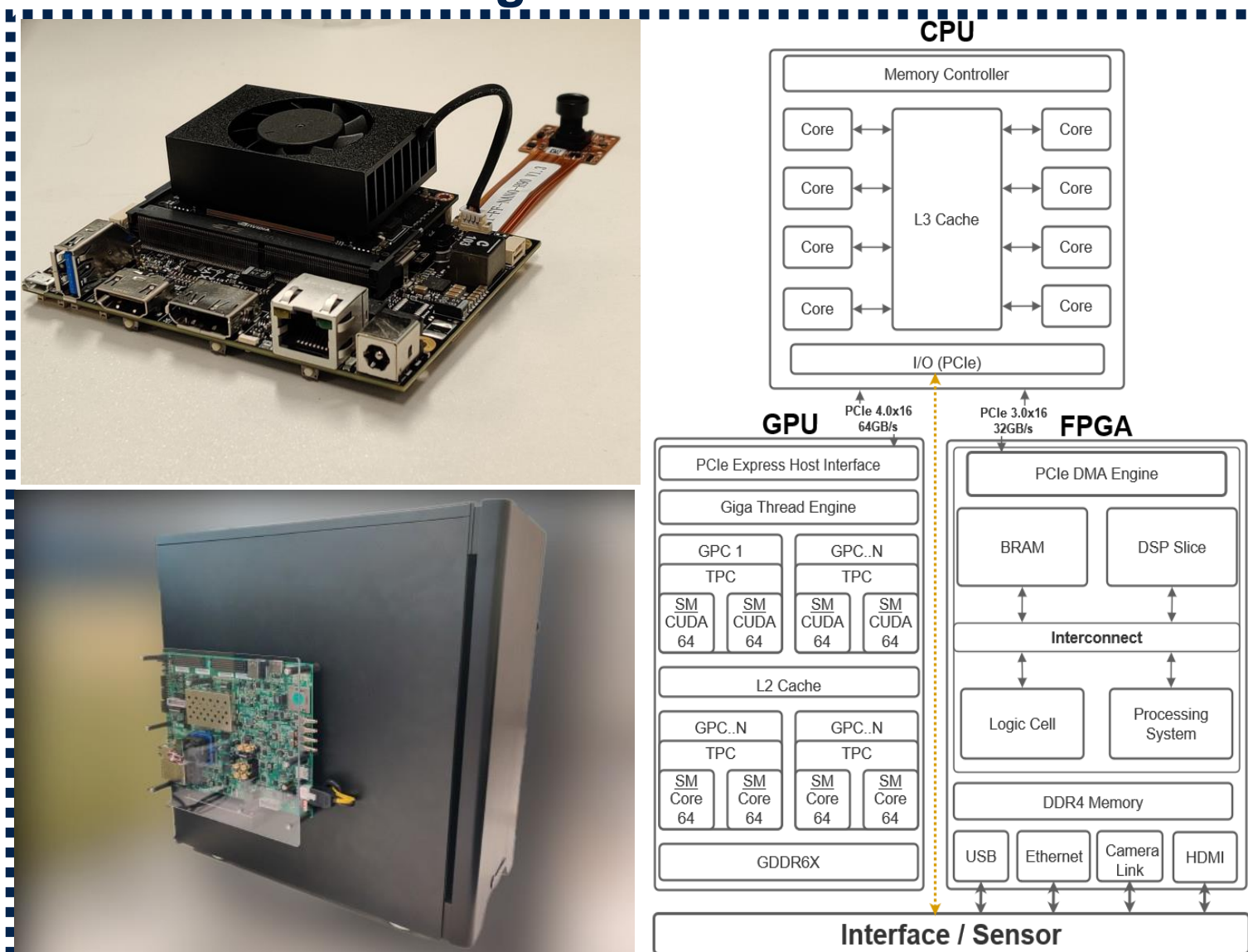# Domain-specific Optimisations of Image Processing Algorithms on Heterogenous Architectures.

Teymoor Ali[1,2,3], Deepayan Bhowmik[2], Robert Nicol[3], University of Stirling[1]/Newcastle University[2]/STMicroelectronics[3]

Email: teymoor.ali@newcastle.ac.uk, deepayan.bhowmik@newcastle.ac.uk, robert.nicol@st.com

UNIVERSITY of STIRLING
Newcastle University
life.augmented

Heterogeneous computing architectures containing CPUs, GPUs, FPGAs and other specialised accelerators on a single interconnected fabric have enabled widespread adoption within many real-time vision applications.
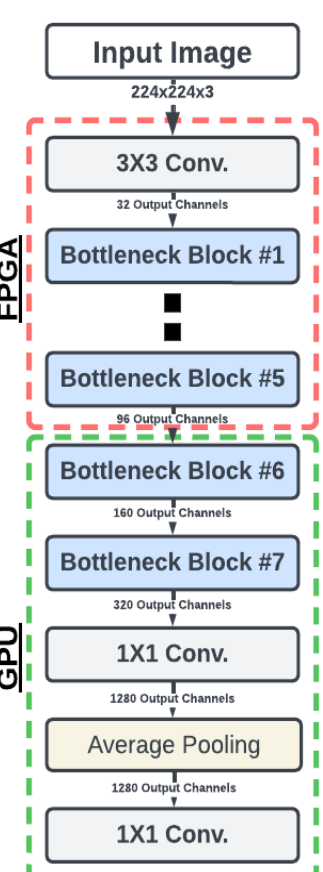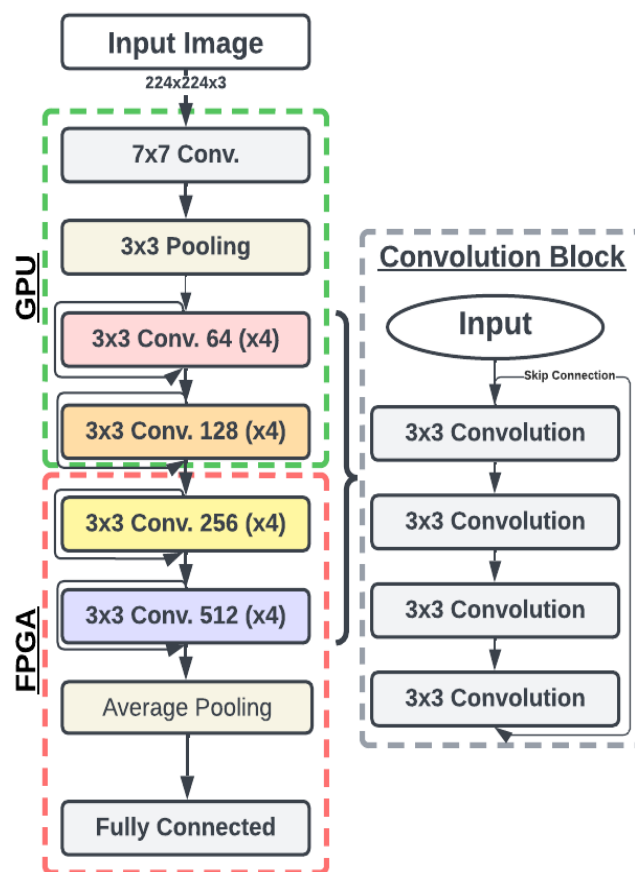
Heterogenous Platforms

Partitioning Strategies

- Designing algorithms on heterogenous architectures requires careful partitioning of operations.
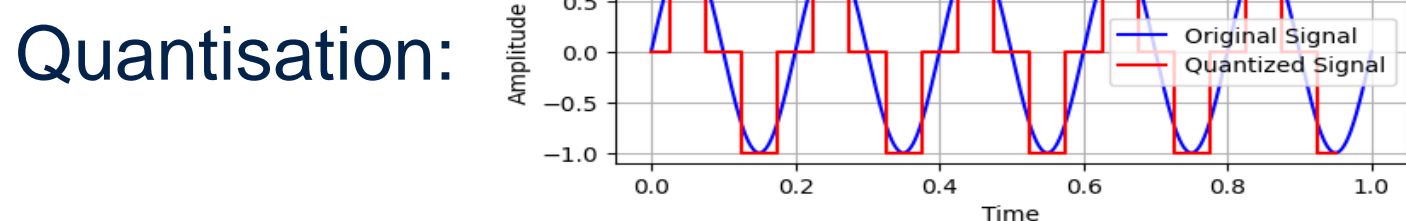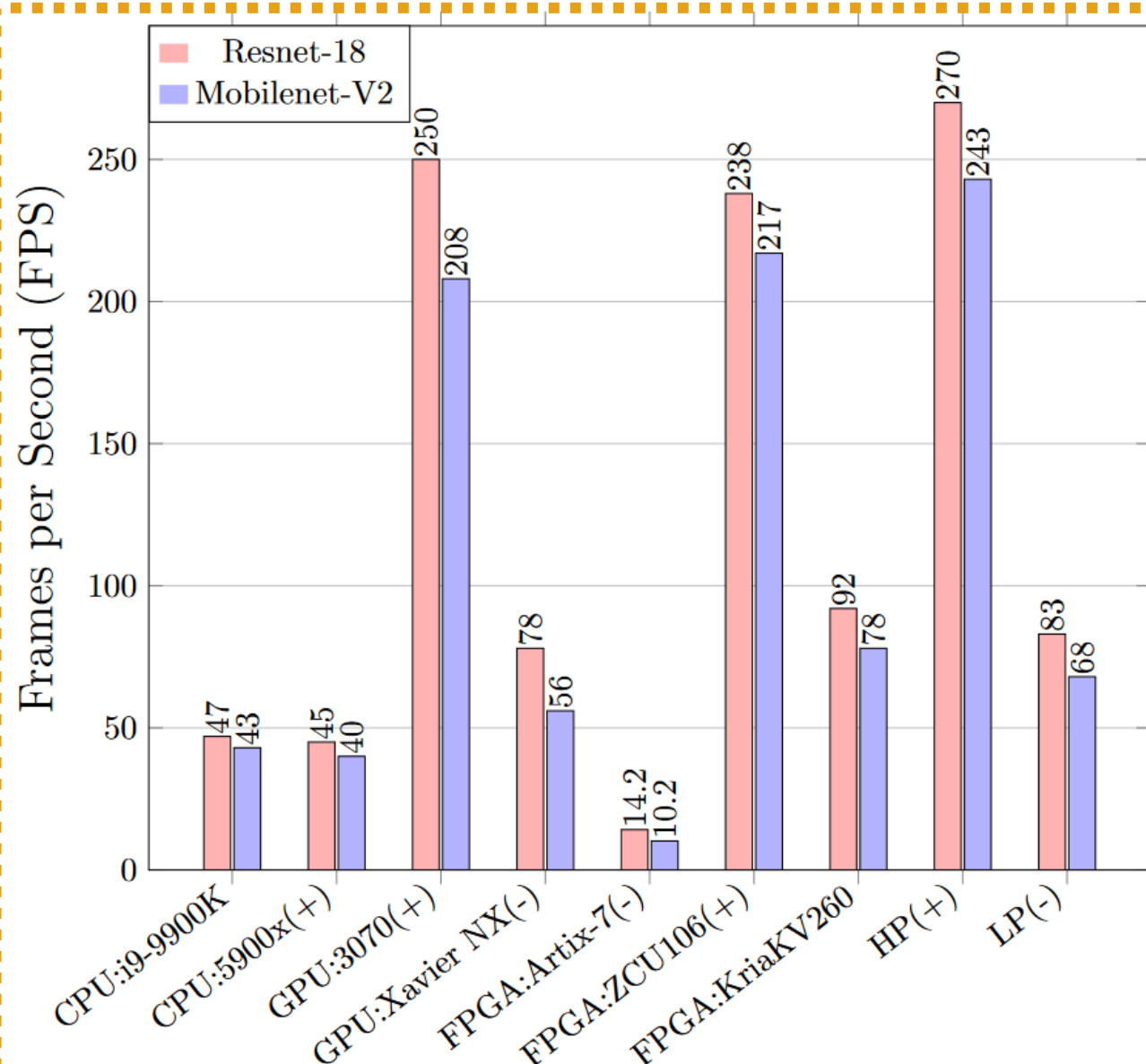

MobileNetV2 / ResNet-18 partitioning diagrams

Domain-Specific Optimisations

Separable Filter:
$$\frac{1}{4}\begin{bmatrix}1\\2\\1\end{bmatrix} \times \frac{1}{4}\begin{bmatrix}1 & 2 & 1\end{bmatrix} = \frac{1}{16}\begin{bmatrix}1 & 2 & 1\\2 & 4 & 2\\1 & 2 & 1\end{bmatrix}$$

Downsampling:

Quantisation:

Kernel Size:
$$\frac{1}{110}\begin{bmatrix}1 & 1 & 2 & 2 & 2 & 1 & 1\\1 & 2 & 2 & 4 & 2 & 2 & 1\\2 & 2 & 4 & 8 & 4 & 2 & 2\\2 & 4 & 8 & 16 & 8 & 4 & 2\\2 & 2 & 4 & 8 & 4 & 2 & 2\\1 & 2 & 2 & 4 & 2 & 2 & 1\\1 & 1 & 2 & 2 & 2 & 1 & 1\end{bmatrix} \Rightarrow \frac{1}{16}\begin{bmatrix}1 & 2 & 1\\2 & 4 & 2\\1 & 2 & 1\end{bmatrix}$$

Benchmarking CNNs on Hardware

Benchmark chart: Frames per Second (FPS) for Resnet-18 and Mobilenet-V2 across platforms. CPU:i9-9900K (47/43), CPU:5900x(+) (45/40), GPU:3070(+) (250/208), GPU:Xavier NX(-) (78/56), FPGA:Artix-7(-) (14.2/10.2), FPGA:ZCU106(+) (238/217), FPGA:KriaKV260 (92/78), HP(+) (270/243), LP(-) (83/68).

Research Outcomes

- **Domain-specific optimisations techniques applied on imaging algorithms, significantly reduce execution time and energy consumption of hardware.**

- **Partitioning algorithms with respect to memory latency onto the most efficient accelerator in a heterogenous domain increases overall performance.**

Benchmarking Results

- **High and low-power heterogenous platforms outperforms their fastest discrete homogenous counterpart architecture (3070)/(XavierNX) in terms of execution time and energy.**

- **Kria FPGA outperforms in FPS and energy than embedded GPU counterpart (XavierNX) at a lower clock speed.**

Newcastle University UNIVERSITY of STIRLING life.augmented