UNIVERSITY of VIRGINIA

Design Space Exploration for Compressed Deep Convolutional Neural Network on SCALE Sim

Peilin Chen, Xinyuan Fu, Hanyuan Gao

6501 Group6

2024/11/07

➢ The hardware platform we use

- Systolic CNN AcceLErator Simulator (SCALE Sim)

Parameter	Value
Array Height	32
Array Width	32
IFMAP SRAM Size	1024
Filter SRAM Size	1024
OFRAM SRAM Size	128
Dataflow	WS

Config file with architecture specs

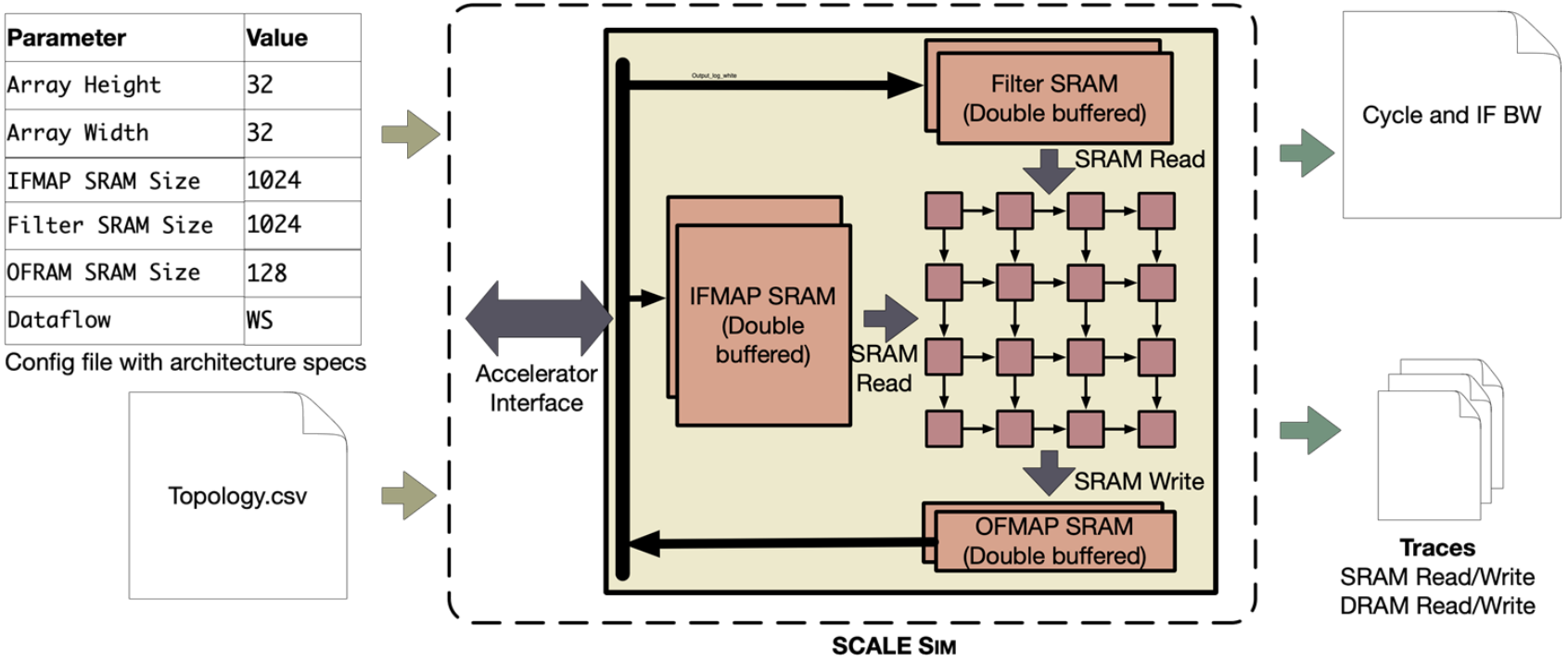

Fig. Overview of SCALE Sim

SCALE Sim takes two files as input from the user: one is the hardware configuration, the other is the neural network topology (workload).

SCALE Sim generates two types of outputs: one is the cycle accurate traces for SRAM and DRAM, the other is the metrics like cycle counts, utilization, bandwidth requirements, total data movement, etc.

https://github.com/scalesim-project/scale-sim-v2?tab=readme-ov-file

Samajdar, Ananda, et al. "A systematic methodology for characterizing scalability of dnn accelerators using scale-sim." 2020 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS). IEEE, 2020.

➢ What we want to do

■ Neural network quantization and hardware design space exploration

	Software part	Hardware part	
The CNNs we use →	CNNs like LeNet-5, VGG, and ResNet are trained using FP32 or FP16 datatype.	Systolic CNN AcceLErator Simulator (SCALE Sim).	← **The hardware we use**
Quantization method →	Weight and Activation co-quantization. W16A16 (weight: int16, activation: int16) W8A8 W4A4 W2A2 Per-tensor, Per-channel, Group-wise……	Hardware design space exploration (DSE). Systolic array size->Height and Width IFMAP SRAM size Filter/Kernel SRAM size OFMAP SRAM size Dataflow-> input stationary, output stationary, and weight stationary Scale up or Scale out ……	← **Design space**
Object →	Compress CNN as much as possible while maintaining the accuracy.	Explore and find the optimal hardware configurations for different CNNs.	← **Object**

UNIVERSITY of VIRGINIA

Thanks