Computing to the Energy and Performance Limits with Heterogeneous CPU-FPGA Devices

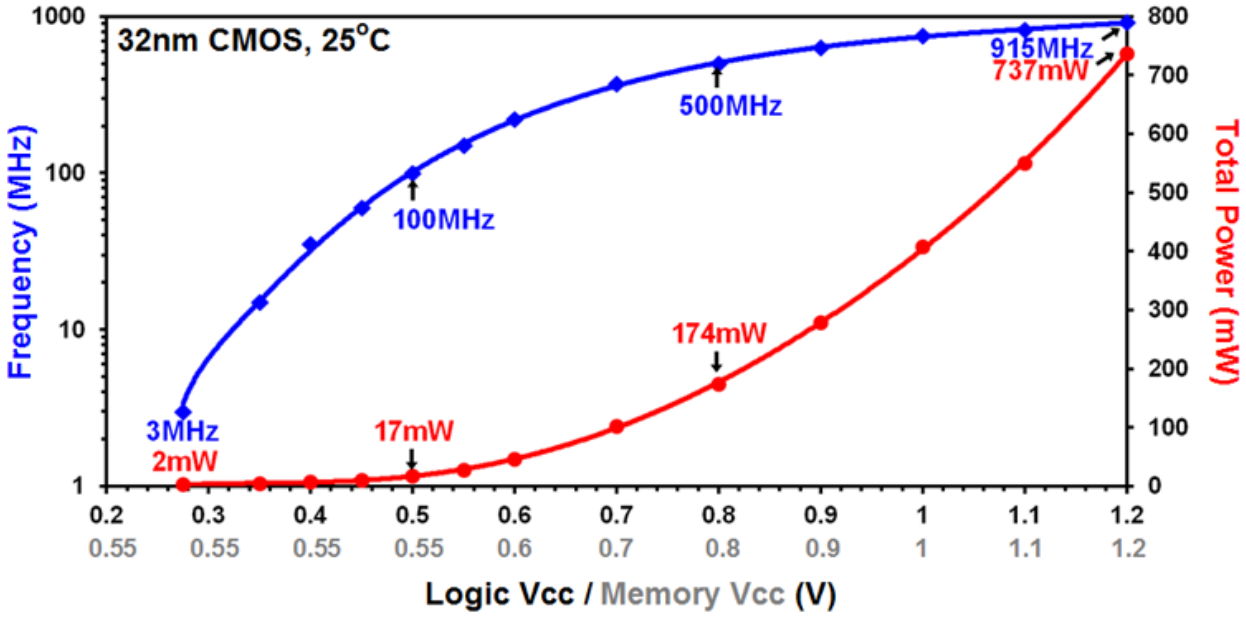
Dr Jose Luis Nunez-Yanez
University of Bristol

Power and energy savings at run-time

Power = $\alpha.C.V^2.f+g1.V^3$

Energy = Power * T

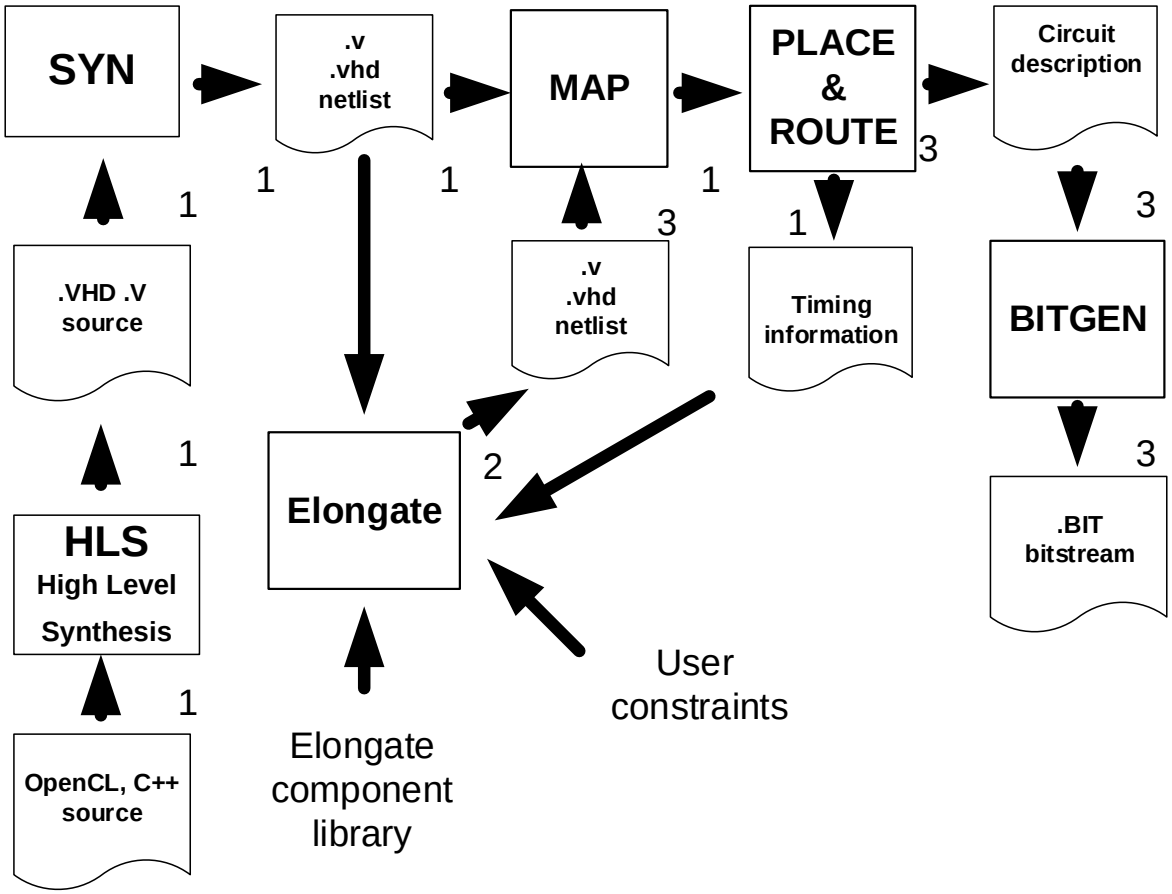
1. DVFS (Dynamic Voltage and Frequency Scaling) : open-loop
   - e.g. AMD cool&quiet, Intel SpeedStep
2. AVS (Adaptive Voltage Scaling) closed-loop
   - e.g ARM Razor
3. AVLS (Adaptive Voltage and Logic Scaling) in reconfigurable chips : FPGAs ?



Power/ Voltage and Frequency relations : Source Intel

Adaptive Voltage Scaling tool flow

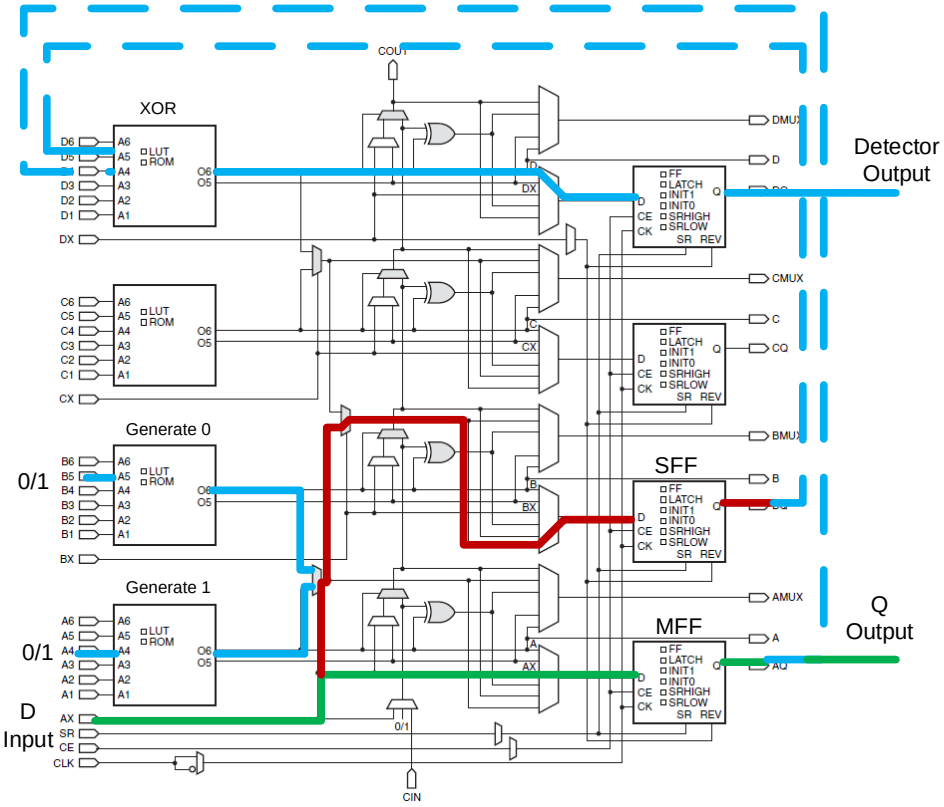
- Tool flow and IP blocks control the frequency and voltage of the device and detect optimal operational points at run-time using in-situ detectors.
- The proposed approach works in a variation-aware closed-loop configuration so it is sensitive to temperature and process variations.
- The flow has been ported to Vivado and it is based on three phases. Phase 3 is done with incremental P&R.



Elongate implementation flow

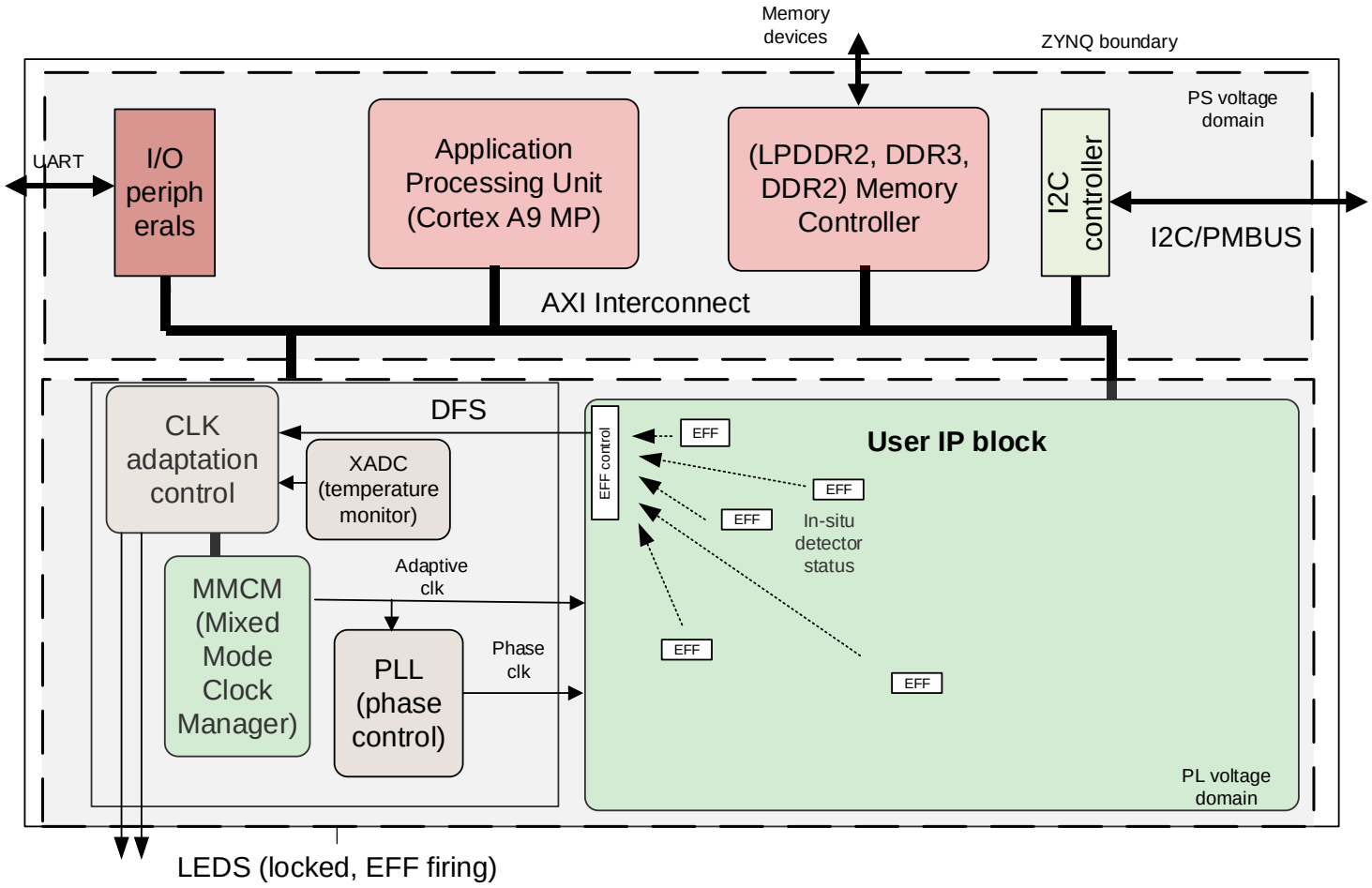
Timing detectors for logic

- 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.



Logic timing detector

System architecture



ARM CPU control the Elongate IP

Case study: Video Fusion

- Video fusion of visible and thermal imagery provides a method to combine complementary information for better data analysis.
- The prototype system uses two cameras interfaced to a Zynq-based board that combines an dual-core ARM processor and a FPGA fabric in the same chip.

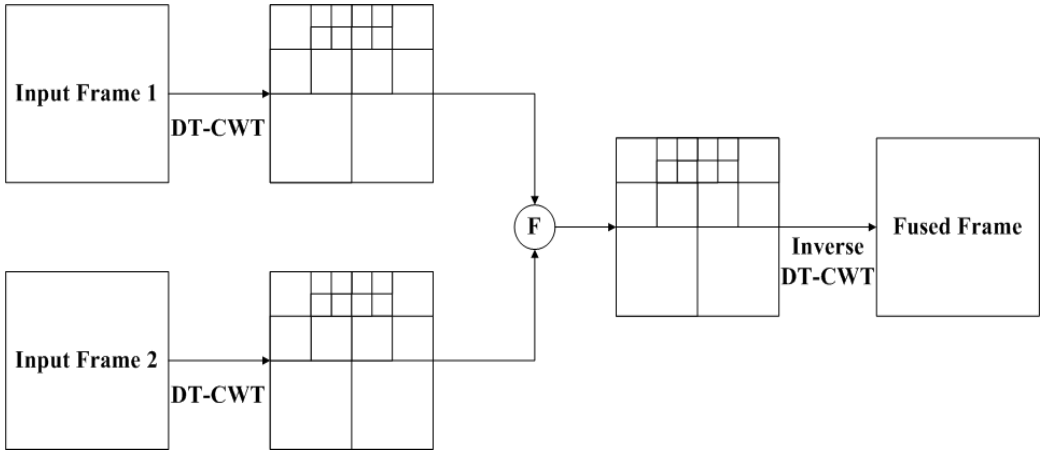


Fig. 1 Fusion with wavelets

Fig. 2 Prototype System

Fusion algorithm with DT-CWT

- DT-CWT algorithm developed in the group as offering good quality of results in presence of noisy input.
- The forward and reversed DT-CWT represent around 70% of total complexity and are selected as candidates for acceleration.
- System operates three times faster (amdahl law) but hardware is too fast and needs to wait for data to be prepared by the processor

Fig. 3 Fusion with DT-CWT

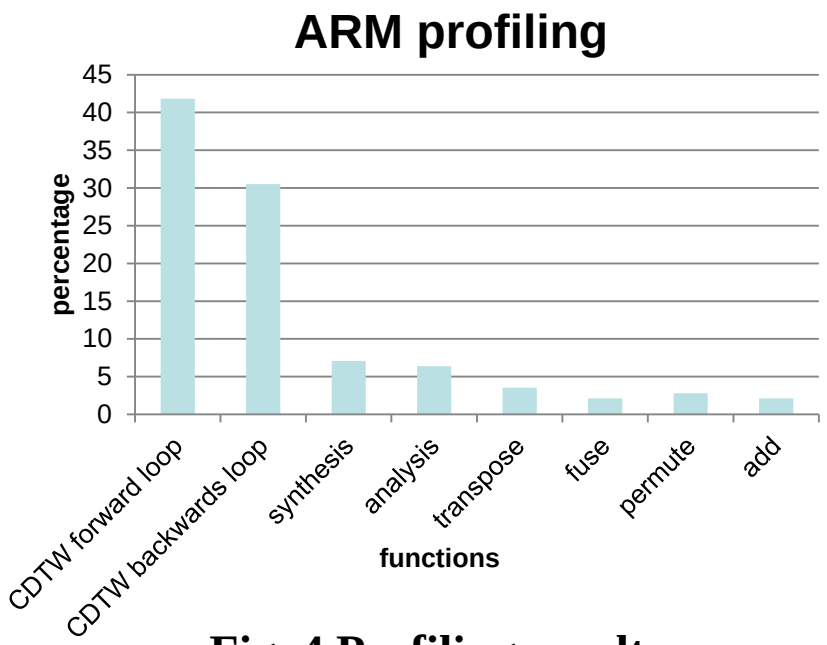
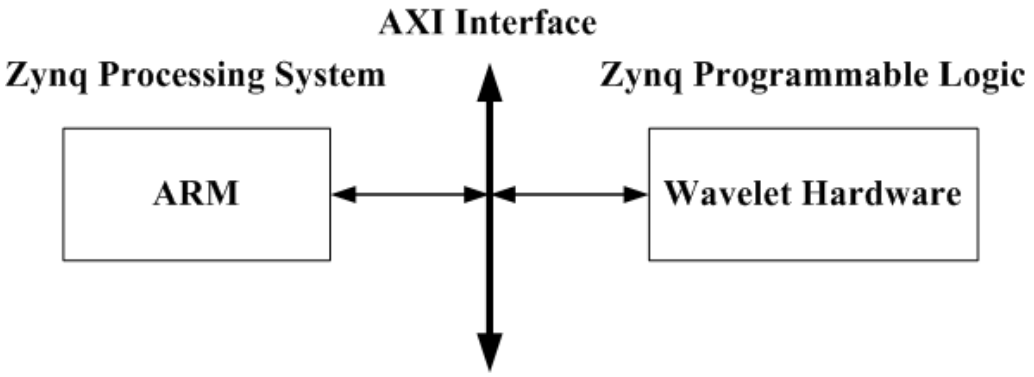


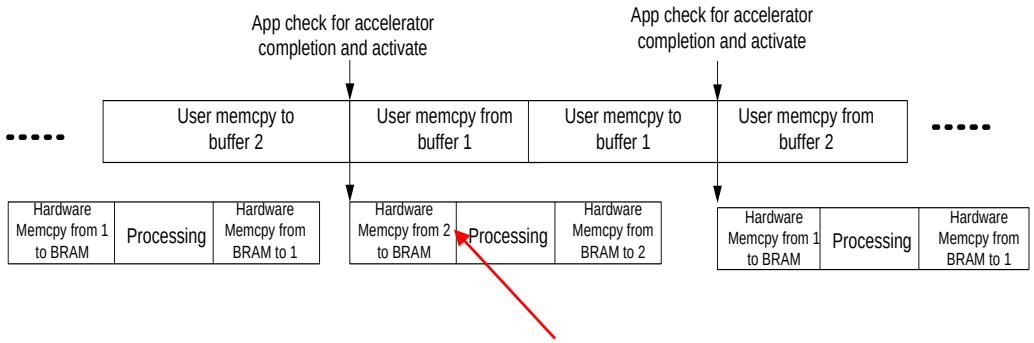
Fig. 4 Profiling results

Hardware and Linux driver development

- Hardware accelerator for wavelets developed using high-level synthesis (e.g. Vivado HLS).
- Processor needs to reserved memory area in kernel space and move pixel data so that it is accessible to the accelerator .
- Double buffering so that data preparation by processor and data processing by accelerator overlap.



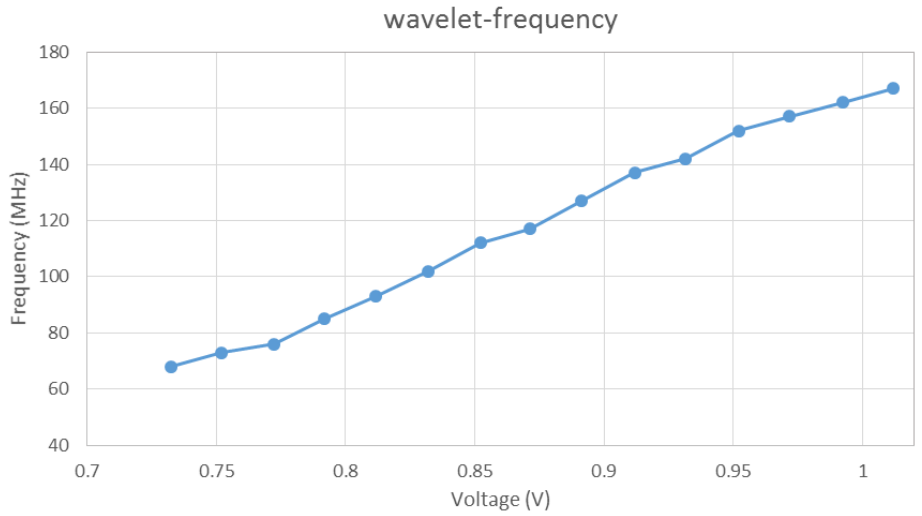
Heterogeneous hardware overview

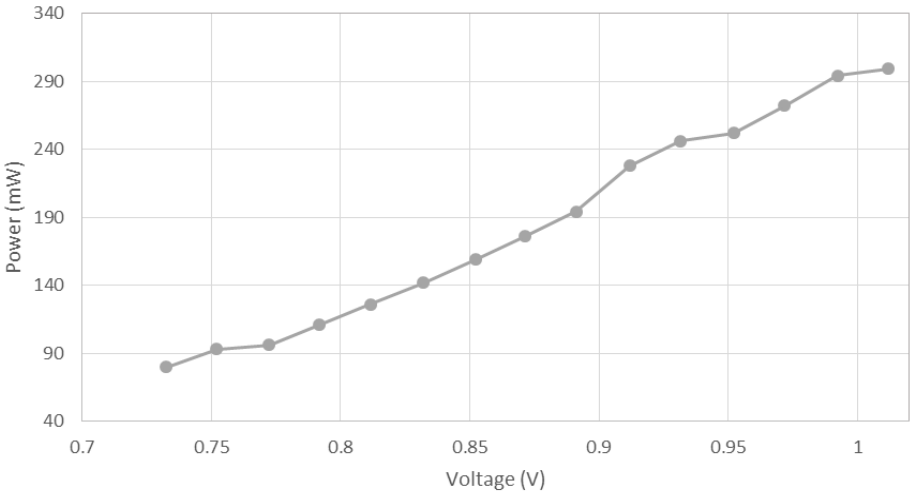


Accelerator is too fast !!

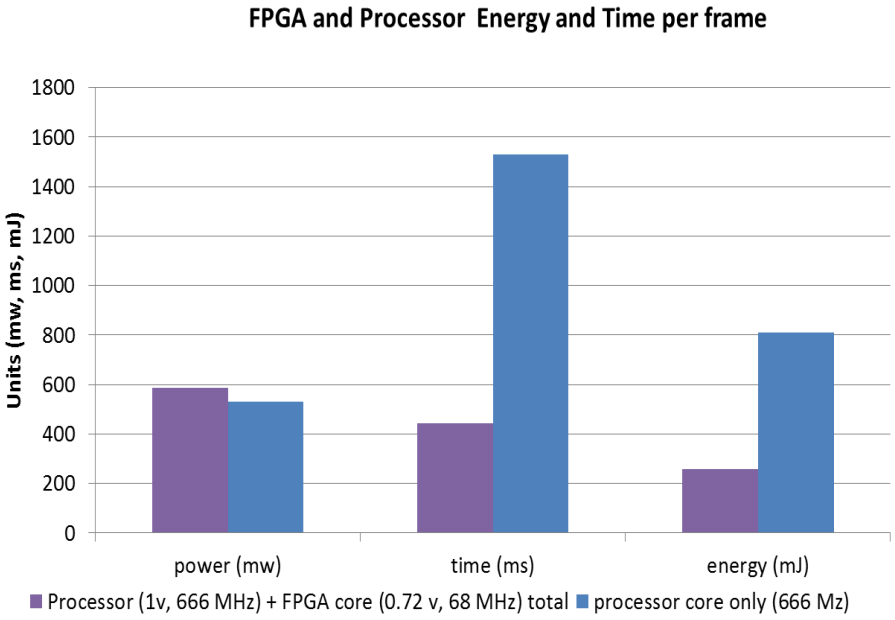
Power and performance analysis

- Valid voltages range from 1 volt to 0.72 volt and frequencies range from 170 MHz to 68 Mhz.
- Most energy efficient point with negligible impact in performance occurs at 0.72 volts and 68 MHz.
- Running the FPGA at nominal 1 v and 68 MHz results in 70-85% higher energy .

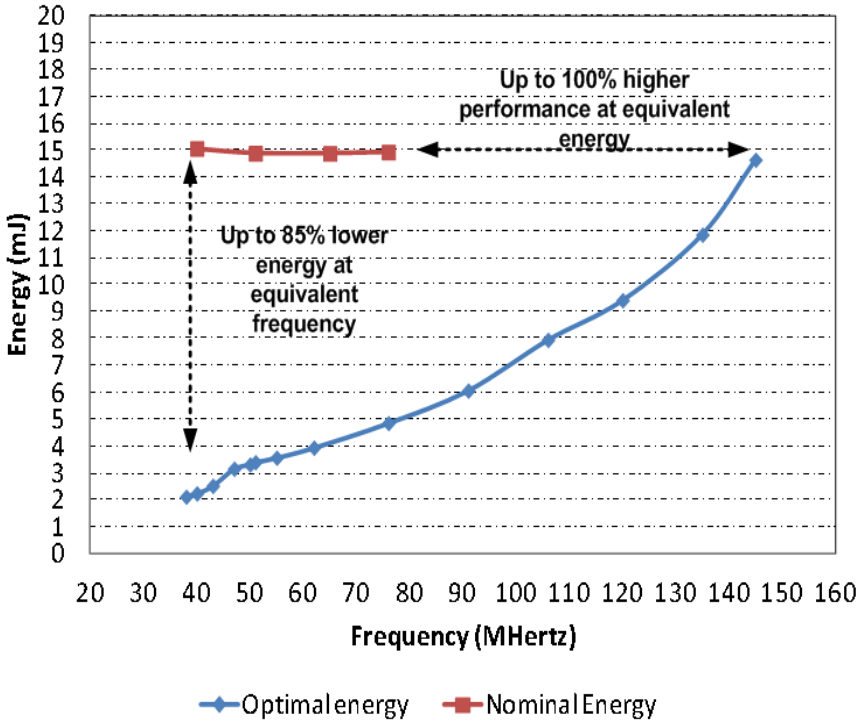




Energy analysis

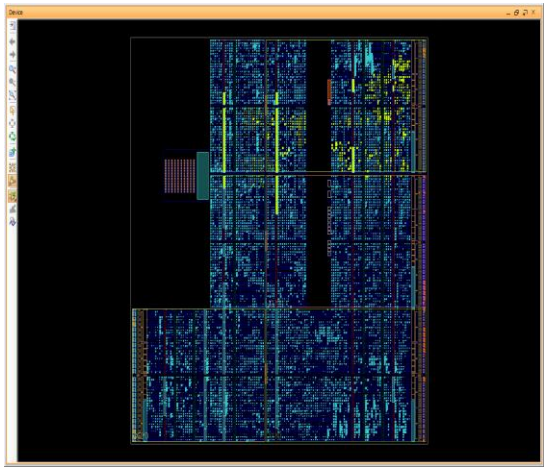


Processor and FPGA fabric energy analysis

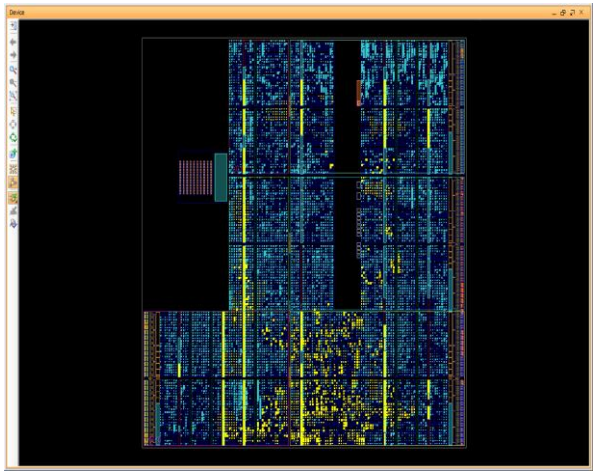


FPGA fabric energy analysis

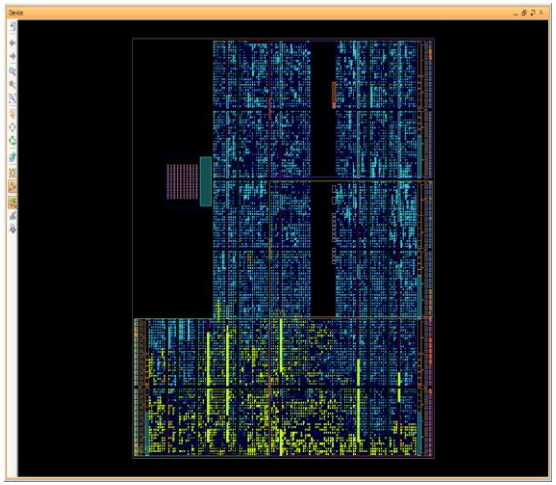
Combining voltage, frequency and logic scaling in AVLS.



Configuration 1



Configuration 2



Configuration 3

- A configurable core can change the effective capacitance affecting both dynamic and static power : Power = $\alpha.C.V^2.f+g1.V^3$
- Hardware configurations with different levels of complexity easy to obtain with high-level synthesis.
- Possible at run-time in modern FPGAs that enable partial dynamic reconfiguration.

Next steps

- Case study based on fusion application shows that adaptive voltage scaling with in-situ detectors can significantly improve performance or reduced energy by computing to the limit of correctness.
- Can we go beyond this limit and how fast can we find these points ?
- AVLS extends AVS by adjusting the logic (i.e. capacitance) of the design at run-time: big/slow cores or small/fast cores ? Application dependent.
- The AVLS concept could be controllable in a energy-aware run-time system under an OpenCL framework.
- The in-situ detector idea could be extended (by the manufacturer) to the hardened processor and memory subsystem.