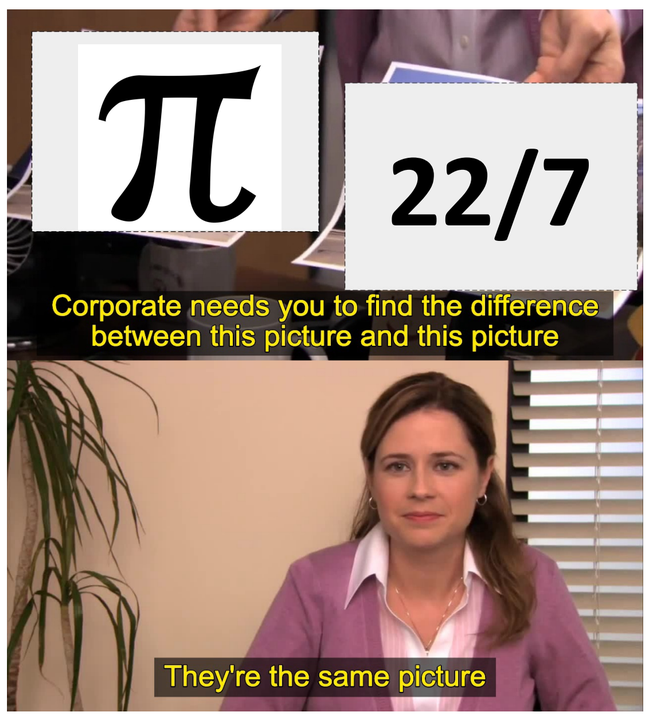
Accelerating the backbone of Artificial Intelligence

Jie Lei
Marie Skłodowska Curie ESR
Universitat Politècnica de València

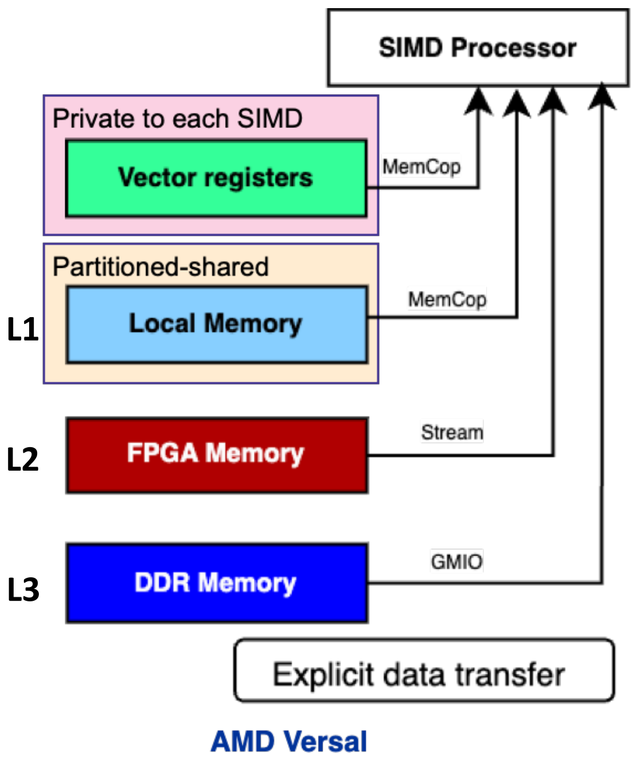
Work of Matrix Multiplication on AMD Versal

Cache hierarchy

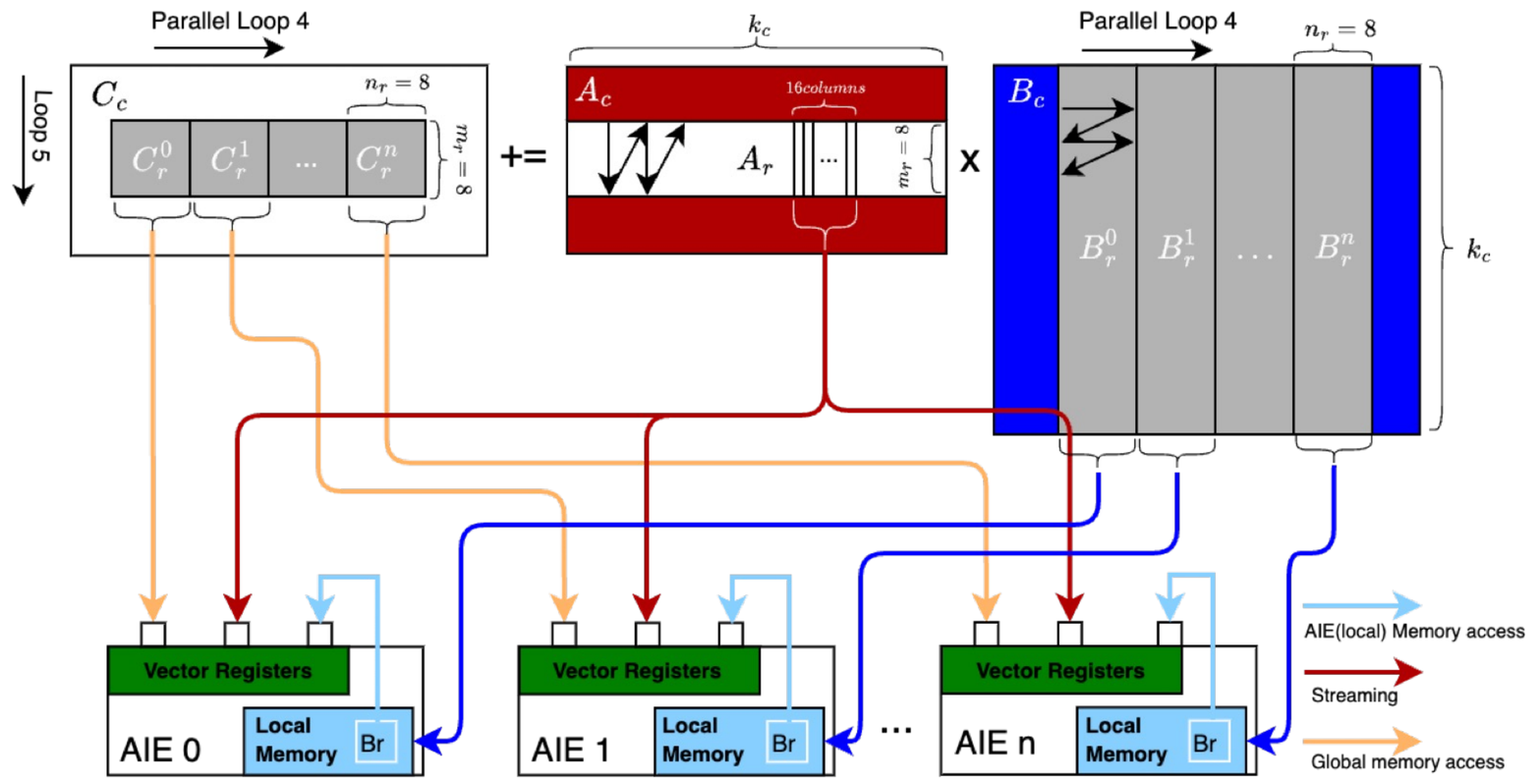
SIMD Compute accelerations



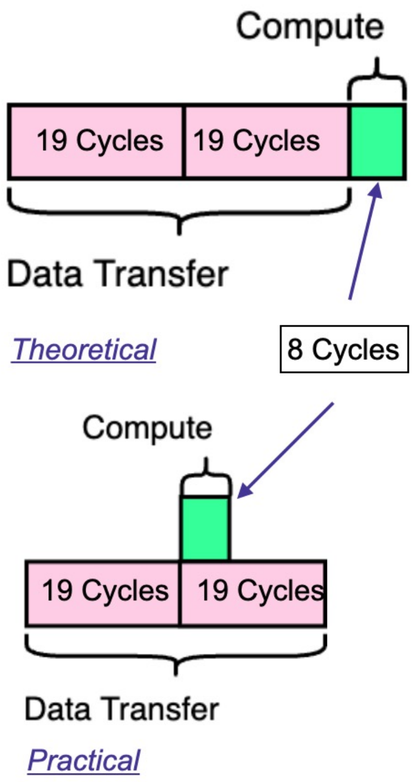
Approximating computing: precision scaling



Proposed memory architecture



Multi-AIE matrix multiplications



Memory bound analyse

#AIE tiles	Instruction Cycles Copy C_r	Arithmetic	Total	Performance/tile (in MACs/cycle)
1	40	4,110	$3,694.1 \cdot 10^3$	31.5
2	58	4,110	$1,916.0 \cdot 10^3$	31.4
4	63	4,110	$958.1 \cdot 10^3$	31.3
8	84	4,110	$498.9 \cdot 10^3$	31.2
16	157	4,110	$275.3 \cdot 10^3$	30.7
32	282	4,110	$162.9 \cdot 10^3$	29.8

Table 2: Distribution of execution time (in cycles) and performance of the parallel design for GEMM when varying the number of AIE tiles between 1 and 32, for a problem of fixed dimension $(m_c, n_c, k_c) = (256,\ 256,\ 2,048)$.

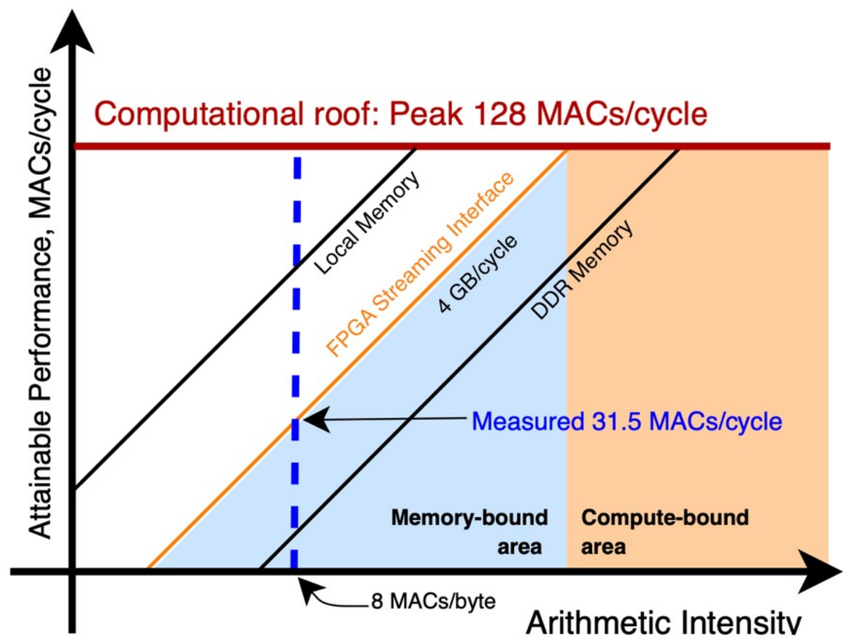


Figure 8: Simplified visualization of proposed work using a roofline model with UINT8 based 8 by 8 micro-kernel of single AIE.

Full Text & Poster

The authors gratefully acknowledge funding from European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska Curie grant agreement No. 956090 (APROPOS: Approximate Computing for Power and Energy Optimisation, http://www.apropos-itn.eu/).