Energy proportional binarized neural networks with adaptive voltage and frequency scaling

Dr Jose Luis Nunez-Yanez
Reader in Energy Efficient and Adaptive computing
University of Bristol

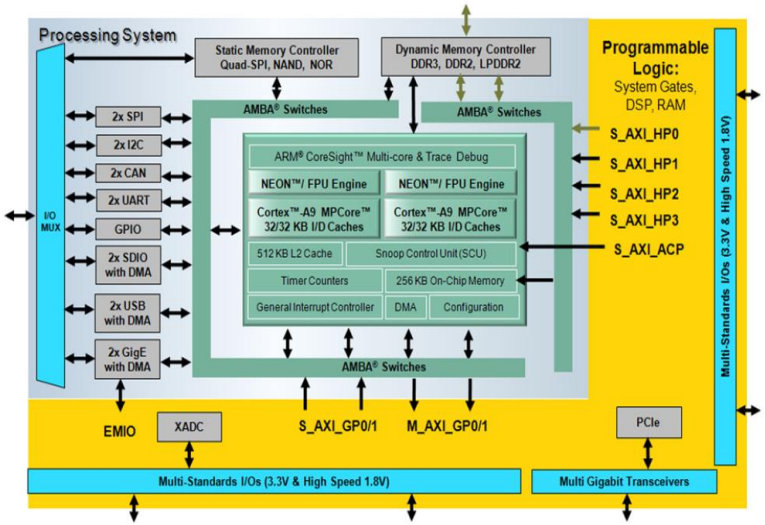
Talk structure

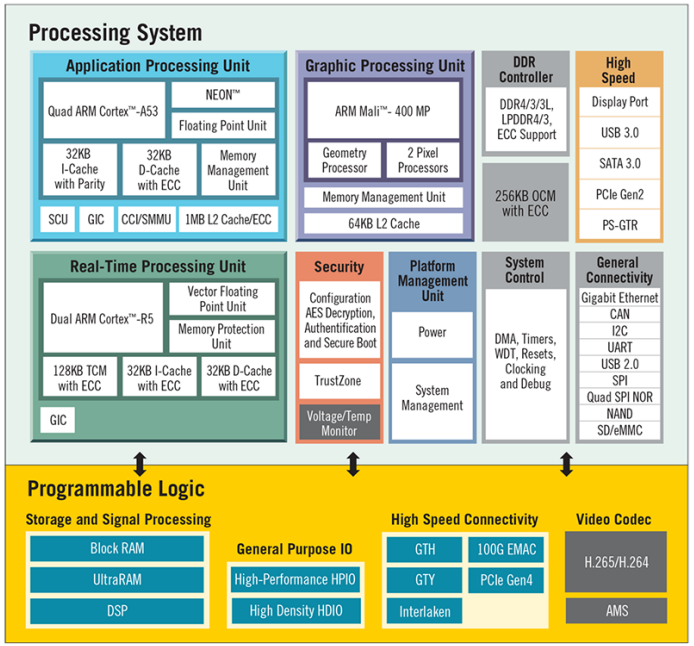
- Adaptive Voltage Scaling framework presentation
- Binarized neural network application.
- Conclusions

Embedded hard processor : Zynq and ZynqMP (ultrascale) families => (ARM + FPGA)

• The Zynq and Zynq Ultrascale processing platform are system on a chip (SoC) processors with embedded programmable logic : processing system (PS) + programmable logic (PL).

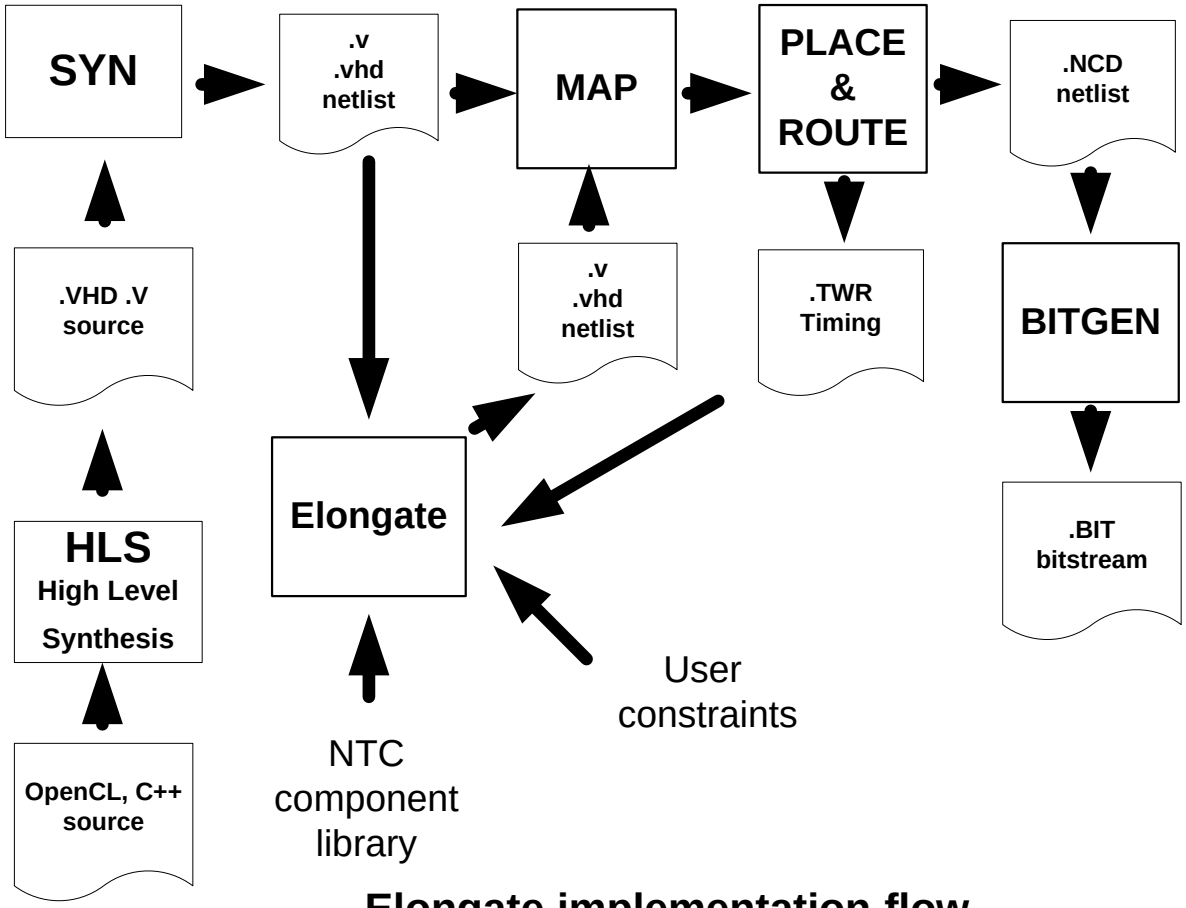
• New programming models for this type of devices favour C/C++ flow with frameworks such as Xilinx SDx replacing traditional RTL design.





Improving the energy efficiency of the FPGA with Adaptive Voltage Scaling.

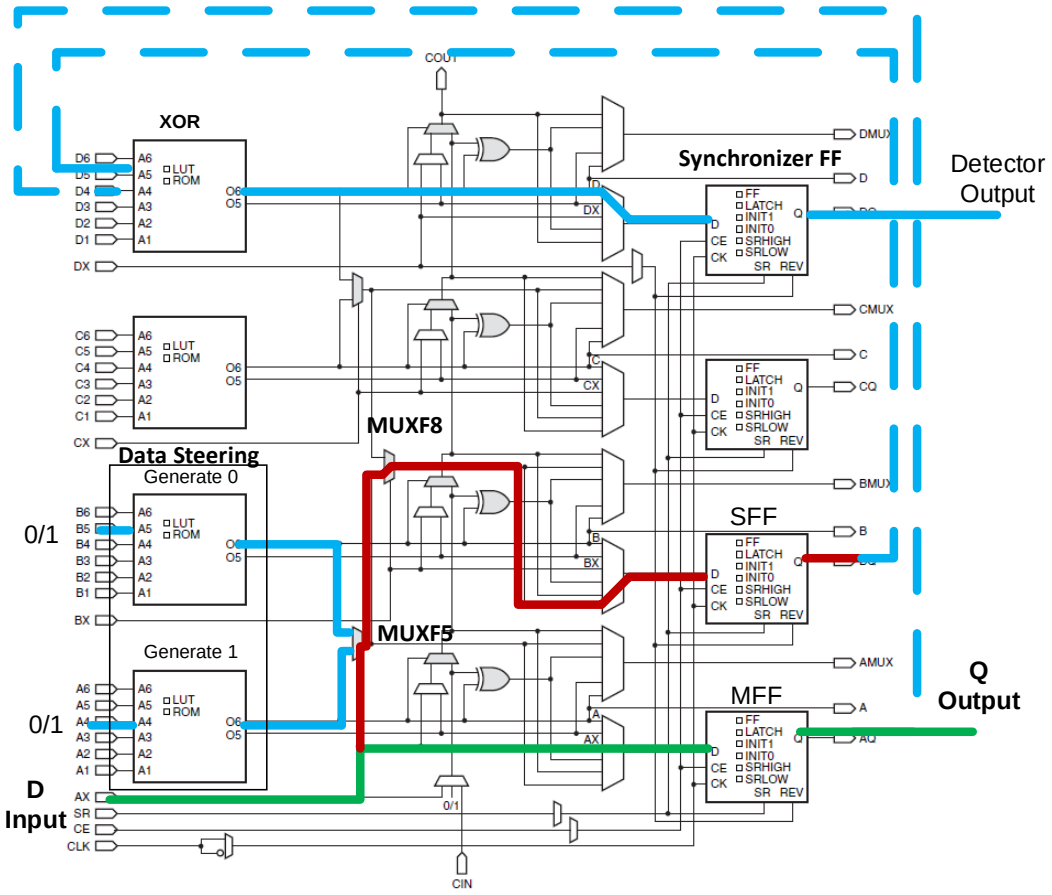
- Slowing down the FPGA device if it is too fast is only energy beneficial if voltage is reduced.
- We have been working on a tool flow and IP blocks to control the frequency and voltage of the device and detect optimal operational points using in-situ detectors.



Elongate implementation flow

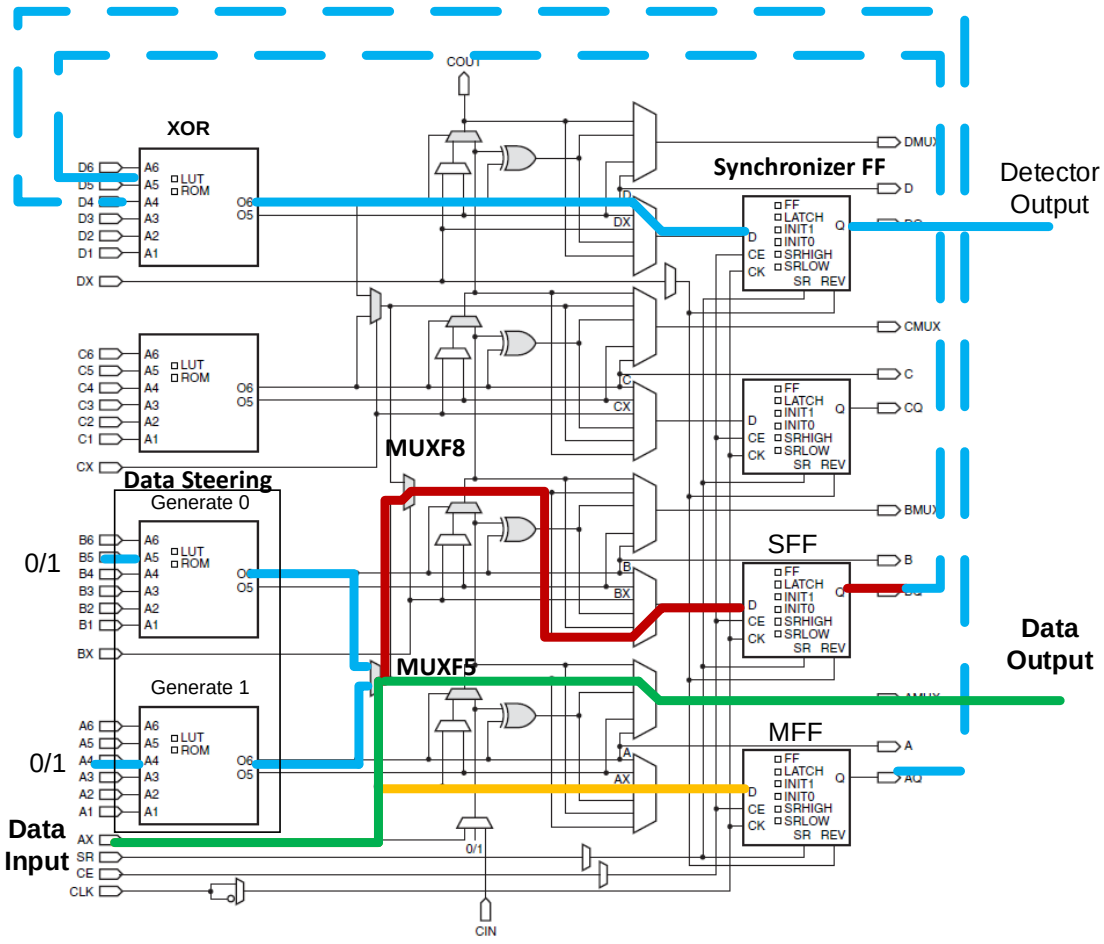
Example of timing detector for logic

- Soft-macro Detectors guarantee that the path of the slow flip-flop (SFF) slightly longer than main flip-flop (MFF).
- Discrepancies between MFF and SFF are detector in XOR and communicated to DFS (Dynamic Frequency Scaling) unit.
- MFF replicates the functionality of the original flip-flop in the critical path.



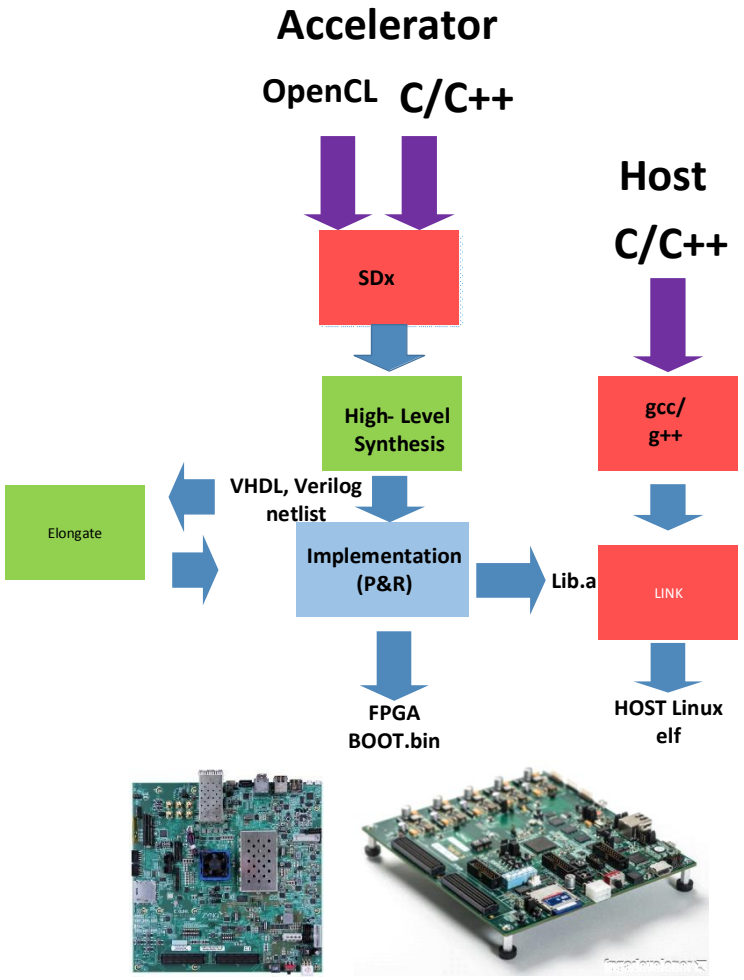
Example of timing detector for memory

- A different type of detector is needed for critical paths that are not directly observable such as memories or DSP blocks
- These elements use a clock with different phase than the main clock. The phase differences is selected to create a critical path in the detector.
- The disadvantage is that both FFs represent an overhead and approach is not as accurate as the logic detectors.



Elongate framework for Zynq and Zynq ultrascale with SDx

- We use the Xilinx SDx tools to generate a hardware library that links with the host code and a BOOT.bin file for the FPGA.
- To use elongate we intercept the VHDL netlist and re-insert the modify netlist directly in Vivado and produce a new BOOT.bin .
- The Zynq ultrascale and Zynq devices use different detector libraries but the flow is very similar.



Application to Neural network inference with fully binarized CNN

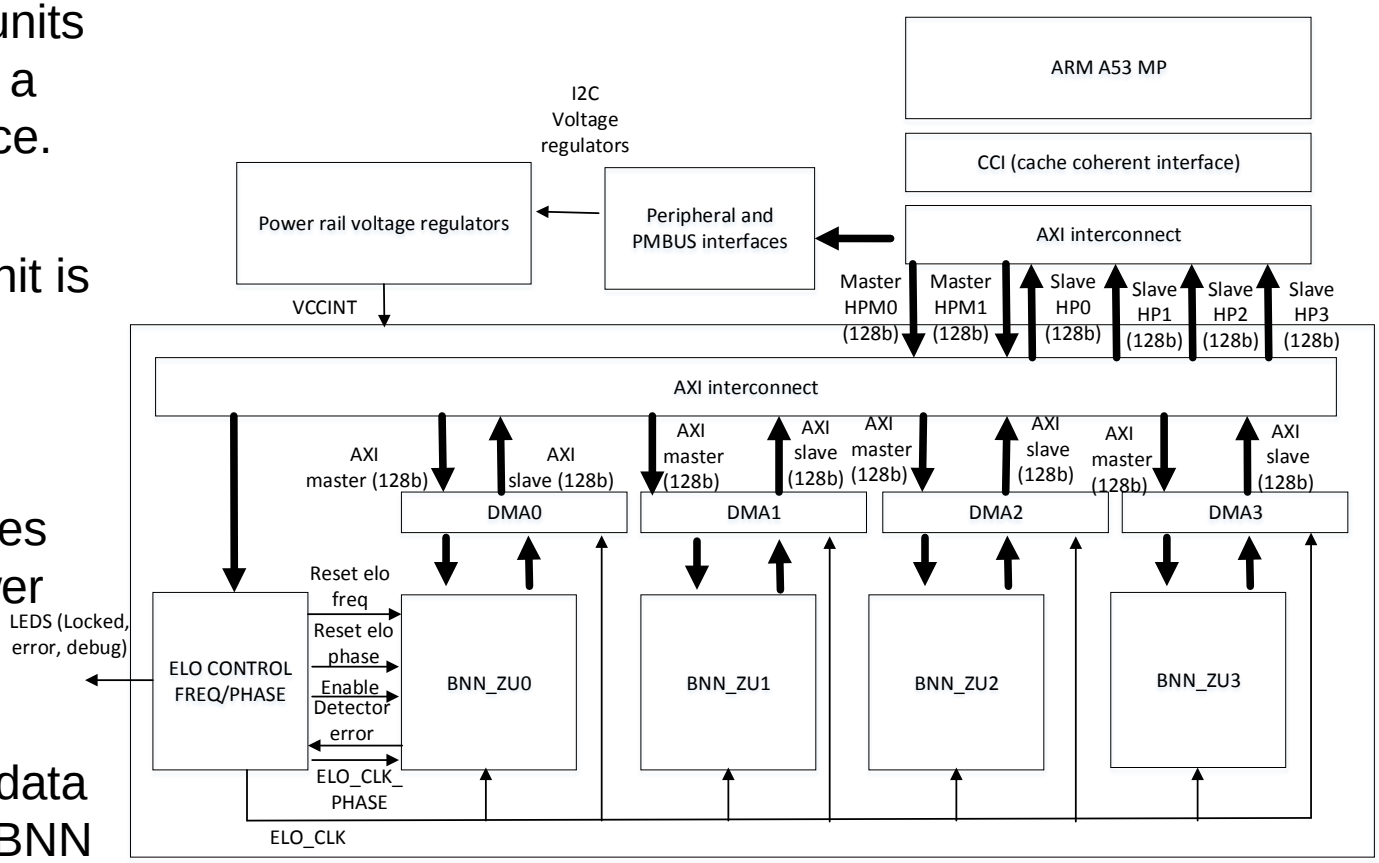
- Based on the Xilinx FINN BNN that uses single bits for weights and activations and ported to SDSoC.
- Accuracy on CIFAR-10 is around 80% while state-of-the-art accuracy is around 90%.
- Each layer uses a configurable number of processing elements (PEs) and each PE uses a variable number of SIMD lines.

FINN
input (32×32 RGB image)
3×3-conv-64
3×3-conv-64
pooling
3×3-conv-128
3×3-conv-128
pooling
3×3-conv-256
3×3-conv-256
FC-64
FC-64
FC-64 (no activation)

FINN layers: convolutional, pooling and fully connected

BNN architecture with voltage and frequency scalability

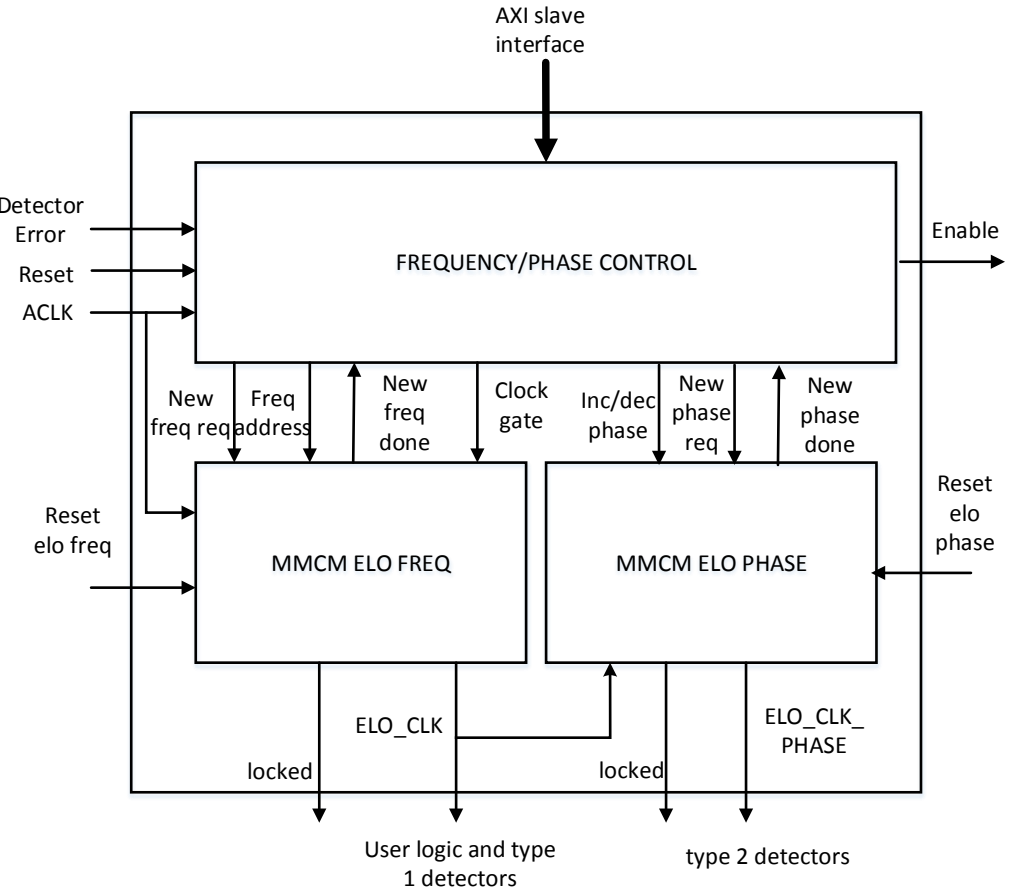
- Up to four compute units working in parallel in a Zynq ultrascale device.
- Only one compute unit is instrumented with Elongate detectors.
- The Zynq version uses ony one CU with fewer PE/SIMD.
- DMA engines move data from memory to the BNN hardware



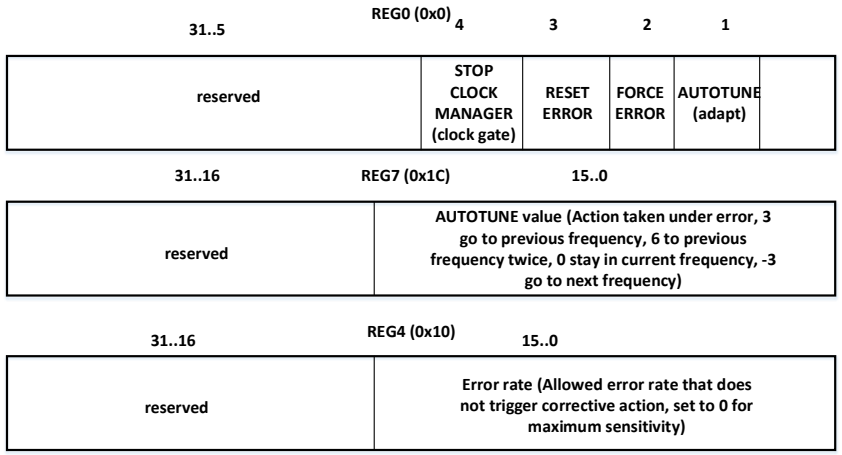
Elongate BNN architecture

Frequency and phase control IP component

- Based on a FSM control unit with two MMCM (Mixed Mode Clock Managers) with locked frequencies and phases.
- AXI slave interface to access control registers for configuration and enabling.



Frequency/phase control



Elongate control registers

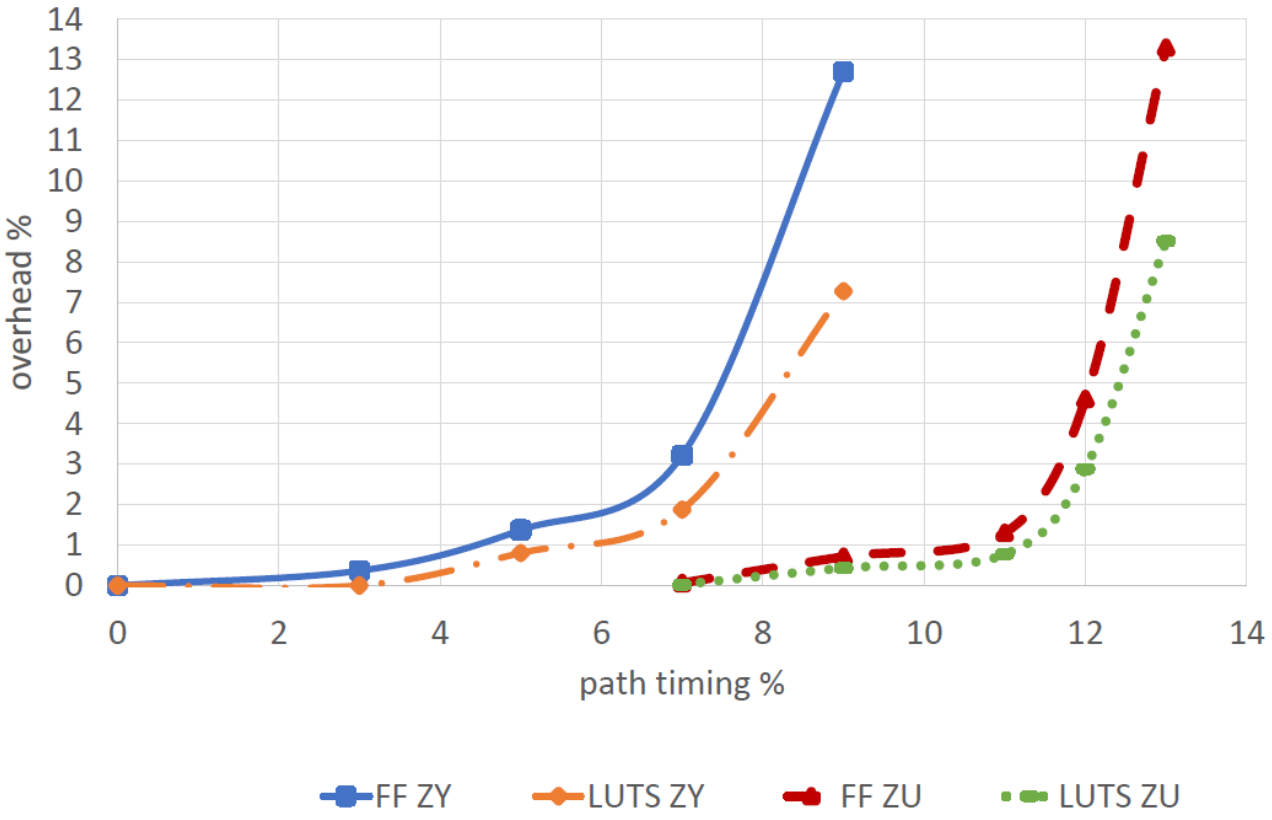
System complexity in Zynq and Zynq Ultra devices

- The Zynq Ultrascale device enables a larger hardware configuration and 2x nominal frequency.
- Resource utilization is close to 100% for both configurations.
- Ultrascale ~12x times more energy efficient at nominal.

	Zynq Device z7020 28nm, dual A9	**ZynqMP Device XCZU9-EG 16nm, quad A53**
LUTs (K)	32	224
FFs (K)	36	209
BRAMs	131	740
CUs	1	4
PEs	91	832
SIMDs	176	1488
KFPS /Watt Nominal	3.2 (1v, 100 MHz)	37.9 (0,85 v, 200 MHz)

Detector overheads

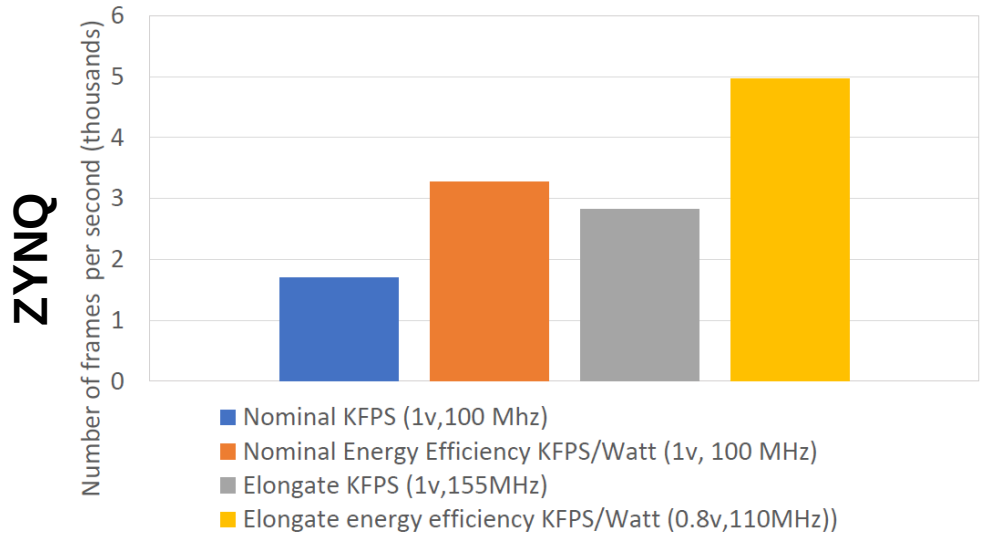
- The number of inserted detectors is user controllable.
- The total number of inserted detectors and protected paths oscillates between 100 to 300 depending on design timing.
- This covers paths with around 10% better timing than the critical path.

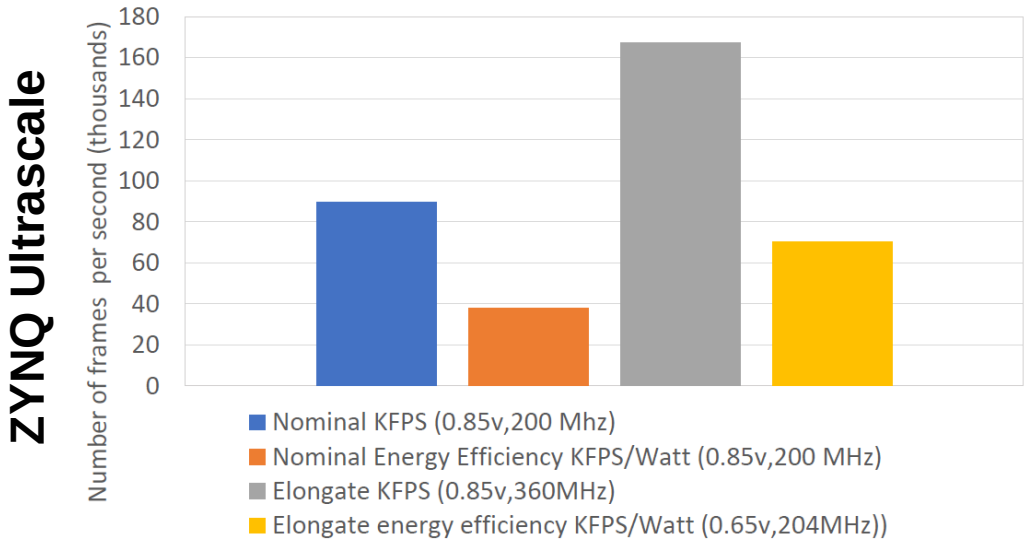


Elongate FF/LUT overhead and path coverage relation

Performance and energy efficiency

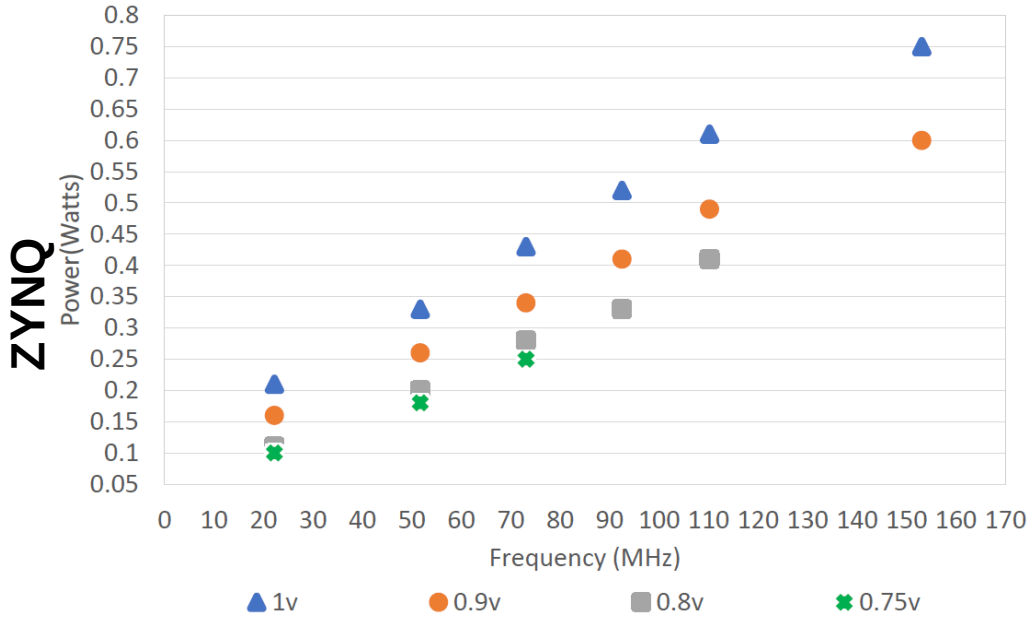
- Performance in the Ultrascale device is 60x better and energy efficiency is 17x than Zynq.
- The Elongate versions improve performance and energy efficiency by up to 80% without affecting accuracy.
- The smaller Zynq device can be more energy efficient that ultrascale if the processing requirements are lower than 1 KFPS.

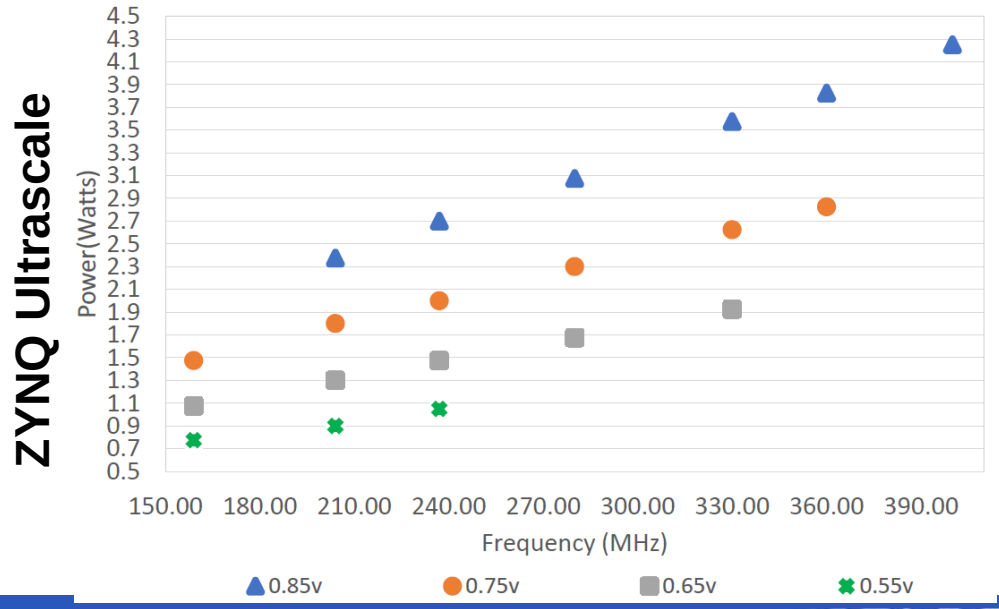




Power scalability

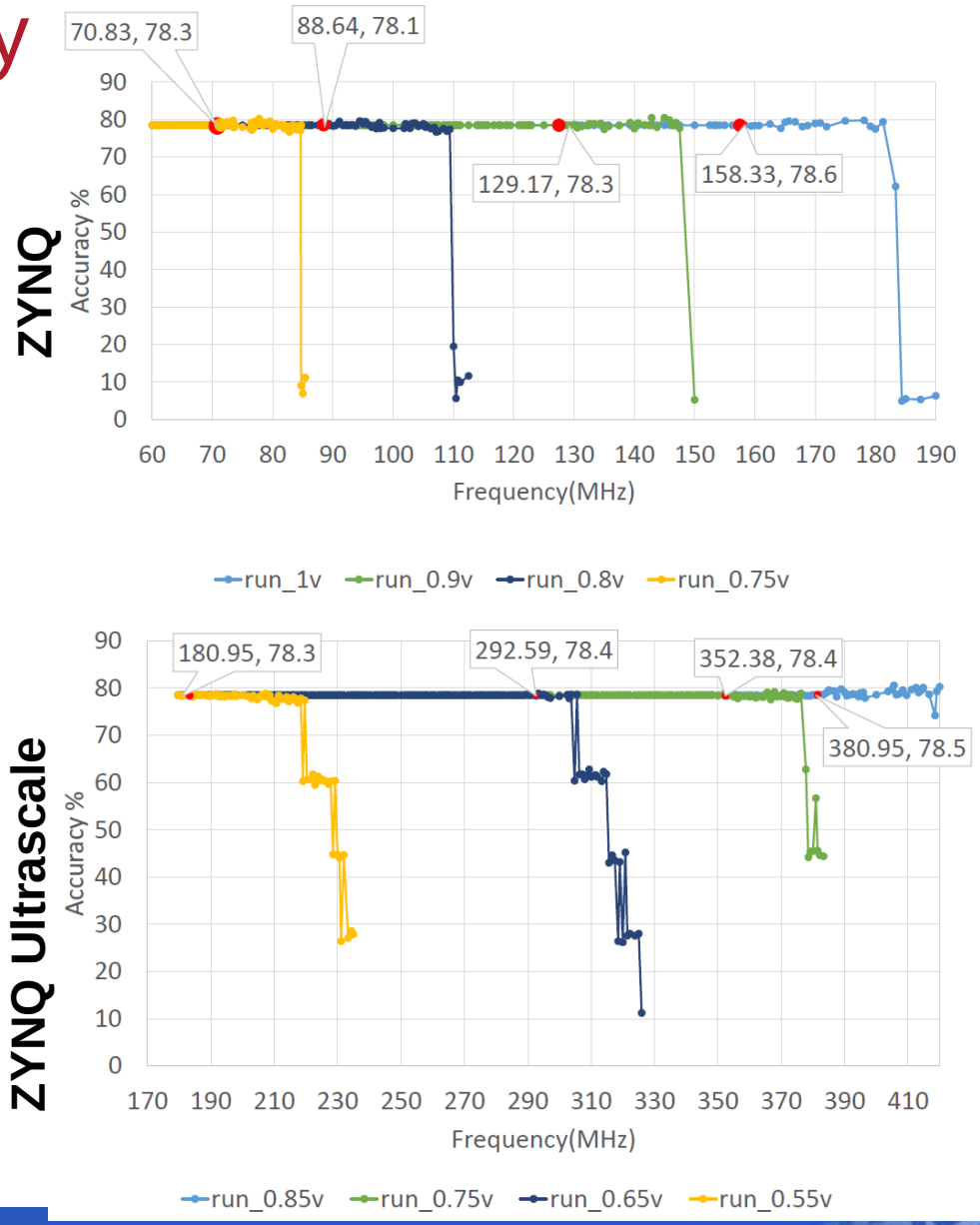
- Valid voltage levels range from 0.55 v to 0.85 v for the 16nm device and 0.75v to 1v for the 28 nm device.
- Both devices obtain good power/frequency scalability that is linear for each voltage level.
- Absolute power requirements are higher in the 16nm Ultrascale device due to its significantly larger size (~6x).





Inference accuracy robustness

- The classification accuracy of the neural network remains within a 1% margin if frequency increases after the first timing errors are detected.
- This ‘noise’ robustness is present in both devices and for all voltage levels.
- Consequently, better energy efficiency and performance are possible if in 1% variability in accuracy is acceptable.



Conclusions

- Adaptive voltage scaling (AVS) in FPGAs with in-situ detectors shows that significant improve performance or reduced energy are possible exploiting margins.
  - Up 80 % lower energy or better performance.
  - Elongate measured 96 KFPS/Watt better than the energy efficiency of IBM TrueNorth (6.1 KFPS/Watt) on the same application
- Elongate portable between the 16 nm Zynq Ultrascale and the 28nm Zynq FPGA technology
- The binarized neural network application is specially suitable since it offers good scalability and robustness after the first timing errors are detected.
- Future work involves making the whole system controllable in a energy-aware run-time system connected to video cameras extracting a variable number of regions of interest in frames before the inference process.

Acknowledgements

- Thanks to Xilinx for access to the FINN neural network and SDx development tools.

- Thanks to EPSRC for supporting this research with the ENPOWER/ENEAC projects.

- If you want to know more or cite please check: **Nunez-Yanez, J, 2017, 'Adaptive voltage scaling in a heterogeneous FPGA device with memory and logic in-situ detectors'. Microprocessors and Microsystems, vol 51., pp. 227-238**

- Happy to answer any questions ?