

Benchmarking Floating Point Performance of Massively Parallel Dataflow Overlays on AMD Versal FPGA Compute Primitives

Mohamed Bouaziz, Suhaib A. Fahmy

King Abdullah University of Science and Technology (KAUST), Saudi Arabia

Motivation

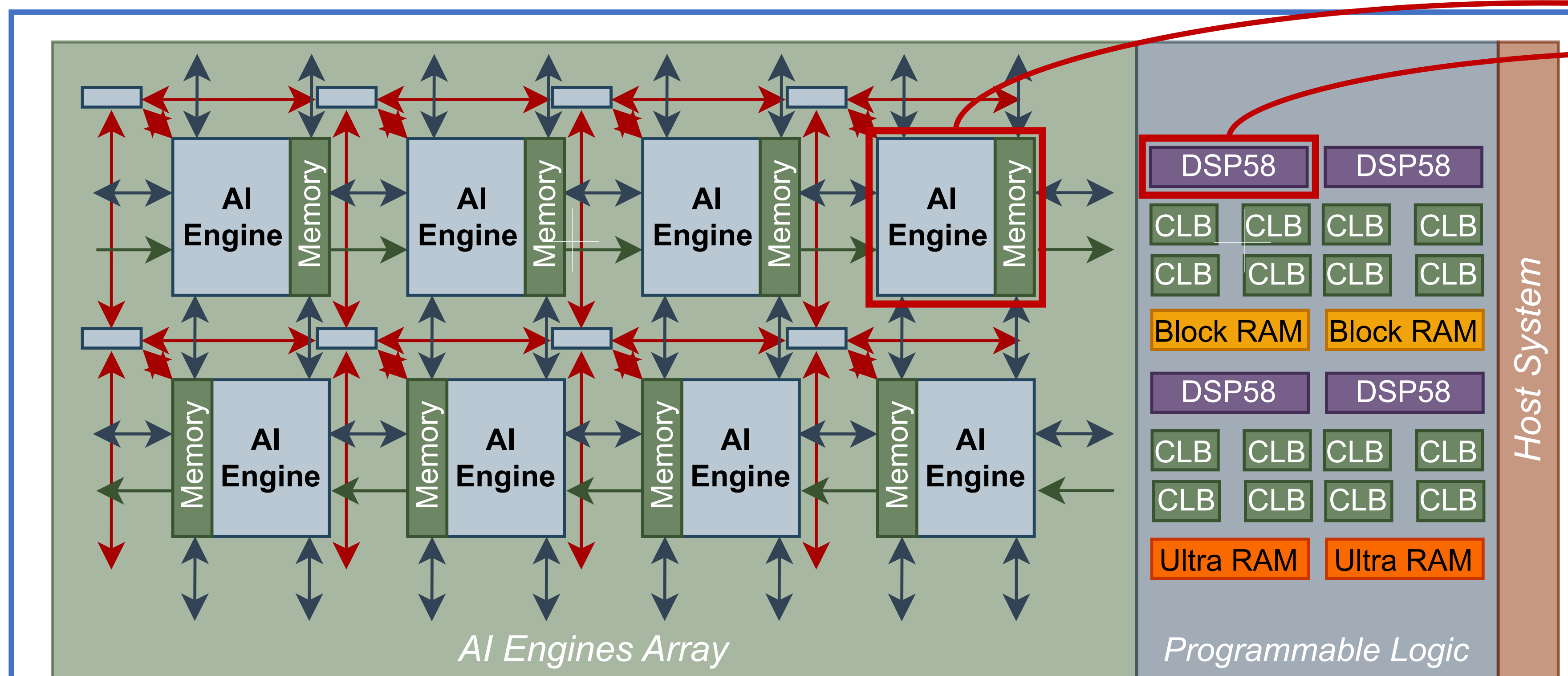


Fig. 1: AMD Versal SoC architecture.

- Implement floating-point operations.
- Pipelined execution.
- ✓ • Allow SIMD operations.
- Allow configurable designs.
- Different data movement patterns.
- ✗ • Different prog./config. models.
- Different HW constraints.



Need for a model to benchmark the distilled performance of these floating-point primitives.

Proposed Approach

- Implement 1xFP operation.
- Configurable pipeline stages.
- Variable frequency.
- Fully custom data movement.

DSP58

- Implement 8xFP operations.
- Fixed pipeline stages.
- Fixed frequency.
- Comm. through available streams.

AI Engine
Memory

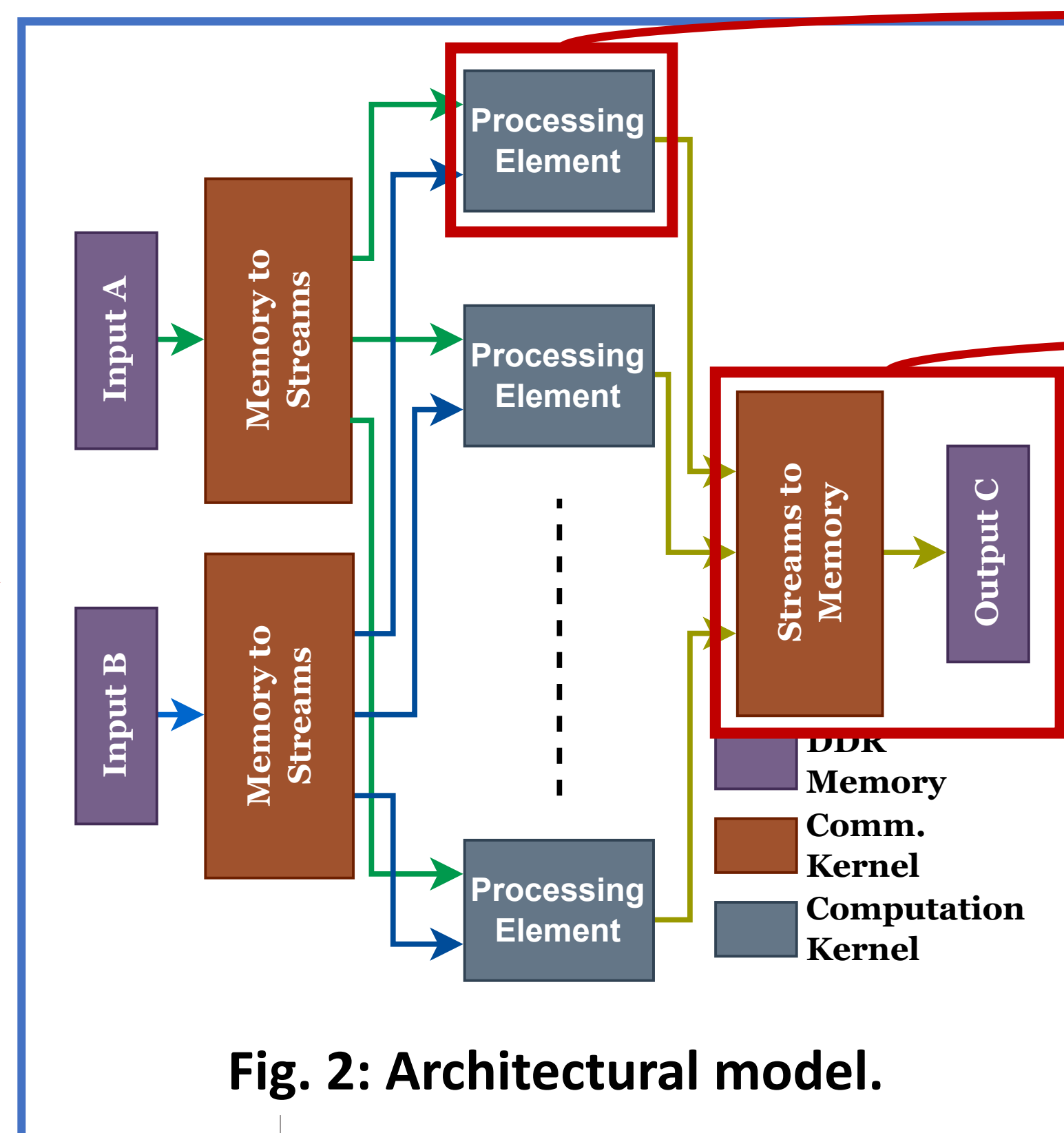


Fig. 2: Architectural model.

8-lane SIMD FP Processing Element.

PE ↔ Off-chip communication.

Key Features:

- Feed-forward streams of data.
- DDR (no HBM) off-chip comm.
- No synchro. among FP primitive.

Proposed Approach

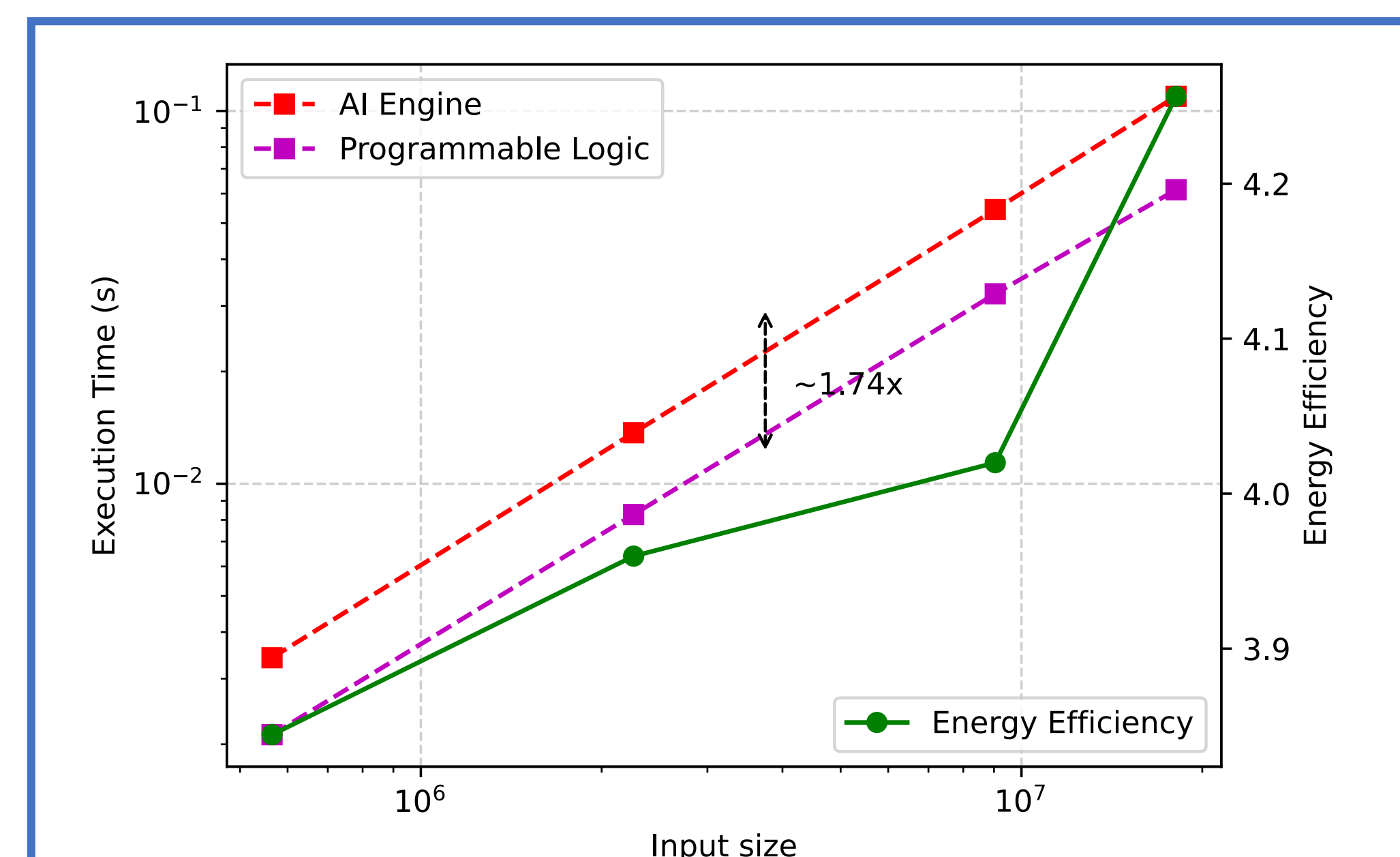
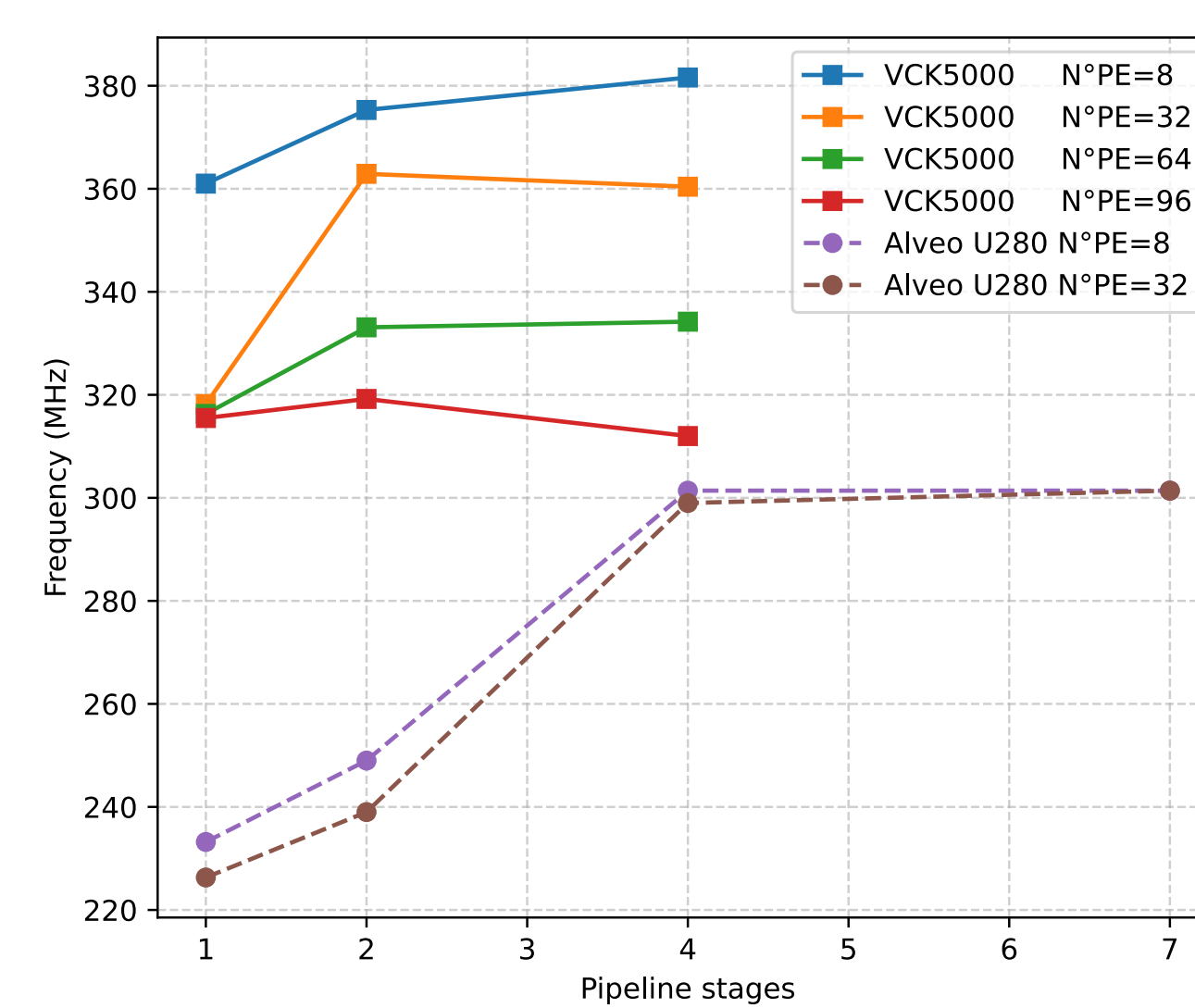
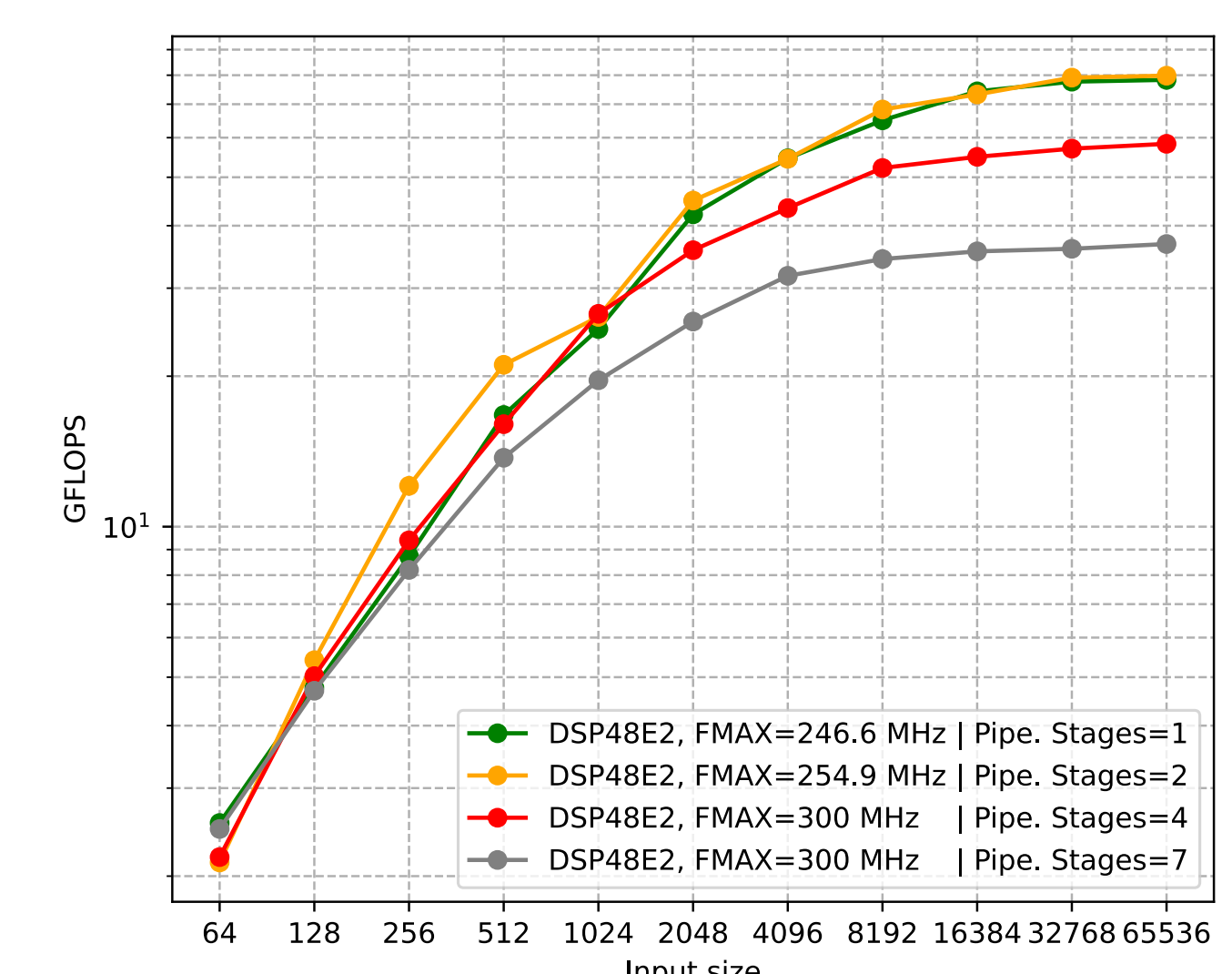


Fig. 3: Execution time and energy efficiency.

Feed-forward applications: faster execution on PL and more energy efficient.



(a) Max. achievable frequency



(b) Throughput on DSP48E2

Fig. 4: Pipelining effect on DSP58 v.s. DSP48E2.

- Drop in freq. due to congestion increases with more pipelining.
- No increase in performance after 4 stages with DSP48E2.