Novel Memory-Efficient Computer Architecture Integration in RISC-V with CXL

Presenter: Sarmad Adeel
Authors: Peter Marosan and Sarmad Adeel, *Blueshift Memory*
Aysa Davey, *ITDev*

Agenda

- Introduction
- Problem Statement
- Architecture
- FPGA Accelerator Card
- Design Performance
- AI/ML & Image Recognition
- Applications
- Conclusion

Introduction

- **Blueshift Memory's** technology optimizes the CPU-memory architecture for more efficient handling of large data sets and time-critical data (e.g. AI and Big Data)
- **Cambridge Architecture™** – the next generation Stored Program Machine
  - Designed to overcome the constraints of the '*Von Neumann bottleneck*'
  - Up to **1,000 times faster** memory access
  - Up to **50% reduction** in energy consumption
  - Gives protection from memory-centric cybersecurity attacks
- This project has been funded by the UK Government, with a $0.5 million Innovate UK Smart Grant

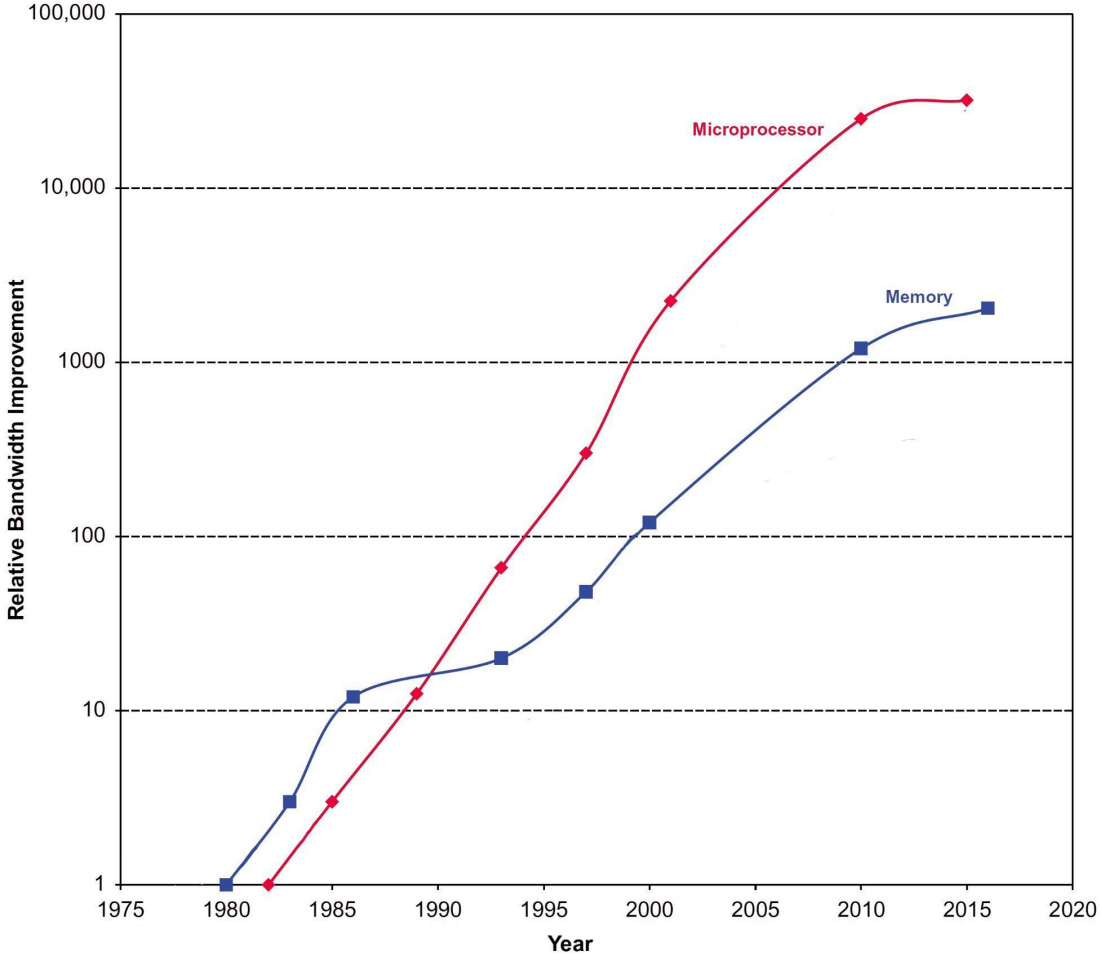
Problem Statement

Modern computing workloads:

- Exhibit high memory intensity
- Are increasingly structured

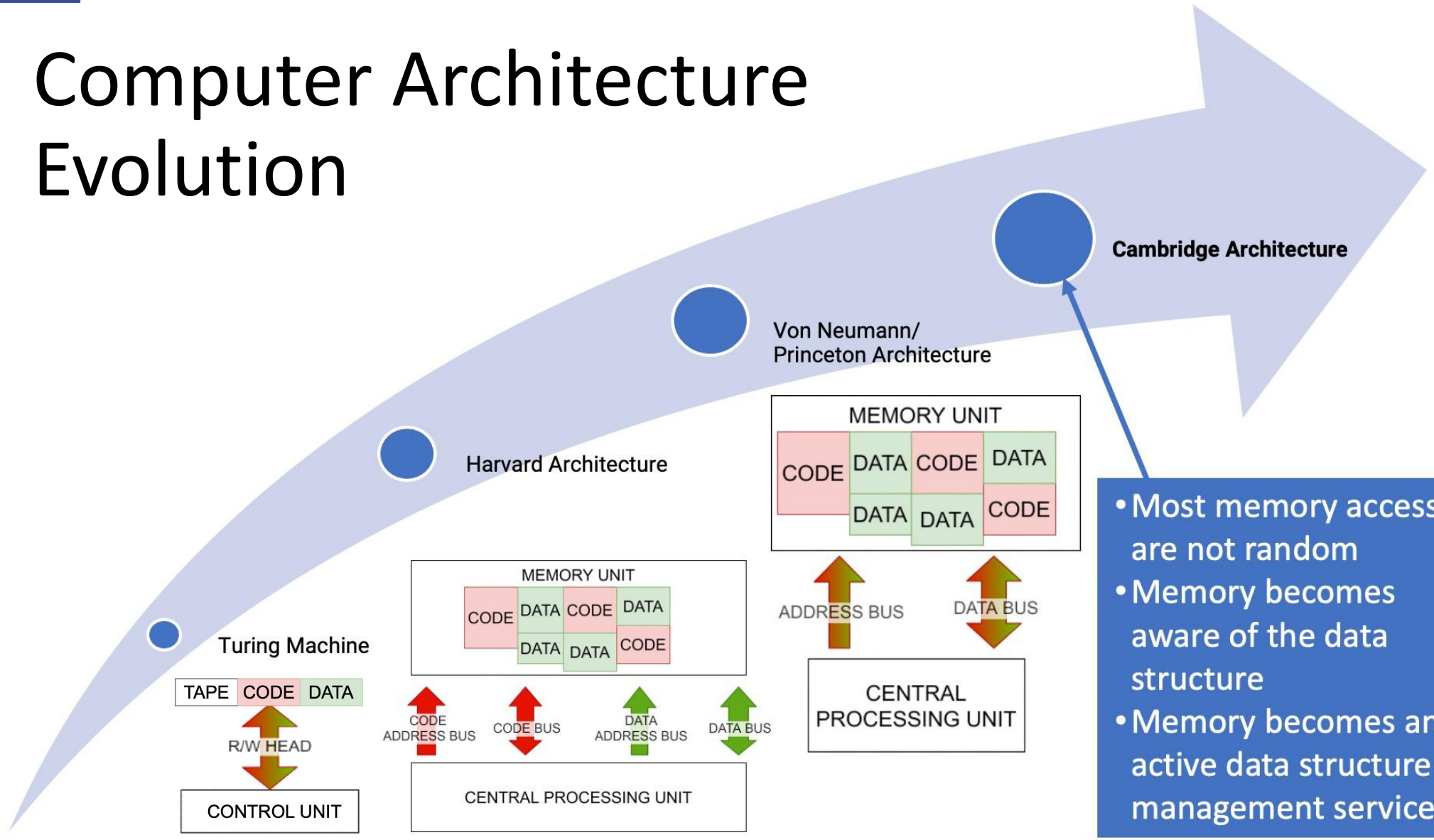
Therefore:

- Memory access is becoming the dominant consumer of energy
- Performance has hit a ‘Memory Wall’, inhibited by the *Von Neumann bottleneck*



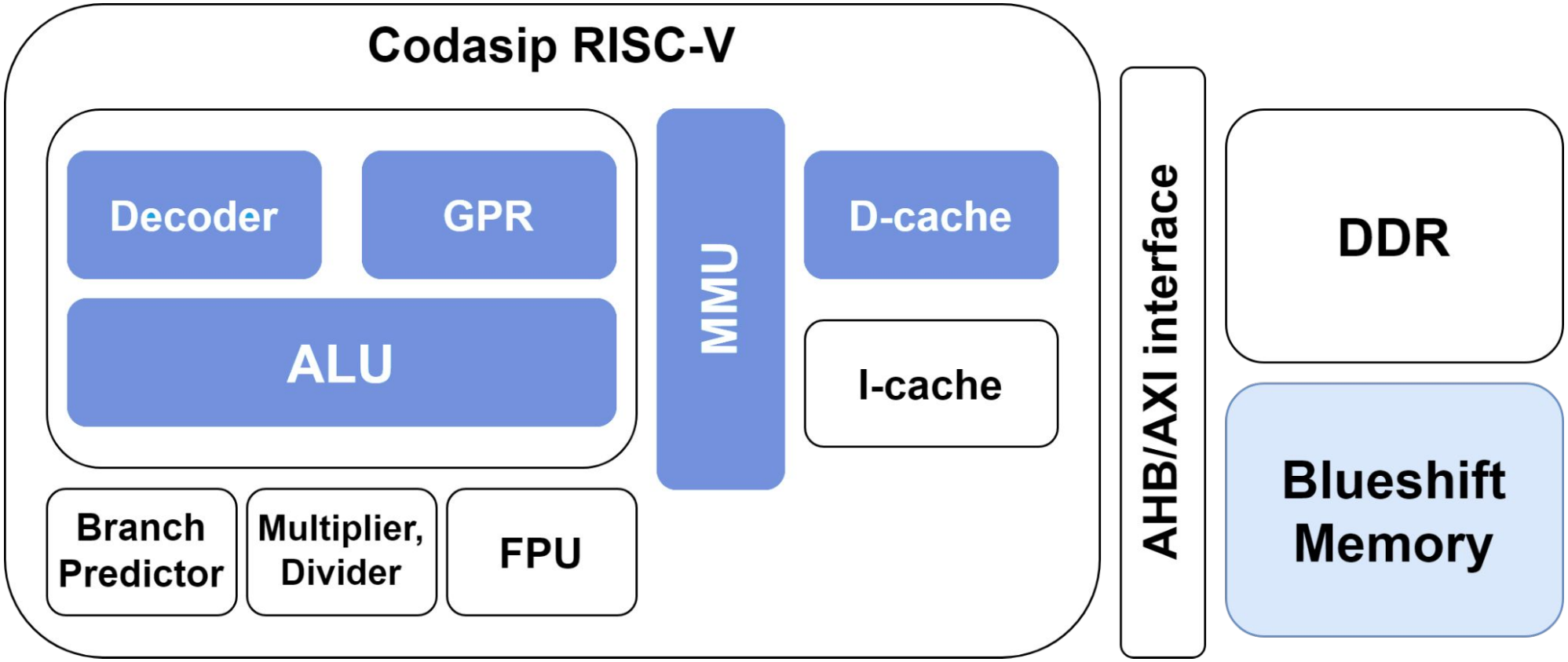
(Source: Computer Architecture: A Quantitative Approach, Hennessy, J.L. and Patterson, D.A., 6th Edition 2019)

Computer Architecture Evolution



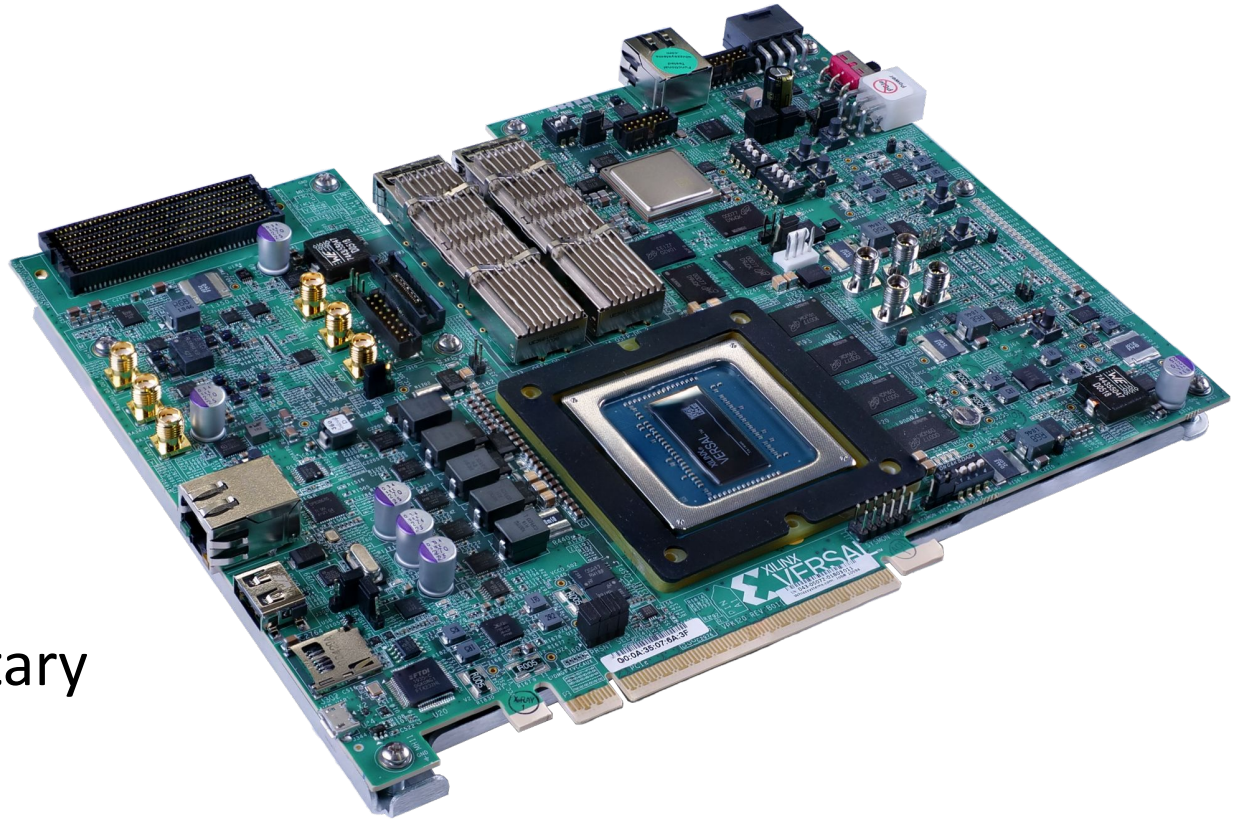
- Most memory accesses are not random
- Memory becomes aware of the data structure
- Memory becomes an active data structure management service

Architecture



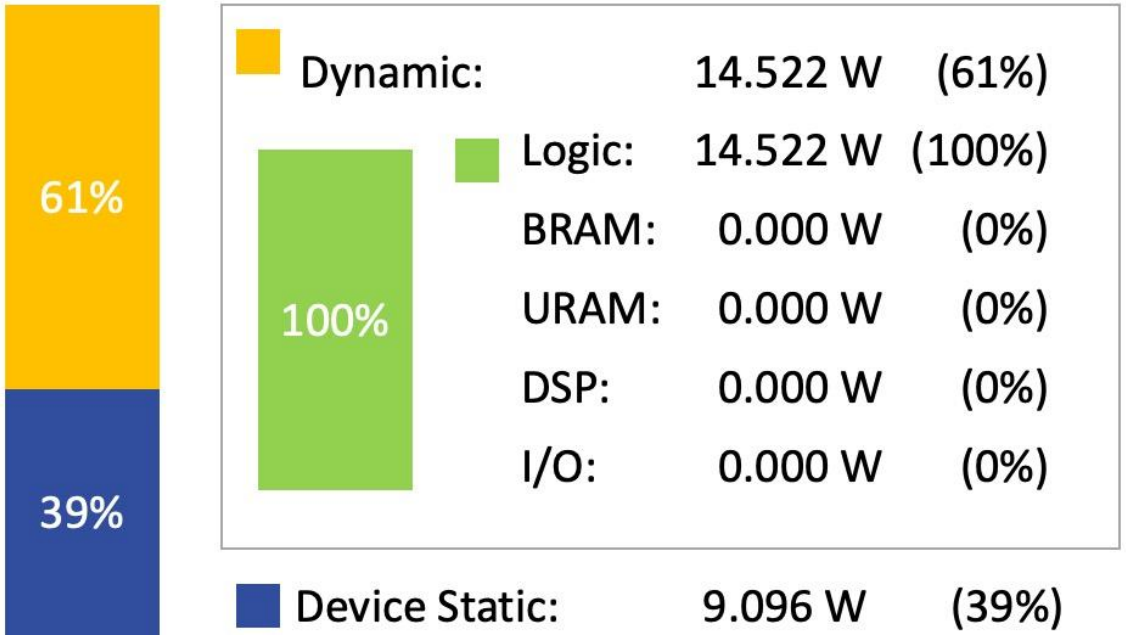
FPGA Accelerator Card

- PCIe® w/ DMA & CCIX (CPM5)
  2 x Gen5x8
- PCIe®
  2 x Gen5x4
- Can be connected to any host to increase memory speed and performance
- Flexible Blueshift Memory proprietary software and driver available



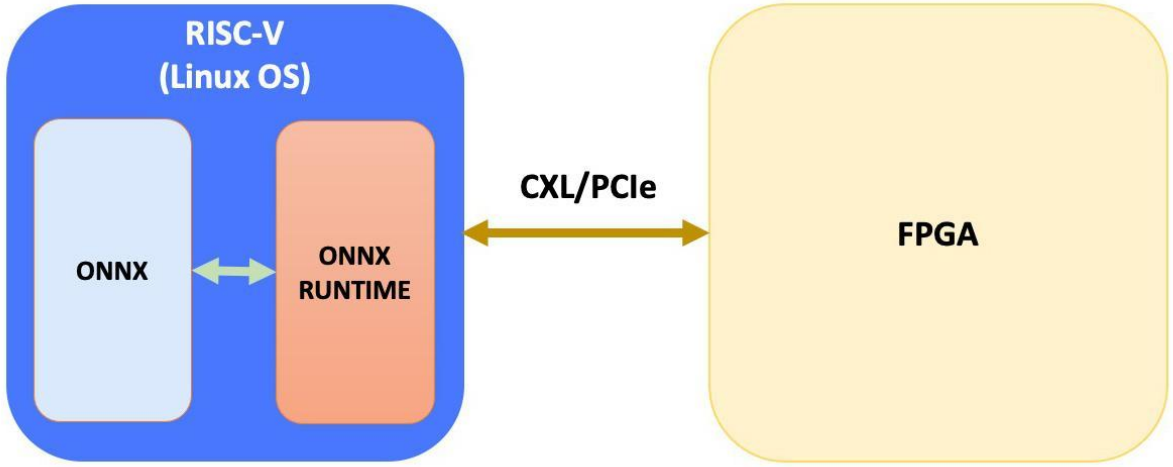
FPGA Performance

- FPGA Clock Frequency 250 MHz
- FPGA Power Consumption 14.5 W
- Linux-Capable SoC
- RV64-GC ISA based RISC-V



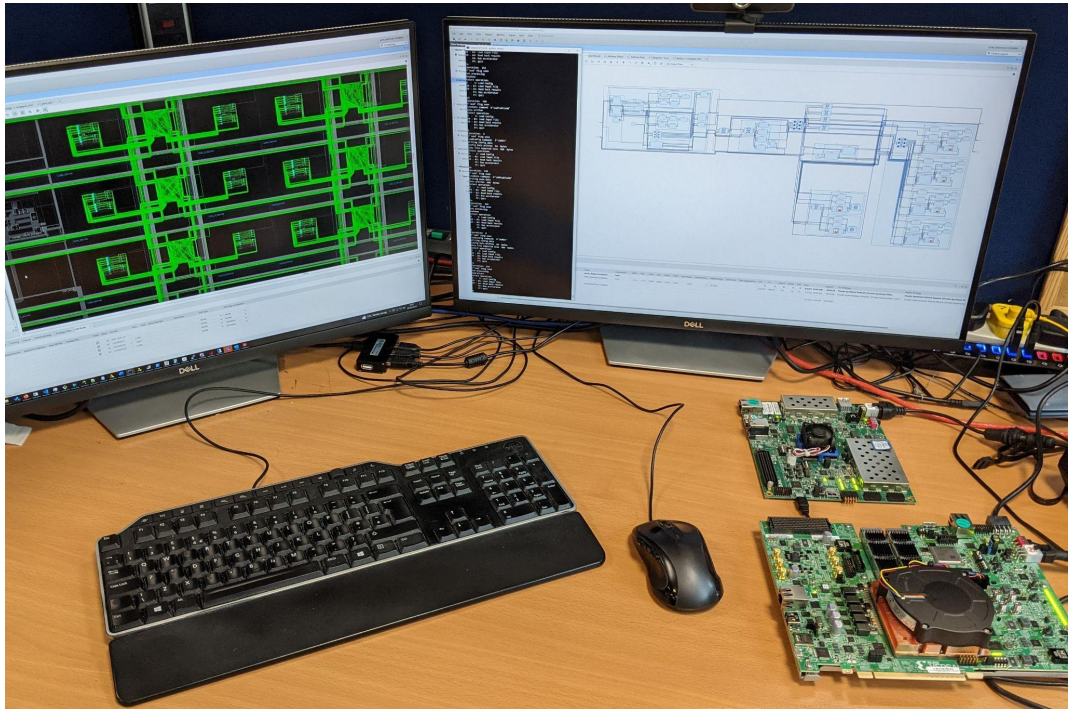
AI/ML & Image Recognition

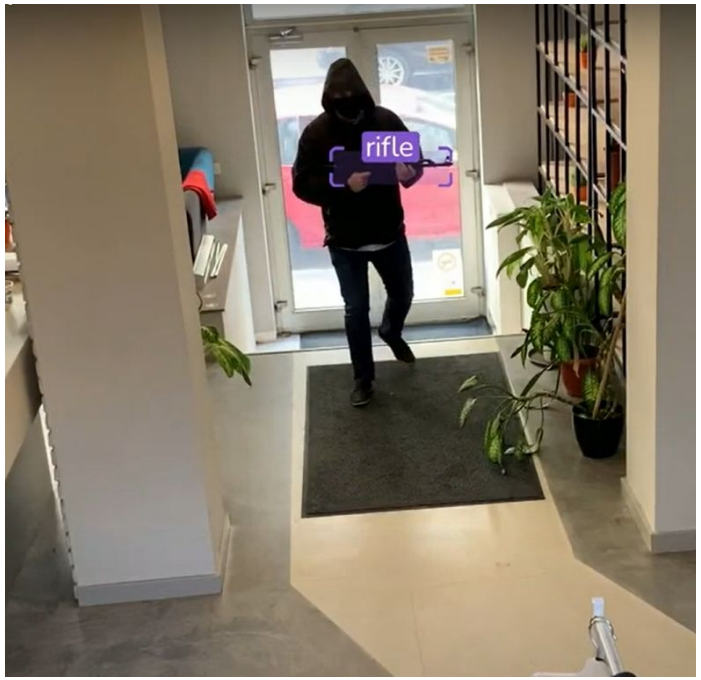
- Open Neural Network Exchange (ONNX)
- CXL and PCIe options
- Faster ML algorithm execution
- Dedicated Hardware Access
- Doubling the CPU performance by reducing the algorithm execution on the CPU

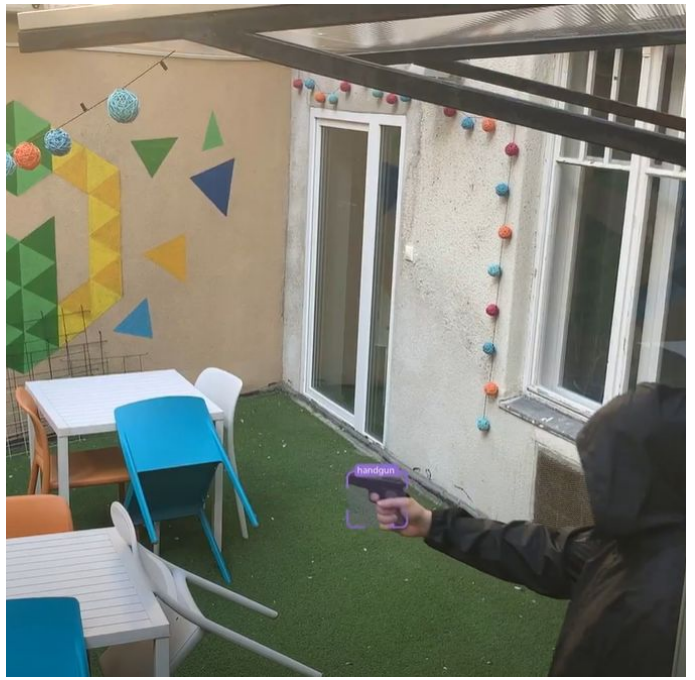


AI/ML & Image Recognition

- ONNX Runtime (ORT) runs on Linux (OS) used as frontend
- PCIe commands are used to communicate with the FPGA from ORT
- Pre-trained YOLO algorithm v5 is used for object detection
- Currently YOLO v5 is tested on the CPU and is working
- Convolution consumes a lot of CPU resources
- To speed up and enhance performance, convolution is implemented on the FPGA







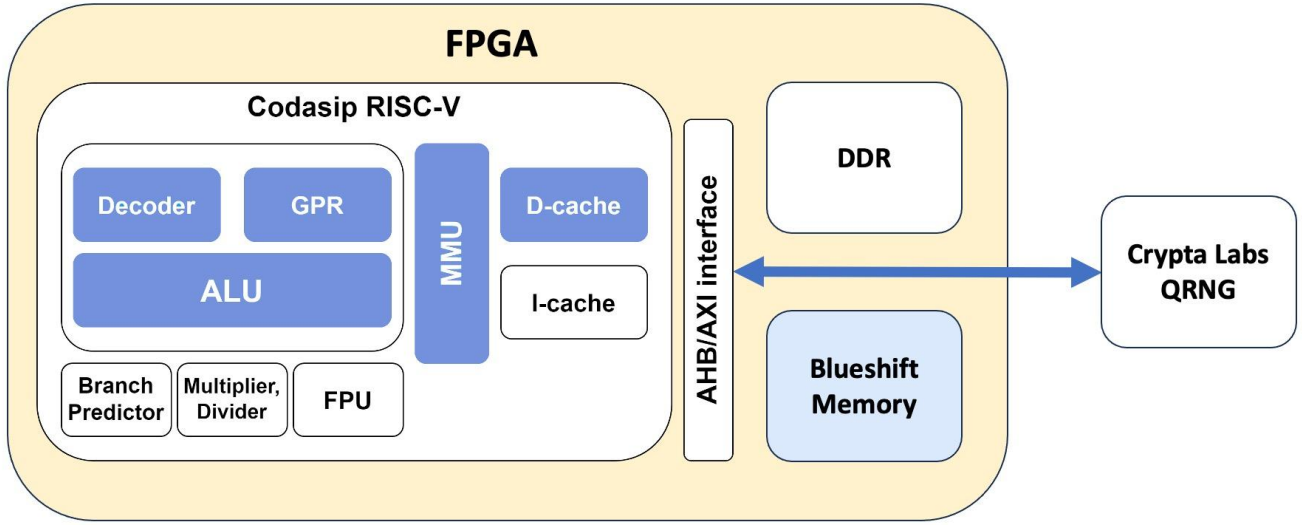
Design Performance

- Total 6 Accelerator Blocks
- 4 Blocks Work in Parallel
- The 1x1 matrix can run for a 20x20 image
- ~80 µs on hardware ~520 µs on the microblaze
- The 3x3 matrix can run for a 40x40 image
- ~192 µs on hardware ~1250 µs on the microblaze
- FPGA Frequency 266.5 MHz
- Additional small latency due to Buses, DDR and software overhead
- FPGA Power Consumption 7.6W

Size	Calculations (Clock Cycles)
160x160	30402
80x80	7682
40x40	1962
20x20	504

Cryptography and Security

- Cybersecurity Memory solution
- Quantum resilient encryption with NIST-compliant entropy health checks
- CXL and PCIe options
- High speed, performance and bandwidth with Quantum Security
- Working towards a higher level of integration



Conclusion

- Complete RISC-V SoC solution with next generation of stored program computer, and non-Von Neumann memory and data intelligence
- Especially well suited to Big Data and other data intensive computing
- High speed, performance and bandwidth with CXL and PCIe Gen 5
- Zero latency
- Reduces energy consumption by **30 - 50%**
- Performance Improvements **2x -1000x**
- **Reduces** the CPU code by providing data specific instructions
- **Doubles** the CPU performance by reducing the memory instructions load on the CPU

Thank You To Our Partners:

- Innovate UK: Smart Grant funding
- Codasip: RISC-V development
- Crypta Labs: Quantum Random Number Generator
- Codee: HPC compilers
- Recog.AI: AI computer vision
- Cambridge University: incubator programme

Our Team

Peter Marosan
CEO/CTO

Sarmad Adeel
Senior Embedded Design Engineer

Sarah Bayliss
Project Manager

And the rest of our development team

Aysa Davey
Contract FPGA Design Engineer

Helen Duncan
CMO

Guillaume d'Eyssautier
Chairman