Deploying Neural Networks on RISC-V with VPU

UAB Universitat Autònoma de Barcelona

Eugeni Casadesús*, Michelle Vargas*, Aleix Valdivieso*, Màrius Montón*±
*Institut d'Estudis Espacials de Catalunya (IEEC)
±Universitat Autònoma de Barcelona (UAB)

IEEC

Basic operations

SAXPY

$$Z[i] = X[i] * a + Y[i]$$

Reduce

$$a = \sum_i X[i] * Y[i]$$

Vector Add

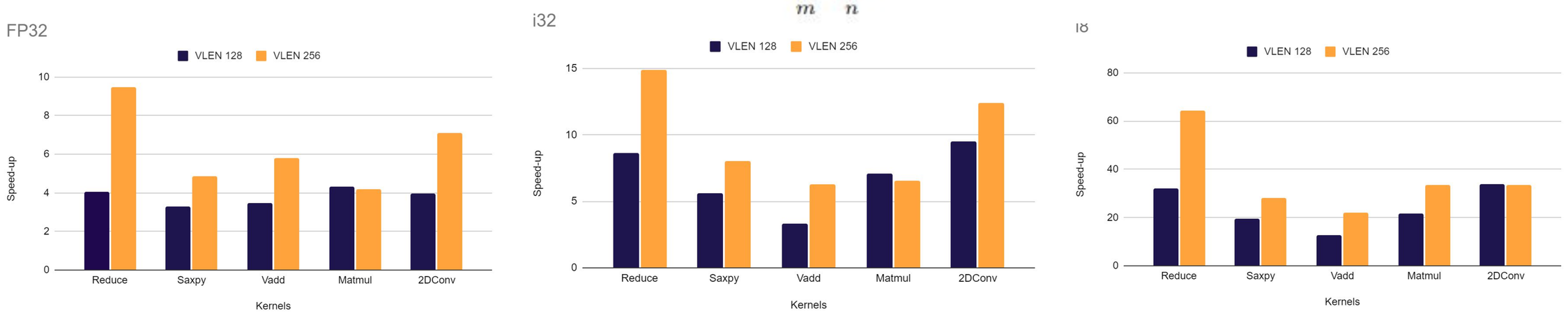
$$Z[i] = X[i] + Y[i]$$

MatMul

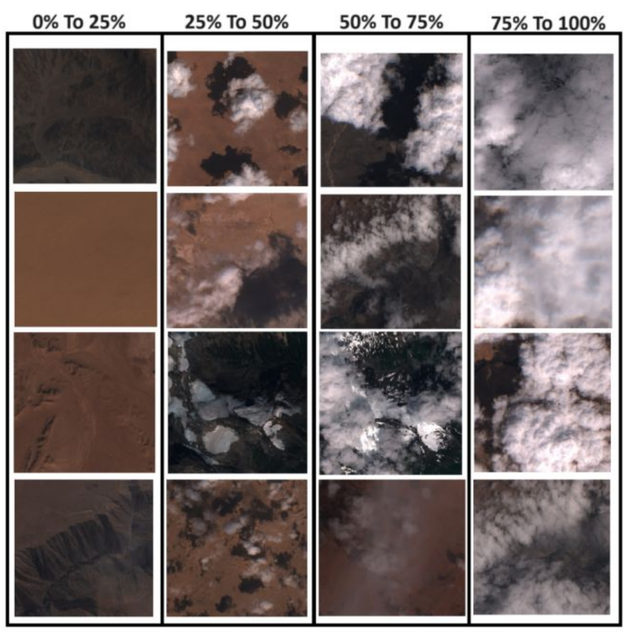
$$Z[i][j] = \sum_k X[i][k] * Y[k][j]$$

2D Convolution

$$Z[i][j] = \sum_m \sum_n X[i][j] \cdot Y[i-m][j-n]$$

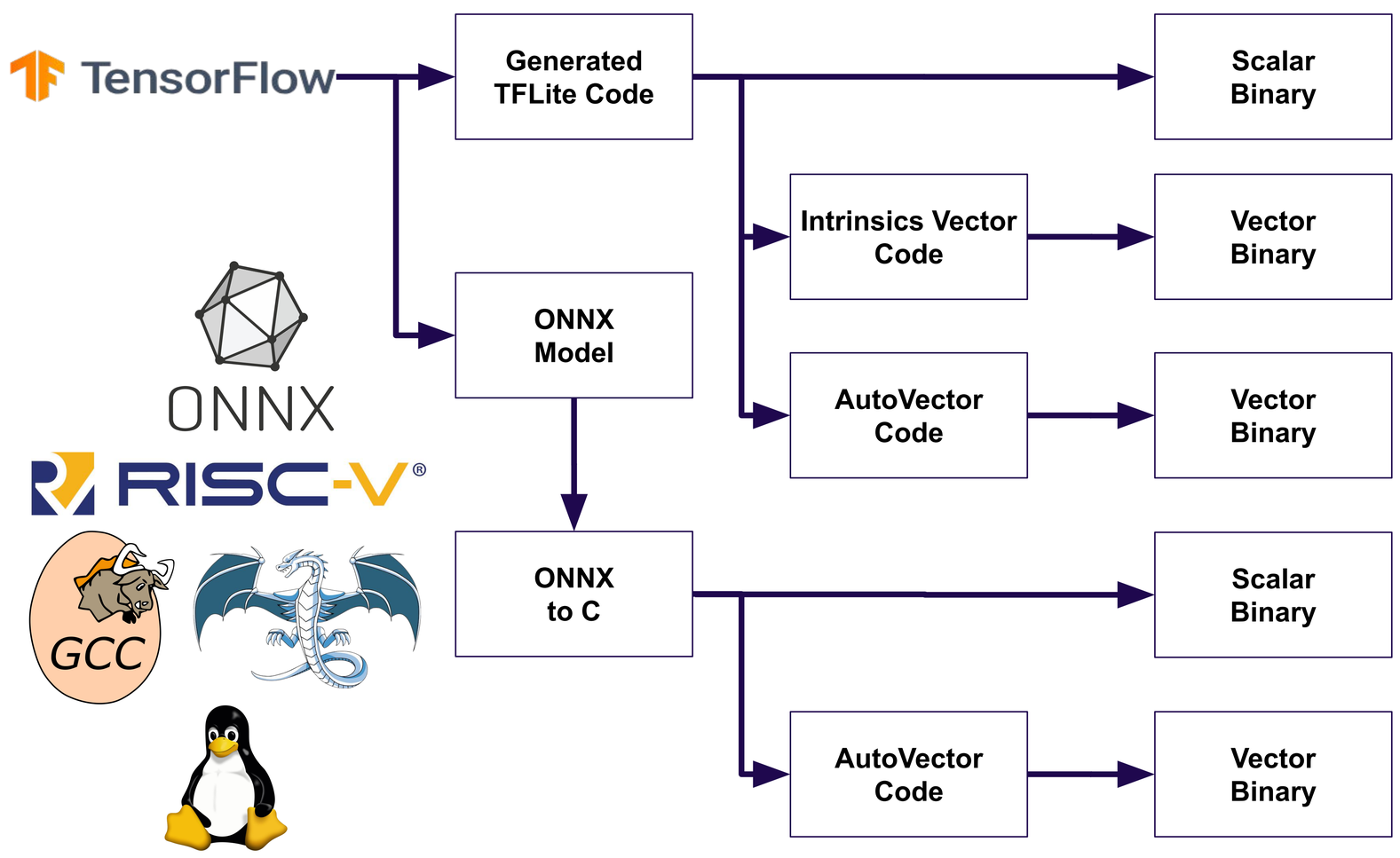


Vectorization process



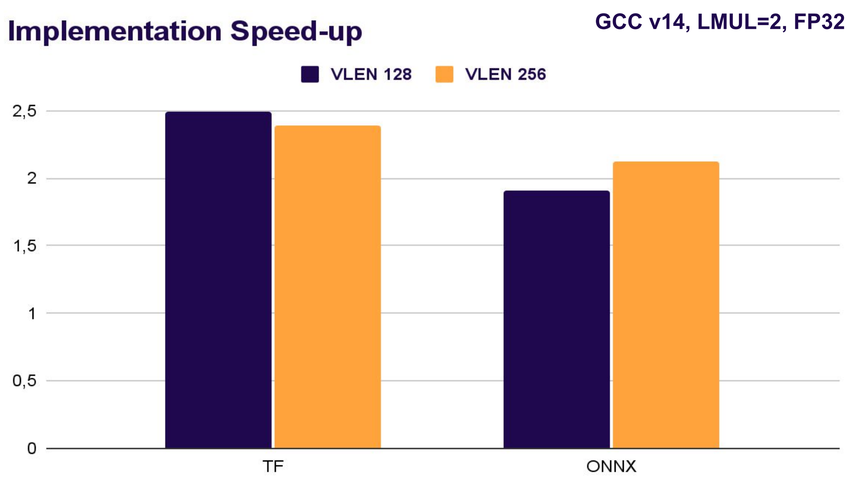
CNN MODELS PERFORMANCE

Model	Loss	Precision	Size (MB)
MobileNet	0.6398	0.8482	16.8
ResNet50	1.0511	0.8622	92.6
InceptionV3	0.8622	0.8502	86.1
VGG16	0.5807	0.8582	512
U-Net	0.1651	0.6874	51.1



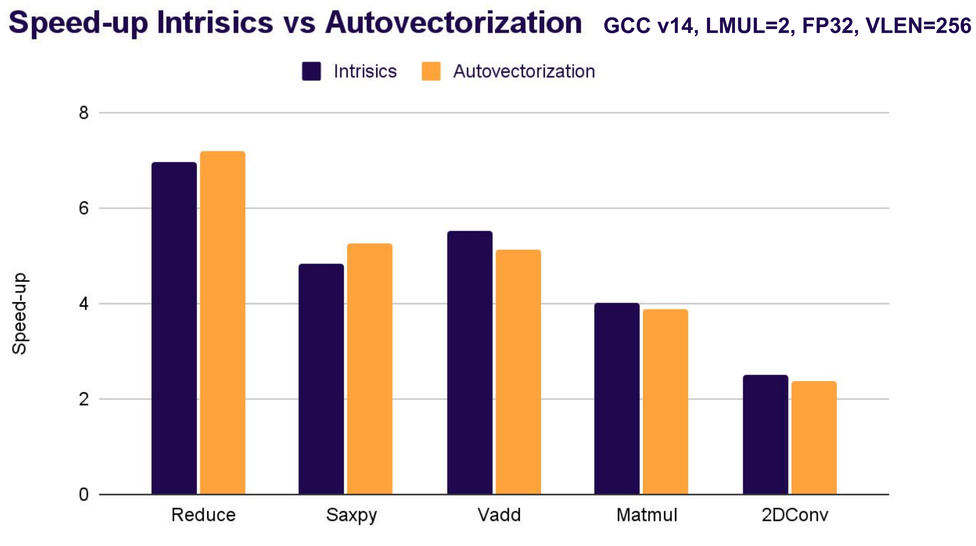
CNN Results

- Canaan MV-K230 & Banana Pi BPI-F3
- GCC autovec / RVV-1.0 / 128, 256 VLEN
- TF / ONNX on GNU/Linux



Autovectorization

- Compilers auto-vectorization works!
- Sweet point for VLEN, LMUL, Caches?



Contact
Màrius Montón
marius.monton@uab.cat