Space DPU*: Constructing a Radiation-Tolerant, FPGA-based Platform for Deep Learning Acceleration on Space Payloads

*DPU = Deep Learning Processing Unit

Presenter:

Jason Vidmar
System Architect - MILCOM, Satcom & Machine Learning
Xilinx Aerospace & Defense

Presentation Contributions By:

Dr. Pierre Maillard, Troy Jones, Minal Sawant, Giulio Gambardella, Nicholas Fraser
Xilinx

June 15, 2021

XILINX

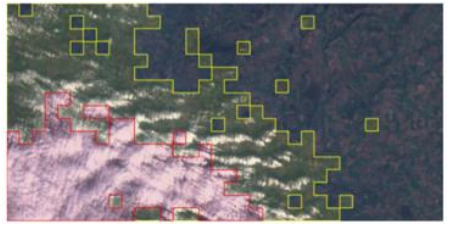
OBDP 2021
European Workshop on On-Board Data Processing
cnes DLR esa
14 - 17 June 2021 | Online Event

Agenda

- What is the “Space DPU” Platform and What Problem Are We Addressing?
- Main Architectural Elements and Design Methodology
- Experiment & Test Results
- Key Take-aways & What’s Next

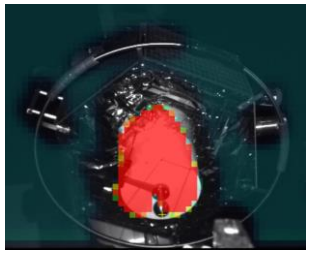
Machine Learning for Space Applications

- Dramatic advancements in Artificial Intelligence (AI), Machine Learning (ML) and Deep Learning (DL) capabilities are disrupting many industries.
- AI/ML/DL being targeted for Space across multiple areas:
  - Computer Vision (e.g., cloud detection, earth observation, weather intelligence, docking/landing assist, etc.),
  - Robotics (e.g., autonomous systems),
  - Diagnostics and Predictive Maintenance (e.g., anomaly detection),
  - Communications (e.g., RF link optimization),
  - Flight Control (e.g., space debris avoidance),
  - Planning and Scheduling,
  - Scientific Analysis



Cloud Detection
(Craft Prospect, Ltd.) [1]

Object Detection in
Satellite Imagery [2]



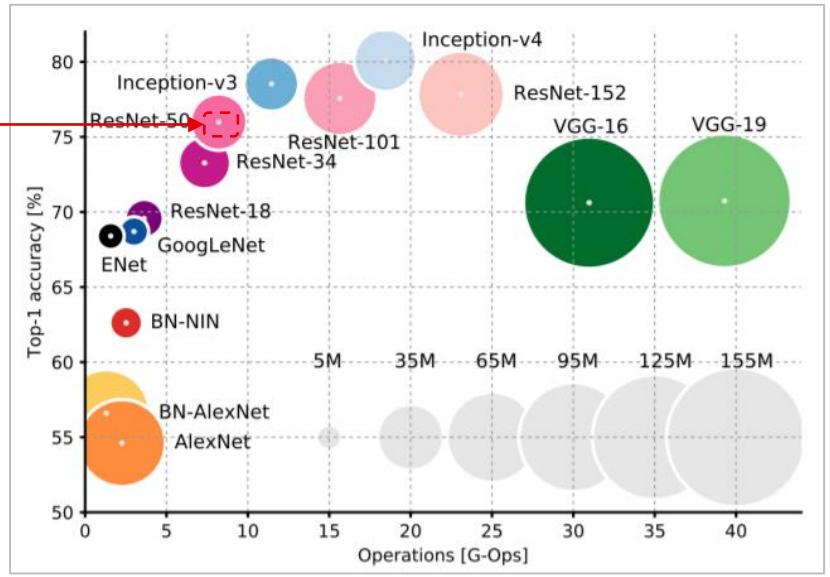
ML-aided Docking Assist [3]

Industry Push to Scale Terrestrial AI & ML Capabilities to Space-based Platforms

XILINX.

FPGAs for Space-based Deep Learning Inference

CNNs Require Lots of Compute and Memory:



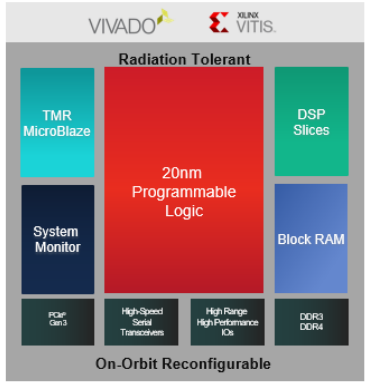
Comparison of CNN Compute and Memory Costs. Source: A. Canziani, et al, 2017. https://arxiv.org/abs/1605.07678 [4]

For ResNet-50 (unpruned):
70 layers, ~7.7 GOPS, ~36 MB static storage (INT8)

MODEL	CONV [GOPS]	FC [GOPS]
ResNet50	7.712	0.004
AlexNet	1.332	0.044
VGG16	30.693	0.247

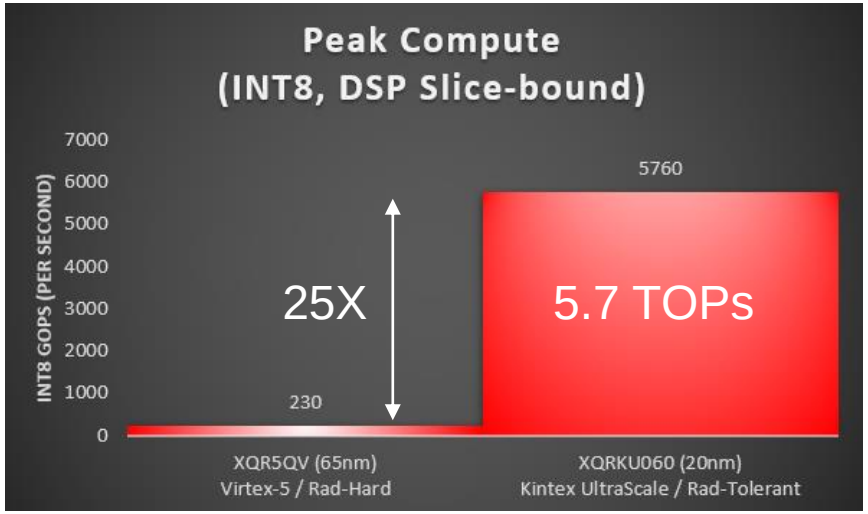
(a) Convolutional layers dominate total compute requirements in CNNs. (Table shows unpruned GOPS.)

RT XQRKU060
Space-grade FPGA is radiation-tolerant with high compute density suitable to DL



- 2760 DSP48E2 Slices: Multi-precision fixed and floating point modes
- 32 High Speed SERDES (12.5Gbps): 400Gbps aggregate BW
- Radiation Tolerance across all orbits: TID >100 Krad/si, SEL >80 MeV-cm^2/mg
- 40x40 mm Ceramic Column Grid Array Packaging
- Class B, Class Y (QML-B, QML-Y Equivalent)

(b) RT XQRKU060 Device (20nm) features [6].



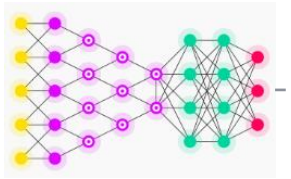
(c) Theoretical peak compute comparison (DSP Slice-bound) between recent Xilinx Space-grade FPGAs.*

RT XQRKU060 is a Viable Target for High Performance CNN Inference in Space…Now Let's Develop a Platform

*Assumes DSP Fmax in -1 Speed Grade (360 MHz @ 65nm; 594 MHz @ 20nm) [6]. 3.5 OPS/Clock/Slice [7].

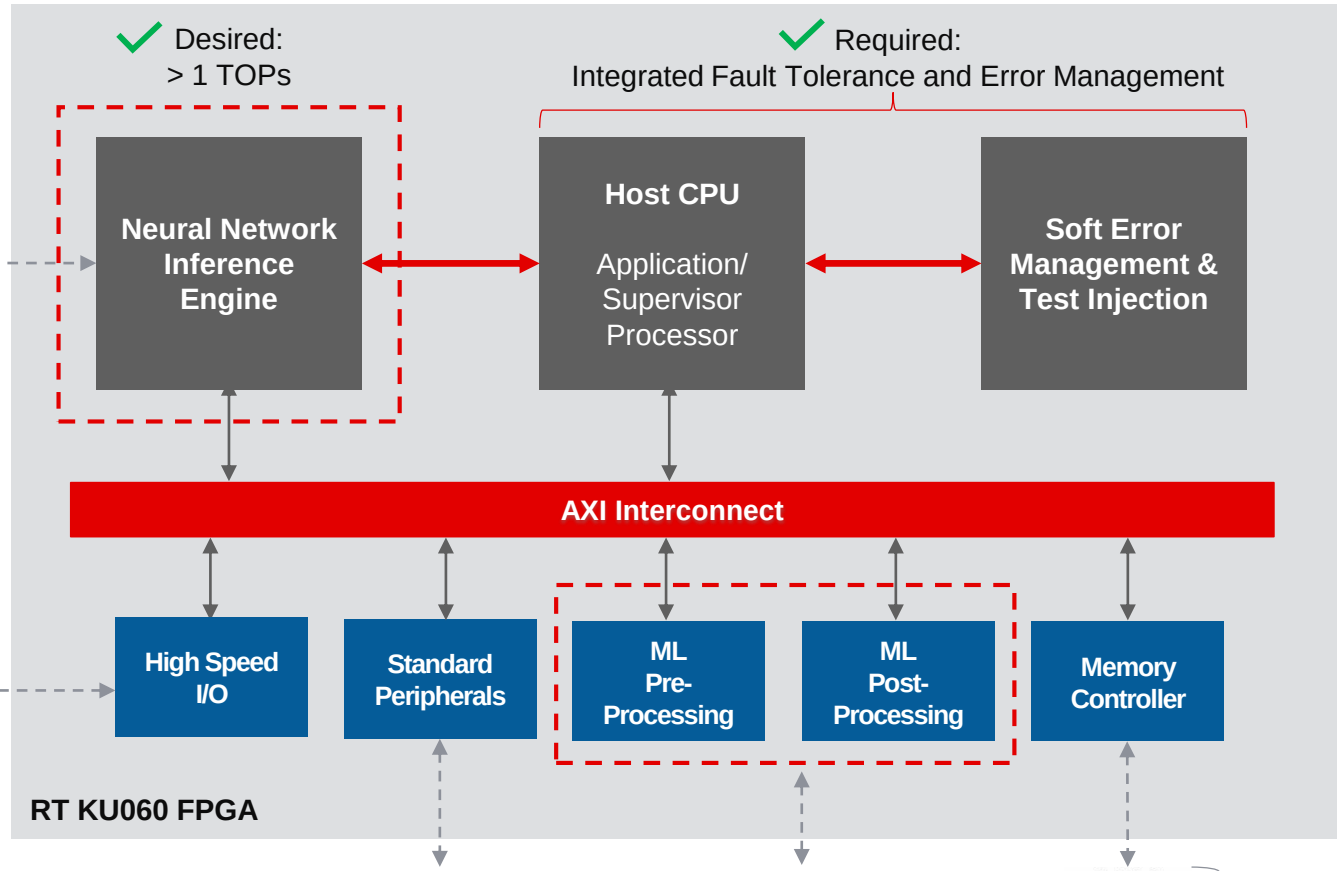
Space DPU Platform: Design & Architecture Criteria

✓ Required:
Direct compilation from standard ML frameworks (Caffe. TF, PyTorch, etc.)



✓ Required:
Efficient tensor processing mapped to 20nm FPGA. Many options: FINN, DPU, HLS4ML (Dataflow, systolic, etc.)

Optional:
High-Speed System Connectivity for Sensor, Comms, Scientific Payloads (Up to 400Gbit/s)



✓ Desired: Leverage Free IP

✓ *Implemented on Current Platform*

Dynamically Reconfigurable Regions via DFX (Optional)

✓ Required:
Boot, Networking, UART, GPIO, Debug, etc.

Optional (PL Offload):
Color Space Conversion, Cropping, Tiling, Norm, NMS, etc.

DRAM

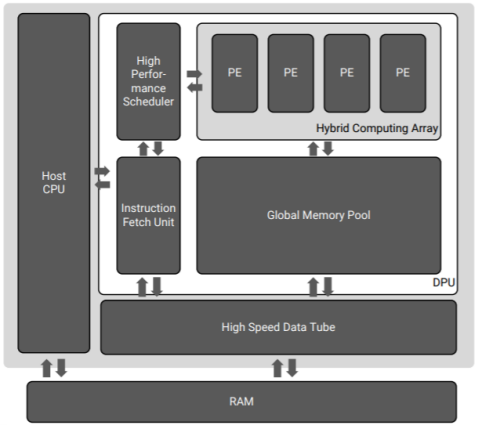
✓ Required:
External Mem for:
Model Weights & Activations, Data Buffering, Host OS

Space DPU Platform Implements a Domain Specific Architecture for Radiation/Fault-Tolerant CNN Inference on XQRKU060

XILINX.

Space DPU Platform: Key Components

Neural Network Inference Acceleration (non-TMR)



Ref: [8][9]

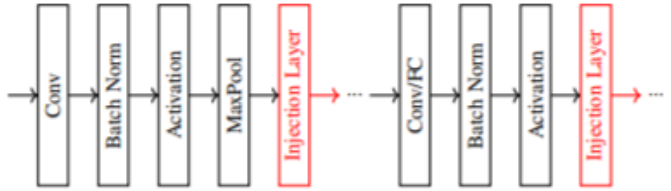
Caffe TensorFlow Keras PyTorch tvm

Xilinx DPU

High-performance Deep Learning Processing Unit (tensor processor) optimized for INT8 precision; DPUCZDX8G ported to 20nm Kintex UltraScale. Programmable with Vitis AI. Open source runtime.

+

Model-level Fault Tolerance (Adds Datapath Resiliency)



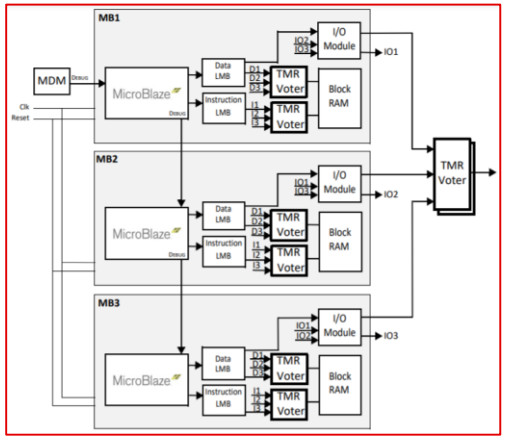
Ref: [5]

Fault Aware Training (FAT)

Probabilistic injection with ad-hoc layer in training framework. Extensible to a variety of architectures and fault models. Increases datapath resiliency to errors induced by SEUs.

+

Control Plane Fault Tolerance & Device-wide Soft Error Mitigation



Ref: [10]

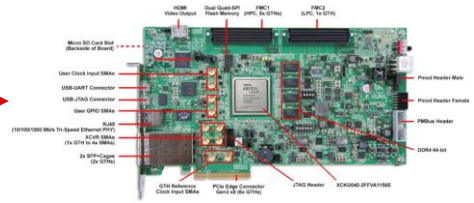
Xilinx TMR MicroBlaze Subsystem

Fault Tolerant - Fail Safe (FT-FS) application processor. Bare-metal / Rich OS support. Integrates Xilinx SEM IP for fast, device-wide background scrubbing (detect/correct soft errors).

Deployment (Custom Platform)

RT XQRKU060
Space-grade Device

Scalable/portable platform components to various hardware targets



Development & Prototyping (Eval Kits & COTS Boards)

Example: Xilinx
KCU105 Evaluation Kit
(XCKU040 Device)

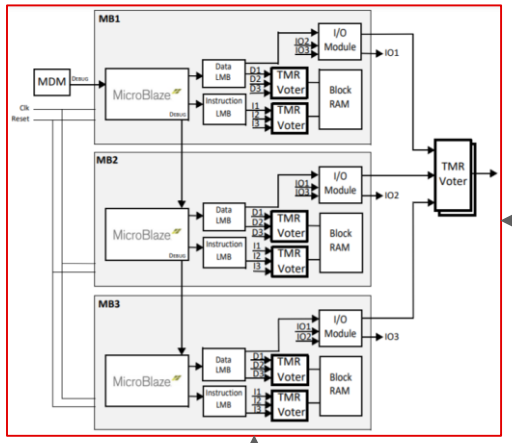
XILINX

Xilinx DPU: Anatomy & Configuration

DPU Architecture: The CNN DPU leverages DSP Slices for compute acceleration within an array of Processing Elements (PEs). Precision is INT8. SPFP32 models are quantized with Vitis AI tools [8] [9].

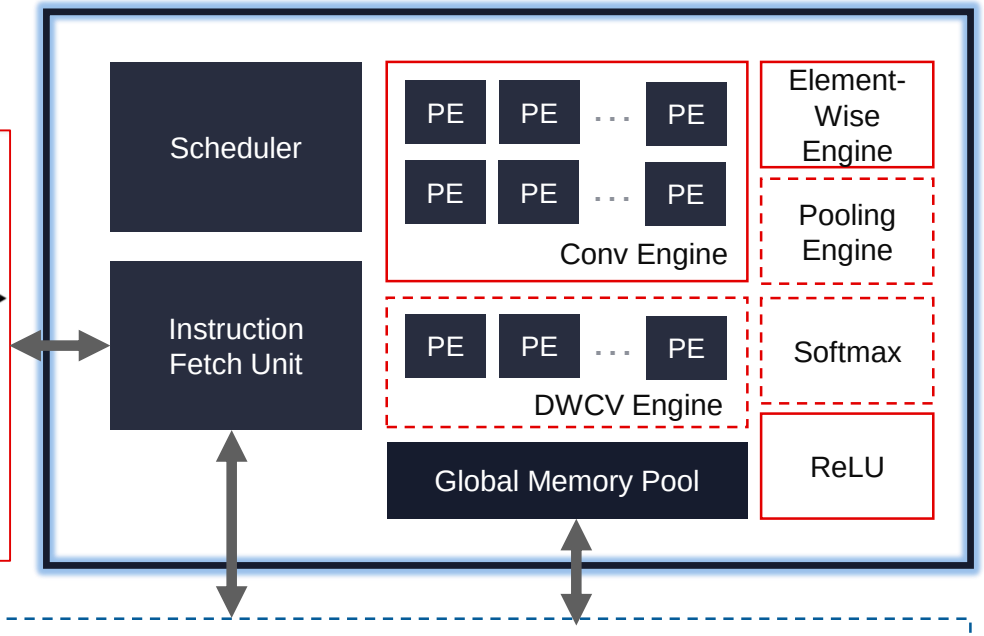
Host CPU (TMR MB)

CNN-enabled Applications



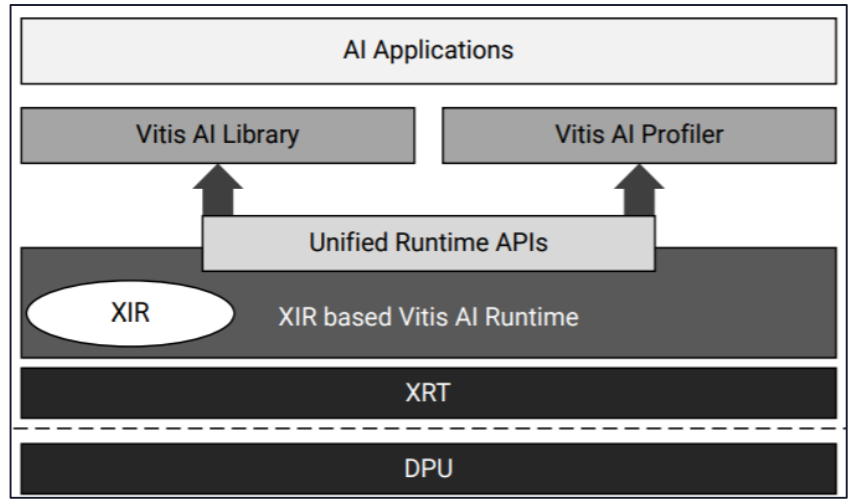
CNN DPU

Co-processor / Accelerated Layers & Operations



Vitis AI Runtime (VART) Stack

Fully Open Source



Key aspects of VART stack optimized for MicroBlaze-based Space DPU Platform.

Space DPU Performance on KCU105 Board: **1.43 TOPs***

ResNet-18: ~70 FPS, Tiny-YOLOv2: ~38 FPS

*Peak TOPs with DPUCZDX8G (single instance), B4096 @ 350 MHz

Xilinx DPU is a family of highly-configurable, efficient tensor co-processors scalable to FPGA, SoC, ACAP and Alveo platforms.

XILINX

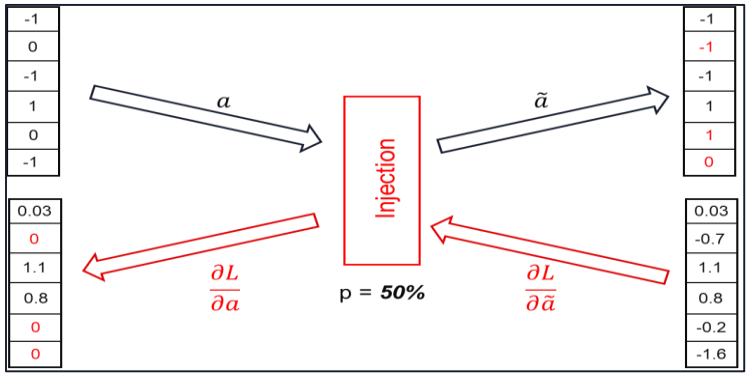
Fault Aware Training (FAT)

Xilinx Research Labs

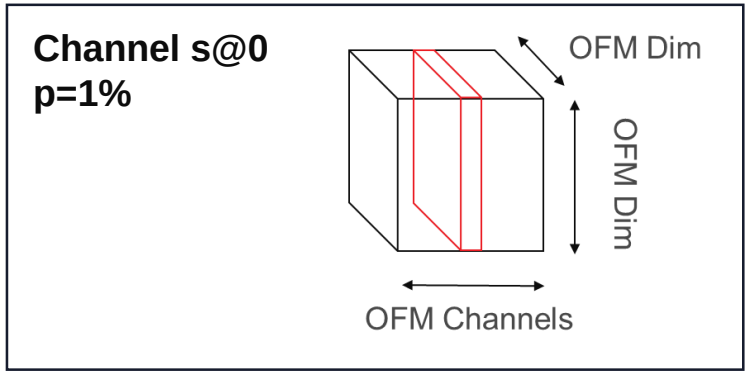
- Main Idea: Hardware Errors (e.g., SEU-induced faults) can be modelled during training and tailored to the chosen hardware inference engine (e.g., DPU)
  - Probabilistic injection with ad-hoc layer in training framework -> FAT: Fault Aware Training [5]
  - No increase in training time nor hardware cost
  - Provides significant increase in minimum accuracy under stuck@ error models
  - No compromise in error-free accuracy
    - Error-free accuracy improved in all of our experiments
  - FAT has been tested on a variety of topology/dataset/precisions

Terminology:

FAT = Fault Aware Training
SAT = Standard Training (no fault model)



(a) **FAT Concept:** PyTorch-developed error injection layer allows injecting errors with a particular error model during training. Forward and backward pass of injection layer shown. Errors are injected with a global probability p, which is a hyper-parameter. [5]



(b) **FAT Implementation for Space DPU Testing:** During training, after every activation function is computed, random channels in the output feature map (OFM) are set to zero, with a probability of 1%.

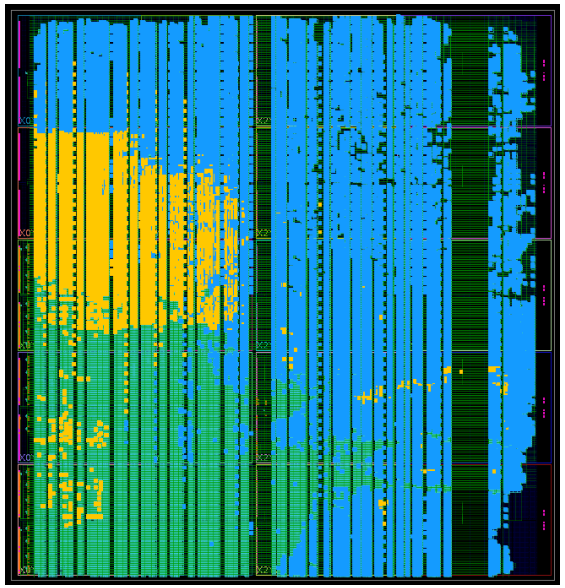
[5] U. Zahid, G. Gambardella, N. J. Fraser, M. Blott, and K. Vissers, "FAT: Training Neural Networks for Reliable Inference Under Hardware Faults," *arXiv:2011.05873 [cs]*, Nov. 2020, Available: http://arxiv.org/abs/2011.05873

FAT is a Training Time Method for Improving Fault Tolerance of Neural Networks Without Added Hardware Cost

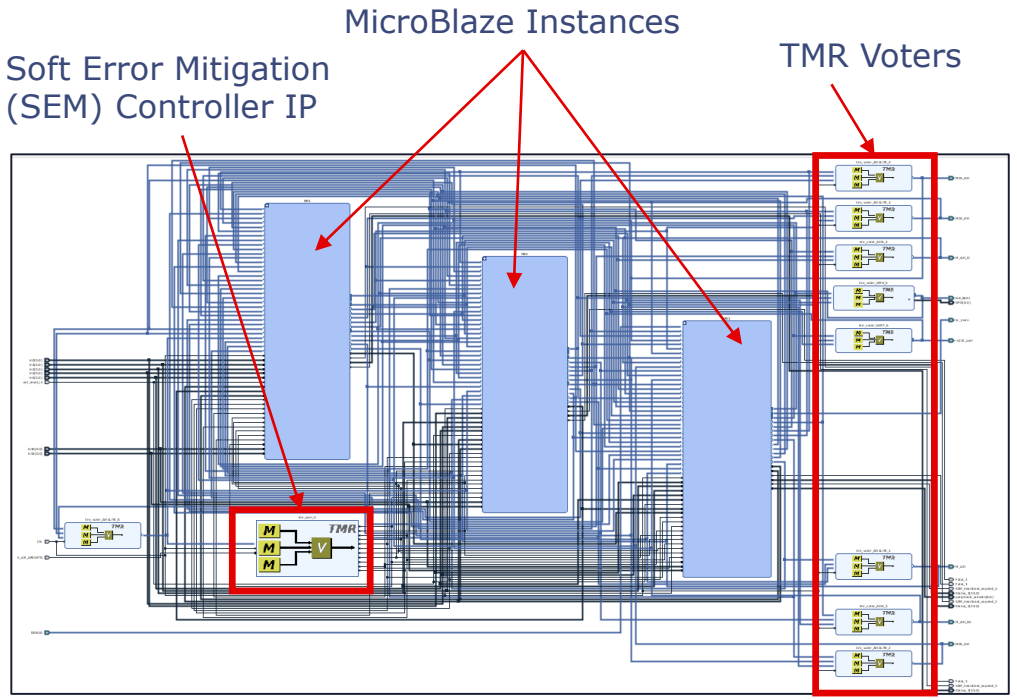
XILINX

Space DPU Platform: Putting It All Together

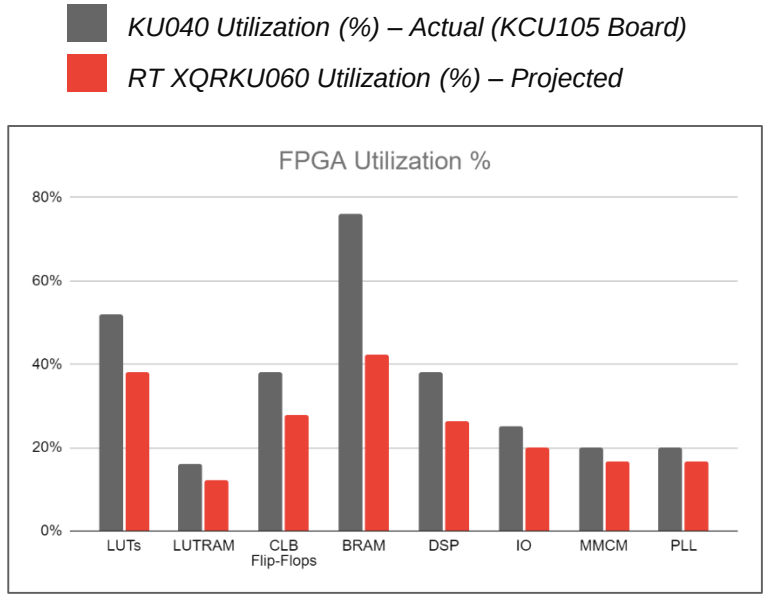
FPGA Implementation



(a) FPGA Implemented Device View (KU040 on KCU105). Placement of DPU and TMR MicroBlaze subsystem is highlighted.



(b) TMR Subsystem Block Diagram as constructed in Vivado IP Integrator. (DPU is one level of hierarchy above TMR Subsystem, along with other peripherals.)



(c) Actual FPGA resource utilization for overall design implemented in the KU040 device (KCU105 board) compared against projected utilization in the larger, Space-grade XQRKU060 device.

TMR MicroBlaze Sub-system

DPU Instance

Other logic

DPU Resource Utilization: ~58k LUTs, ~105k FFs, 261 BRAMs, 704 DSP Slices

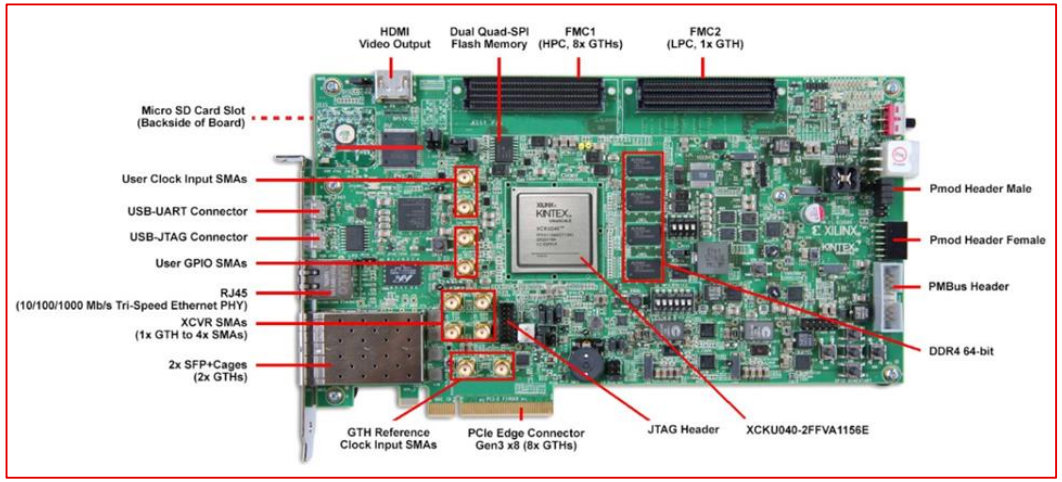
"As Is" Placement Derived from Vivado 2019.2 Defaults; Design Has Room to Grow, Particularly in KU060

XILINX.

Experiment and Results

Beam Test Experiment Configuration

- Beam
  - >64 MeV protons beam test
- Platform
  - TMR MicroBlaze @ 175 MHz; Linux kernel 4.9 (Petalinux)
  - Single DPUCZDX8G instance (B4096) @ 350 MHz
  - “Ground-up” build in Vivado IPI (2019.2)
  - Simplified XRT + Vitis AI 1.2 Runtime (VART)
- Models
  - **ResNet-18** (Image Classification)
    - ~3.65 GOPS
    - SAT and FAT Models; Compiled/quantized with Vitis AI
  - **Tiny-YOLOv2** (Object Detection)
    - ~7 GOPS
    - SAT Model only; Darknet model converted then quantized with Vitis AI
- Dataset(s) & Image Preparation
  - ILSVRC2012 (ResNet-18)
    - 224x224 crops; mean-centered (pre-processed offline)
  - Pascal VOC 2012 (Tiny-YOLOv2)
    - 416x416 crops; mean-centered (pre-processed offline)



(a) Board Platform for Beam Test: Xilinx KCU105 Eval Kit with XCKU040 Kintex UltraScale FPGA.

Objective: Deploy Space DPU Platform in Emulated Space Env.; Characterize Effectiveness of Fault Mitigations

Test Methodology (ResNet-18)

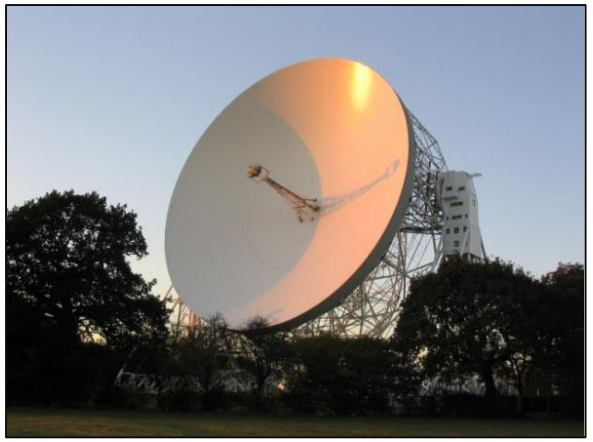
Example Classification Output Logging

```
==========================================
Load image : ILSVRC2012_test_00083589
Iteration 1 of 10
Fri Aug 07 13:02:36 2020
Start @  1462218218.49011
top[0] prob = 0.998448 name = radio telescope, radio reflector
top[1] prob = 0.001501 name = solar dish, solar collector, solar furnace
top[2] prob = 0.000027 name = analog clock
top[3] prob = 0.000005 name = sundial
top[4] prob = 0.000004 name = airship, dirigible
End @ 1462218218.59011
```

Top-1 Probability

Top-1 Class Prediction

Top-5 Model Class Predictions and Probabilities Captured for Each Test Image



(a) Example image from validation dataset (prior to pre-processing) (not displayed during test script execution)

>500 CRAM soft errors Detected and Corrected Using UltraScale SEM IP for each collective group of SAT and FAT Test Runs.

Experiment Design

Background Scrubbing:
SEM IP with correction and essential bit classification enabled. Detects and corrects soft errors occurring in configuration RAM.

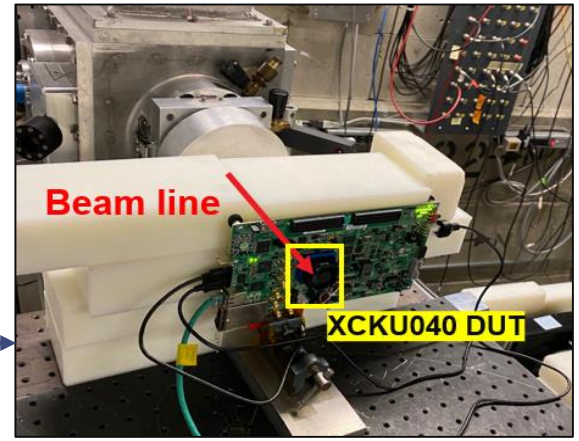
TMR Subsystem Fault Management:
TMR Manager implements voting and fault monitoring of TMR MB Subsystem internal state. After first failure, moves to lockstep mode.

Baseline Configuration (All Tests)

SAT Model Executes on DPU

FAT Model Executes on DPU

Run-specific Testing (A/B)



(b) Kintex XCKU040 DUT mounted on Xilinx KCU105 test board during proton beam testing at Crocker Nuclear Lab (CNL).

XILINX.

Observed Error Types: Accuracy

ResNet-18

Example: Incorrect Classification

(Accuracy Degradation vs. Error-free Model)

Top-1 Error-free Prediction

komondor (sheep dog) (87.71%)

Top-1 Actual Prediction Under Beam (Error)

window shade (61.65%)

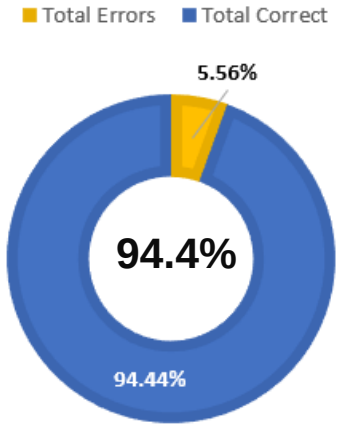
Can Be False Positive or False Negative

Image: ILSVRC2012_val_00000383
Model: resnet-18 (SAT)

Note: Images are for illustration; actual images were cropped to 224x224 and mean-centered prior to model training and classification

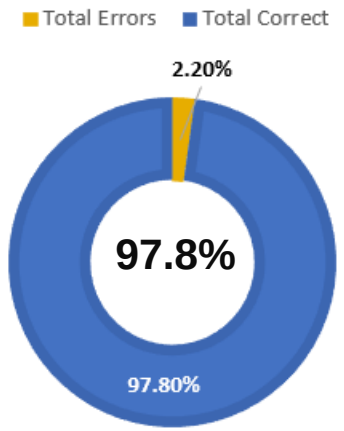
Overall Results (Beam Test): Top-1 Accuracy

(Ground Truth = Error-free Predictions)



SAT Model

-5.56% Degradation



FAT Model

-2.20% Degradation

+3.4% Accuracy Improvement vs. SAT
>50% Error Reduction

Soft Errors Can Induce Faults That Reduce Prediction Accuracy; FAT Shows Meaningfully Better Fault Tolerance vs. SAT

Observed Error Types: Certainty

ResNet-18

Example: Probability Error

(Certainty Degradation vs. Error-free Model)

Top-1 Error-free Prediction

water ouzel, dipper (95.76%)

Top-1 Actual Prediction Under Beam

(Correct Classification, Different Probability)

water ouzel, dipper (74.71%)

Primary concern is False Negatives due to probability degradation…but False Positives also possible

Image: ILSVRC2012_val_00024059
Model: resnet-18 (SAT)

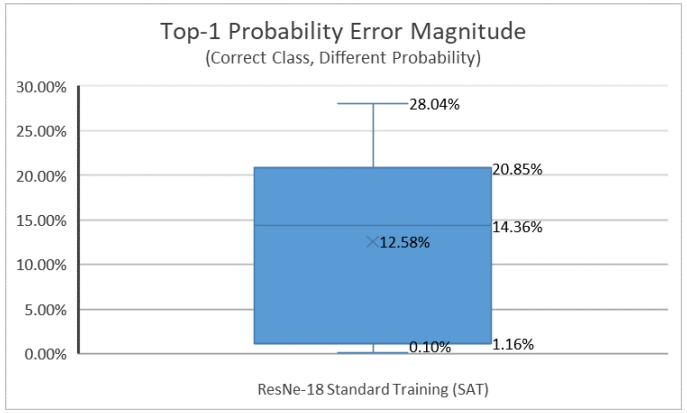
Note: Images are for illustration; actual images were cropped to 224x224 and mean-centered prior to model training and classification

Overall Results (Beam Test): Top-1 Probability Error

(Correct Class, Different Probability)

SAT Model

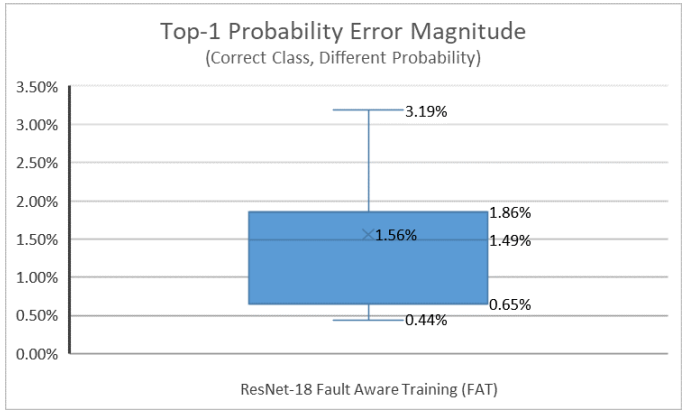
6.7% of correct predictions exhibit prob. errors



Mean Error = 12.58%, σ = 10.07%, IQR = 19.69%

FAT Model

7.9% of correct predictions exhibit prob. errors



Mean Error = 1.56%, σ = 0.84%, IQR = 1.21%
(FAT Exhibits >10X Lower Error Dispersion)

Soft Errors Can Induce Faults That Reduce Prediction Certainty; FAT Shows Meaningfully Lower Probability Error vs. SAT

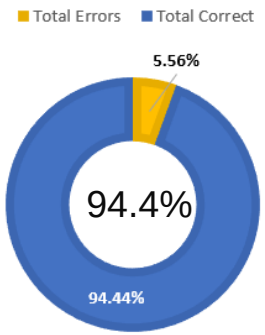
XILINX.

Statistical Analysis: Summarized Results from Beam Test (ResNet-18)

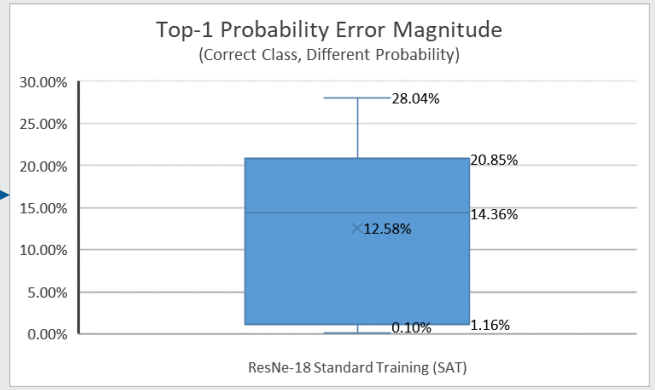
Top-1 Accuracy*
(Ground Truth = Error-free Predictions)

Top-1 Certainty (Probability Errors)
(Correct Class, Different Probability)

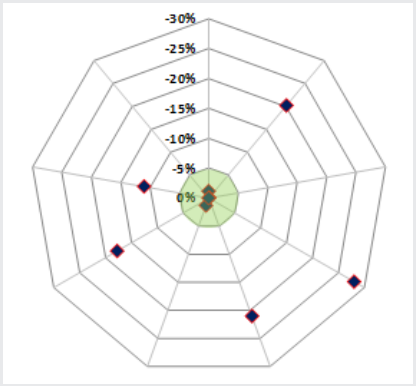
Standard Training (SAT)



6.7% of correct predictions exhibit prob. errors



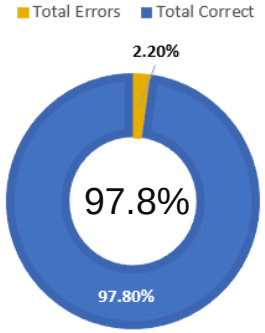
SAT Model Probability Errors
Mean Error = 12.58%, σ = 10.07%, IQR = 19.69%



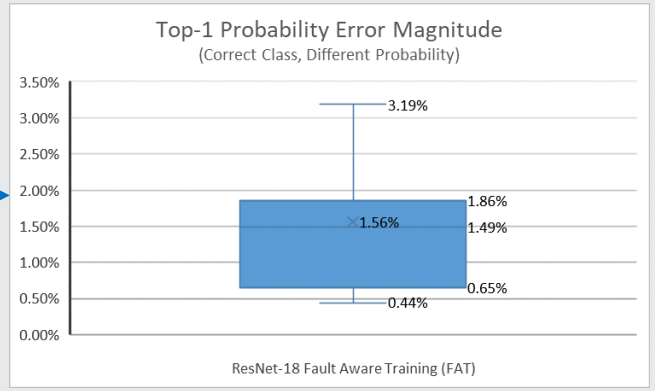
SAT Model Probability Error Plot

Fault Aware Training (FAT)

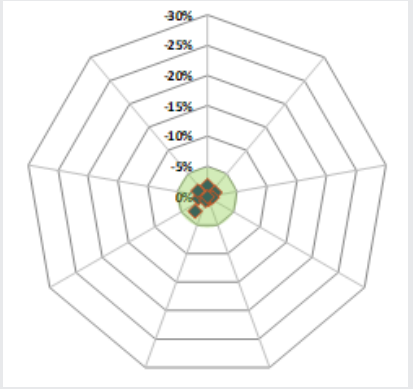
+3.4% Accuracy Improvement vs. SAT
>50% Class Error Reduction



7.9% of correct predictions exhibit prob. errors



FAT Model Probability Errors
Mean Error = 1.56%, σ = 0.84%, IQR = 1.21%



FAT Model Probability Error Plot
(>10X Lower Error Dispersion vs. SAT)

Detailed SEU/SEFI analysis forthcoming (IEEE).

*Ground Truth Reference = 100% (ResNet-18 Model Predictions for Test Images in Non-Irradiated Environment, aka., Error-free Predictions)

Fault Aware Trained Model Exhibits Superior Classification Accuracy and Lower Probability Error Under Beam-Induced Faults

XILINX.

Summary

Key Take-aways

- “Space DPU” Platform is a viable base architecture for radiation-tolerant deep learning on Xilinx 20nm FPGAs in Space environments
  - ✓ Key Components are Freely Available (DPU, Vitis AI, TMR MicroBlaze)
- Architecture is extensible between different targets for development and deployment
- Fault Aware Training (FAT) meaningfully reduces neural network misclassifications and probability degradation in presence of SEU-induced faults without adding hardware cost or complexity
  - ✓ Broadly-applicable technique scalable across a variety of DNNs and hardware backends

What’s Next?

- Multiple Improvements Being Considered:
  - Mainstream Vitis / Vitis AI support for “Space DPU”
  - Acceleration-enabled platform(s)
  - Error Mitigation enhancements
  - What Else Would You Like to See? Let us know…

Acknowledgements

- The Space DPU project brought together a diverse and talented team spanning more than 5 countries:
  - Dr. Pierre Maillard, Yanran Chen and Paula Chang (Xilinx Radiation Effects Team)
  - Nicholas Fraser, Giulio Gambardella and Dr. Michaela Blott (Xilinx Research Labs)
  - Minal Sawant, Jason Vidmar and Troy Jones (Xilinx Aerospace & Defense)
  - Numerous additional Xilinx colleagues for their skill and contributions: Shuai Zhang, Hong Luo, Ye Yang, Bingqing Guo, Srikanth Erusalagandi, Jana G, Yashu Gosain, Roland Petersson, Stefan Asserhall, Kamran Khan, Mark Harvey, Ramine Roane, Sudip Nag, Yi Shan and many others
  - Numerous members of Space community for concept validation and suggestions

Special Thanks to ESA (Gianluca Furano, David Steenari & many others) for their insight, inspiration & encouragement during the development of the Space DPU project.

References

[1] (Craft Prospect Ltd) Karagiannakis, Phil, Ireland, Murray, and Greenland, Steve, “Case Study: Prototyping CNNs on Zynq for Space Applications, Using PYNQ,” in *Exploring Zynq MPSoC: With PYNQ and Machine Learning Applications*, Glasgow, Scotland, UK: University of Strathclyde, 2019. [Online]. Available: https://www.zynq-mpsoc-book.com

[2] I. Ahmed, S. Din, G. Jeon and F. Piccialli, "Exploring Deep Learning Models for Overhead View Multiple Object Detection," in IEEE Internet of Things Journal, vol. 7, no. 7, pp. 5737-5744, July 2020, doi: 10.1109/JIOT.2019.2951365, Available: https://ieeexplore.ieee.org/document/8891768

[3] N. Evers, “Deep learning in Space,” *Medium: Towards Data Science*, Apr. 30, 2019, Available: https://towardsdatascience.com/deep-learning-in-space-964566f09dcd

[4] A. Canziani, A. Paszke, and E. Culurciello, “An Analysis of Deep Neural Network Models for Practical Applications,” *arXiv:1605.07678 [cs]*, Apr. 2017, Available: http://arxiv.org/abs/1605.07678

[5] U. Zahid, G. Gambardella, N. J. Fraser, M. Blott, and K. Vissers, “FAT: Training Neural Networks for Reliable Inference Under Hardware Faults,” *arXiv:2011.05873 [cs]*, Nov. 2020, Available: http://arxiv.org/abs/2011.05873

[6] Xilinx, “Radiation Tolerant Kintex UltraScale XQRKU060 FPGA Data Sheet,” 2020. Available: https://www.xilinx.com/support/documentation/data_sheets/ds882-xqr-kintex-ultrascale.pdf

[7] Y. Fu, E. Wu, A. Sirasao, S. Attia, K. Khan, and R. Wittig, “Deep Learning with INT8 Optimization on Xilinx Devices,” Xilinx, 2017. Available: https://www.xilinx.com/support/documentation/white_papers/wp486-deep-learning-int8.pdf

[8] “Zynq DPU v3.3 Product Guide,” Xilinx, Feb. 2021. Available: https://www.xilinx.com/support/documentation/ip_documentation/dpu/v3_3/pg338-dpu.pdf

[9] “Vitis AI User Guide,” Xilinx, 2021. Available: https://www.xilinx.com/support/documentation/sw_manuals/vitis_ai/1_3/ug1414-vitis-ai.pdf

[10] “Triple Modular Redundancy (TMR) Subsystem v1.0 Product Guide,” Xilinx, 2019. Available: https://www.xilinx.com/support/documentation/ip_documentation/tmr/v1_0/pg268-tmr.pdf

Thank You

Xilinx Mission

Building the Adaptable, Intelligent World