NEULINK SEMICONDUCTORS

NEUVERGE INFERENCE FRAMEWORK

WHITEPAPER

WWW.NEULINKSEMI.COM

NeuVerge

NeuVerge enables multiple downstream tasks using state-of-the-art convolutional neural network (CNN) models which can compress raw pixels into feature maps at high speed. These downstream tasks include computational vision (e.g. stereo correspondence, optical flow), object detection and segmentation, and image encoding for multimodal large language model (LLM) tokens [1-6].

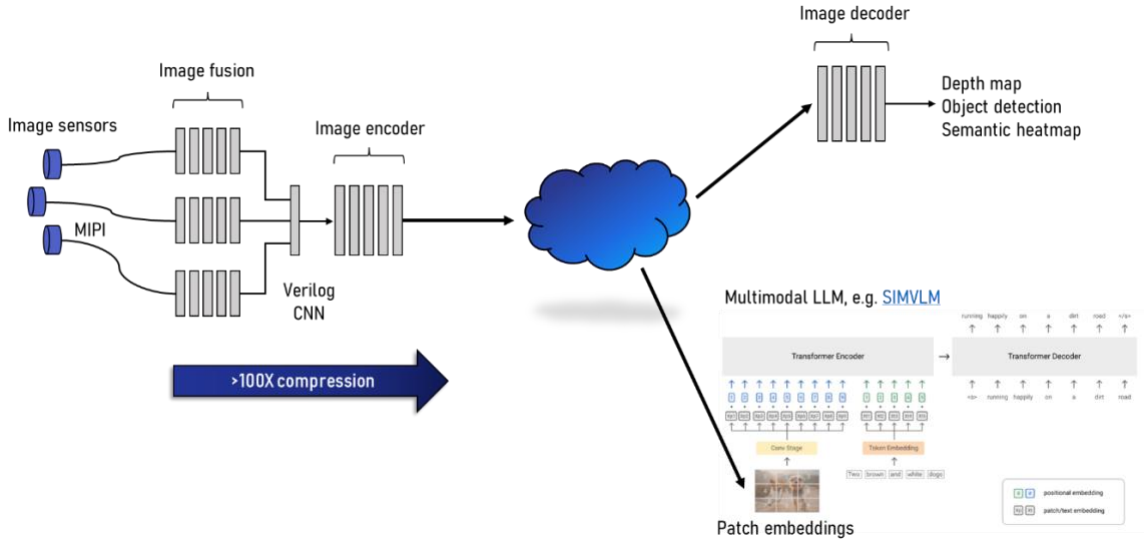

Figure 1

Current generation image sensors can generate high pixel rates e.g. 1600x1400@90fps=201M pixels/s per camera. To efficiently evaluate a CNN at this data rate, we use pure Verilog code to achieve three degrees of parallel implementation. First, in each layer the dot products for all output channels are computed in parallel using dedicated multiply/accumulate (MAC) hardware. Second, the incoming tensors to each layer are evaluated in parallel using vertical stripes, again with dedicated hardware. Third, all layers are computed in parallel using pipelining. The effective throughput of the Verilog CNN is proportional to the product of these three terms, which can easily produce 10K+ MAC units which all run in parallel. To keep that many MAC units busy, we must eliminate memory bottlenecks by moving the model weights from monolithic, off chip memories to distributed, on chip memories.

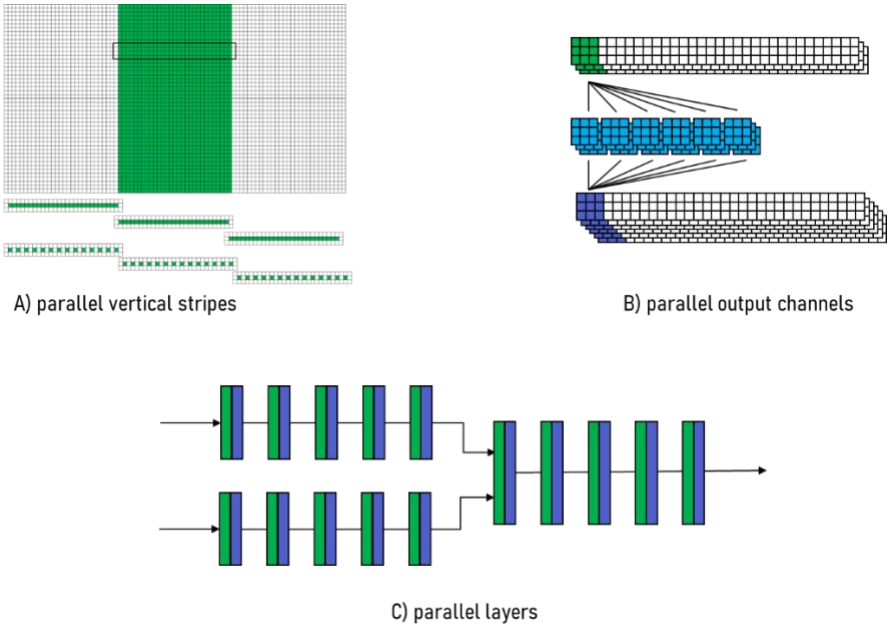

Figure 2

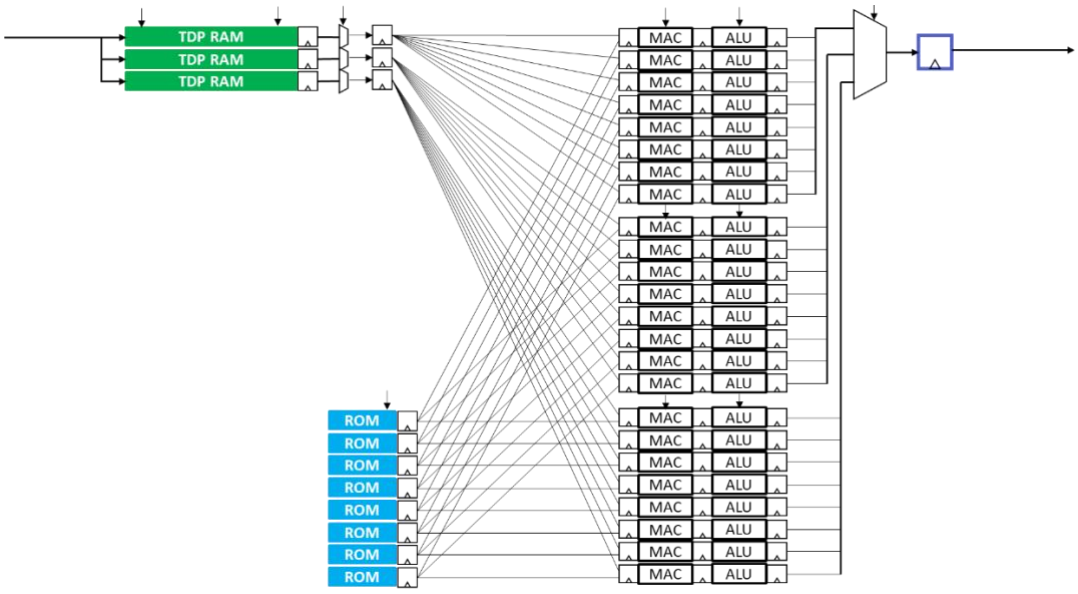

Figure 3

As part of NeuVerge, we have developed a compiler called Andromeda which converts a trained TFLite CNN model into pure Verilog code, using the parallel implementation described above. The compiler takes a target Verilog clock rate and input frame rate as parameters, adjusting the generated Verilog code as needed to achieve real time performance. The generated Verilog code uses a streaming architecture which uses only row buffers and operates on a fixed schedule without flow control. This results in a solution with very high throughput and low latency, with no external components. The generated Verilog code is a module level IP block with streaming AXI-S input and output interfaces. This IP can be simulated, compiled to an FPGA bitstream, or hardened into an ASIC block.

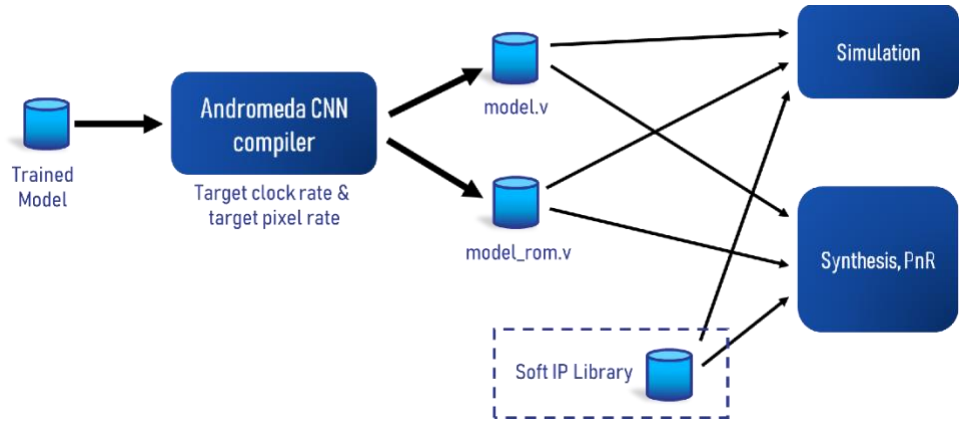

Figure 4

A typical application for this IP is a real time image encoder with the following benefits:

1) real time performance at maximum image sensor resolution and frame rate

2) supports fusion of multiple input frames into a single output feature map

3) can achieve >100X compression converting from raw pixels to features

4) can drive multiple downstream tasks, including traditional machine vision and multimodal LLM

5) larger models can be ASIC hardened using higher target clock rate and ROM based weights

Model training can follow different strategies. One approach is to use continuous learning with reconfigurable FPGA bitstreams, enabling the model to learn an unknown distribution over time. Another approach is to collect a large dataset and train a foundation model with static weights and hardened ASIC implementation. This will result in the highest performance and smallest area, for high volume applications.

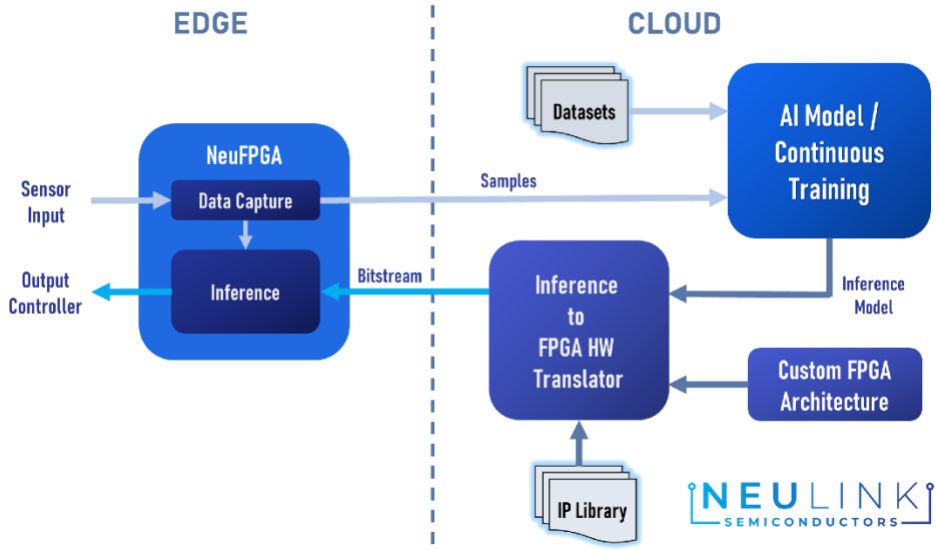

Figure 5

Reference	
1	SIMVLM
2	A ConvNet for the 2020s
3	CLIP
4	Self-supervised Visual Feature Learning
5	AoCStream: All-on-Chip CNN Accelerator
6	GPU Utilization and CNN Inference