Acceleration by Separate-Process Cache for Memory-Intensive Algorithms on FPGA via High-Level Synthesis

Master’s Degree Thesis

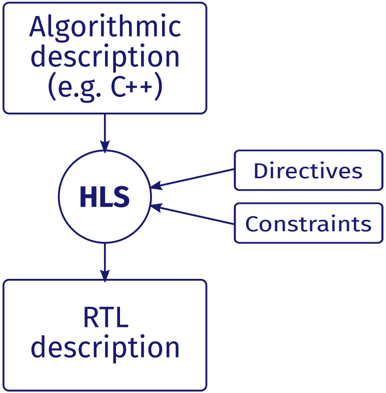
Brignone Giovanni

Supervisor: Prof. Lavagno Luciano

October 21, 2021

Politecnico di Torino

High-Level Synthesis



Primary goals:

- Productivity
- Design space exploration

Limitations:

- Low-level control

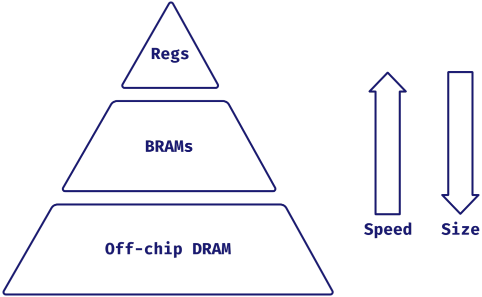
Memory management

Scratchpad:

- Manual memory selection
- HLS state of the art

Cache:

- Automatic full hierarchy exploit
- Thesis work objective



Previous work

Architecture:

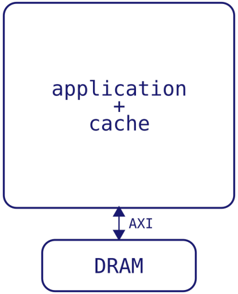
- Cache inlined in application

Implementation:

- `C++` class

User friendliness:

- *Integrable*: `operator[]` overload
- *Configurable*: template parameters
- *Observable*: hit ratio reports



Development

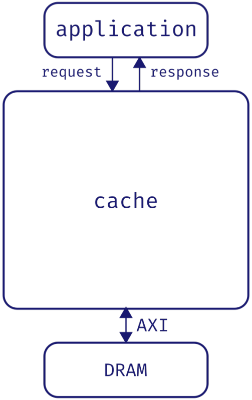
Basic architecture

Objective:

- Limit application cluttering

Proposed solution:

- Cache: separate process
  - Modeling: threads (SW); dataflow (HW)
  - Communication: FIFOs
- Application:
  1. Send *request*
  2. Receive *response*



Basic architecture

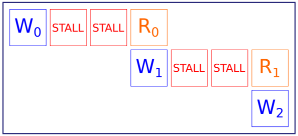
Cache process:

- Optimal pipeline (II=1)

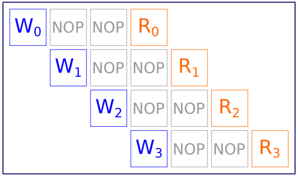
Interface:

- Scheduler unaware of latency between *request* and *response*
- Workaround: forced clock cycles

Original schedule



Schedule with workaround



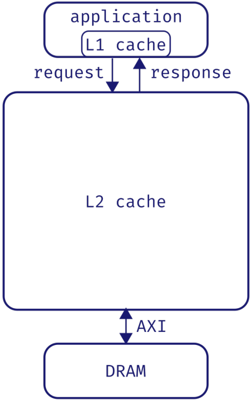
Multi-levels architecture

Objective:

- Improve read performance

Proposed solution:

- L1 cache inlined in application
  - Write-through
  - Direct-mapped



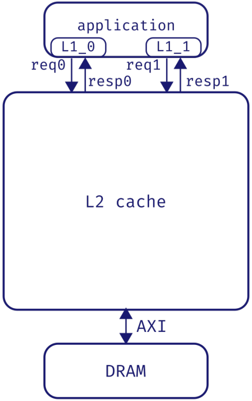
Multi-ports architecture

Objective:

- Efficient application loop unrolling

Proposed solution:

- Single L2 cache with multiple ports
- Multiple L1 caches



Results

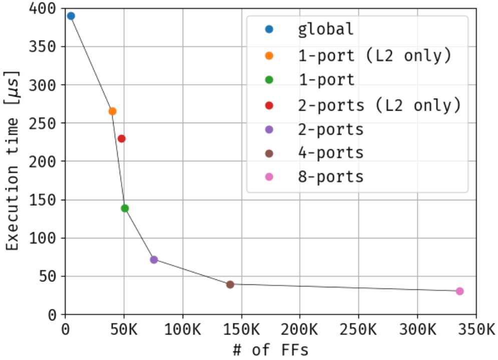
Matrix multiplication

Algorithm:

$$C = A \times B, \quad A, B, C \in \mathbb{R}^{32 \times 32}$$

Caches configuration:

- A: 1 line of 32 elements
- B: 32 lines of 32 elements (direct mapped)
- C: 1 line of 32 elements



Summary

Summary

Achieved results:

- Multi-process modeling for HLS
- Design space extended by cache

Future work:

- Fix interface schedule: RTL implementation

References

Ma, L., Lavagno, L., Lazarescu, M., & Arif, A. (2017). Acceleration by inline cache for memory-intensive algorithms on fpga via high-level synthesis. *IEEE Access*, *PP*, 1–1. https://doi.org/10.1109/ACCESS.2017.2750923

Xilinx Inc. (2021). *Vitis high-level synthesis user guide*. https://www.xilinx.com/support/documentation/sw_manuals/xilinx2021_1/ug1399-vitis-hls.pdf