National Science Foundation's Industry/University Cooperative Research (I/UCRC) Program

B1-24: Fault-Tolerant Techniques for Heterogenous Computing Architectures

Mission-Critical Computing
NSF CENTER FOR SPACE, HIGH-PERFORMANCE, AND RESILIENT COMPUTING (SHREC)

SHREC Annual Workshop (SAW23-24)

University of Pittsburgh
BYU Brigham Young University
Virginia Tech
UF University of Florida

January 17-18, 2024

Dr. Mike Wirthlin & Dr. Jeff Goeders

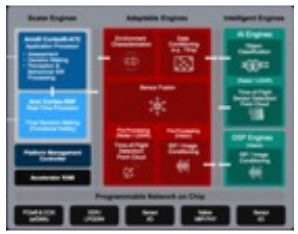
Nathan Baker, MS
Tyler Ricks, MS
Garrett Smith, BS
Jacob Brown, BS
Ethan Campbell, BS
Zach Driskill, BS
Julia Hansen, BS
Caleb Price, BS
Sam VanDenBerghe, BS
Rami Arafah, BS
Collin Lambert, BS

Number of requested memberships ≥ 4

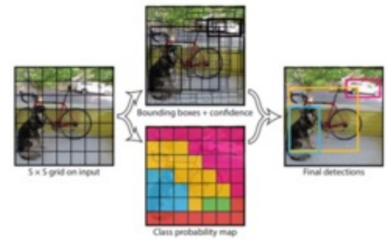
B1

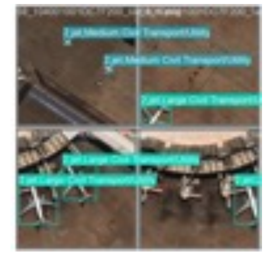
Project Tasks

Task 1: Versal ACAP Reliability

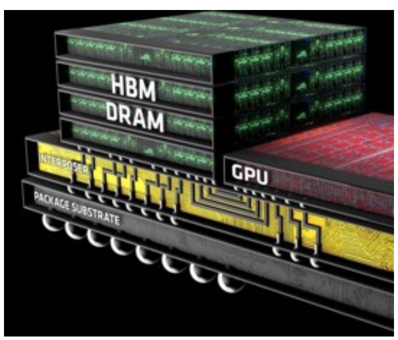


Task 2: Reliable Deep Learning

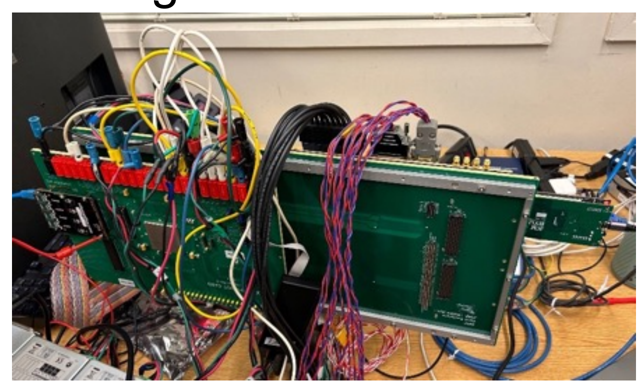




Task 3: High Performance Memory Reliability



Task 4: Radiation Testing of Heterogenous Devices

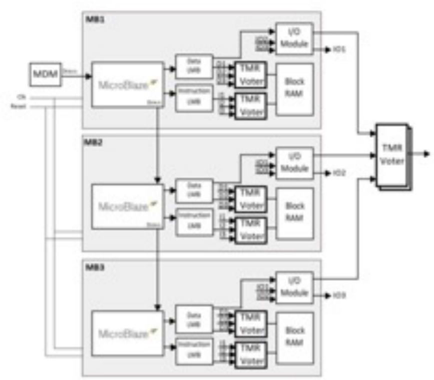


Task 1 – Versal ACAP Reliability

B1

- AMD announced plans for Versal ACAP devices for space and military applications (XQR Versal)
  - Machine Learning Inference
  - On-board data processing
  - High-speed I/O interfaces

- Support reliability of Versal ACAP Platform
  - Provide documentation and member support for Versal
  - Support scrubbing modes
    - Versal SMAP scrubber
    - XilSEM scrubbing
  - Fault tolerant firmware
  - Support XRTC Versal radiation testing



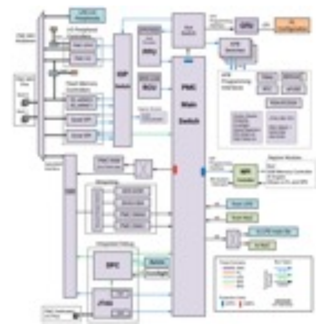
	XQRVC1902	XQRVE2302
AI Engines	100	17
DSPs	1,968	464
Logic Cells (K)	1,968	329
DDR Controllers	4	1
PL Memory (Mb)	191	86
Gigabit Tx/Rx	44	8

Dual core A72, Dual core R5F, 256 KB OCM

B1

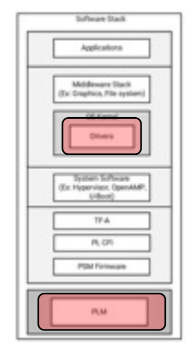
Fault Tolerant Versal PLM Firmware

- Versal PLM firmware *essential* to Versal reliability
  - Failures in PPU/PLM will bring down entire system
  - Fault tolerance features must be enabled *and* tested
- Versal Firmware Enhancements
  - *Active* scrubbing of PPU and PLM RAM and registers
  - Improved XilSEM scrubbing (SEFI handling)
  - Watchdog management of key PMC functions
  - Fault tolerant PSM firmware (using PPU approach)
- Fault Tolerant Linux for Versal
  - Integrate enhanced Versal firmware into Linux image
  - Provide Linux hooks to support reliability features
    - PL CRAM scrubbing and logging
    - Priority memory scrubbing



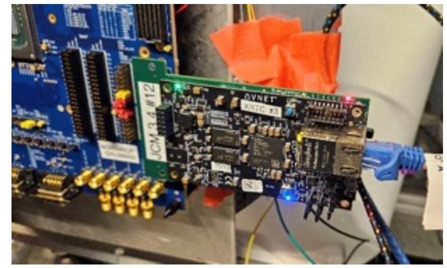
Versal PMC
- Booting
- Security
- Configuration
- Scrubbing

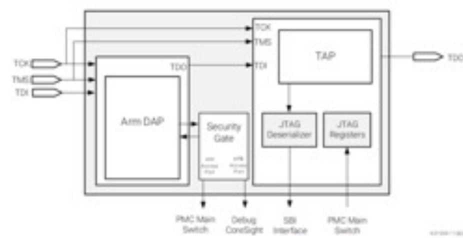
Firmware stack
- Scrubbing
- Memory ECC
- Watchdog

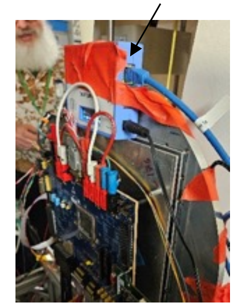


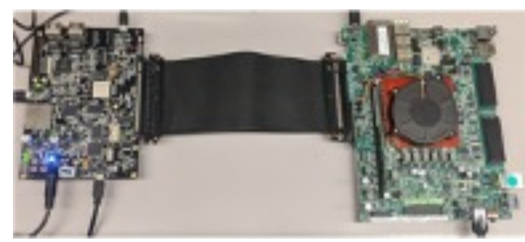
Versal JCM and Scrubbing Support

- Improved SMAP JCM scrubbing
  - Scrubbing of SMAP/SBI registers
  - Support PLM SMAP timeout and recovery
  - Improve SMAP operating speed
- JCM support for Versal DAP port access
  - Extract processor state much quicker than SmartLynqs
    - Efficient Versal memory extraction essential for radiation testing
    - Read internal memory, processor registers, and PLM state
  - Implement AMD/Xilinx "Hardware Server" in JCM
- High-seed PCIe Scrubbing
  - Perform PCIe scrubbing in bare metal on PolarFire
    - Previously required Linux on PolarFire
  - Integrate PCIe scrubbing on Versal XRTC board





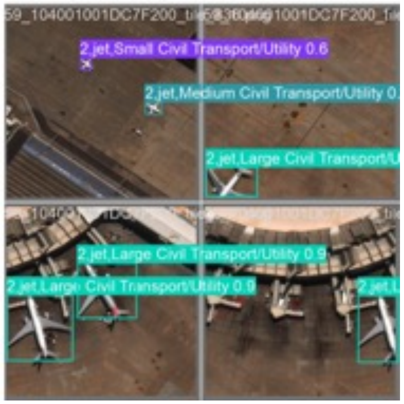




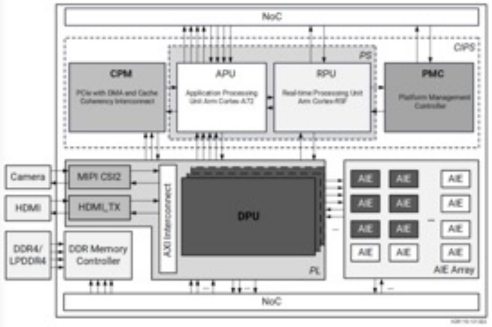
B1

Task 2 - Reliable Deep Learning

In 2023 we began exploring using the Versal ACAP devices with AI Engines to run machine learning benchmarks. This year:



- Experiment with different DPU configurations
  - DPUCVDX8G has several parameters to determine utilization of AI engines and other hardware resources
  - Evaluate the effect of DPU configurations on throughput and latency of several Yolo models
- Fault Injection & Radiation testing
  - Investigate whether faults can be injected directly into AI engines
  - Use fault injection to investigate the impact on deep learning behaviors
  - Measure the impact to YOLO prediction accuracy
- Investigate custom HDL implementation of YOLO
  - So far we have been using Xilinx/AMD's DPU, which is a configurable hardware IP.
  - We plan to investigate whether we can obtain better performance with a custom hardware design.
- Generate predictions from live camera feed
  - To date, we have been running on static images in memory
  - Creating a full system with a live feed will allow us to perform reliability experiments on a more complete system.



Bare Metal AI-Engine Designs

Goal: Create bare-metal designs that generate AI-engine traffic

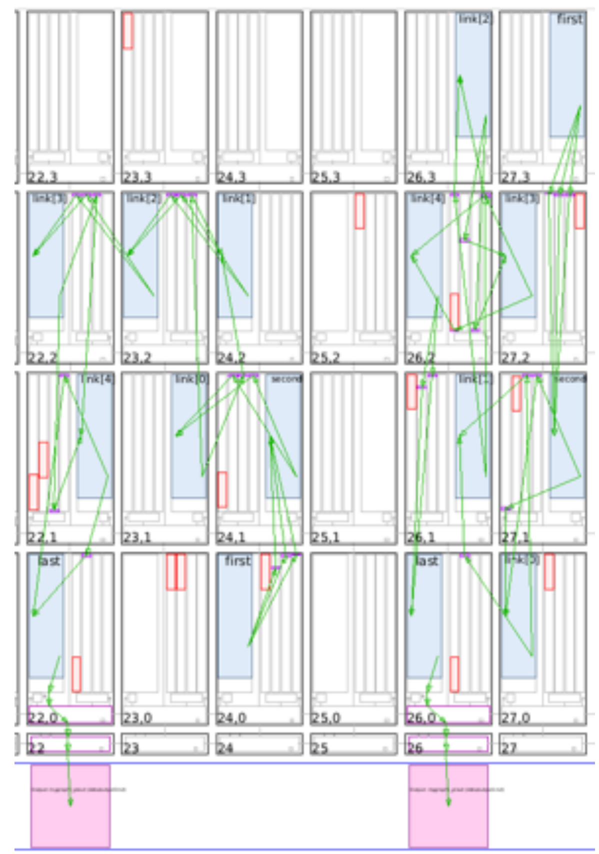
- Provides more fine-tuned control over the AI Engines
- Useful for radiation testing

Approach:

- Incorporate AI Engines into designs with PL and ARM cores
- Use bare metal C/C++ toolchain in Vitis
- HLS is used for hardware kernels in the PL
- Generate designs with various access patterns, throughput, and latency.
- Data can be chained through multiple AI engines to create complex data movement patterns.

We have started testing small designs, and are working to scale to larger systems.

- Example: On the right is a set of two identical and parallel designs, each consisting of a hardware kernel (*pink*) and resources from 8 AI Engines (*blue*). Visuals generated by Vitis Analyzer.



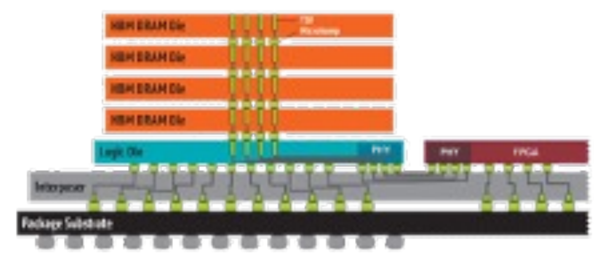
Task 3 – HBM Reliability and Performance

B1

In 2023 we focused on techniques to warm up HBM FPGAs to use them in below freezing environments. This year:

1. Continue exploring HBM in harsh environments
   - Investigate data integrity at different temperatures
2. Benchmarking FPGA HBM performance
   - Investigate performance at different clock rates, sequential vs random accesses, access patterns through the crossbar, in the presence of contention, etc.
   - Test with benchmark applications

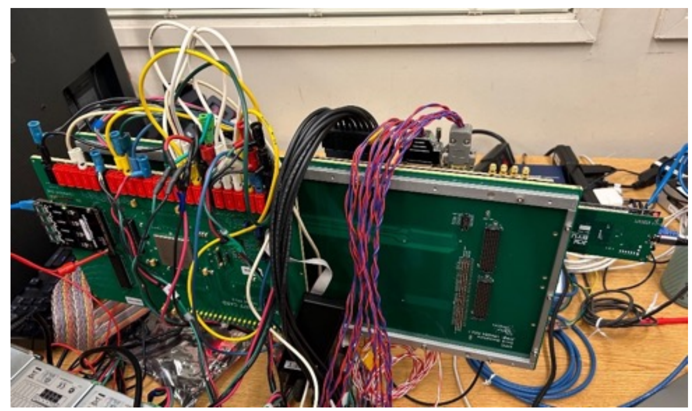
- Investigate HBM reliability
  - Perform fault injection on FPGA HBM controller
  - Radiation testing with high-throughput HBM benchmark
    - Capture error rates of HBM
    - Investigate impact of basic parameters (traffic level, number of channels used, etc)
    - Determine whether failure modes observed during fault injection are reproduced in radiation beam



B1

Task 4 – Radiation Testing

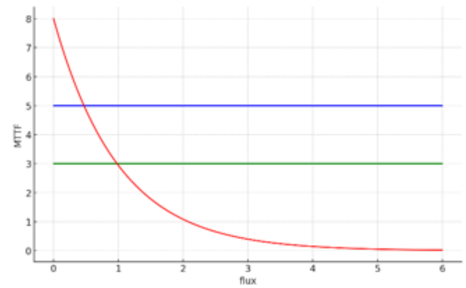
- Radiation testing necessary for understanding complex device failures
  - Identify failure mechanisms and single-event functional interrupts
  - Measure improvement of fault tolerant techniques
- Novel radiation testing methodologies needed for complex heterogeneous devices
  - High-flux testing approaches
  - Simultaneous device testing strategies
  - Low cross-section technologies
- Dedicated tests for Versal
  - Improved Versal firmware
  - Versal AI bare metal testing
  - Linux Versal



B1

Complex Device Testing Strategies

- High-Flux Processor Testing
  - High-flux testing distorts reliability of ”mitigated” systems
    - Relationship between failure rate and flux increases non-linearly
    - Mitigated systems with “repair” have flux limits
  - Increased “repair” rate needed for processor testing
    - High memory scrub rate
    - Fast TMR recovery approaches

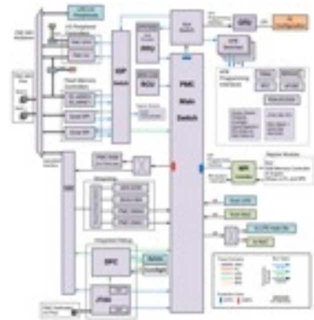


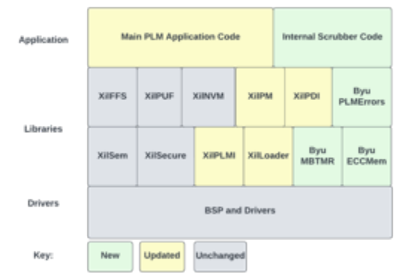
- Parallel System Component Testing
  - Complex device have *many* components requiring testing
  - Multiple components need to be tested simultaneously
    - Processors, Programmable Logic, AI Engines, etc.

- Post Radiation Fault Injection
  - Extract additional information from radiation test through “replay”
  - Compare fault injection “replay” with test behavior

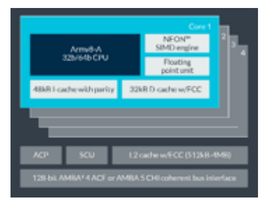
Versal Radiation Test Experiments

- Versal Firmware Test
  - PLM memory scrubbing
    - Reduce/eliminate PPU hang/failures
    - Facilitate failure recovery
  - PLM Watchdog recovery testing
  - XilSEM "unrecoverable error" recovery
- Versal Processor Testing
  - Active on-chip memory scrubbing
  - DAP controller failure analysis
  - Reliable Linux testing
- Component Testing
  - AI engine reliability
  - NOC reliability
  - DDR controller reliability

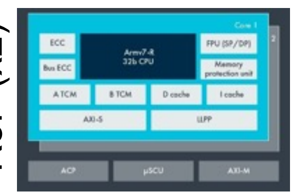


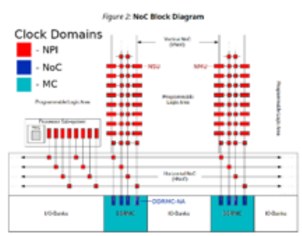


A27 (x2)

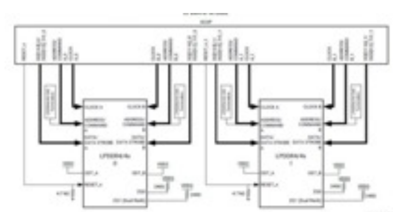


R5F (x2)

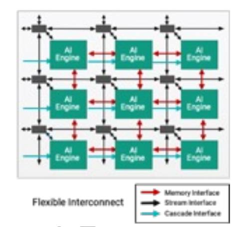




NOC



DDR



AIE

B1

Anticipated Radiation Test Experiments

- Berkeley National Laboratory (Heavy Ion)
  - Versal Reliability (multiple components tested)
    - F/T PLM Firmware, F/T Processor support
    - AI engine reliability, NOC reliability
    - High-flux processor testing methodologies
  - Anticipated date: February (likely others)

- ChipIR, UK (Neutron)
  - HBM controller reliability
  - Versal Neutron testing
  - Processor testing methodologies
  - Post-radiation fault injection
  - Anticipated date: June (pending proposal)

B1

Questions?