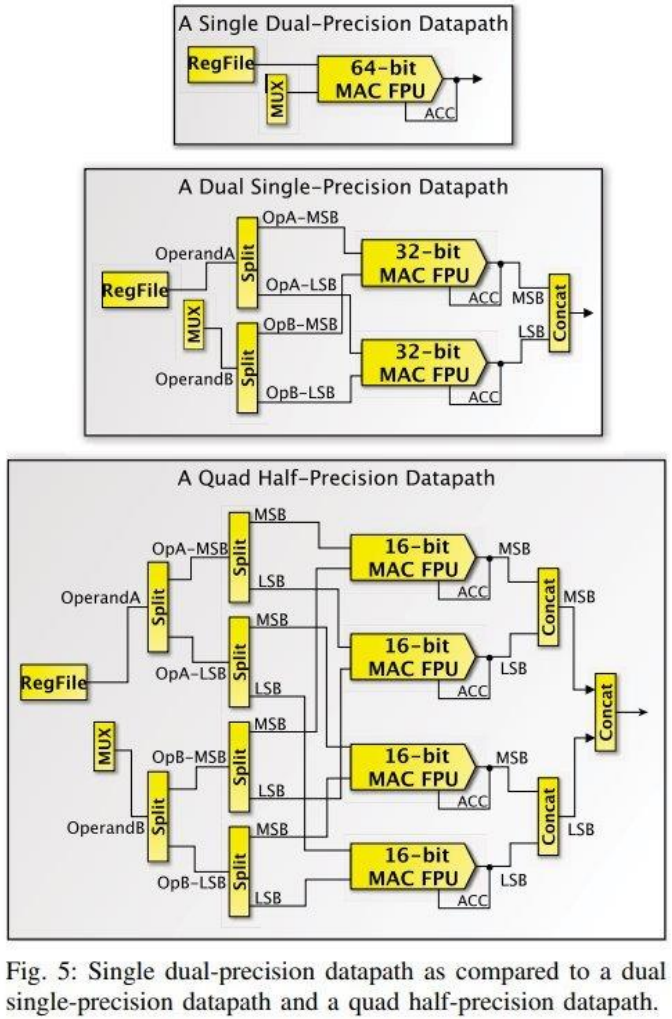


Fig. 5: Single dual-precision datapath as compared to a dual single-precision datapath and a quad half-precision datapath.

Figure 1 and 2 : Packed SIMD datapath and register packing from Abdelhamid, R.B. and Yamaguchi, Y., 2022, December. Packed SIMD Vectorization of the DRAGON2-CB. In *2022 IEEE 15th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC)* (pp. 85-92). IEEE.

Proposed Methodology

SPONSORED BY CEDA sigda

Architectural Novelties

Pipelined Vector Floating Point Unit (FPU)

Reconfigurable Precision: Supports IEEE 754 Single (SP-32), Double (DP-64), Tensorfloat (TF-32), Bfloat (BF-16), Quarter (QP-8), and custom precision formats.

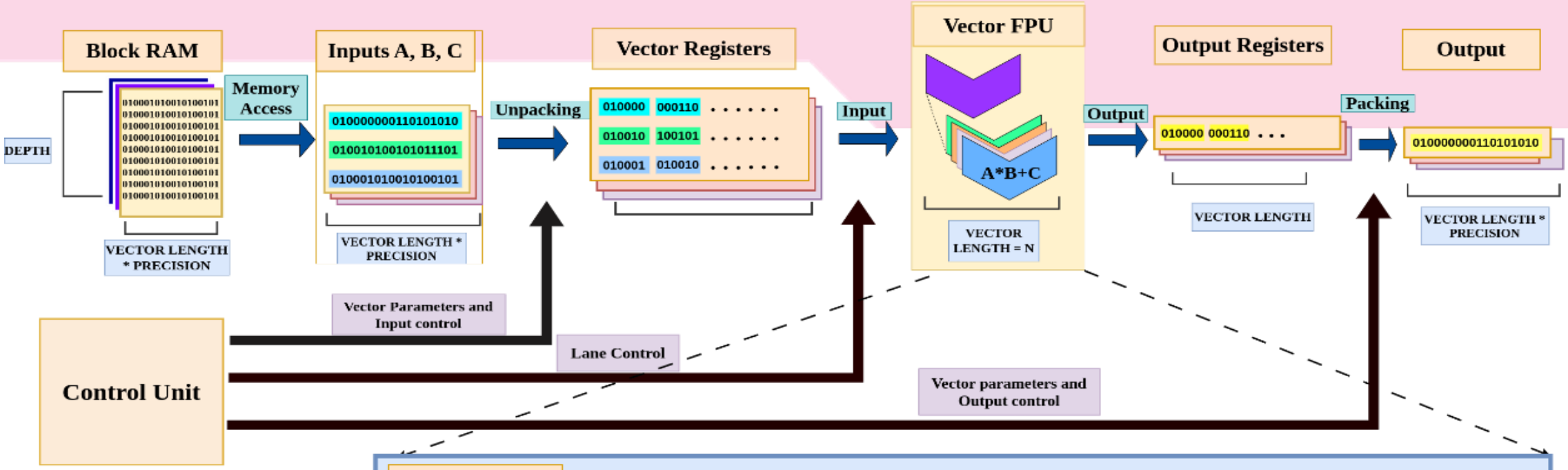
Reconfigurable Lane Adjustment: Lane widths based on requirement

Vectorization via Packing and Unpacking

Adaptability for Multiple Precision Formats and Custom Precision Flexibility

Zynq FPGA platform: Enables real-time reconfiguration for efficient vectorization and hardware acceleration.

Vector FPU Architecture – FPGA Based IP



- Reconfigurable FP Hardware IP
- Reconfigurable Lane Width
- Custom Floating Point Formats
- Scalable Parallel Processing
- Pipelined Architecture
- FPGA based prototype of the IP

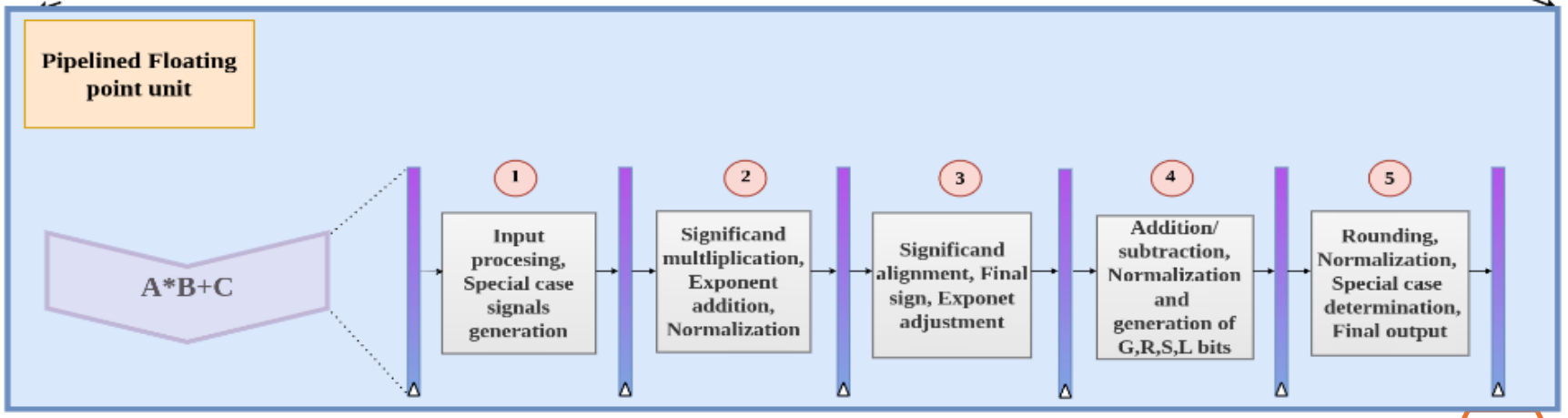


Figure 4: Block Diagram of proposed architecture

Acceleration: Tiled FPGA FP Vector MAC Units

- 1 tile with vector lane = 6, can support loop unroll factor 6.
- 4 such tiles can compute 24 FP multiply-add operations in parallel which provides high level of parallelism. Image shows FPGA implemented layout with multiple vector MAC FP units
- Our design provides reconfigurability of width of lane according to unroll factor.

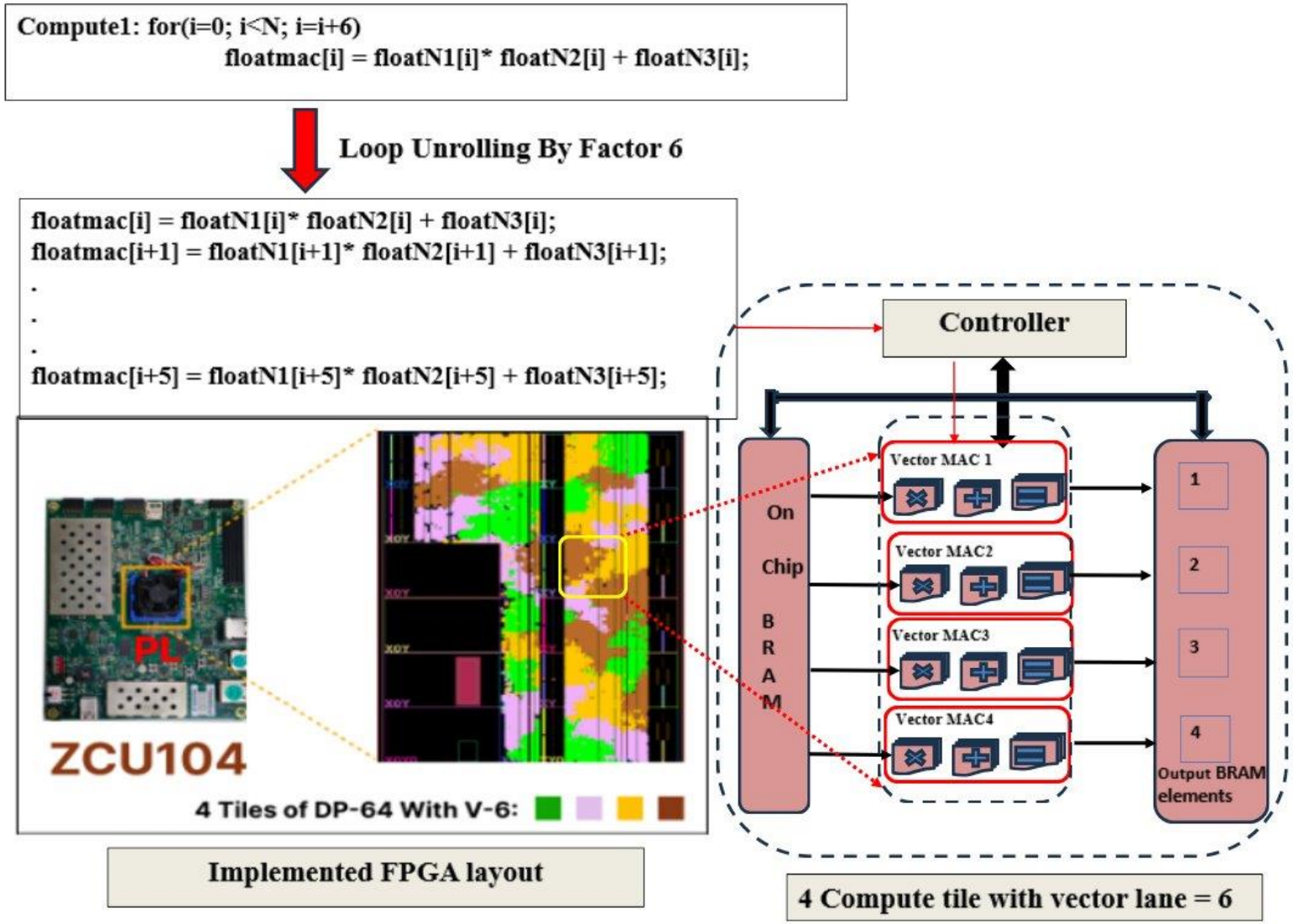
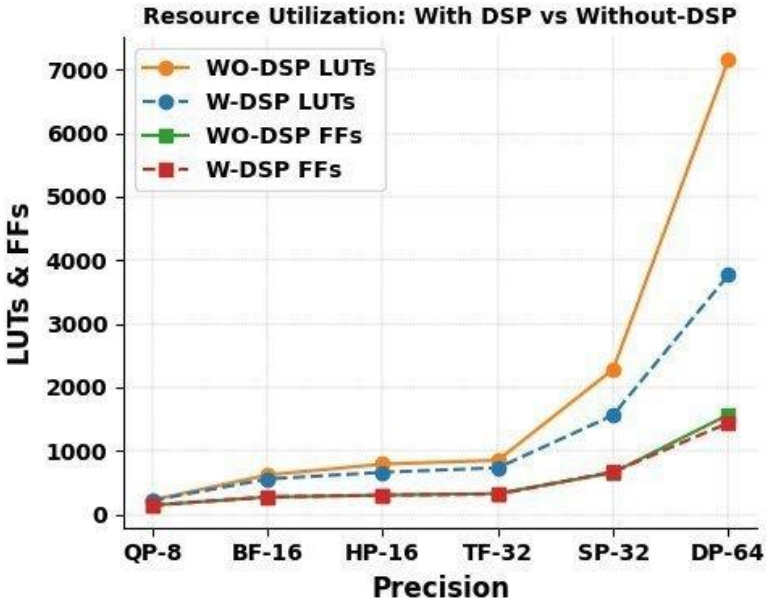


Figure 5: Acceleration by lane width reconfiguration

Results, Applications and Future Work

SPONSORED BY CEDA SIGDA

FPGA Based Results on Xilinx ZCU104



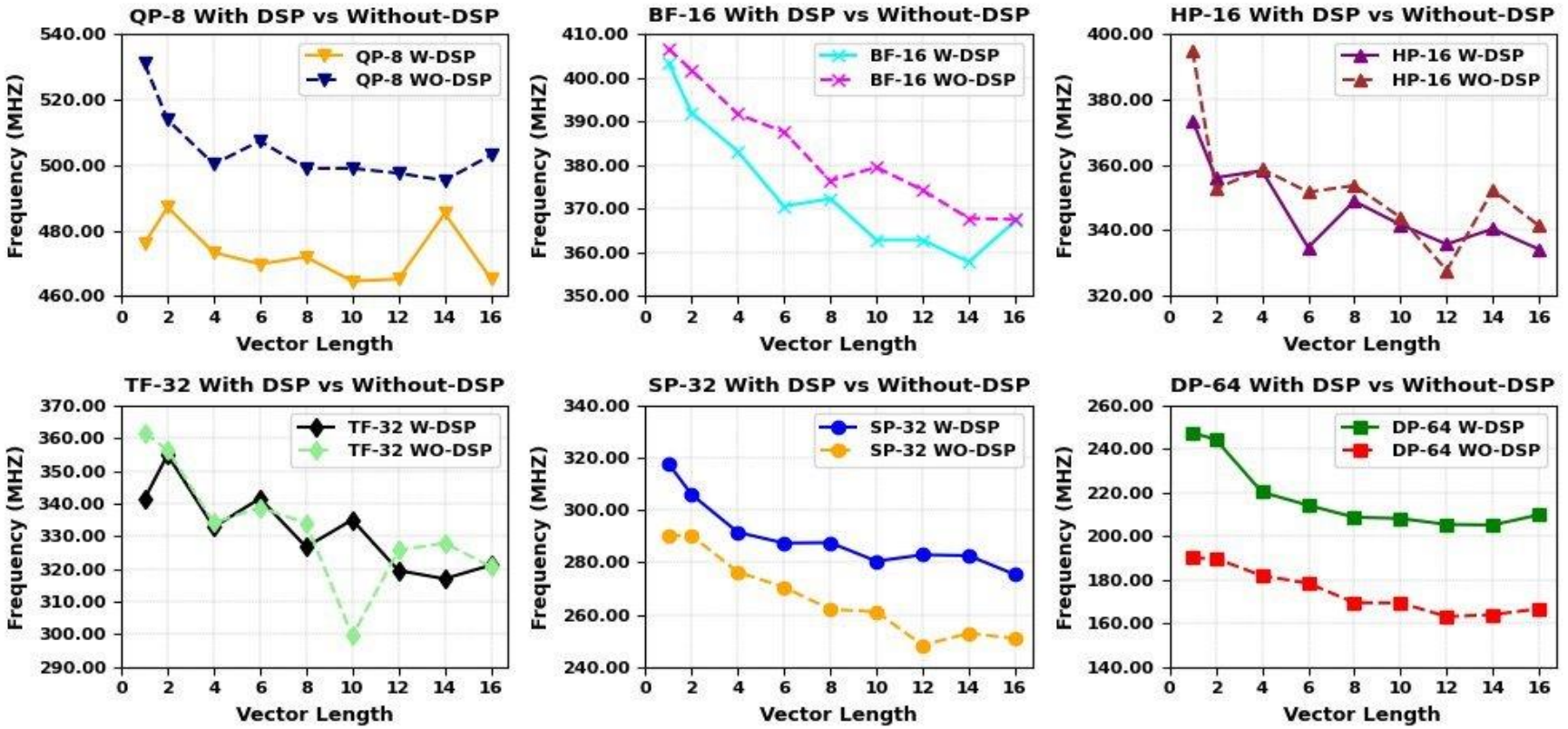


Figure 6 and 7: Resource utilization and frequency observations

- **DSP based designs:** Resource efficient for all precisions, energy efficient for higher precisions
- **SP-32 + DSP:** 31.6% fewer LUTs, 9.3% higher frequency, 24.7% lower power
- **DP-64 + DSP:** 47% fewer LUTs, 30% higher frequency, 41% lower power
- **Recommendation:** Use DSPs for higher precisions; LUT-based designs suit lower precisions

FPGA Based Results on Xilinx ZCU104

- Highest Energy efficiency of 34.26 GLOP/S-Watt achieved by QP-8 without DSP.
- DSP based designs are more energy efficient for higher precisions as they consume less power

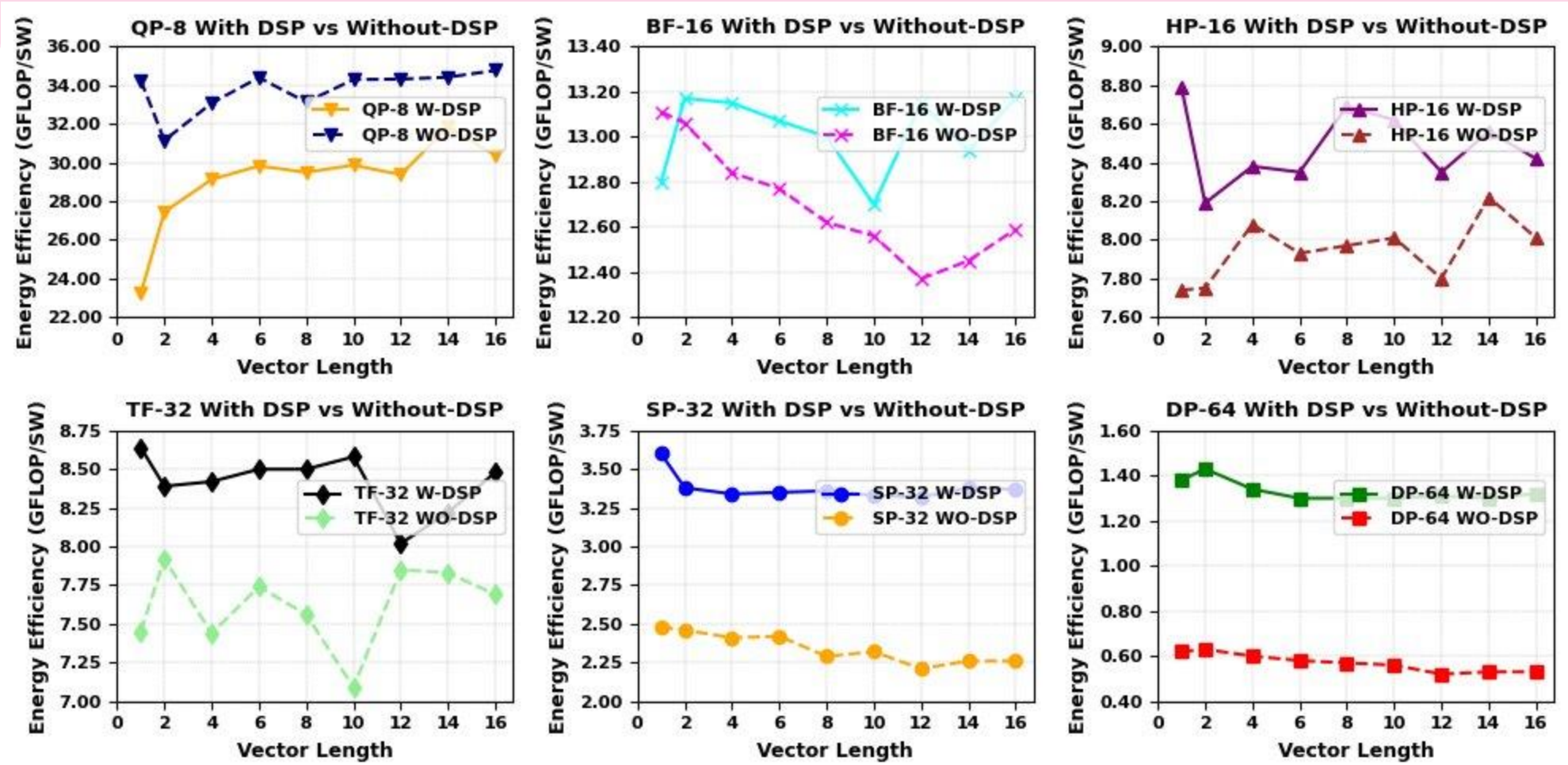


Figure 8: Energy Efficiency observations

Applications and Future Work

Future Work:

- Integration with Neural Processing Units (NPUs)
- AI-Driven Optimization
- Collaborative Architectures for AI Training
- Evaluation of Alternative Memory Architectures
- Interoperability with Other Hardware Accelerators

Applications

- High-Performance Computing (HPC)
- AI Model Compression and Acceleration
- Workload based Resource Allocation
- Data-Driven Applications
- Edge Computing

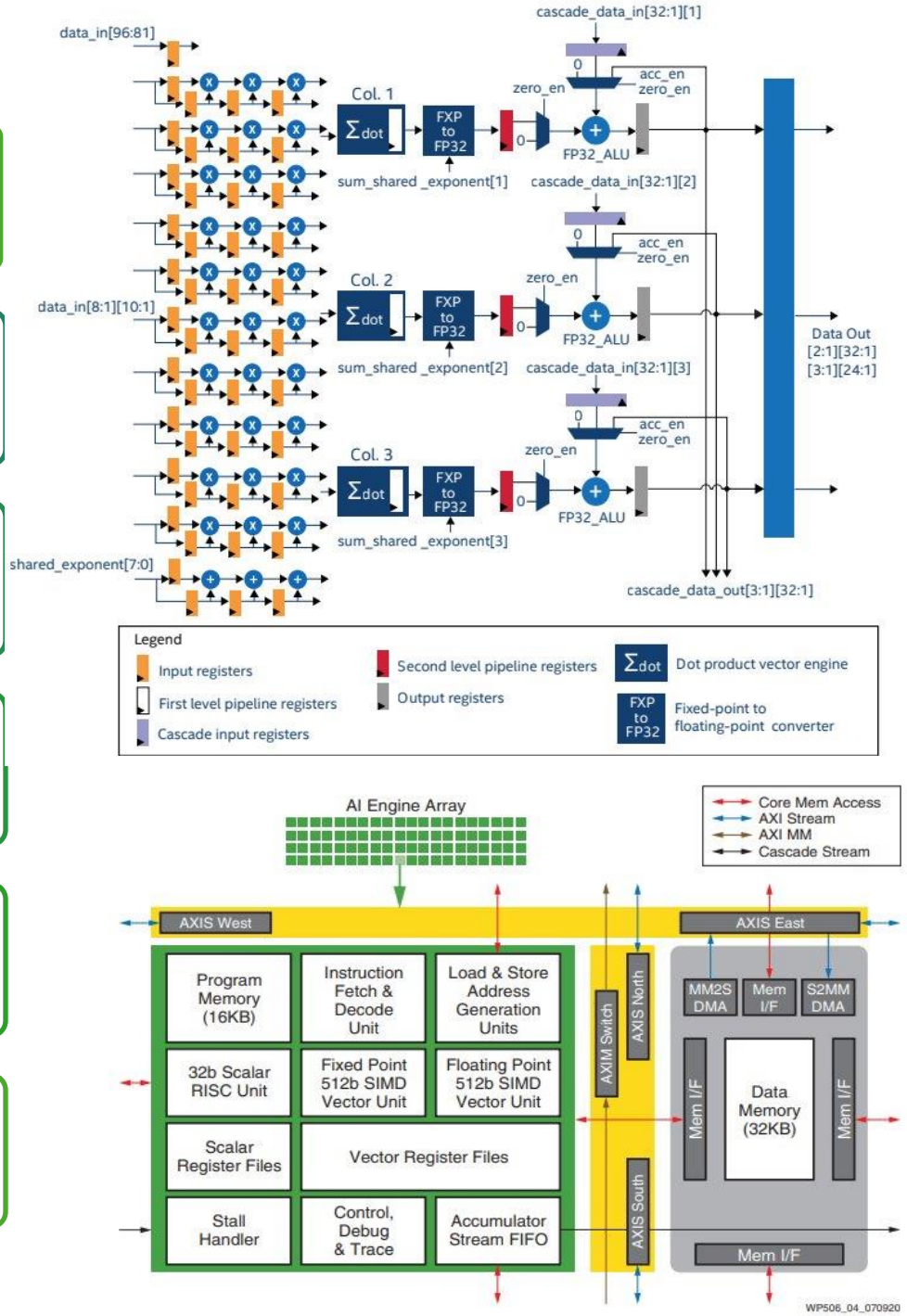


Figure 9: Dot Product Engine in ML (https://www.bittware.com/resources/fpga-neural-networks/)

Figure 10: AI engine array, including scalar/vector units and floating-point vector units (https://www.bittware.com/resources/fpga-neural-networks/)

Summary

Architectural Novelty: Pipelined Vector FPU:

- Fully pipelined
- Packing and unpacking for vectorization
- Supports multiple IEEE 754 precision and custom formats.

Reconfigurable Lane Width Adjustment:

- Lanes adjust based on the requirement
- Allows flexible unrolling factors and parallelism (up to chosen FP multiply-add operations)

Performance Gains on ZCU104 FPGA: DSP boosts throughput by 8% for SP32 and 19.8% for DP64 across lane configurations.

Energy Efficiency:

- Highest energy efficiency (34.26 GLOP/SW) achieved by QP-8 without DSP
- DSP-based designs are more power-efficient overall (better for higher precisions).