http://synergy.ece.gatech.edu

SCALE-Sim
Systolic CNN Accelerator Simulator

Open sourced at

https://github.com/ARM-software/SCALE-Sim

Project website

https://scalesim-project.github.io/

ASPLOS 2021 **April 16, 2021**

Kind notice

Please note that all the sessions of this tutorial is being recorded

Outline

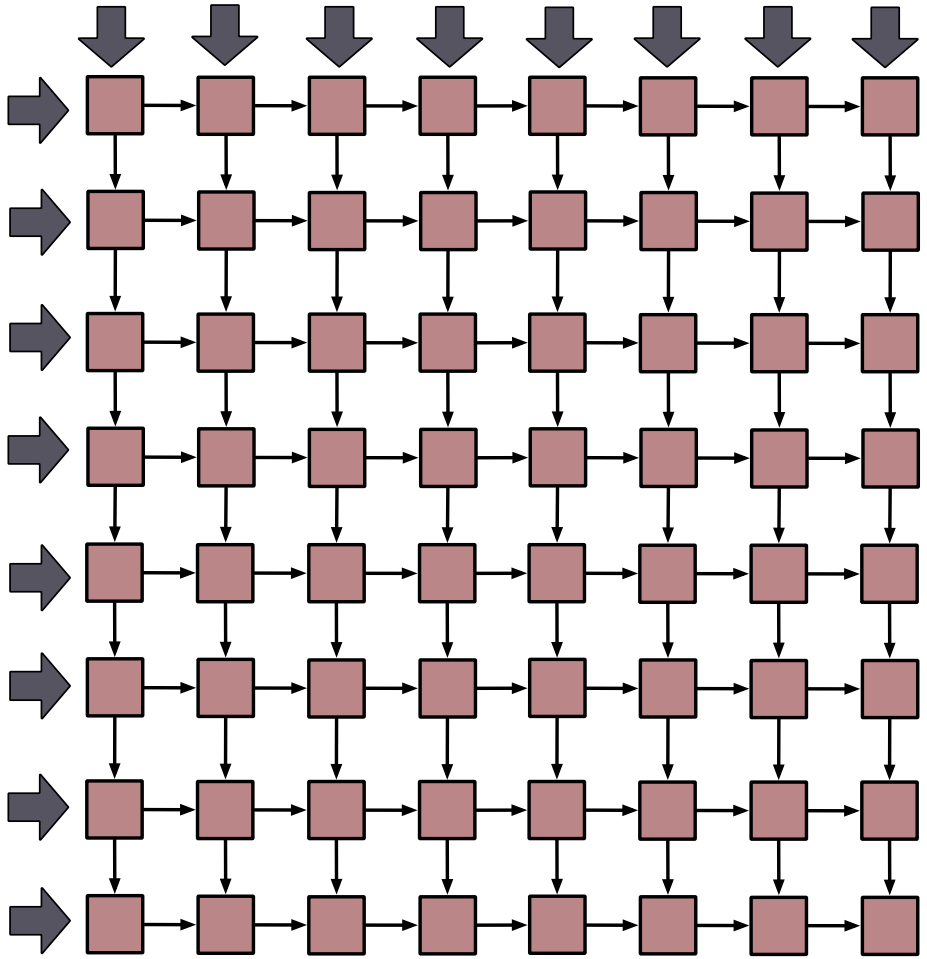
1. **Simulating for DNN accelerator**
   - Motivation
   - Metrics of interest

2. **SCALE-Sim**
   - Overview
   - Modelling compute, memory, and interface
   - Modelling GEMM
   - Modelling Convolutions
   - Dataflows
   - Outputs

3. **Demos**

Systolic arrays

High parallelism



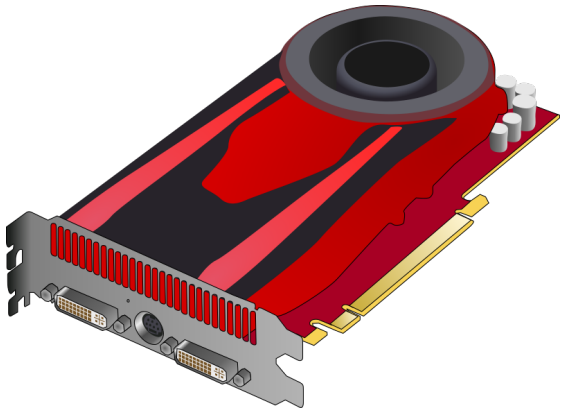
Efficient data reuse

Simple implementation

High scalability

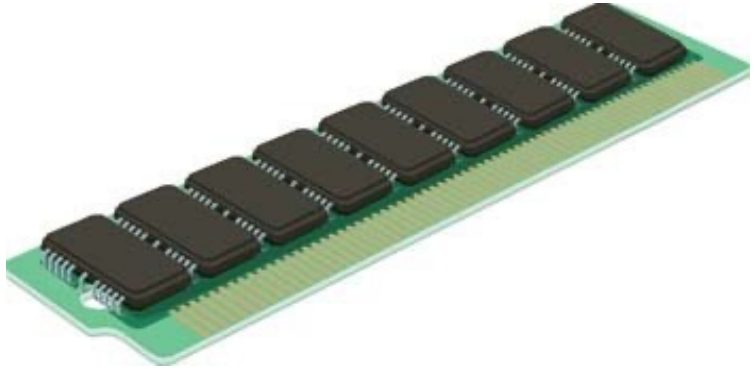
Metrics of interest

Modeling compute



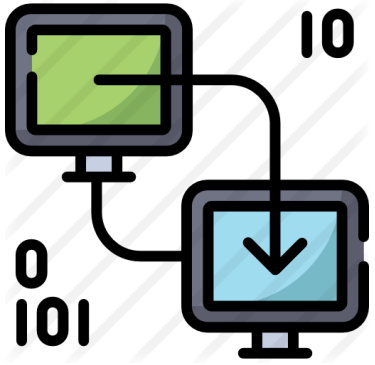
- Performance
- Efficiency
- Scalability

Modeling memory



- Reuse
- Performance
- Efficiency

Modeling interface



- System level implications
- Performance

Outline

1. **Simulating for DNN accelerator**
   - Motivation
   - Metrics of interest

2. **SCALE-Sim**
   - Overview
   - Modelling compute, memory, and interface
   - Modelling GEMM
   - Modelling Convolutions
   - Dataflows
   - Outputs

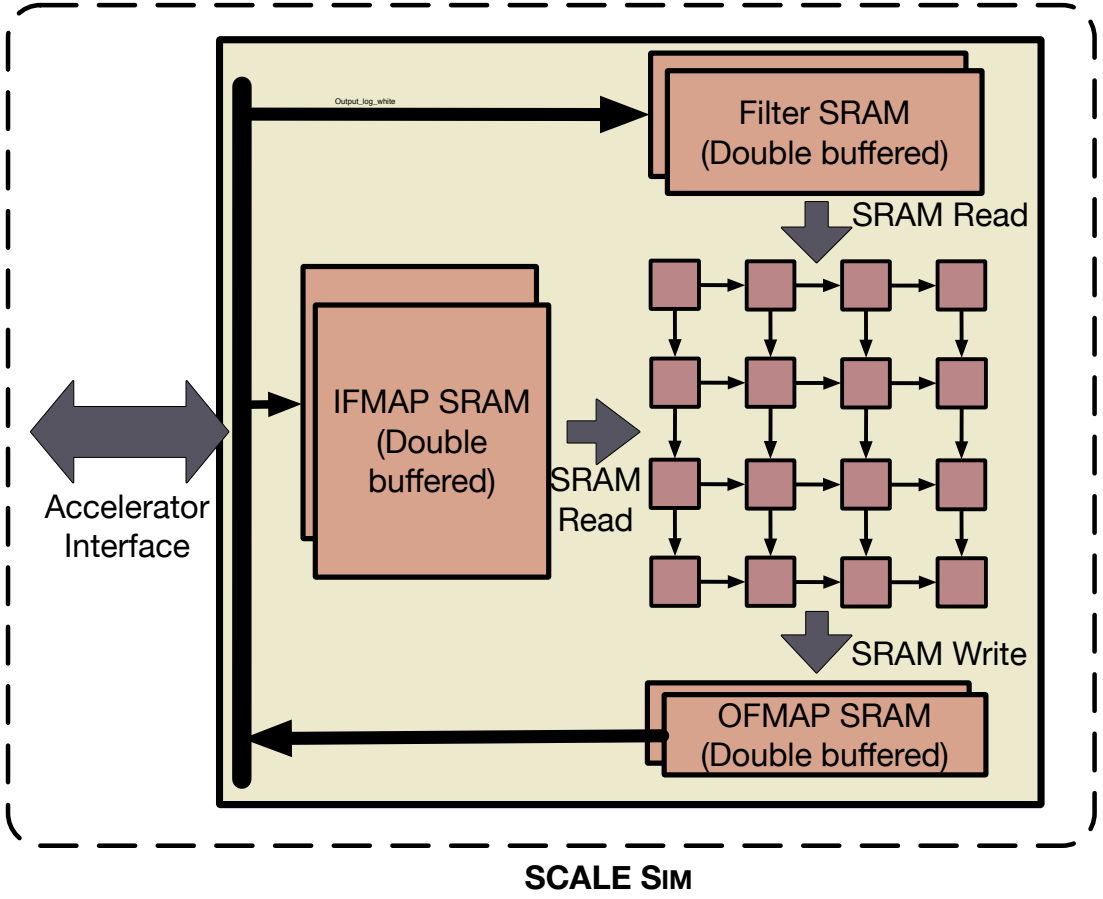
3. **Demos**

SCALE-Sim: Overview

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

Topology.csv



Cycle and IF BW

Traces
SRAM Read/Write
DRAM Read/Write

Inputs

Outputs

Inputs: Config file

```
1   [general]
2   run_name = "GoogleTPU_os"
3
4   [architecture_presets]
5   ArrayHeight:    256
6   ArrayWidth:     256
7   IfmapSramSz:    8192
8   FilterSramSz:   8192
9   OfmapSramSz:    8192
10  IfmapOffset:    0
11  FilterOffset:   10000000
12  OfmapOffset:    20000000
13  Dataflow:       os
```

Array microarchitecture

Memory Sizes

Matrix offsets

Data flow

Inputs: Topology CSV file

	Layer name	IFMAP Height	IFMAP Width	Filter Height	Filter Width	Channels	Num Filter	Strides
2	Conv1	224	224	7	7	3	64	2
3	Conv2red	56	56	1	1	64	64	1
4	Conv2	56	56	3	3	64	192	1
5	Inc3a_1x1	28	28	1	1	192	64	1
6	Inc3a_3x3red	28	28	1	1	192	96	1
7	Inc3a_3x3	28	28	3	3	96	128	1

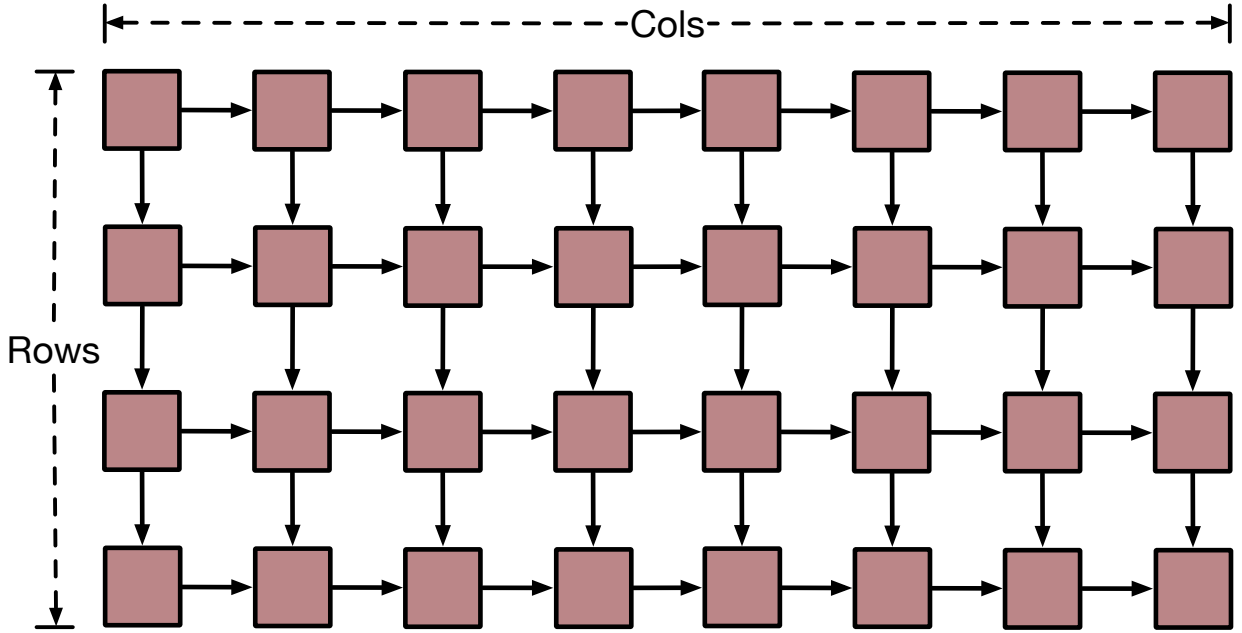
Layer-by-layer configuration

Network hyperparameters

SCALE-Sim

Modelling compute

- Support for non-square arrays
- Support for multiple dataflows
- Fast execution by tracking the edges
- Layer by layer execution
- Intrinsic folding of big compute



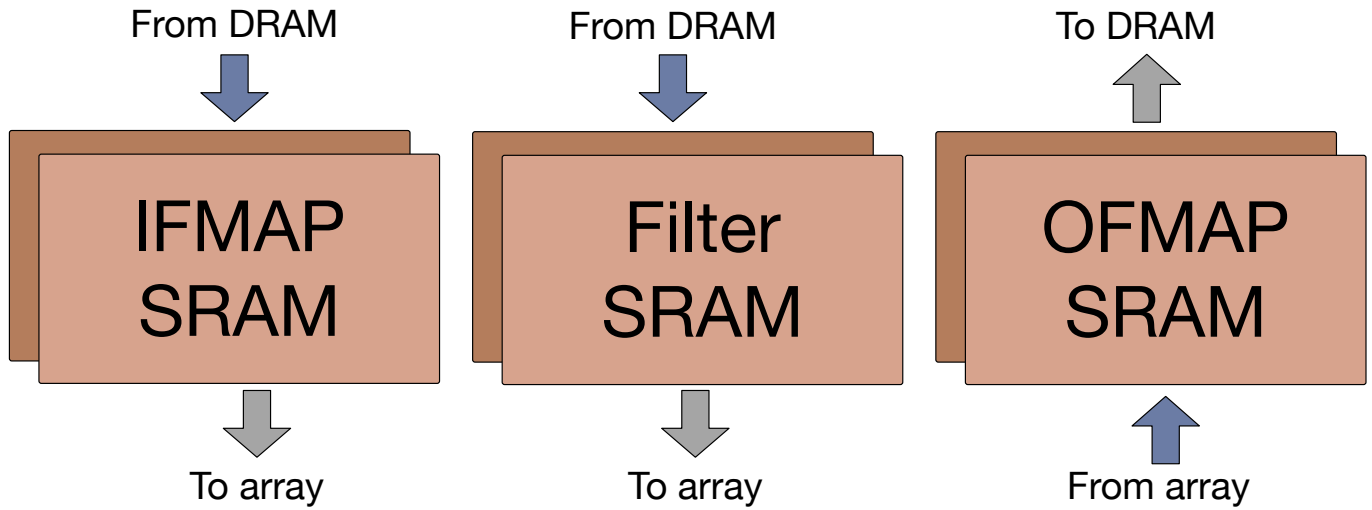
SCALE Sim

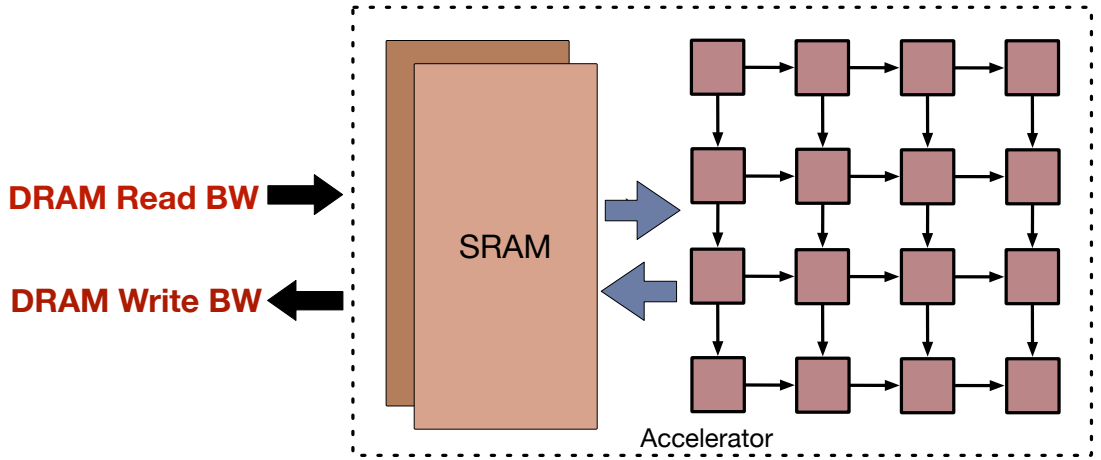
Modelling on-chip memory

- Double buffered memories
- Models three memory regions one for each matrix
- No replication of matrix elements in SRAM buffers

Modelling system interface

- Tool outputs the required DRAM read and write bandwidth requirements





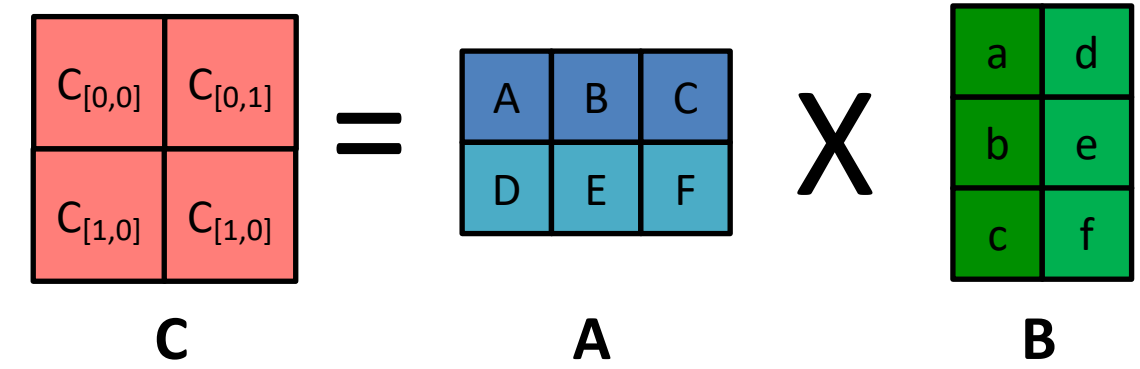
Outline

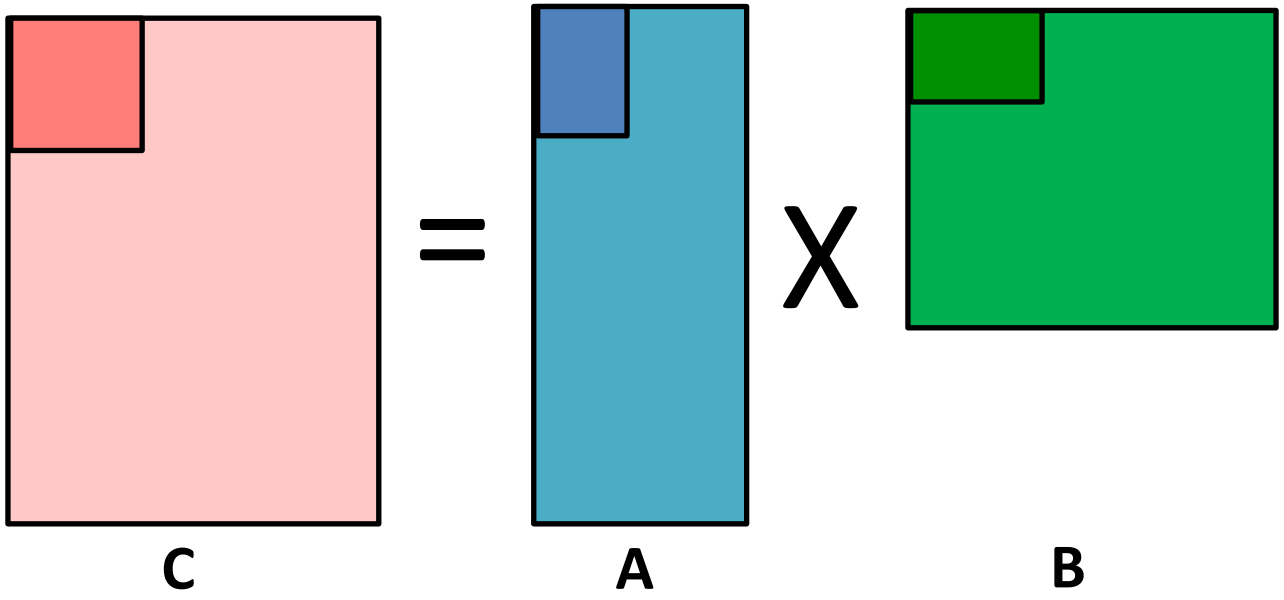
1. **Simulating for DNN accelerator**
   - Motivation
   - Metrics of interest

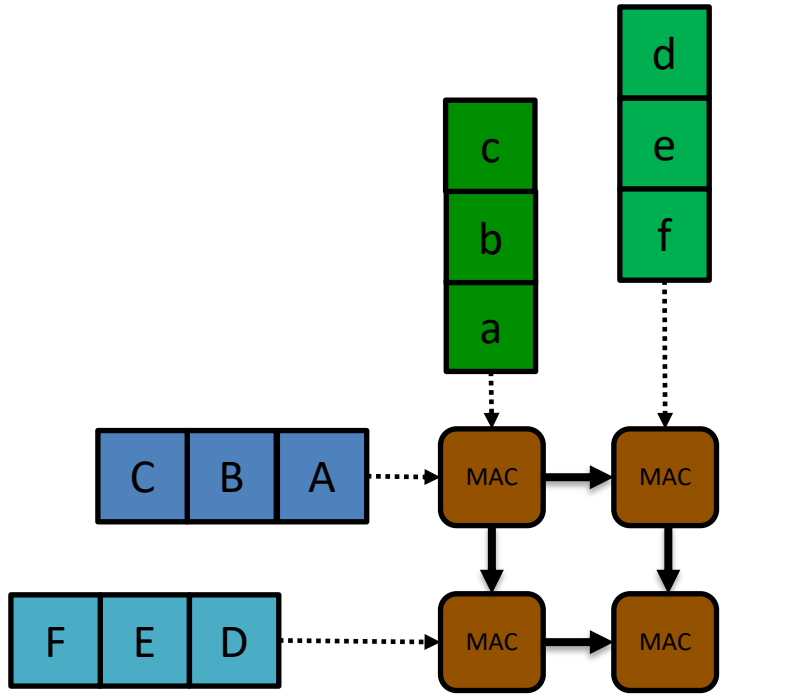
2. **SCALE-Sim**
   - Overview
   - Modelling compute, memory, and interface
   - Modelling GEMM
   - Modelling Convolutions
   - Dataflows
   - Outputs

3. **Demos**

Modelling GEMM

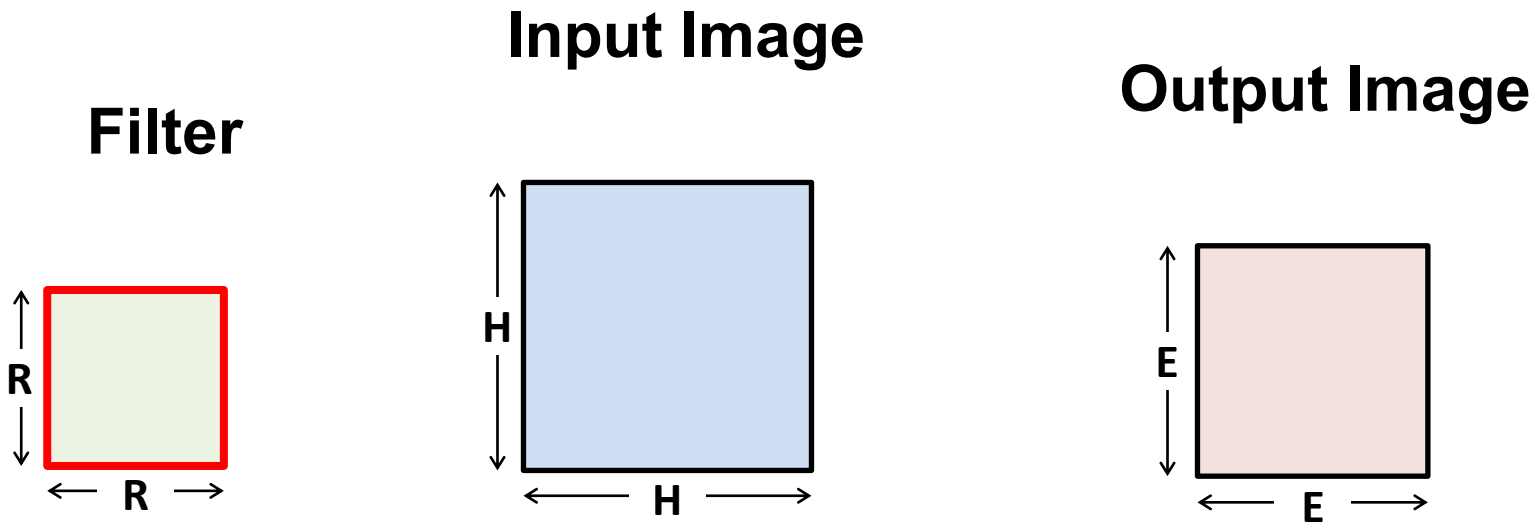




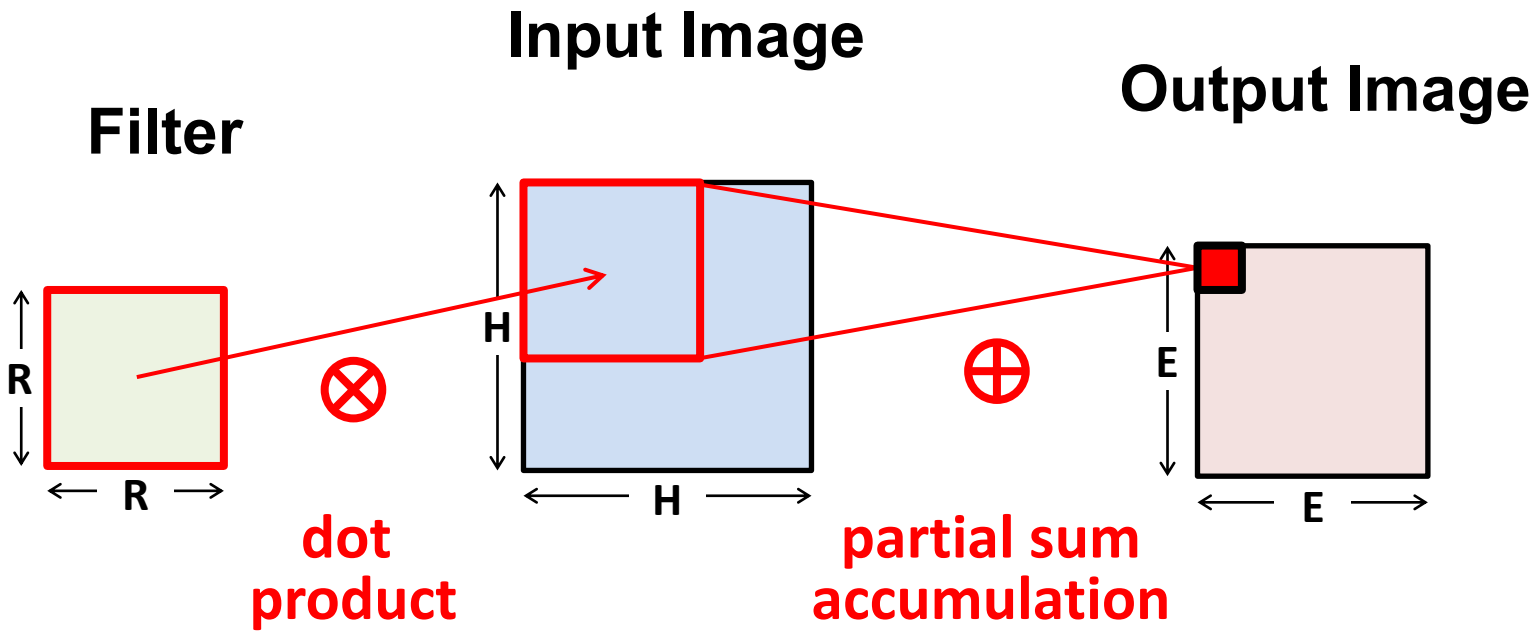


2x2 systolic array

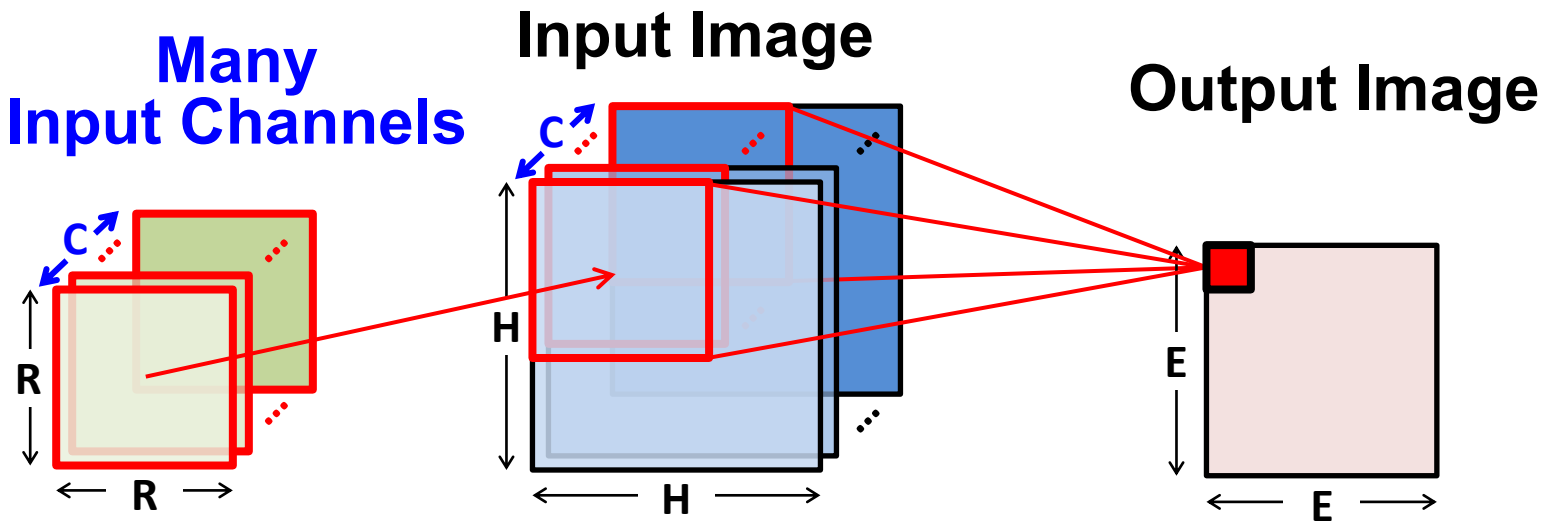
Convolution in CNN



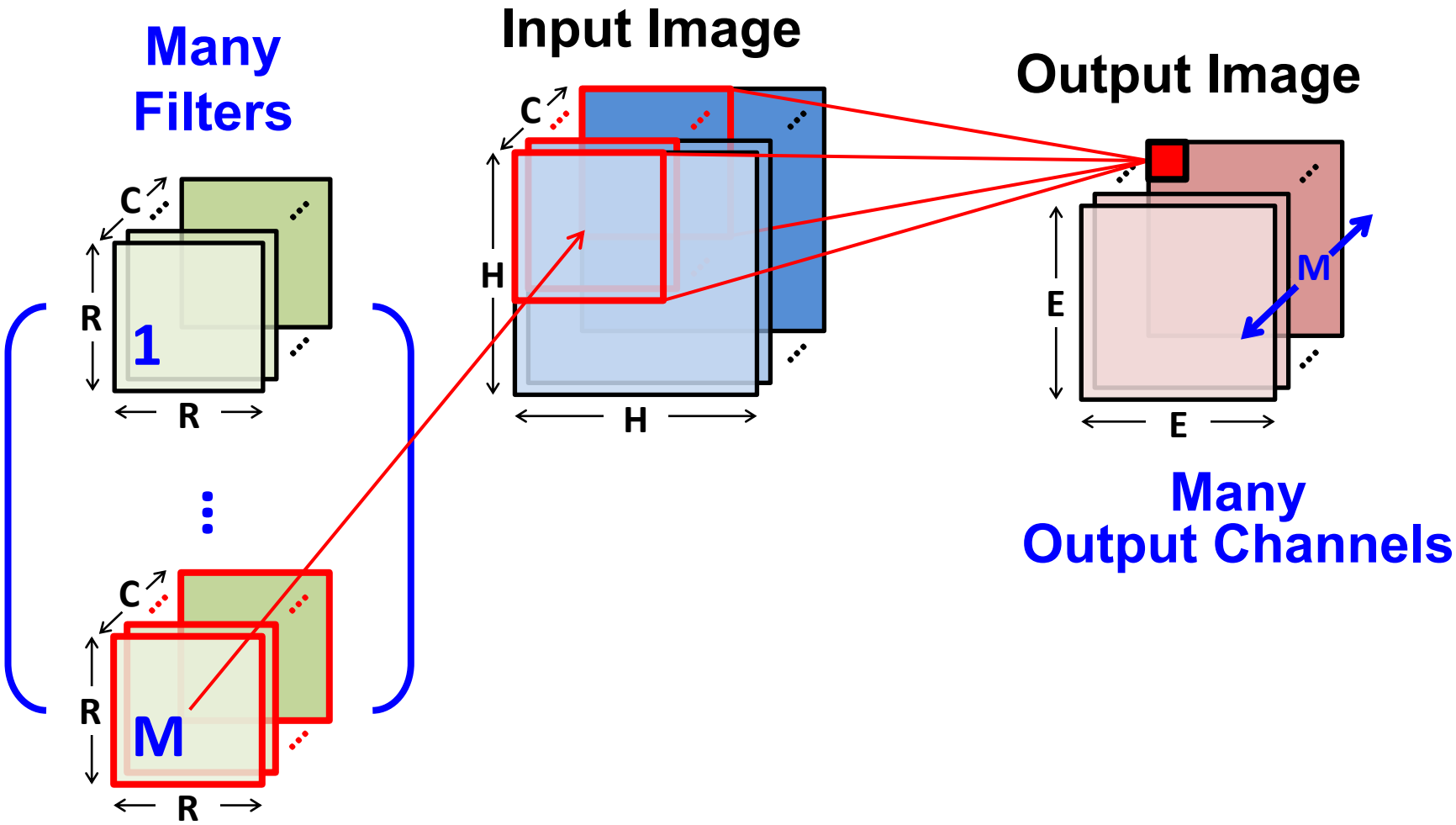
Convolution in CNN



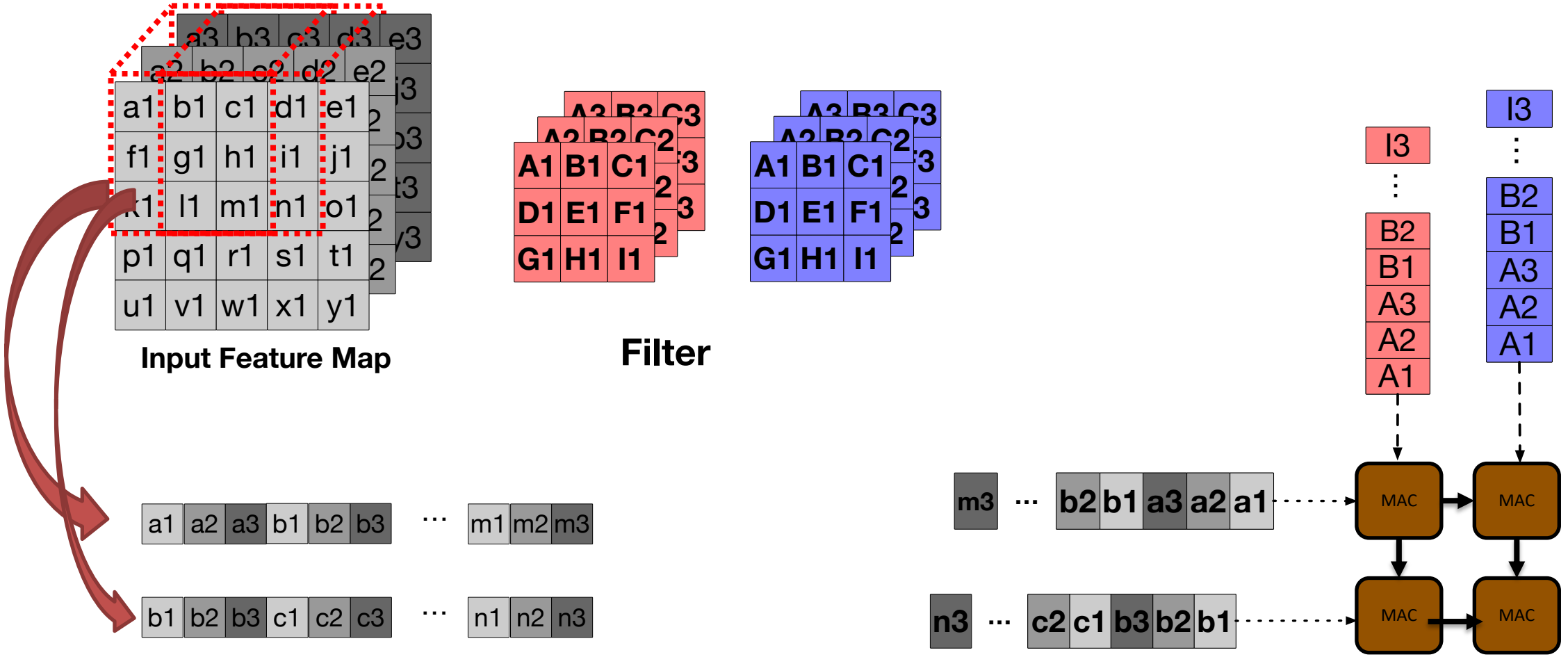
Convolution in CNN



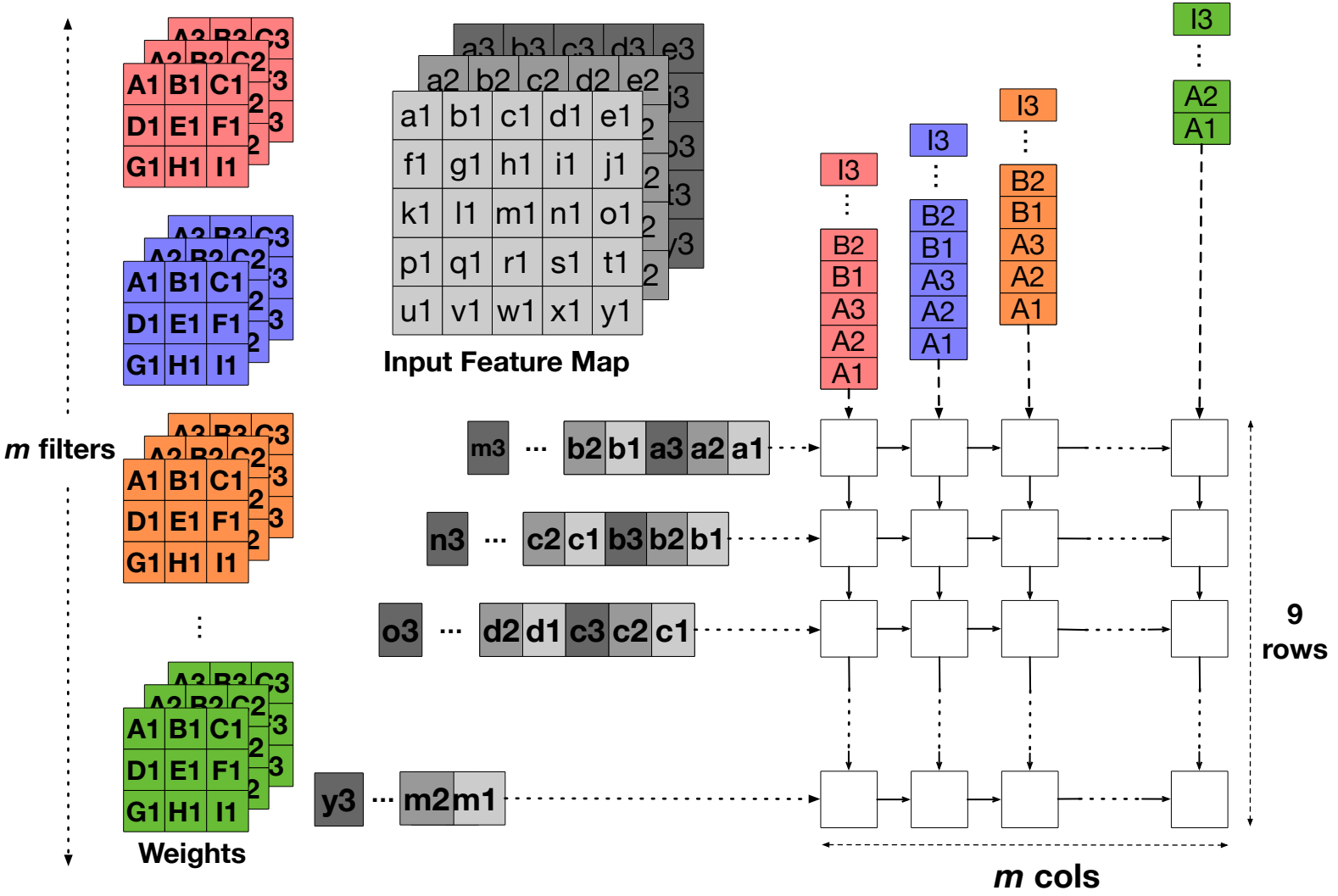
Convolution in CNNs



SCALE-Sim: Modelling convolutions



Dataflows: Output Stationary



Each MAC unit responsible for particular output pixels

Accumulation of partial sums done locally

Each column generates pixels from different output channel

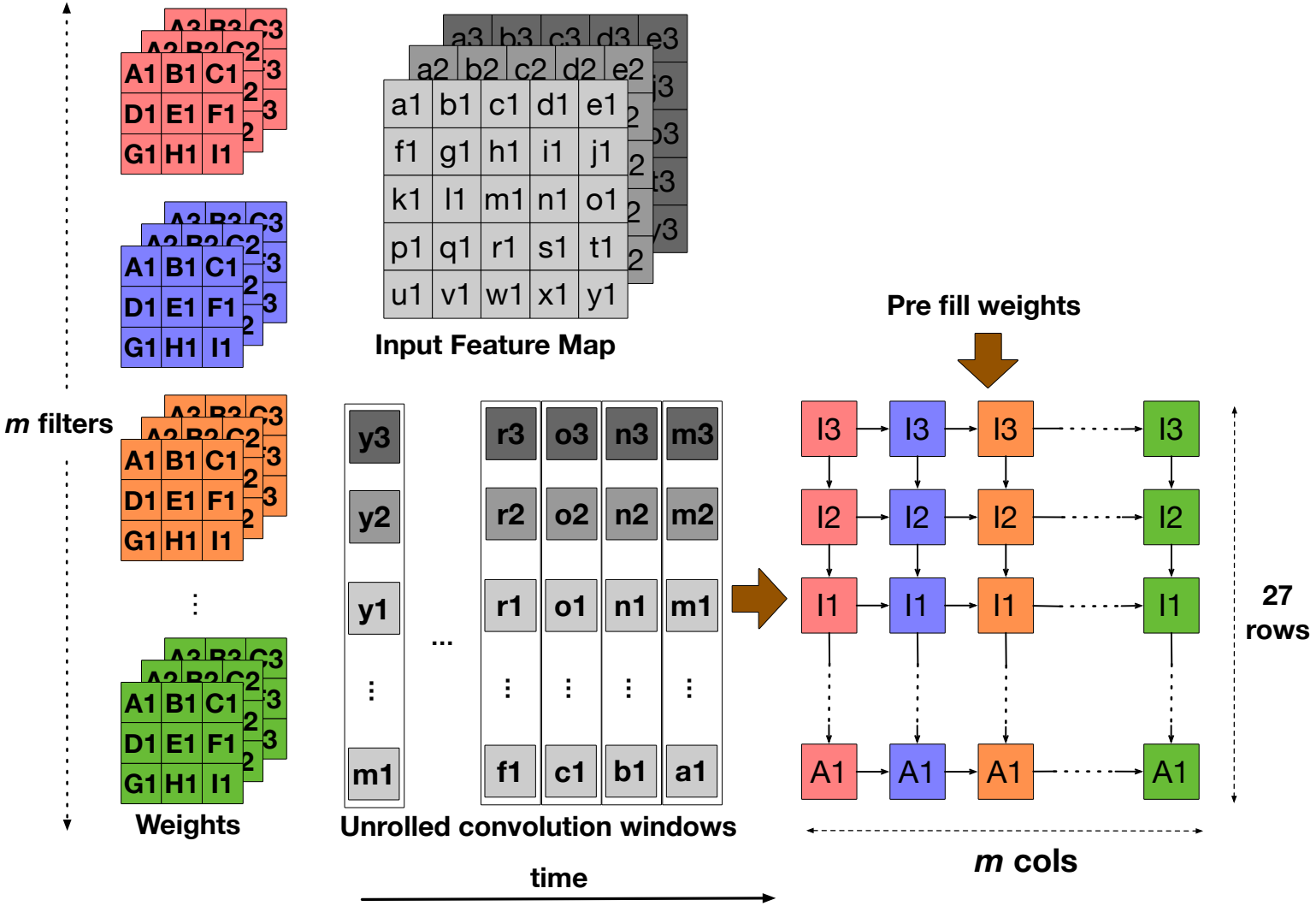
SCALE-Sim assumes output collection is not on critical path

Maximum usable dimensions

Rows: Pixels per output channel

Cols: Number of filters

Dataflows : Weight Stationary



Elements of filters are pre-filled into MAC units

Every column is assigned unique filter

Reduction is done across the rows within a column

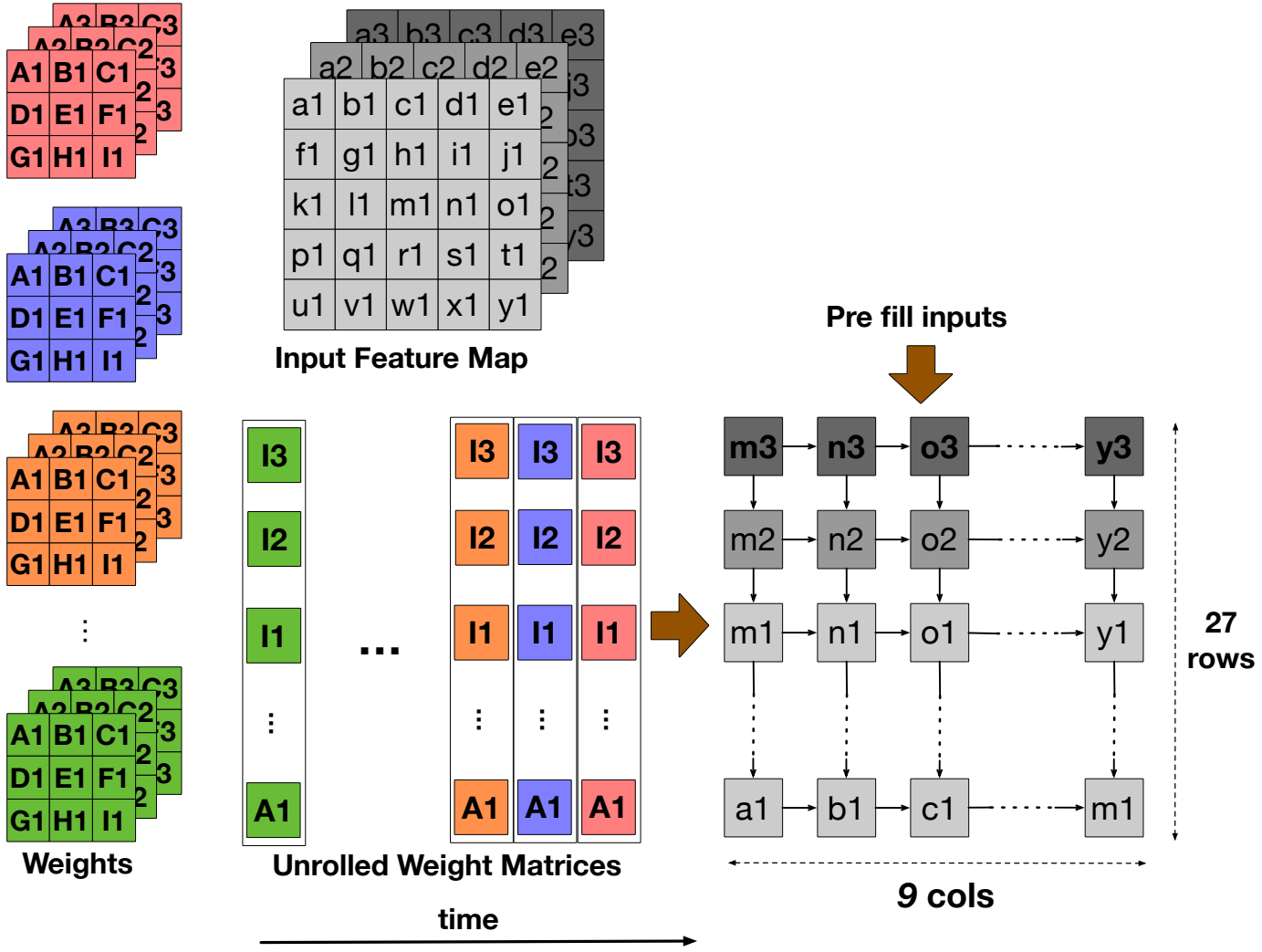
Critical path contains time to fill in weights, partial sum generation and reduction time

Maximum usable dimensions

Rows: Partial sums per pixel

Cols: Number of filters

Dataflows : Input Stationary



Elements of convolution windows are pre-filled into MAC units

Every column is assigned output pixel

Reduction is done across the rows within a column

Critical path contains time to fill in input elements, partial sum generation and reduction time

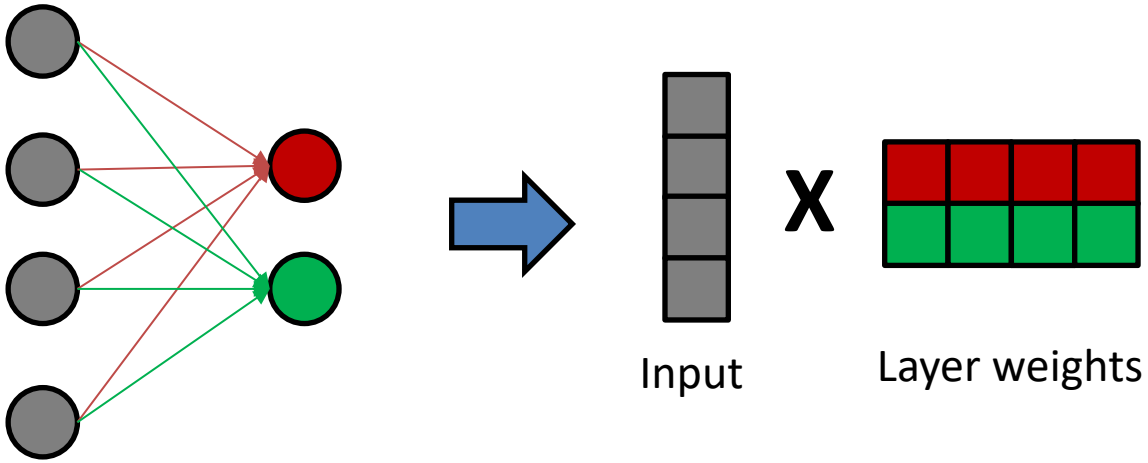
Maximum usable dimensions

Rows: Partial sums per pixel

Cols: Number of output pixels per output channel

Supporting other layer types

Fully connected



- **Can be modelled as matrix vector multiplication**
- **SCALE-Sim models as convolution with input dimension same as weight dimension**

✓ **LSTMs**

✓ **Attention**

} Modelled as **Matrix-Vector** or **Vector-Vector**

✗ **Softmax**

✗ **Pooling**

} **Elementwise operations** Not efficient on systolic arrays

Outline

1. **Simulating for DNN accelerator**
   - Motivation
   - Metrics of interest

2. **SCALE-Sim**
   - Overview
   - Modelling compute, memory, and interface
   - Modelling GEMM
   - Modelling Convolutions
   - Dataflows
   - Outputs

3. **Demos**

Console Output

```
====================================================
****************** SCALE SIM **********************
====================================================
Array Size:     32x32
SRAM IFMAP (kB):        64
SRAM Filter (kB):       64
SRAM OFMAP (kB):        64
Dataflow:       Weight Stationary
CSV file path:  /content/drive/MyDrive/scalesim_resources/topologies/conv_nets/alexnet_part.csv
Number of Remote Memory Banks:  1
Bandwidth:      10
Working in USE USER BANDWIDTH mode.
====================================================

Running Layer 0
100%|██████████| 112284/112284 [00:59<00:00, 1893.41it/s]
Compute cycles: 439609
Stall cycles: 327326
Overall utilization: 23.42%
Mapping efficiency: 94.53%
Average IFMAP DRAM BW: 9.997 words/cycle
Average Filter DRAM BW: 9.998 words/cycle
Average OFMAP DRAM BW: 7.907 words/cycle
Saving traces: Done!
************ SCALE SIM Run Complete ****************
```

1

1 **Summary of input configurations**

Console Output

```
====================================================
****************** SCALE SIM **********************
====================================================
Array Size:     32x32
SRAM IFMAP (kB):        64
SRAM Filter (kB):       64
SRAM OFMAP (kB):        64
Dataflow:       Weight Stationary
CSV file path:  /content/drive/MyDrive/scalesim_resources/topologies/conv_nets/alexnet_part.csv
Number of Remote Memory Banks:  1
Bandwidth:      10
Working in USE USER BANDWIDTH mode.
====================================================

Running Layer 0
100%|██████████| 112284/112284 [00:59<00:00, 1893.41it/s]
Compute cycles: 439609
Stall cycles: 327326
Overall utilization: 23.42%
Mapping efficiency: 94.53%
Average IFMAP DRAM BW: 9.997 words/cycle
Average Filter DRAM BW: 9.998 words/cycle
Average OFMAP DRAM BW: 7.907 words/cycle
Saving traces: Done!
************ SCALE SIM Run Complete ****************
```

2

1 **Summary of input configurations**

2 **Run and stall cycles**

Console Output

```
==========================================================
******************** SCALE SIM **********************
==========================================================
Array Size:     32x32
SRAM IFMAP (kB):        64
SRAM Filter (kB):       64
SRAM OFMAP (kB):        64
Dataflow:       Weight Stationary
CSV file path:  /content/drive/MyDrive/scalesim_resources/topologies/conv_nets/alexnet_part.csv
Number of Remote Memory Banks:  1
Bandwidth:      10
Working in USE USER BANDWIDTH mode.
==========================================================

Running Layer 0
100%|██████████| 112284/112284 [00:59<00:00, 1893.41it/s]
Compute cycles: 439609
Stall cycles: 327326
Overall utilization: 23.42%
Mapping efficiency: 94.53%
Average IFMAP DRAM BW: 9.997 words/cycle
Average Filter DRAM BW: 9.998 words/cycle
Average OFMAP DRAM BW: 7.907 words/cycle
Saving traces: Done!
************ SCALE SIM Run Complete ****************
```

1. **Summary of input configurations**
2. **Run and stall cycles**
3. **Mapping efficiency and compute utilization**

Console Output

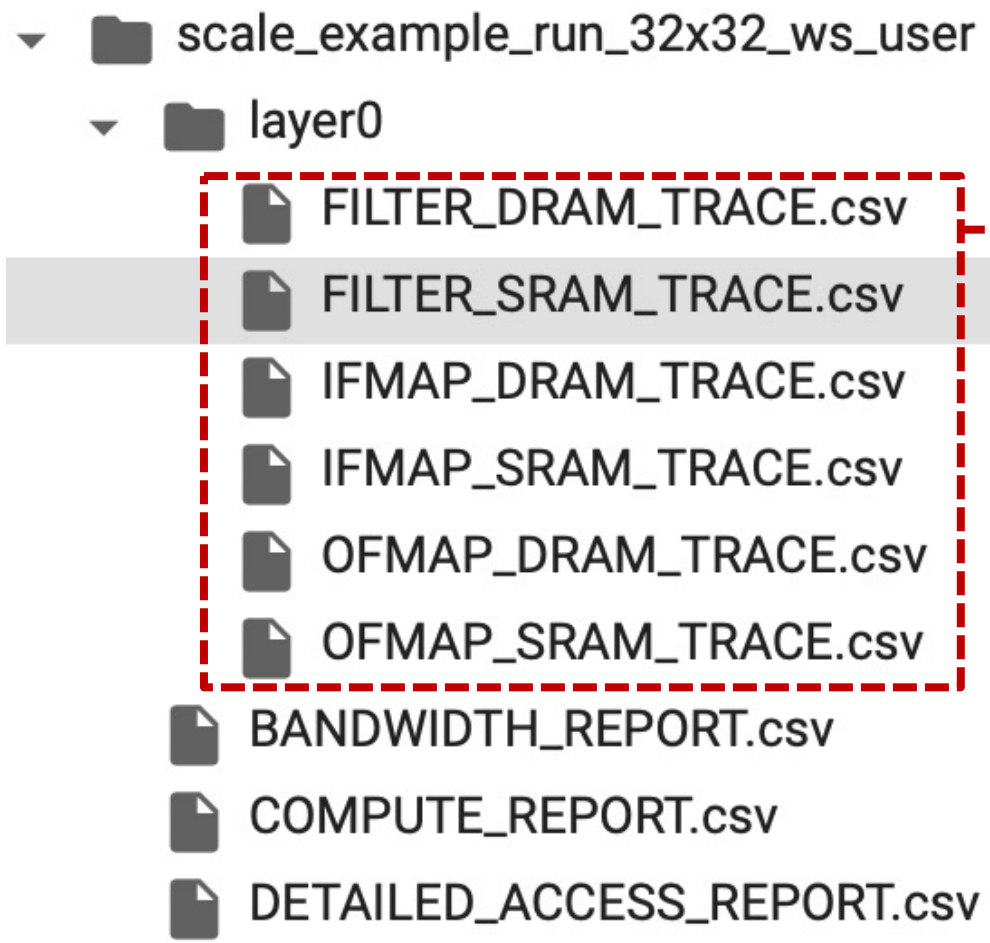
```
==========================================================
******************** SCALE SIM **********************
==========================================================
Array Size:     32x32
SRAM IFMAP (kB):        64
SRAM Filter (kB):       64
SRAM OFMAP (kB):        64
Dataflow:       Weight Stationary
CSV file path:  /content/drive/MyDrive/scalesim_resources/topologies/conv_nets/alexnet_part.csv
Number of Remote Memory Banks:  1
Bandwidth:      10
Working in USE USER BANDWIDTH mode.
==========================================================

Running Layer 0
100%|██████████| 112284/112284 [00:59<00:00, 1893.41it/s]
Compute cycles: 439609
Stall cycles: 327326
Overall utilization: 23.42%
Mapping efficiency: 94.53%
Average IFMAP DRAM BW: 9.997 words/cycle
Average Filter DRAM BW: 9.998 words/cycle
Average OFMAP DRAM BW: 7.907 words/cycle
Saving traces: Done!
************ SCALE SIM Run Complete ****************
```

4

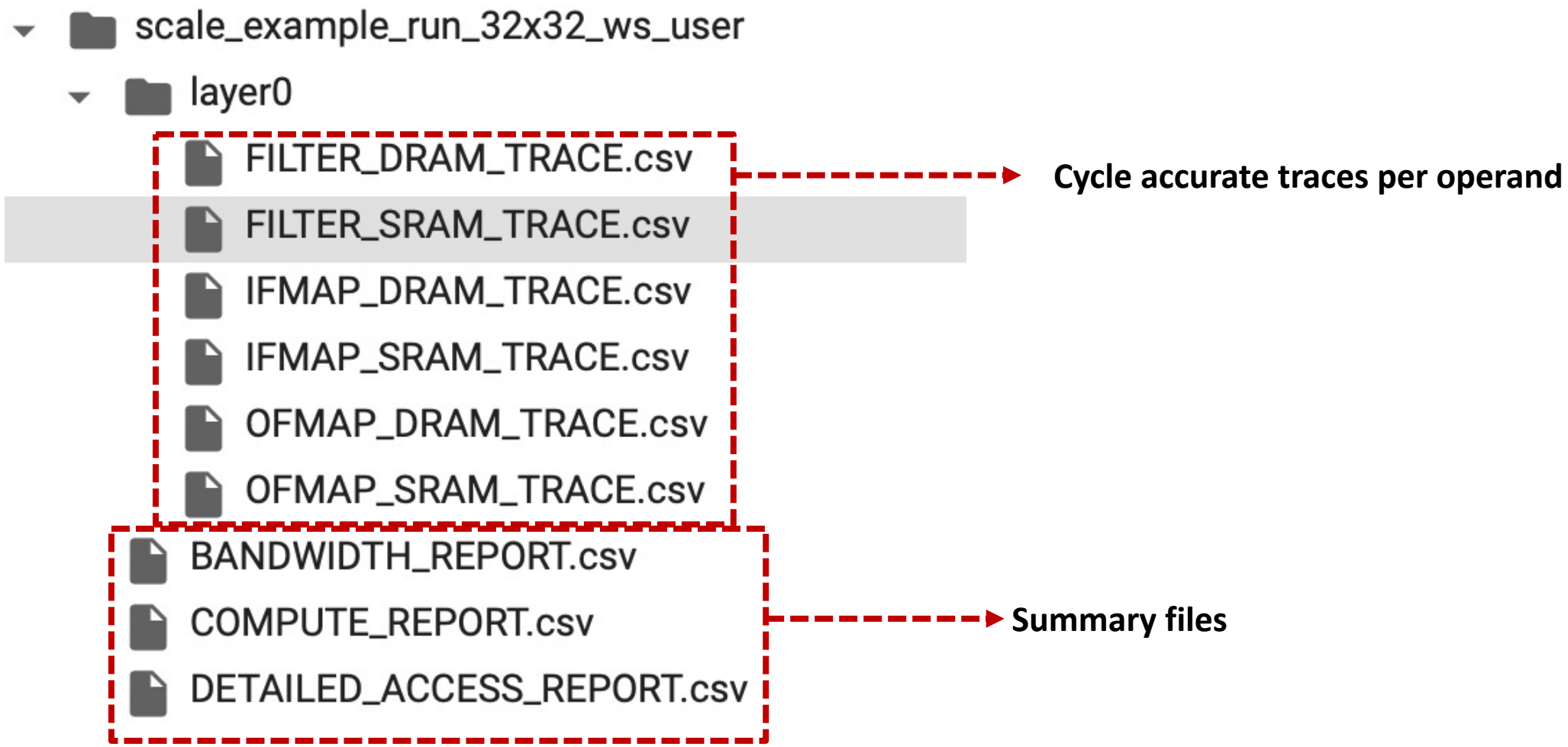
1. **Summary of input configurations**
2. **Run and stall cycles**
3. **Mapping efficiency and compute utilization**
4. **Off chip access bandwidth**

Generated outputs



Cycle accurate traces per operand

Generated outputs



Summary Files

Filename	Attributes
COMPUTE_REPORT.csv	Layer wise compute cycles, stall cycles, mapping utilization etc
BANDWIDTH_REPORT.csv	Layer wise SRAM and DRAM access bandwidths for operands
DETAILED_ACCESS_REPORT.csv	Access counts and timing informataion

Announcement!

SCALE-Sim v2 Release (Beta)

We are releasing a new version of SCALE-Sim : https://github.com/scalesim-project/scale-sim-v2

We will soon have a stable version repo in ARM's Github

Announcement!

SCALE-Sim v2 Release (Beta)

We are releasing a new version of SCALE-Sim : https://github.com/scalesim-project/scale-sim-v2

We will soon have a stable version repo in ARM's Github

New features

1. Tool can be run in both stall free and bandwidth limited mode
2. New metrics like mapping efficiency, stall count added
3. Modular code
4. Available as python package
5. More enhancements in the pipeline!

Announcement!

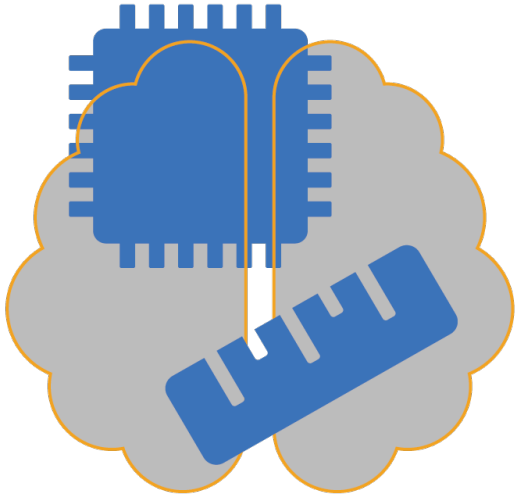
SCALE-Sim v2 Release (Beta)

We are releasing a new version of SCALE-Sim : https://github.com/scalesim-project/scale-sim-v2

We will soon have a stable version repo in ARM's Github

New features

1. Tool can be run in both stall free and bandwidth limited mode
2. New metrics like mapping efficiency, stall count added
3. Modular code
4. Available as python package
5. More enhancements in the pipeline!



We also have a new website

https://scalesim-project.github.io

Outline

1. **Simulating for DNN accelerator**
   - Motivation
   - Metrics of interest

2. **SCALE-Sim**
   - Overview
   - Modelling compute, memory, and interface
   - Modelling GEMM
   - Modelling Convolutions
   - Dataflows
   - Outputs

3. **Demos**

Demos

We will showcase SCALE-Sim v2 capabilities with 3 tutorials

1. **Using SCALE-Sim as a package**
   Design space exploration of a systolic accelerator

2. **Adding new features to Simulator**
   Adding new buffer hierarchies in SCALE-Sim

3. **Using SCALE-Sim as a library to build bigger simulators**
   Building a Scaled-out simulator using SCALE-Sim API