Using the VectorBlox™ SDK to Create Programmable AI/ML Applications in RT PolarFire® FPGAs

A Leading Provider of Smart, Connected and Secure Embedded Control Solutions

SMART | CONNECTED | SECURE

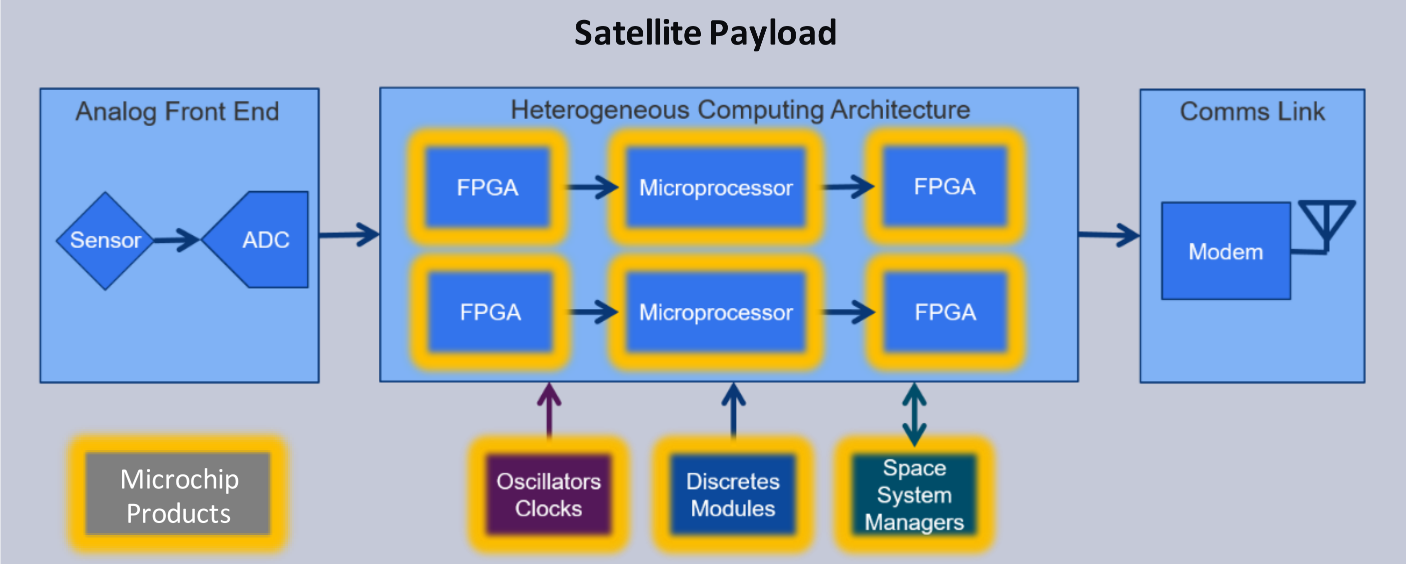
Presented at European Space Agency
On-Board Data Processing Workshop

Ken O'Neill
Aaron Severance
Diptesh Nandi

June 15, 2021

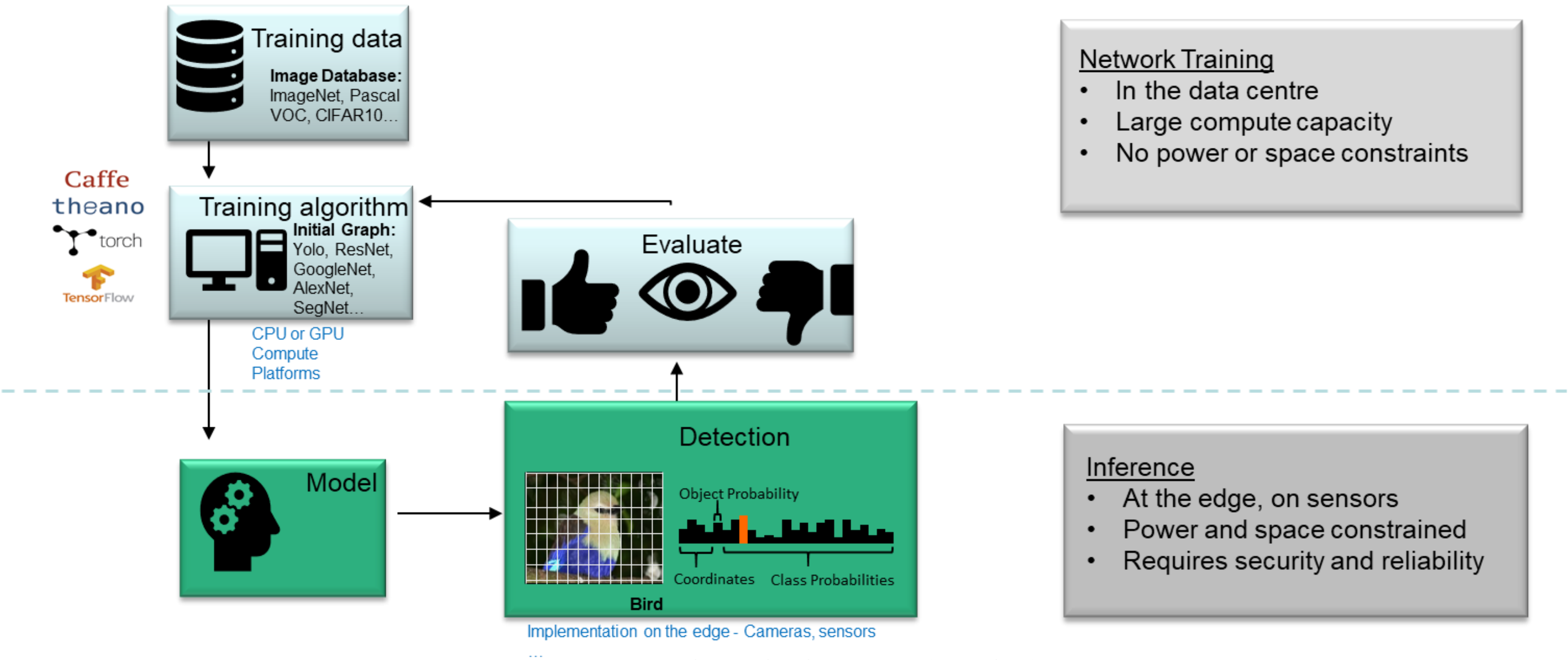
AI/ML in Space

- **Background**
  - Satellite operators demanding more information from space assets
  - Sensor resolution increasing faster than available downlink bandwidth
  - Faster frame rates, more channels in multi-spectral imaging
- **Artificial Intelligence and Machine Learning (AI/ML) allows new operating models**
  - Permits more elaborate processing of data on orbit
  - Allows transmission of processed information, instead of raw data
  - Enables autonomous decision-making on orbit – use, discard, prioritize



AI/ML Implementation: Neural Network Training and Inferencing

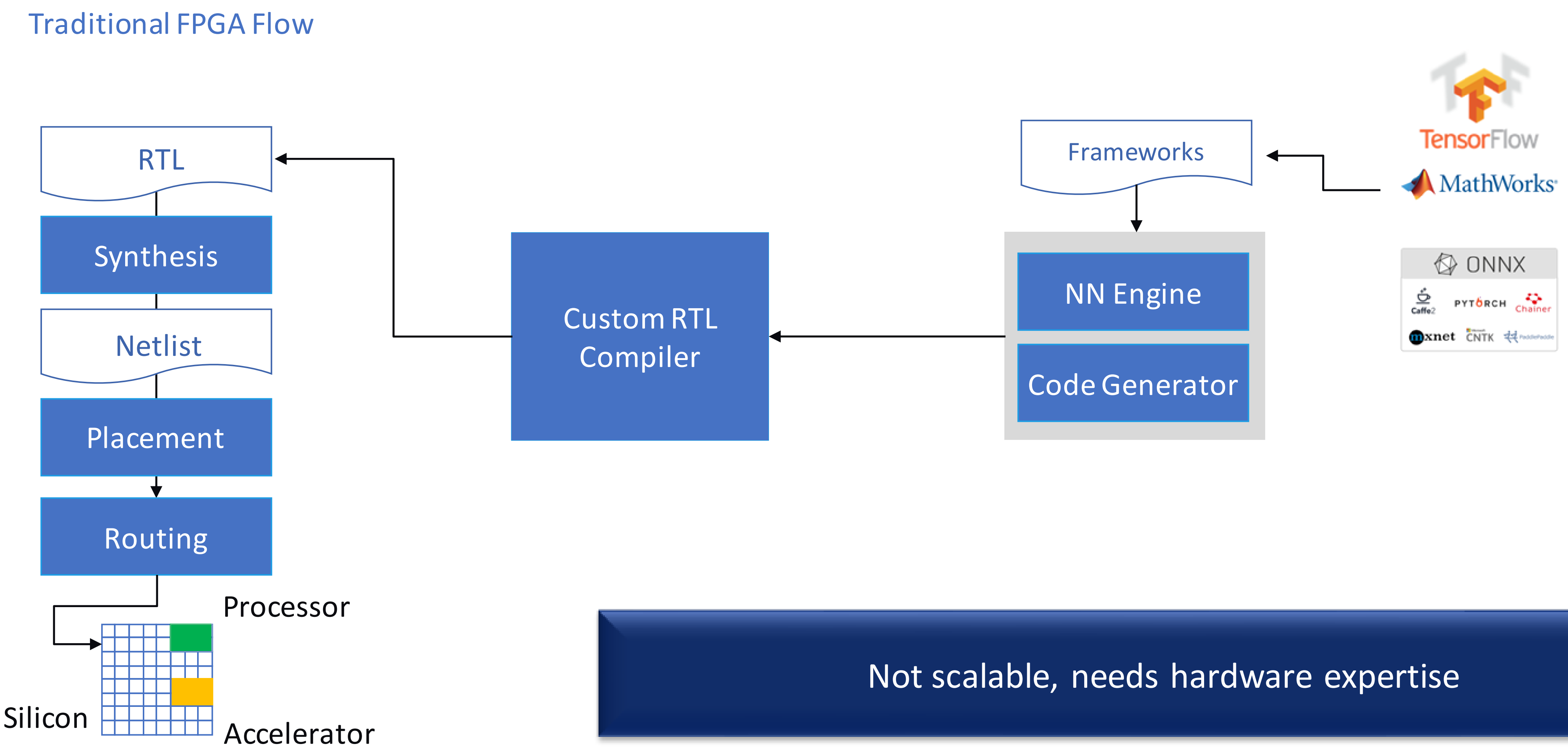
- **Training is performed on the ground with large data sets**
  - Unconstrained compute resources, no power limitations
- **Inferencing is performed in the space-flight application**
  - Limitations on computing throughput, power and thermal constraints



Network Training
- In the data centre
- Large compute capacity
- No power or space constraints

Inference
- At the edge, on sensors
- Power and space constrained
- Requires security and reliability

Traditional FPGA Design Flow Challenges

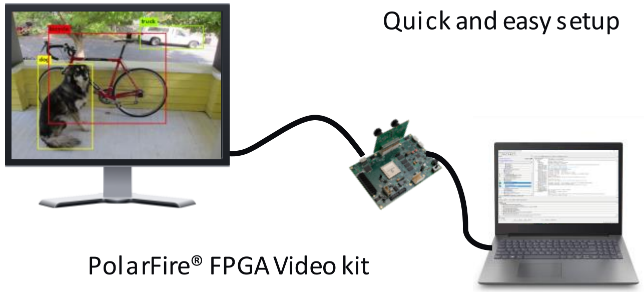


Not scalable, needs hardware expertise

VectorBlox™ Software Development Kit and IP

VectorBlox SDK and NN IP enables:

- Software developers to run Neural Networks (NN) without prior FPGA knowledge
- Utilization of most popular NN software frameworks
- Simulation in software without procuring hardware



Quick and easy setup

PolarFire® FPGA Video kit

CNNs are written into external Non-Volatile memory
No need to reprogram, recompile the FPGA when updating CNNs

TensorFlow
MathWorks
ONNX
PYTORCH
Chainer
mxnet
CNTK

Wide framework interoperability

User CNN or Standard CNN → VectorBlox SDK (Python) → *Write to SPI Flash* → SPI Flash

Embedded Design

SPI Flash

DDR4

RT PolarFire® or PolarFire FPGA

Other Peripherals

FPGA Design (Libero) → *Program the FPGA with the Smart Camera Design* → RT PolarFire® or PolarFire FPGA

From Customer

VectorBlox Scope

Cloud/ Workstation	Linux PC	FPGA/ MPU/ CPU
Training	**Compress/ Calibrate**	**Inferencing**

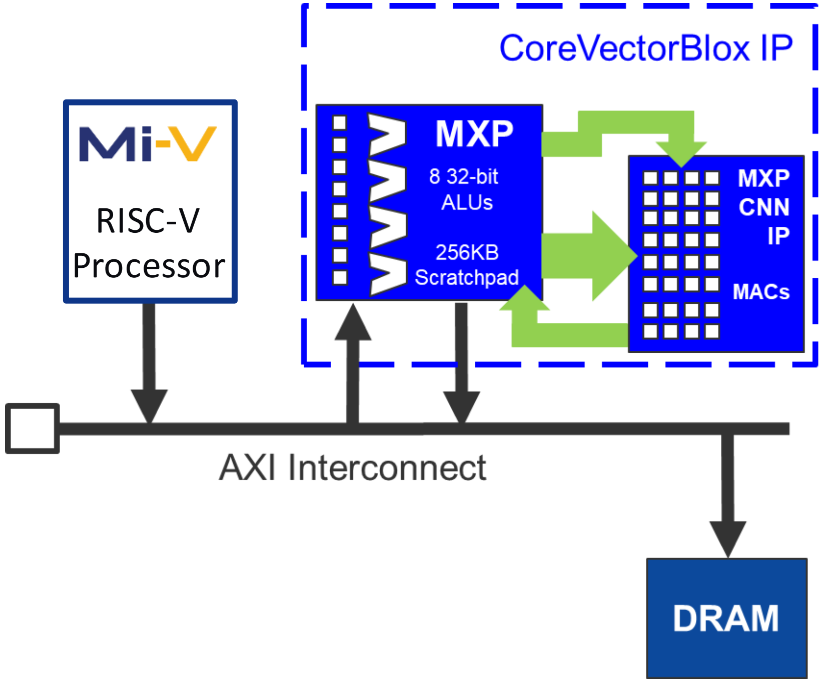
Training Dataset → 32-bit CNN → 8-bit Trained CNN → Execute on FPGA

VectorBlox SDK's scope is in "Compression" and "Inferencing"

MICROCHIP

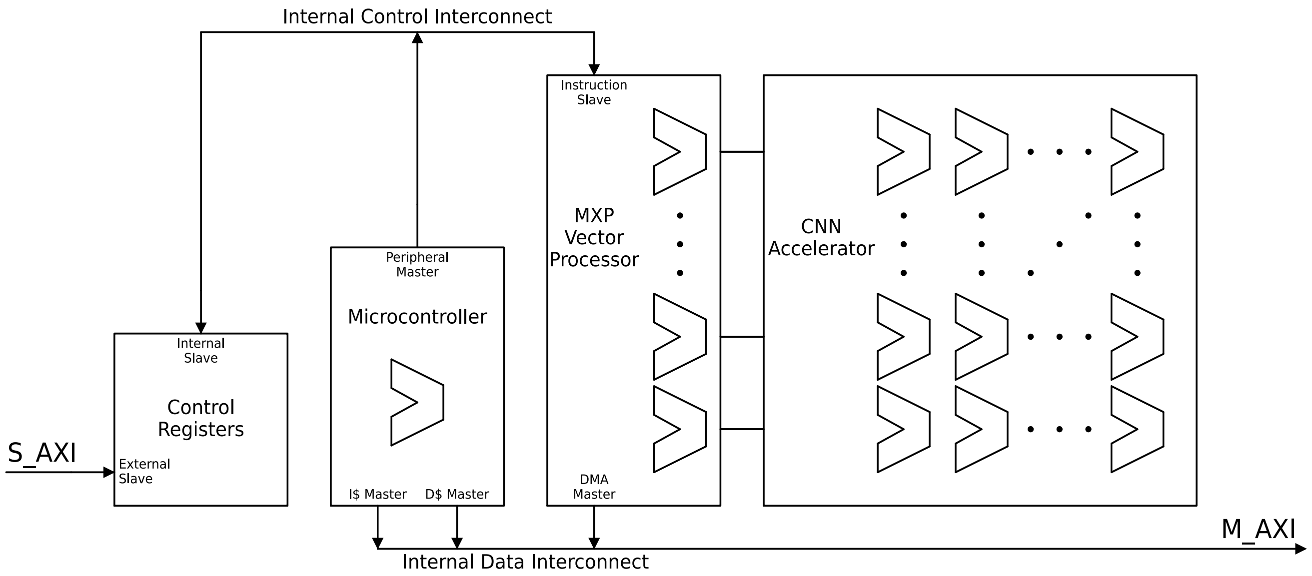
CoreVectorBlox Neural Network Engine

- **Multi-core Software Architecture**
  - Mi-V runs complete software-based application
- **VectorBlox Matrix Processor (MXP)**
  - Elementwise tensor operations
    (add, sub, xor, shift, mul, dotprod …)
  - Up to 8 32-bit ALUs
  - Mixing precisions ok (int8, int16, int32)
  - 256 KB scratchpad memory and DMA controller
- **VectorBlox CNN**
  - Tensor multiply-accumulate operations
  - Fixed at either int8 precision
  - Layer enhancements added with software update



Overlay implementation allows several different networks to run on the same FPGA design without the need to resynthesize

CoreVectorBlox IP



Configuration	Vector Processor Width	Vector Scratchpad	CNN Accelerator Array Size
V250	128-bit	64kB	16x16
V500	256-bit	128kB	16x32
V1000	256-bit	256kB	32x32

- **Control Registers**
  - Control, Status and Error signaling
- **Microcontroller**
  - Parses network structures from BLOBs
  - Controls MXP Vector Processor
- **MXP Vector Processor**
  - Processes general neural network layers and offloads convolution layers to CNN Accelerator
- **CNN Accelerator**
  - Processes convolution network layers

VectorBlox™ 8-bit Accuracy

Input Framework	Model	Accuracy		
		Simple 8-bit	VectorBlox SDK 8-bit	Floating-point 32-bit
Caffe	Squeezenet 1.1	58.8	58.6	59.2
TensorFlow	Mobilenet v1	68.8	71.2	71.6
TensorFlow	Mobilenet v2	69.0	71.6	72.0
ONNX	Resnet18 v1	71.4	72.8	72.8
PyTorch	Resnet50	75.2	75.0	75.0
Darknet	TinyYOLO v2 VOC	54.2	54.4	55.1
Darknet	TinyYOLO v3 COCO	39.5	40.4	40.9

MICROCHIP

VectorBlox™ Performance and Utilization

Without TMR – default implementation

Configuration	Size (no TMR)			fmax (base) MHz	Peak GOPs	mw/GOP	Performance (FPS)	
	kLUTs	MPF500T	RTPF500T				Mobilenet-v1	TinyYOLO-v3
V250	28	6%	6%	154	79	7.0	26.2	9.1
V500	48	10%	10%	143	146	6.4	47.7	16.6
V1000	63	13%	13%	136	279	5.1	68.0	26.5

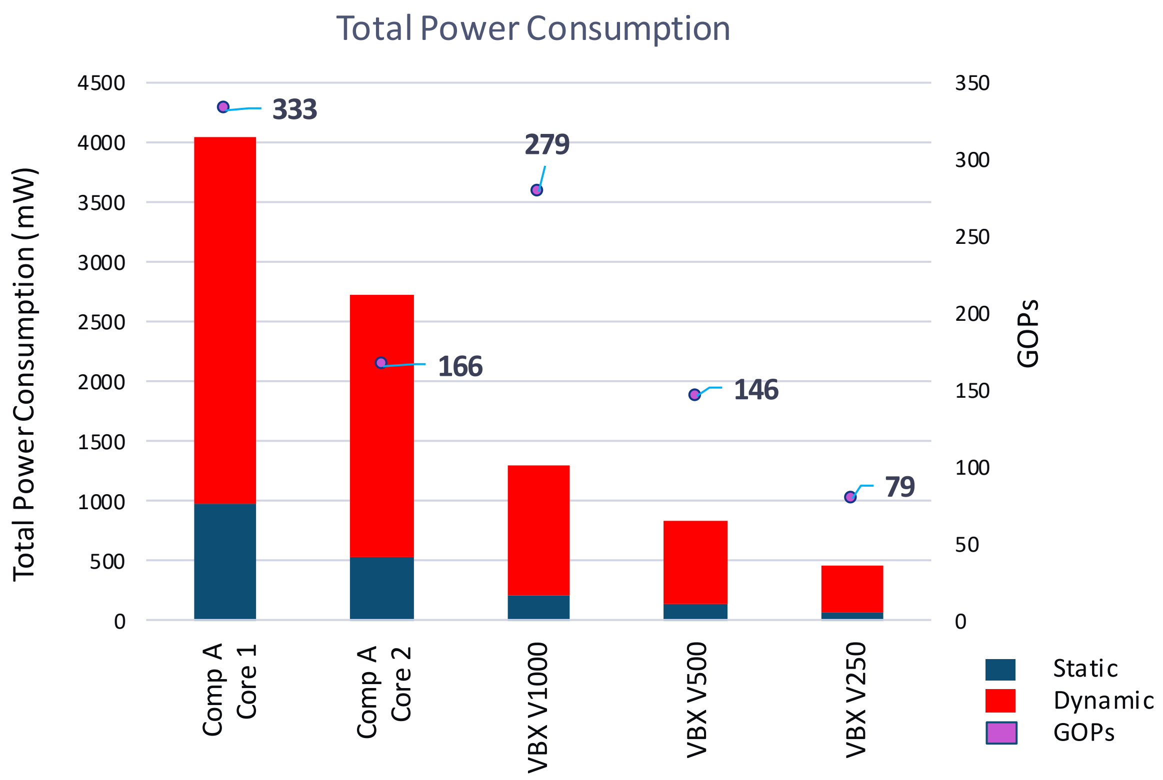
With TMR – synthesized local TMR, preliminary results

Configuration	Size (synthesized local TMR)				fmax (base) MHz	Peak GOPs	mw/GOP	Performance (FPS)	
	kLUTs	kDFF	MPF500T	RTPF500T				Mobilenet-v1	TinyYOLO-v3
V250	56	66	14%	14%	102	52	-	17.4	6.0
V500	97	116	24%	24%	86	88	-	28.7	10.0
V1000	136	161	33%	33%	76	156	-	38.0	14.8

MICROCHIP

2-3x More Power Efficient Inferencing

Core Name	Peak GOPs	Dynamic Power (mW)	Static Power* (mW)	Total Power (mW)	Total Power (mW/GOP)
VectorBlox™ V1000	279	1094	206	1300	**5.1**
VectorBlox V500	146	698	127	825	6.4
VectorBlox V250	79	387	65	452	7.1
Competitor A Core 1	332.8	3072	976	4048	12.2
Competitor A Core 2	166.4	2201	528	2729	16.4



- **50% to 70% power saving for similar inferencing performance**
- **Higher inferencing performance in applications with power constraints**

*Scaled for resource utilization

RT PolarFire® FPGA

RT PolarFire FPGAs

- High density, high performance
- Absence of configuration upsets
- Lowest power consumption in class
- Path to QML qualification

- 481K logic elements
- 33 Mbits SRAM
- 1,480 multipliers
- 24 x 10 Gb/sec serdes

RT PolarFire FPGA Schedule:

- Engineering models available now
- Mil Std 883 B expected in early 2022
- QML class Q expected in early 2022
- QML class V expected in 2023

Available Today:

- RT PolarFire FPGA engineering models
- Commercial MPF300 evaluation kit
- Synthesis support for TMR
- Power calculator

RT PolarFire® FPGA Radiation Summary

- **Robust TID, 100 krad**
- **No configuration upsets**
  - LET 80 MeV-cm^2/mg, fluence more than 5E8 ions/cm^2
- **SEL in GPIO cells**
  - LET_{TH} 80 MeV-cm^2/mg
- **SEFI in reset circuit**
  - 1 in 40 years, in GEO solar min
- **SEU in flip-flops**
  - Unprotected upset rate ~ 1E-7 errors/bit-day, GEO solar min
  - TMR with constrained placement ~ 1E-11 errors/bit-day, GEO solar min
- **Reprogramming on orbit**
  - Reprogramming on orbit is supported – heavy ion and proton test data is available
- **Next tests:**
  - Clock networks, PLL, SERDES, SRAM with EDAC, Mathblocks, SET, . . .

Summary and Conclusion

- **VectorBlox™ Accelerator SDK creates efficient networks for FPGA deployment**
  - Optimization and conversion to 8-bit models with minimal loss of accuracy
  - Bit-accurate simulation environment allows evaluation before programming
- **CoreVectorBlox IP creates flexible, power-efficient neural networks on PolarFire and RT PolarFire FPGAs**
  - Update and modify networks without changing the FPGA design
  - Networks can be designed and modified by engineers with no FPGA experience
- **RT PolarFire® allows greater inferencing performance in space**
  - Higher GOPS/mW in power-constrained space applications
  - Radiation effects suitable for LEO, GEO and deep space

Q & A