Single-Event Effects Laser Testing of a 7nm FinFET System-on-Chip with AI-Acceleration Capabilities

S. Achaq[1,2], V. Pouget[2], L. Artola[1], G. Hubert[1], A. Urena[1], F. Manni[3], A Dufour[3], J. Boch[2]

1, ONERA, Toulouse, France. 2, IES, Univ. Montpellier, CNRS, Montpellier, France. 3, CNES, Toulouse, France

Context & Motivation

- Single-Event Effects (SEE) in FinFET technologies at the 7nm node:
  - Particle beam tests have shown a reduction of per-bit Soft Error Rate until 7nm
  - **Laser testing at 7nm**? Possibility to generate Single-Event Upsets (SEU)? Spot size limitation?

- Recent generation of Systems-on-Chip (SoCs):
  - Importance of the physical organization of memory resources for embedded mitigation mechanisms

- The rapid diffusion of embedded artificial intelligence (AI)
  - The impact of radiation-induced single-event effects (SEE) in edge-AI hardware is a growing concern

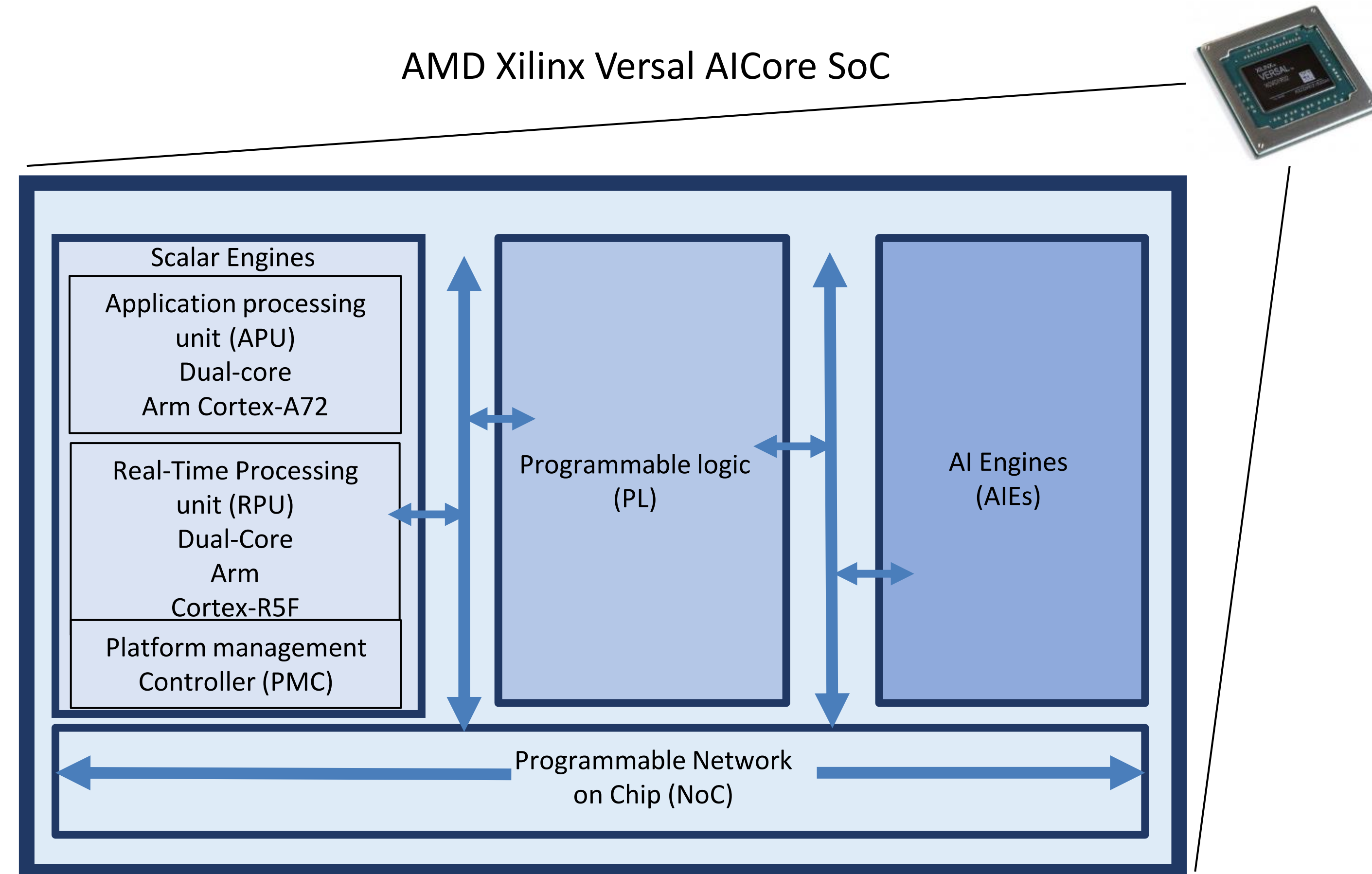
- In the context of the development of an embedded-AI application on the AMD-Xilinx Versal AICore series 7nm FinFET SoC:
  - **Laser testing of SRAM memory resources**
    - **Validate the capabilities of the technique**
    - **Investigate the SRAMs organization and sensitivity**
  - **Laser testing of adaptive intelligence engines (AIEs)**
    - **Investigate the sensitivity of AIE resources**
    - **Study the impact of SEE in a convolutional Neural network hardware accelerator**

Outline

- Device and resources under test
- Hardware & Software test bench
- Methodology
- Results & Discussion
- Summary

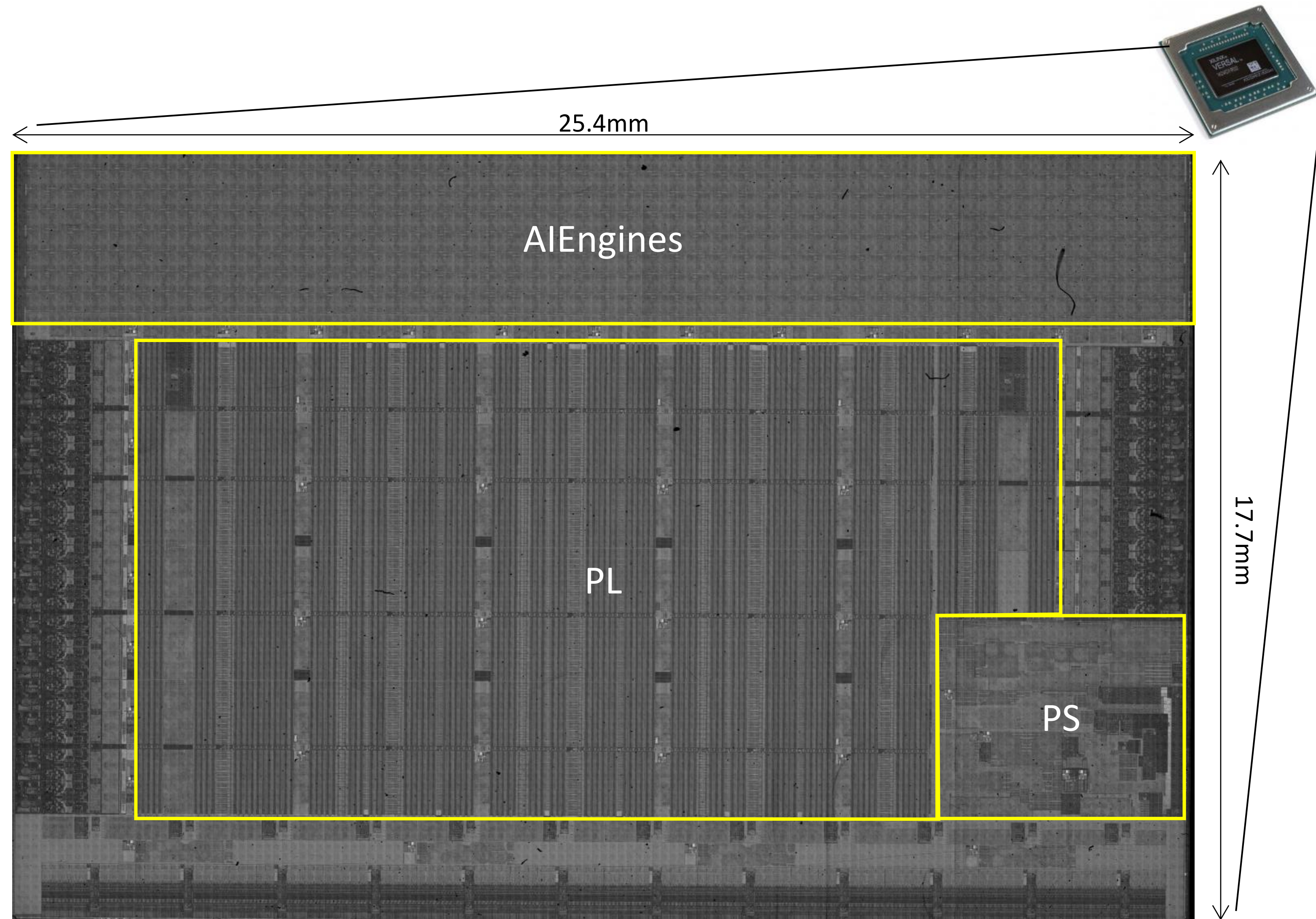
Device and resources under test

- Versal AI-Core System-on-Chip
  - 7nm FinFET technology
  - Complex monolithic SoC, embedding:
    - Multiple processing cores (PS)
    - Configurable FPGA-like logic (PL)
    - Network-on-Chip
    - AI dedicated engines (AIEs): array of 400 cores, each with local dedicated memory
- Resources under test in this work
  - PS
    - On-Chip-Memory (OCM)
      - 256 kB with ECC
  - PL
    - Configuration RAM
      - 363 Mb
    - Block RAM
      - 36 Kb block RAMs with ECC
    - Ultra-RAM
      - 288 Kb URAMs with ECC
  - AIEs including:
    - Instruction memory
      - 16 KB
    - Data buffers
      - 32 KB



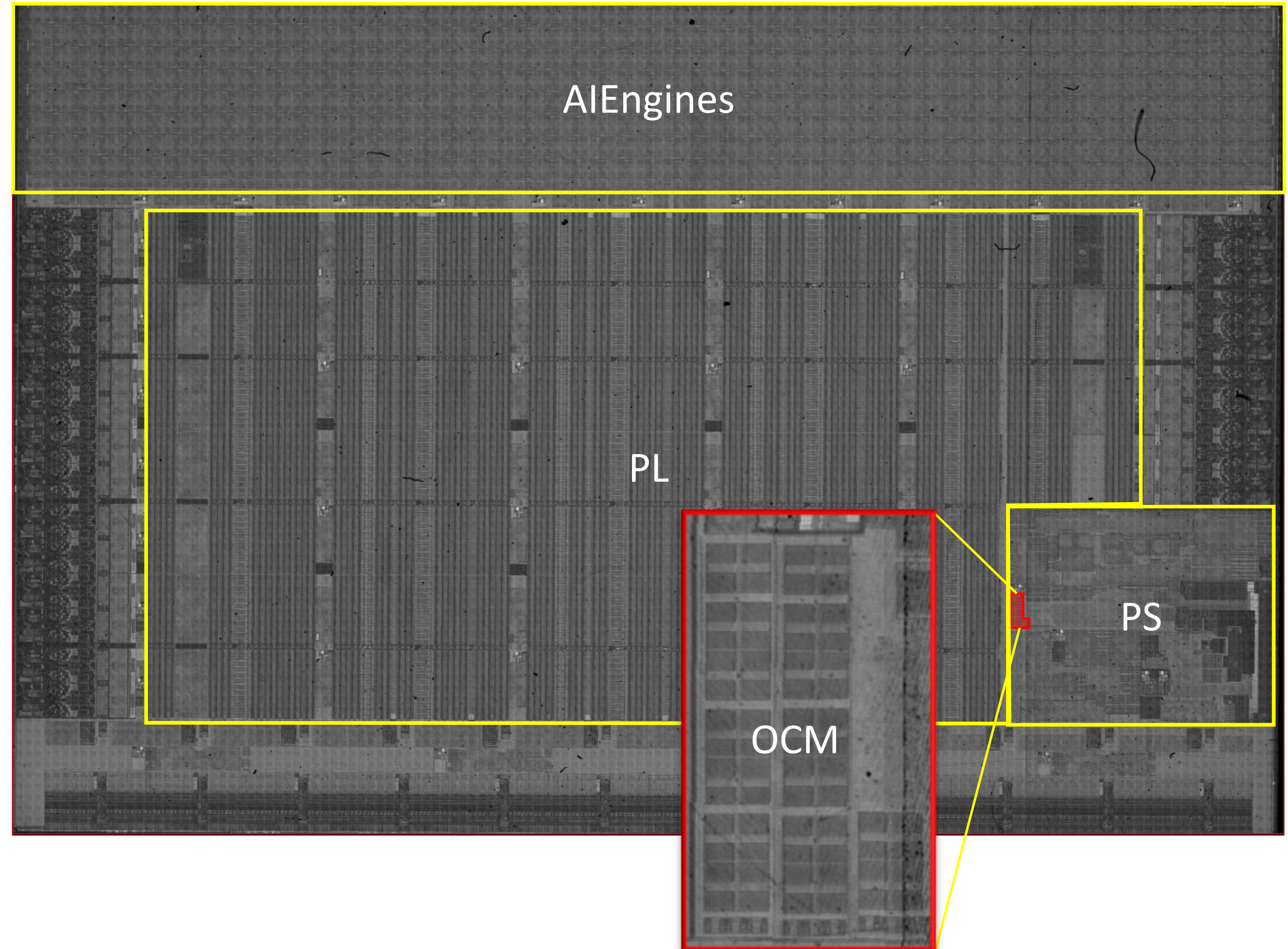
Device and resources under test

- Versal AI-Core System-on-Chip
  - 7nm FinFET technology
  - Complex monolithic SoC, embedding:
    - Multiple processing cores (PS)
    - Configurable FPGA-like logic (PL)
    - Network-on-Chip
    - AI dedicated engines (AIEs): array of 400 cores, each with local dedicated memory
- Resources under test in this work
  - PS
    - On-Chip-Memory (OCM)
      - 256 kB with ECC
  - PL
    - Configuration RAM
      - 363 Mb
    - Block RAM
      - 36 Kb block RAMs with ECC
    - Ultra-RAM
      - 288 Kb URAMs with ECC
  - AIEs including:
    - Instruction memory
      - 16 KB
    - Data buffers
      - 32 KB



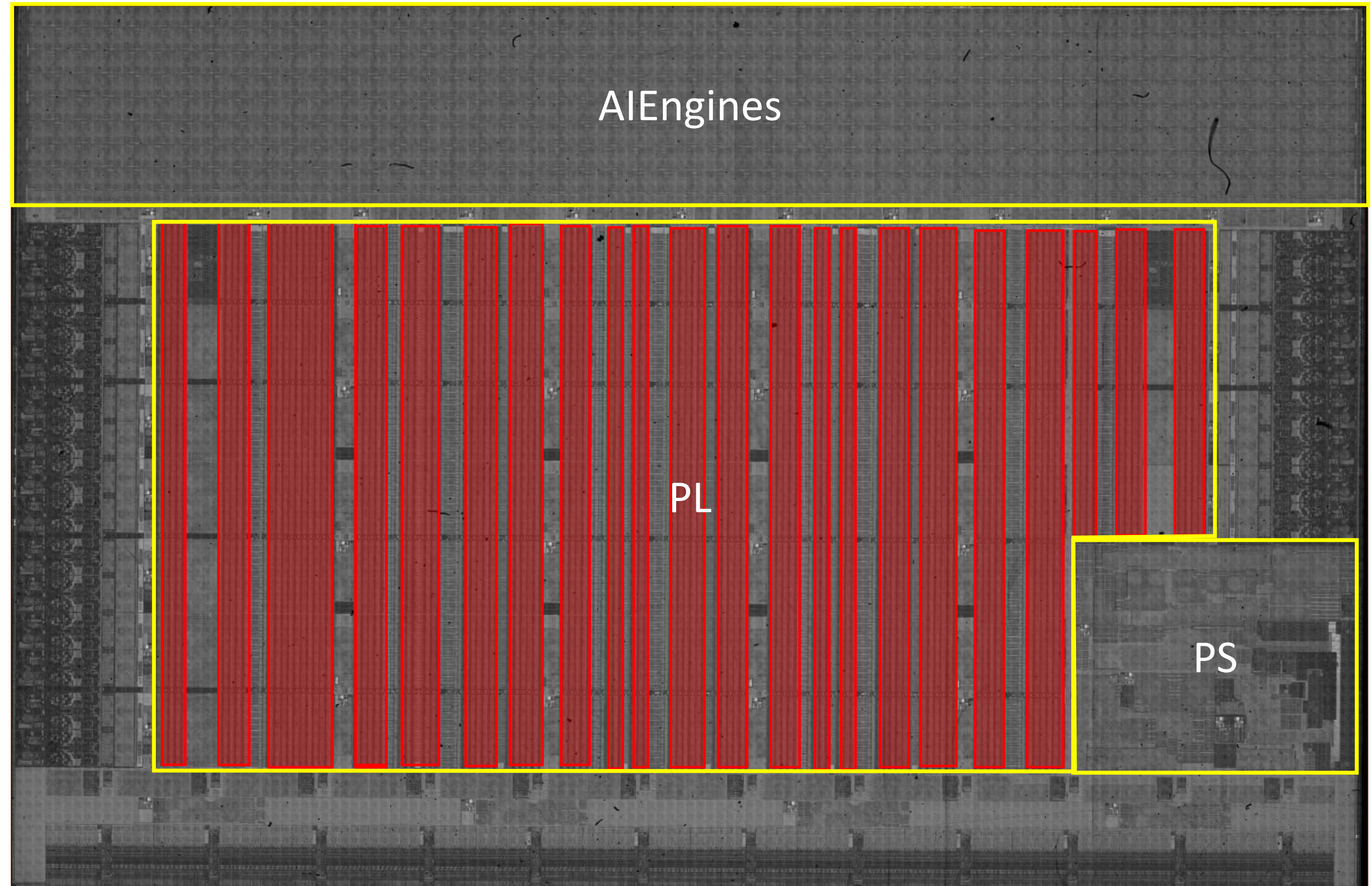
Device and resources under test

- Versal AI-Core System-on-Chip
  - 7nm FinFET technology
  - Complex monolithic SoC, embedding:
    - Multiple processing cores (PS)
    - Configurable FPGA-like logic (PL)
    - Network-on-Chip
    - AI dedicated engines (AIEs): array of 400 cores, each with local dedicated memory
- Resources under test in this work
  - PS
    - On-Chip-Memory (OCM)
      - 256 kB with ECC
  - PL
    - Configuration RAM
      - 363 Mb
    - Block RAM
      - 36 Kb block RAMs with ECC
    - Ultra-RAM
      - 288 Kb URAMs with ECC
  - AIEs including:
    - Instruction memory
      - 16 KB
    - Data buffers
      - 32 KB



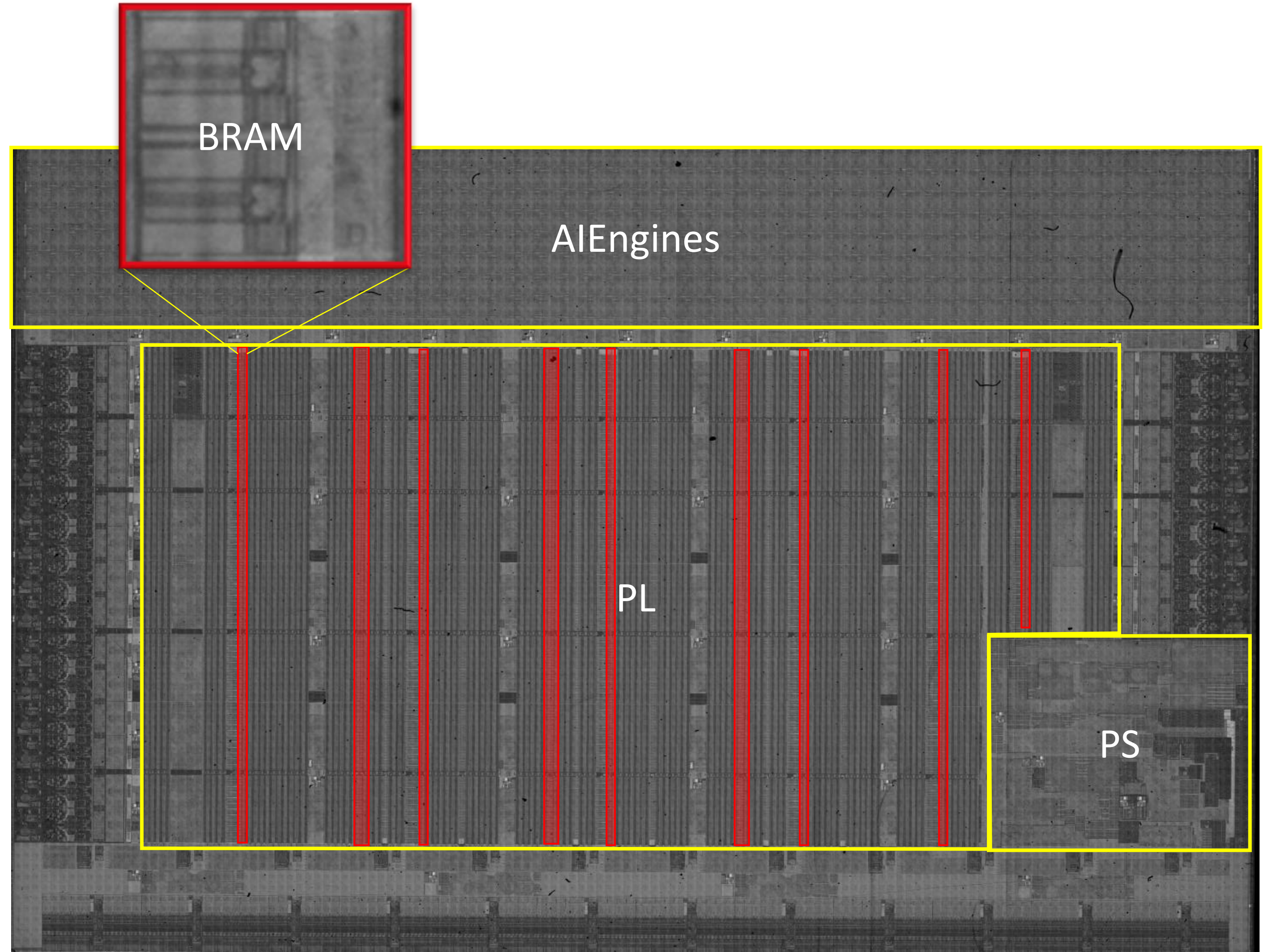
Device and resources under test

- Versal AI-Core System-on-Chip
  - 7nm FinFET technology
  - Complex monolithic SoC, embedding:
    - Multiple processing cores (PS)
    - Configurable FPGA-like logic (PL)
    - Network-on-Chip
    - AI dedicated engines (AIEs): array of 400 cores, each with local dedicated memory
- Resources under test in this work
  - PS
    - On-Chip-Memory (OCM)
      - 256 kB with ECC
  - PL
    - Configuration RAM
      - 363 Mb
    - Block RAM
      - 36 Kb block RAMs with ECC
    - Ultra-RAM
      - 288 Kb URAMs with ECC
  - AIEs including:
    - Instruction memory
      - 16 KB
    - Data buffers
      - 32 KB



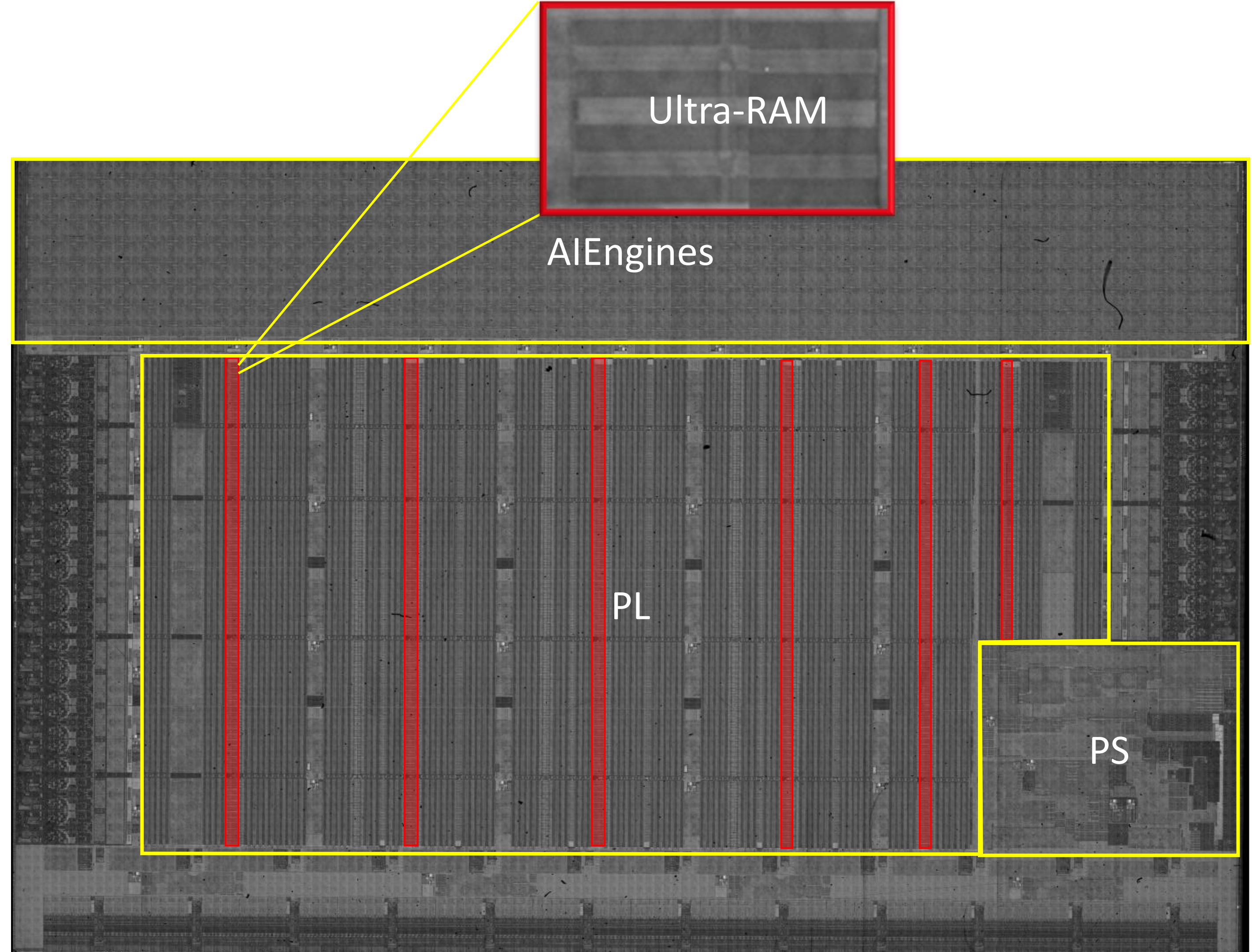
Device and resources under test

- Versal AI-Core System-on-Chip
  - 7nm FinFET technology
  - Complex monolithic SoC, embedding:
    - Multiple processing cores (PS)
    - Configurable FPGA-like logic (PL)
    - Network-on-Chip
    - AI dedicated engines (AIEs): array of 400 cores, each with local dedicated memory
- Resources under test in this work
  - PS
    - On-Chip-Memory (OCM)
      - 256 kB with ECC
  - PL
    - Configuration RAM
      - 363 Mb
    - Block RAM
      - 36 Kb block RAMs with ECC
    - Ultra-RAM
      - 288 Kb URAMs with ECC
  - AIEs including:
    - Instruction memory
      - 16 KB
    - Data buffers
      - 32 KB



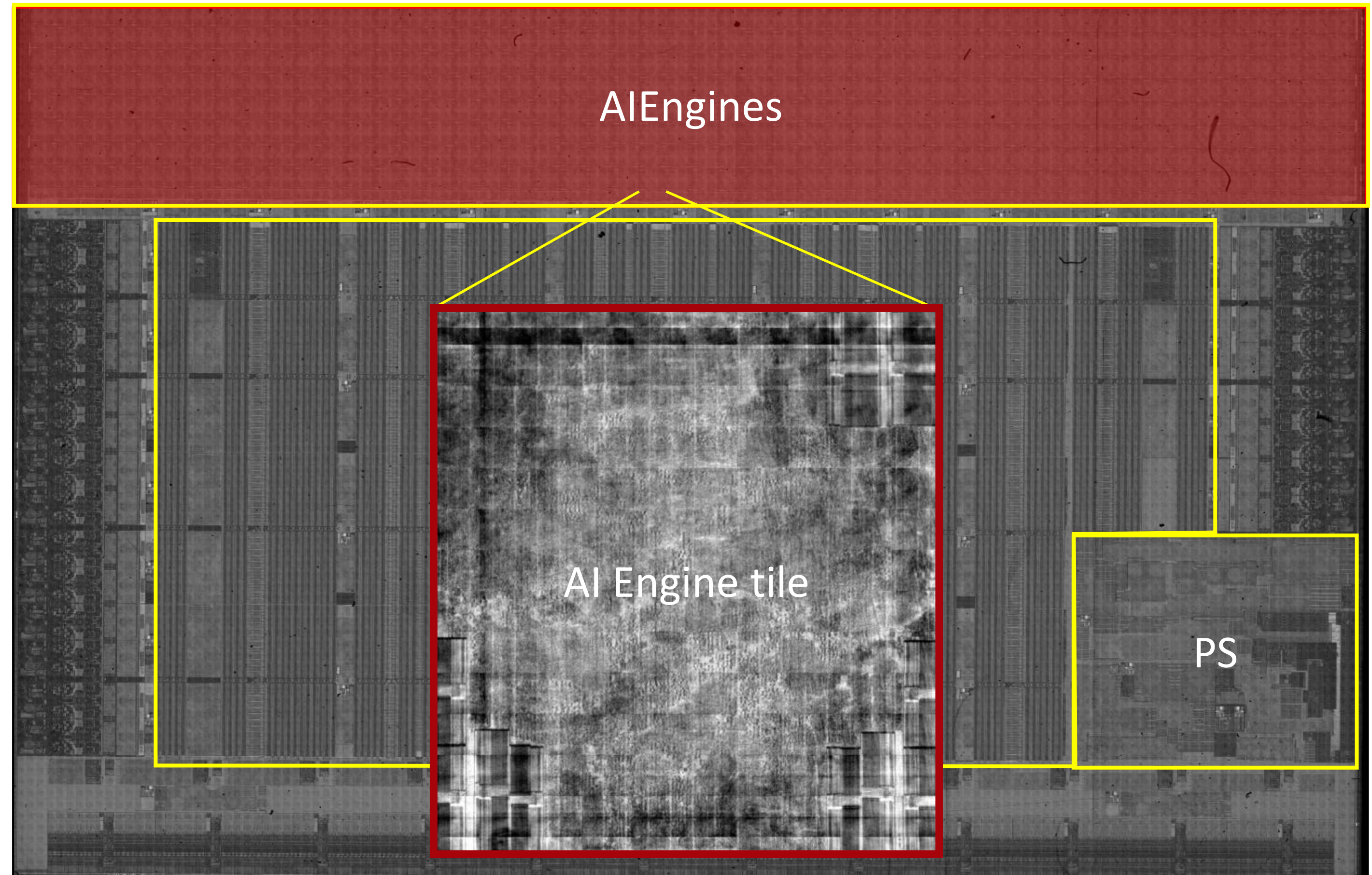
Device and resources under test

- Versal AI-Core System-on-Chip
  - 7nm FinFET technology
  - Complex monolithic SoC, embedding:
    - Multiple processing cores (PS)
    - Configurable FPGA-like logic (PL)
    - Network-on-Chip
    - AI dedicated engines (AIEs): array of 400 cores, each with local dedicated memory
- Resources under test in this work
  - PS
    - On-Chip-Memory (OCM)
      - 256 kB with ECC
  - PL
    - Configuration RAM
      - 363 Mb
    - Block RAM
      - 36 Kb block RAMs with ECC
    - Ultra-RAM
      - 288 Kb URAMs with ECC
  - AIEs including:
    - Instruction memory
      - 16 KB
    - Data buffers
      - 32 KB



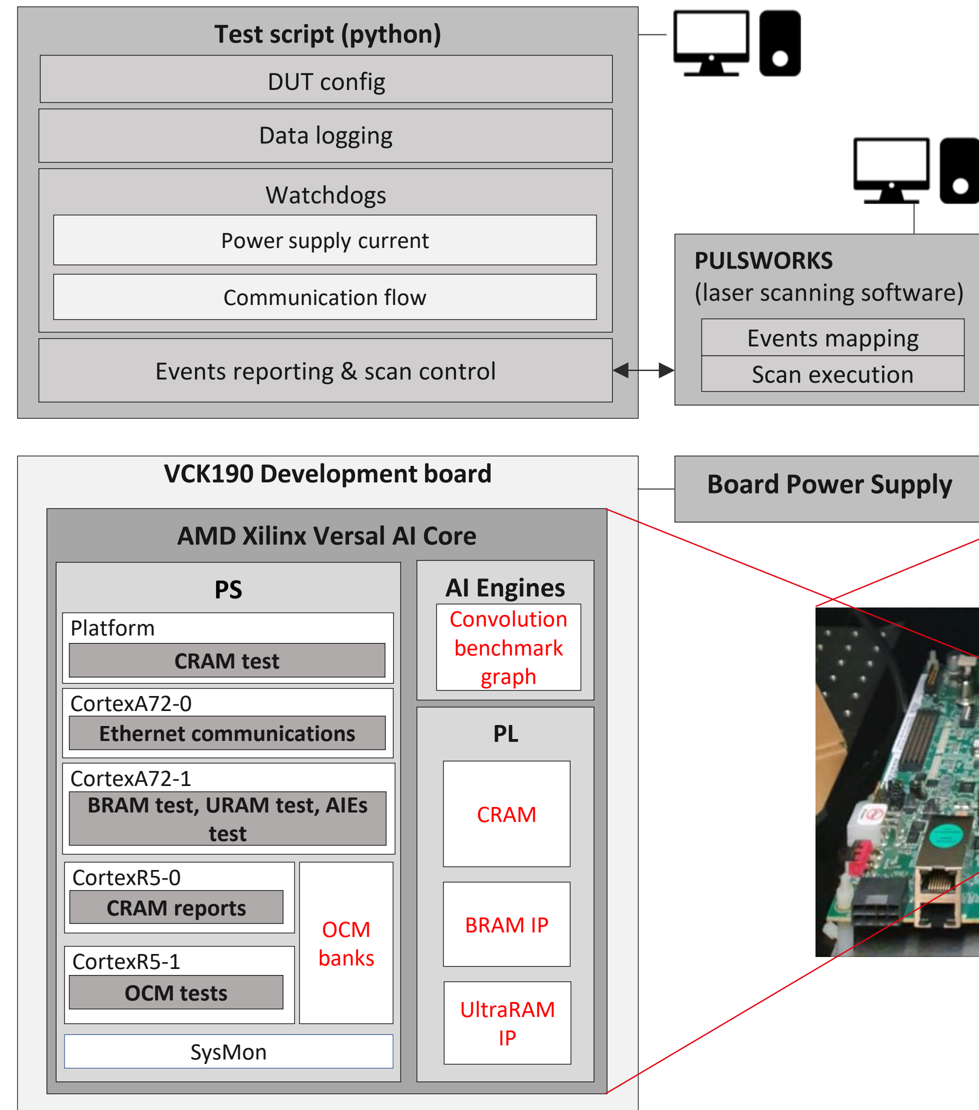
Device and resources under test

- Versal AI-Core System-on-Chip
  - 7nm FinFET technology
  - Complex monolithic SoC, embedding:
    - Multiple processing cores (PS)
    - Configurable FPGA-like logic (PL)
    - Network-on-Chip
    - AI dedicated engines (AIEs): array of 400 cores, each with local dedicated memory
- Resources under test in this work
  - PS
    - On-Chip-Memory (OCM)
      - 256 kB with ECC
  - PL
    - Configuration RAM
      - 363 Mb
    - Block RAM
      - 36 Kb block RAMs with ECC
    - Ultra-RAM
      - 288 Kb URAMs with ECC
  - AIEs including:
    - Instruction memory
      - 16 KB
    - Data buffers
      - 32 KB



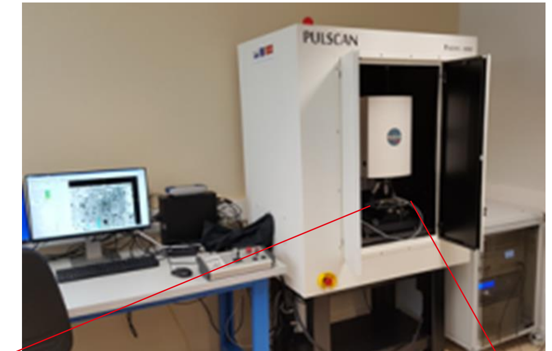
HW & SW Test bench

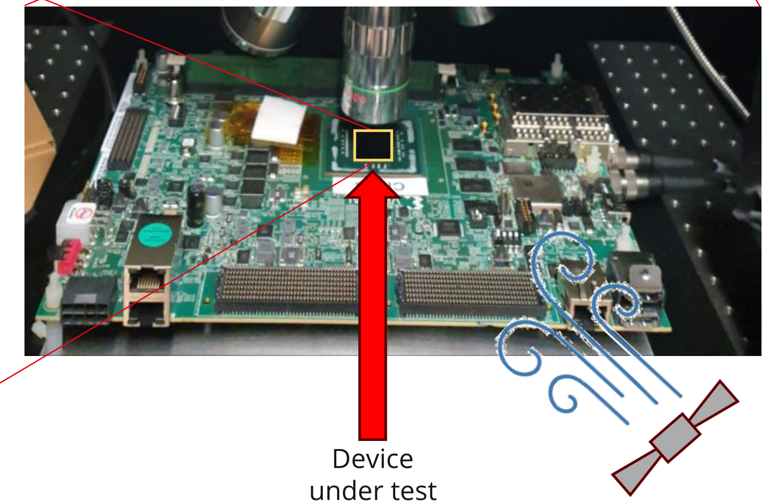
- Laser tests performed at IES using Single-Photon Absorption (SPA) through the substrate backside
- Sample backside preparation:
  - Heatsink removed, substrate thinned down to 85 µm and polished
- DUT self-test approach executed by embedded software running on PS cores
  - ECC in SRAMs is disabled
- The implemented graph consists of 32 AIEs performing a convolution operation in parallel
- Dissipation of DUT self-heating during test operations using a constant air flow at ambient T°
  - Temperature stabilized at 77°C



Laser parameters:

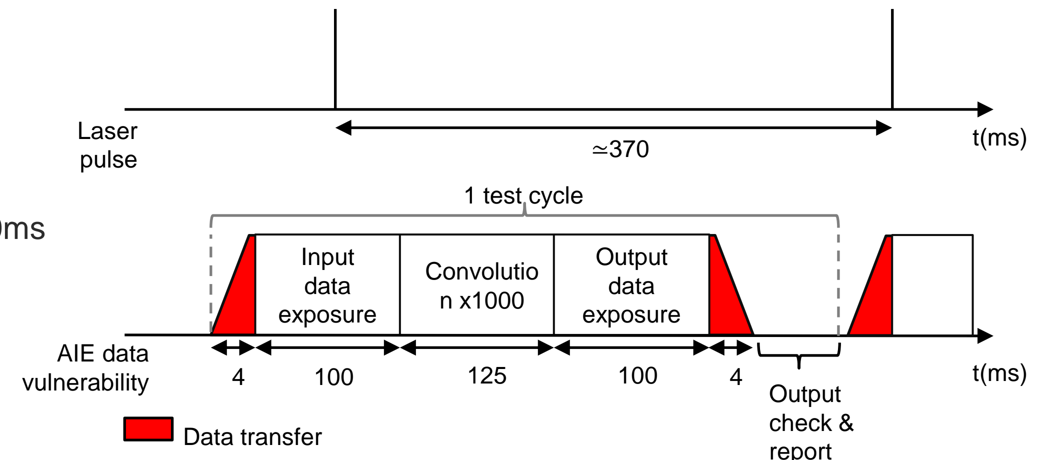
- wavelength: 1064nm (SPA)
- pulse duration: 30ps
- spot size (ϕ 1/e²): 1.1µm





Methodology

- High-resolution imaging of the whole DUT backside (10500 pictures) for navigation
- Laser tests are performed asynchronously, i.e no sync between laser pulses and test clock
- The laser pulse triggering frequency is adjusted in order to get 1 pulse per test cycle and scan position and to prevent cumulative heating
- For each SRAM resource:
  - SEU laser energy threshold measurement
  - Mapping & Laser cross section measurement
- For AI Engines:
  - Increase of vulnerability window to improve events statistics:
    - Delay between end of input data transfer and graph execution: 100ms
    - Delay between end graph execution and beginning of output data transfer: 100ms
    - Calculation repeated 1000 times (so that it takes ~100ms) per test cycle
  - SEU laser energy threshold measurement and mapping
- Typical scan parameters:
  - XY scanning steps: 0.5 x 1µm
  - Dimensions of regions of interest (scanned zones): A few 100s of µm x 100s µm for each resource



OCM

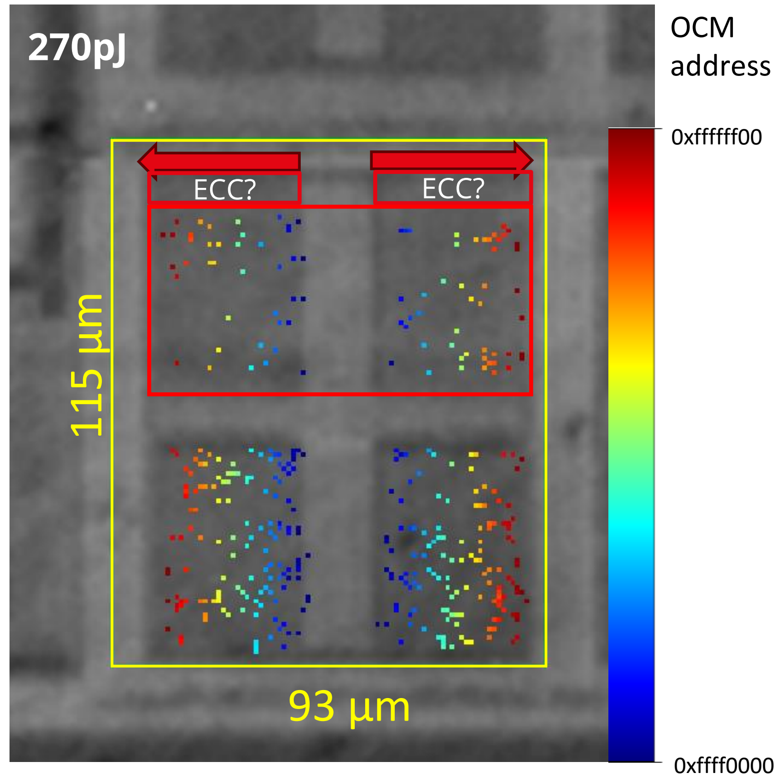
Mapping of 1/4 of Bank 3

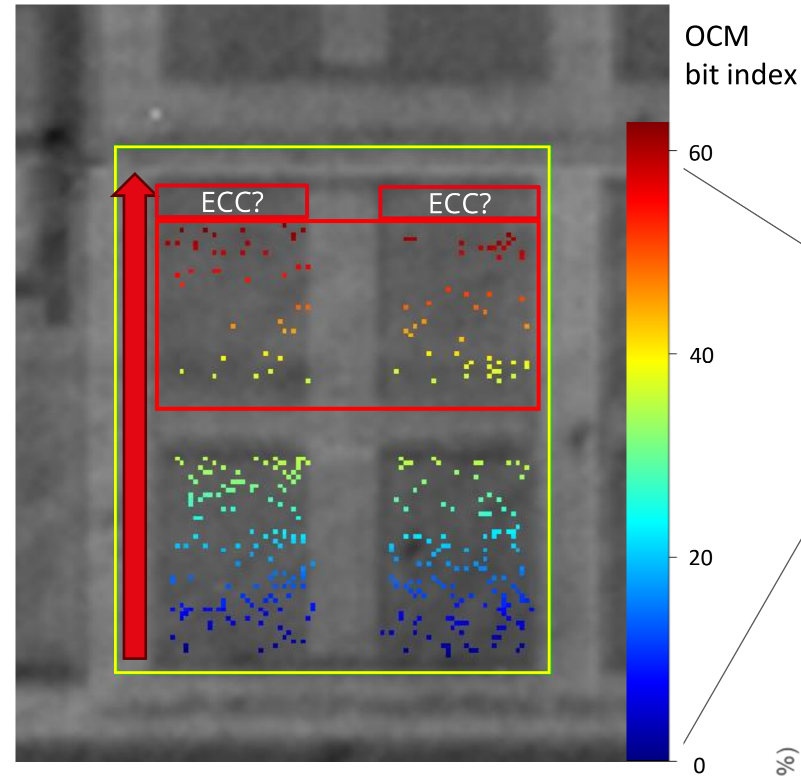
Estimated bit size
0.0363µm²

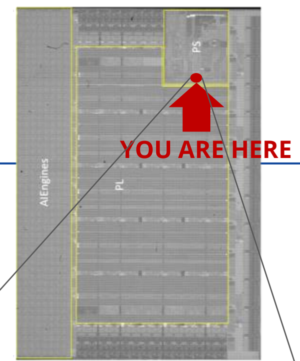
SEU energy threshold
230 ± 10pJ

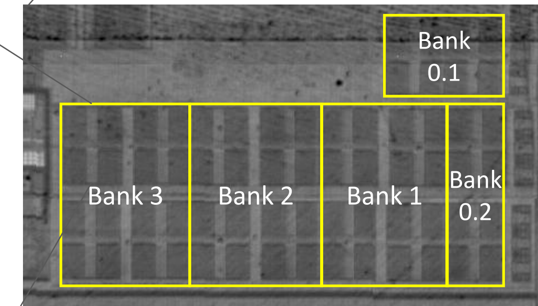
Observed SEUs
0→1 and 1→0

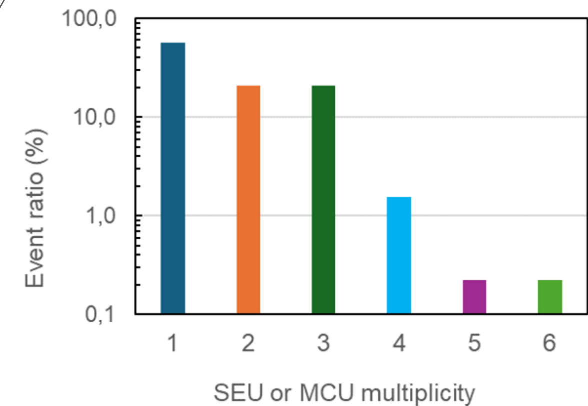
Max MCU multiplicity
6











- Address evolves horizontally in each OCM block
- The bit Index increases vertically from the bottom to the top
- A region without events: reserved for ECC bits?
- Inhomogeneity of the two top OCM blocks mappings due to the thermo-mechanical instability

BRAM

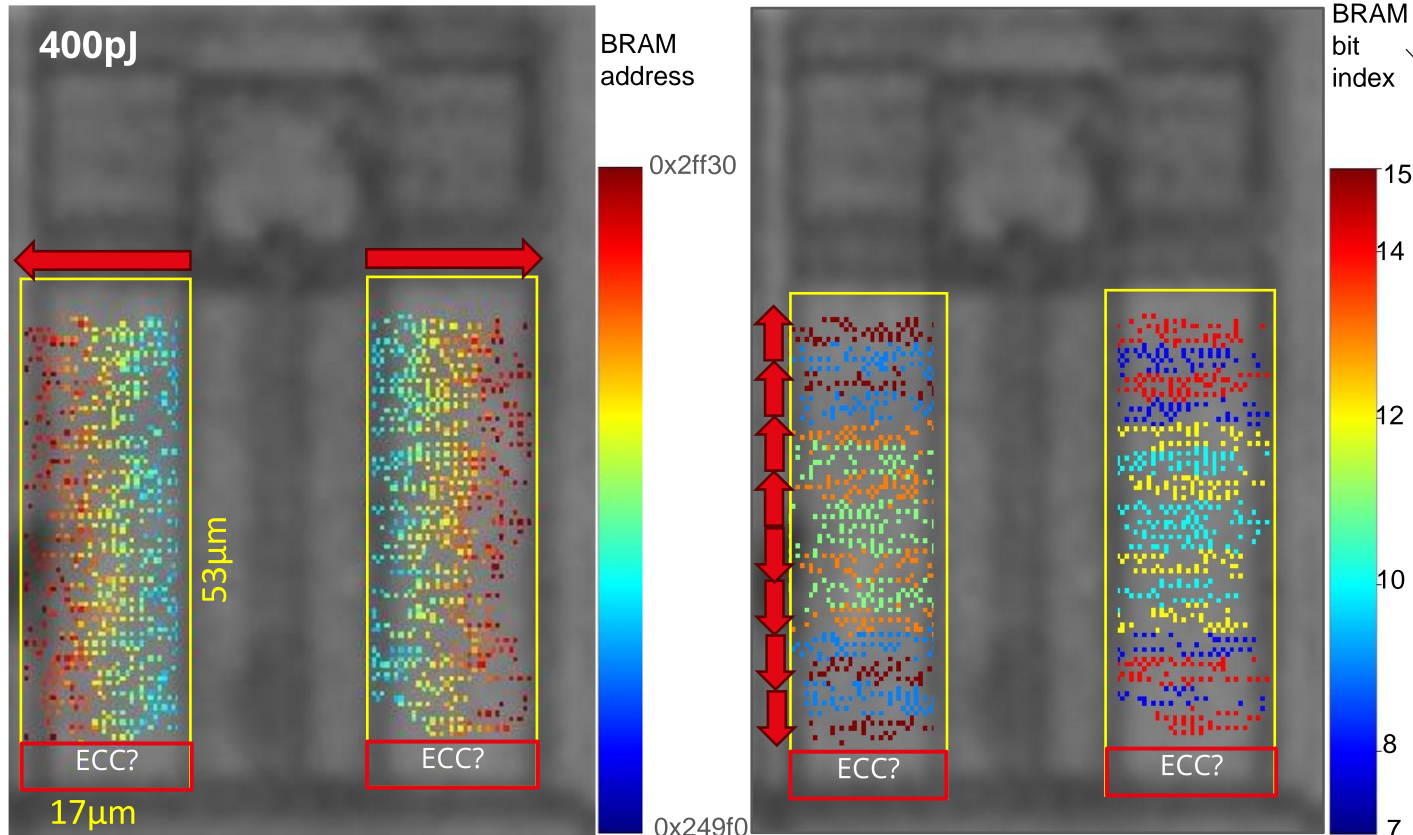
Estimated bit size
0.0978µm²

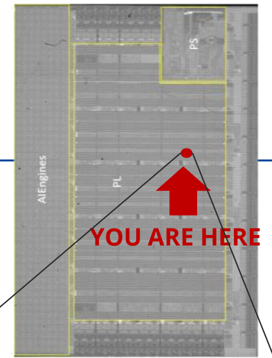
SEU energy threshold
280 ± 10pJ

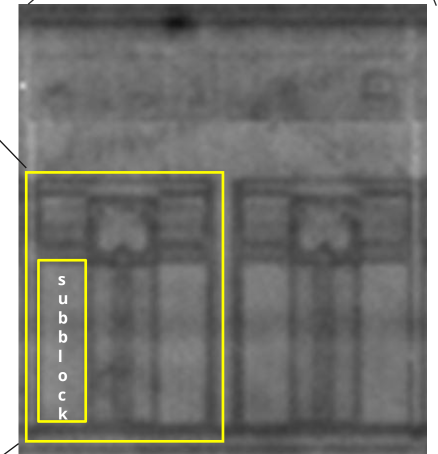
Observed SEUs
0→1 and 1→0

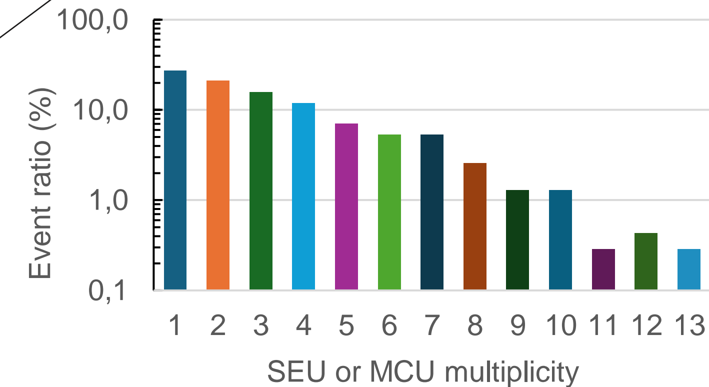
Max MCU multiplicity
13

Mapping of 1/2 of BRAM block





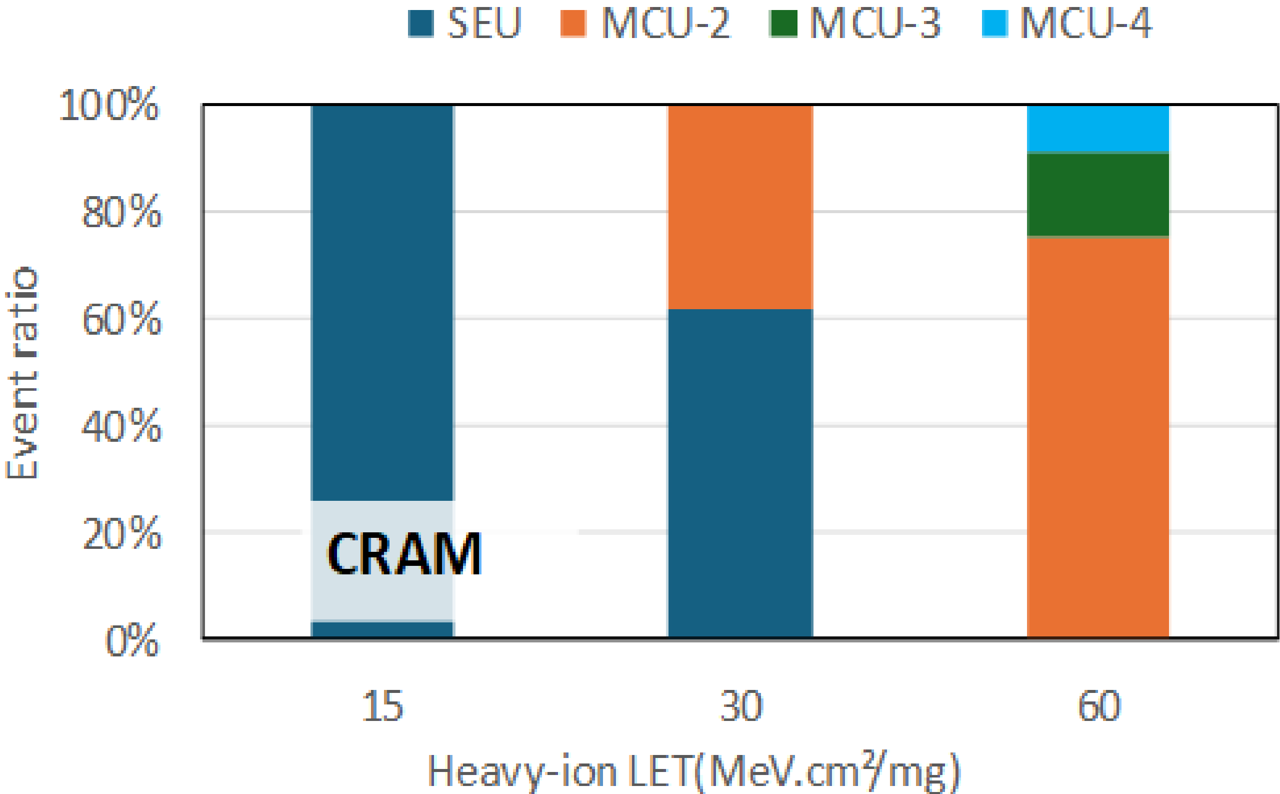


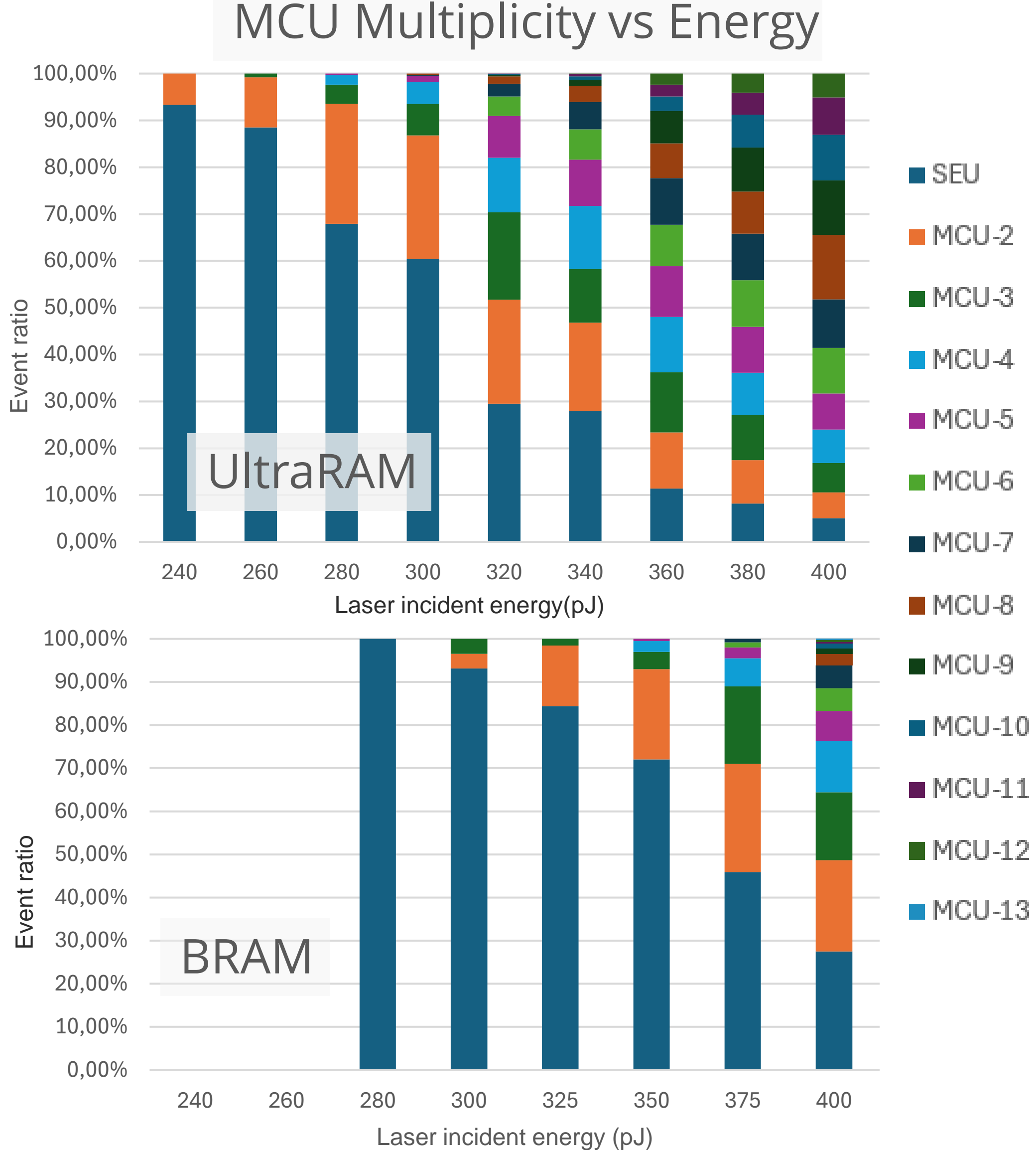


- Address evolves horizontally in each BRAM block
- The bit Index increases vertically in a symetrical pattern in each BRAM sub-block
- A region without events: reserved for ECC bits?

Events multiplicity

- SEUs observed despite the ratio between the laser spot and the elementary SRAM cell dimensions
- As expected, events multiplicity increases as the laser incident energy increases



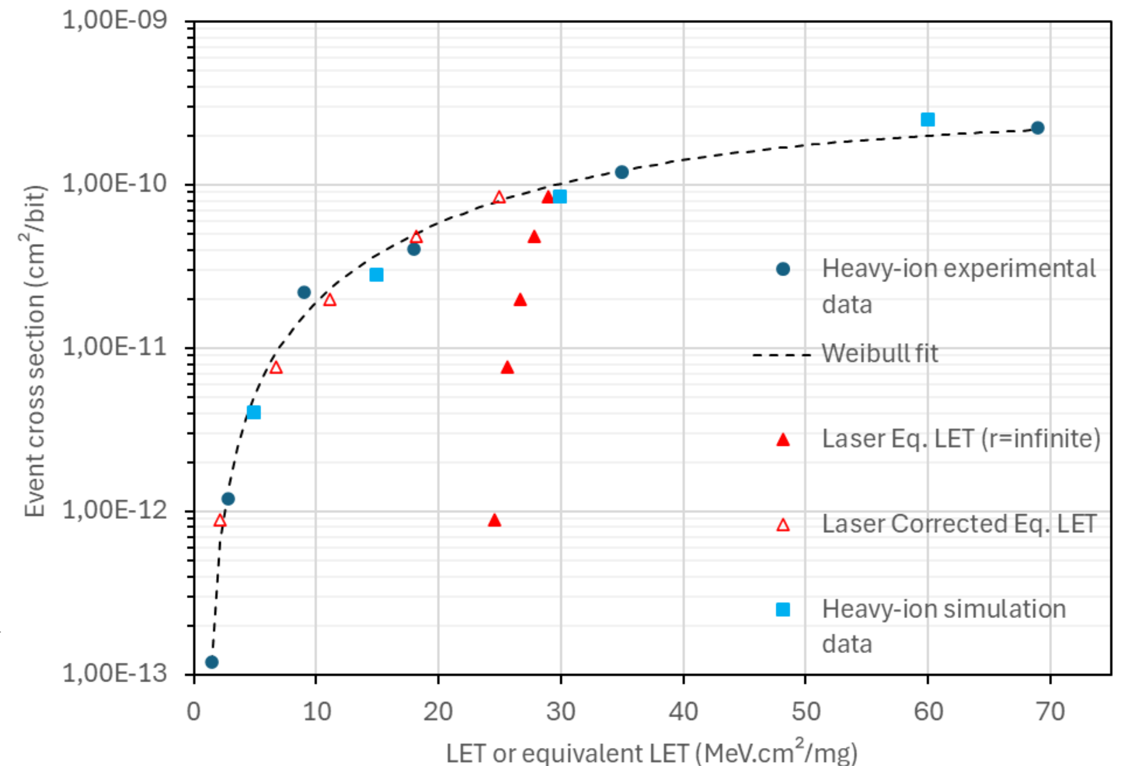


Correlation with heavy-ion data

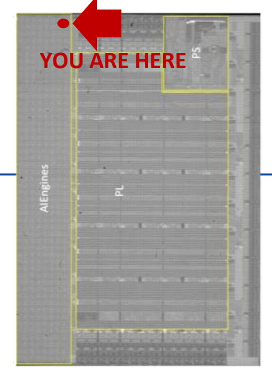
- Correlation of the laser results for the CRAM with heavy ion data from [*A. Dufour et al, RADECS 2023*]
- The laser equivalent LET threshold is much higher than the one measured with heavy ions when using infinite radius of integration
- We introduce a new empirical model to take into account the fact that the laser-generated charge collection efficiency has a limited radial extent
  - Introducing an energy-dependent correction factor

$$LET_{Corrected}(E) = \gamma \sqrt{\sigma_{Laser}(E)}\ LET_{Eq}(E)$$

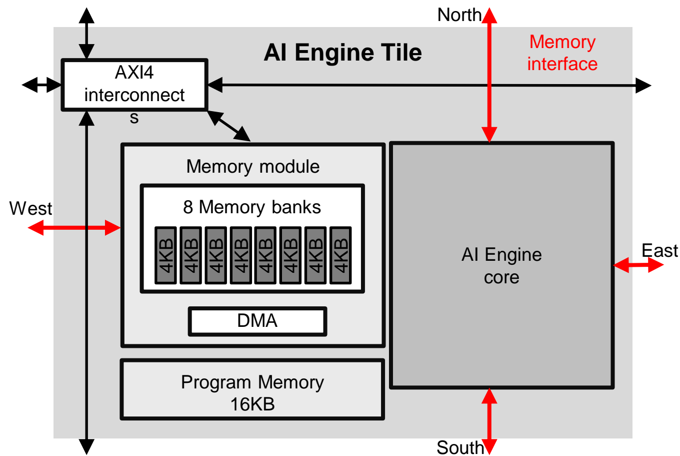
E the laser pulse energy
$\boldsymbol{\sigma_{Laser}}$ the laser cross section
γ a fitting factor



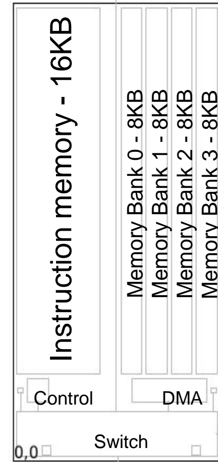
AIEngine – Ressources identification

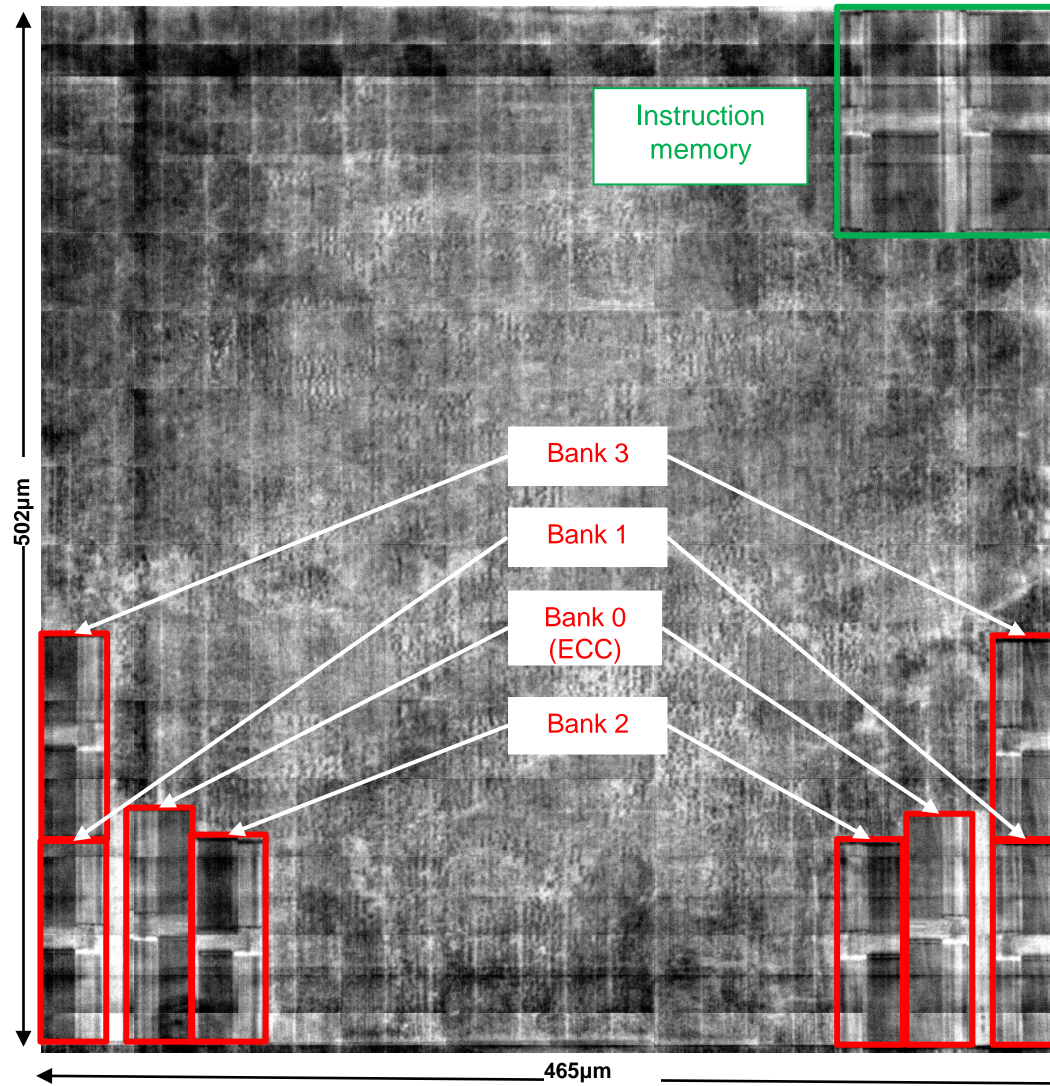


Hardware architecture of an AIEngine



Memory blocks in the manufacturer tool





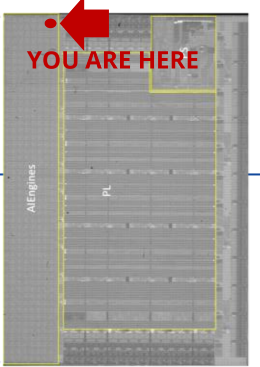
- Locations of the instruction and data memory banks are identified by analyzing laser SEUs and SEFIs fault signatures

INSTRUCTION MEMORY ESTIMATED BIT SIZE
0.039µm²

DATA MEMORY ESTIMATED BIT SIZE
0.033µm²

SEU ENERGY THRESHOLD
260 ± 10PJ

AIEngine – Mapping

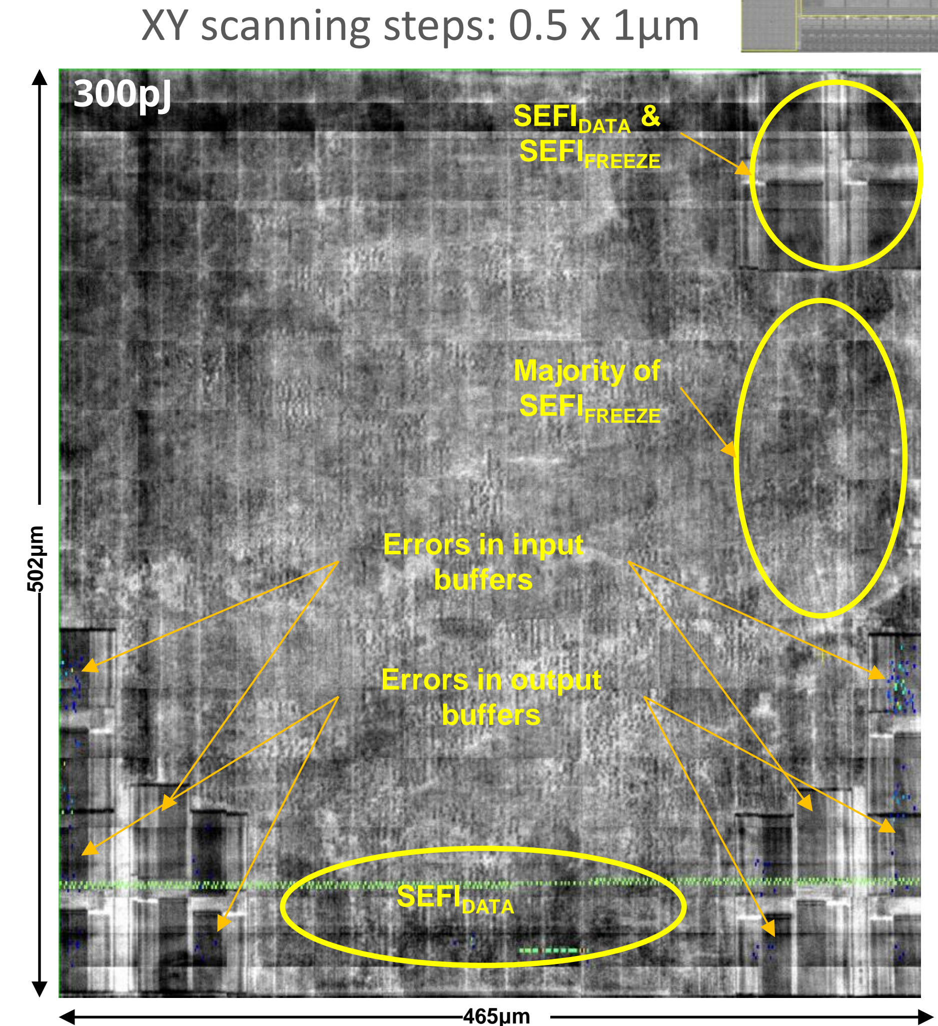


- Main observed events:
  - Single-event functional interrupts (SEFIs)
    - Freezes ($SEFI_{FREEZE}$)
    - Data-related events ($SEFI_{DATA}$) in the form of repeated error patterns over multiple test cycles
  - Upsets in output data buffers
    - Directly detected when comparing output with the golden data
  - Upsets in input data buffers
    - Detected by analyzing erroneous output data
      - Error pattern is a signature of the convolution filter

Example:

Input (1 error) ⊗ filter = Output (5 errors)

1	0	1
0	1	0
1	0	1



Test challenges

- Big chip with many different resources
  - The test and scanning strategy must be well thought
  - Trade-off between mapping resolution and test loop duration
- Embedded test software complexity for stimulating the AI Engines
  - Limited documentation for bare-metal design flow
  - Compilation/synthesis time required for every minor change in the kernels code
- Optimization of the vulnerability window of AI Engines is required for good events statistics
- Thermo-mechanical stability issues
  - Increasing with the device workload especially when increasing the number of active AI Engines
    - Active cooling required for testing a complete application graph
  - Focus variations related to the important initial warpage of the die

Summary

- SEE laser testing of the main resources of the Versal, a 7nm FinFET complex SoC
- The capability of the SPA laser testing technique to generate single-bit and multiple-cell upsets is confirmed @ 7nm
- The address- and bit-maps of various SRAM resources were extracted, providing useful insights on the physical organization of the device
- Locations of the main resources of an AI Engine, the instruction memory and data memory, are identified
- Different types of events in the AIE instruction memory and data buffers were observed
- Future work: test of a ResNet accelerator graph implemented in the AIEngines using active cooling of the die

Single-Event Effects Laser Testing of a 7nm FinFET System-on-Chip with AI-Acceleration Capabilities

S. Achaq[1,2], V. Pouget[2], L. Artola[1], G. Hubert[1], A. Urena[1], F. Manni[3], A Dufour[3], J. Boch[2]

1, ONERA, Toulouse, France. 2, IES, Univ. Montpellier, CNRS, Montpellier, France. 3, CNES, Toulouse, France