uSystolic: Byte-Crawling Unary Systolic Array

Di Wu and Joshua San Miguel
di.wu@ece.wisc.edu and jsanmiguel@wisc.edu

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Executive Summary

- ❑ Identify the challenges in existing low-power GEMM architectures based on binary and unary computing.
- ❑ Propose the uSystolic architecture to combine flexible systolic array with low-power unary computing.
- ❑ Evaluate uSystolic in the context of DNNs to reveal its application accuracy and hardware performance.

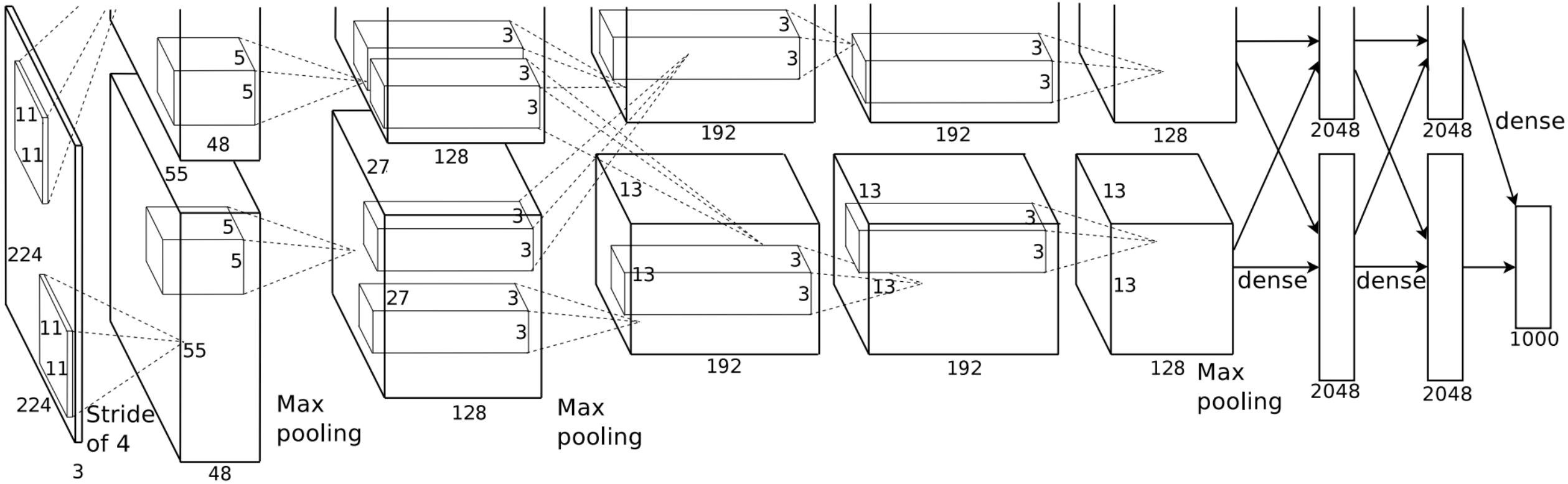
Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Outline

- ❑ Motivation
- ❑ Background
- ❑ Architecture
- ❑ Evaluation
- ❑ Conclusion

Low-power GEMM

- Core of DNNs
  - Over 90% DNN operations are GEMMs

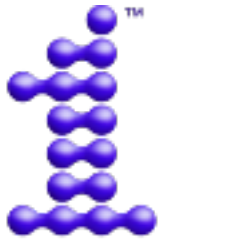


"ImageNet Classification with Deep Convolutional Neural Networks", NeurIPS 2012

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN-MADISON

Low-power GEMM

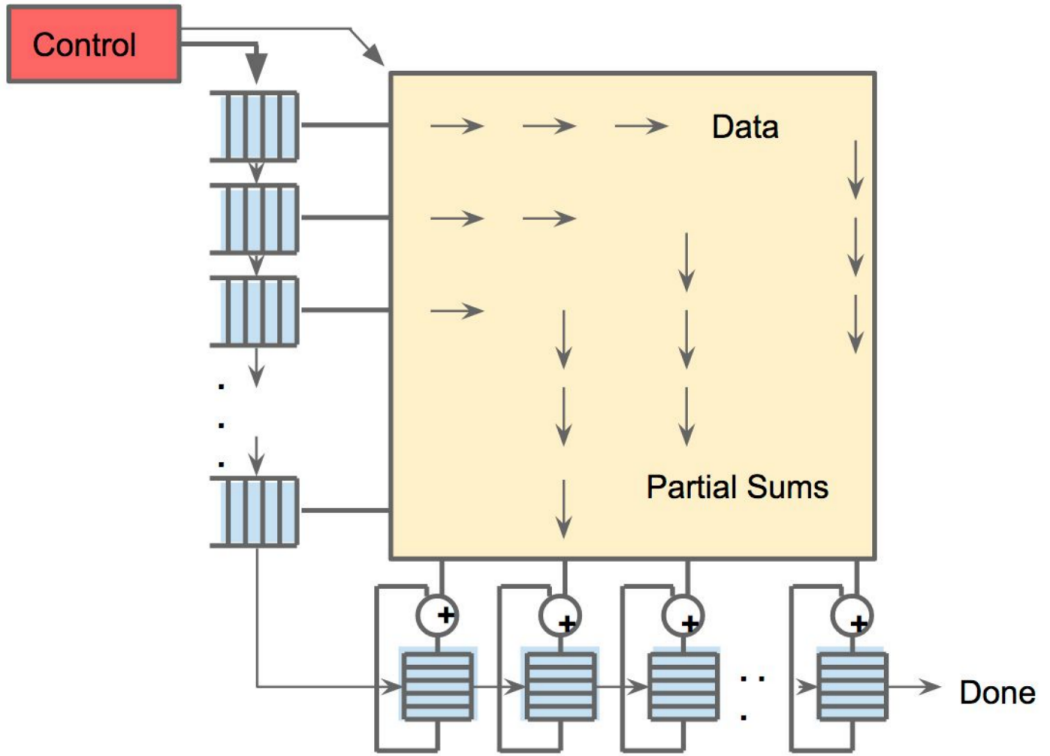
- Core of DNNs
  - Over 90% operations in DNNs are GEMMs
- Existing optimizations
  - Programming language and compiler
    - oneDNN, cuBLAS, FBGEMM, etc.



oneAPI Deep Neural Network Library (oneDNN) from Intel

Low-power GEMM

- Core of DNNs
  - Over 90% operations in DNNs are GEMMs
- Existing optimizations
  - Programming language and compiler
    - oneDNN, cuBLAS, FBGEMM, etc.
  - Architecture and microarchitecture
    - TPU, BFloat16, etc.

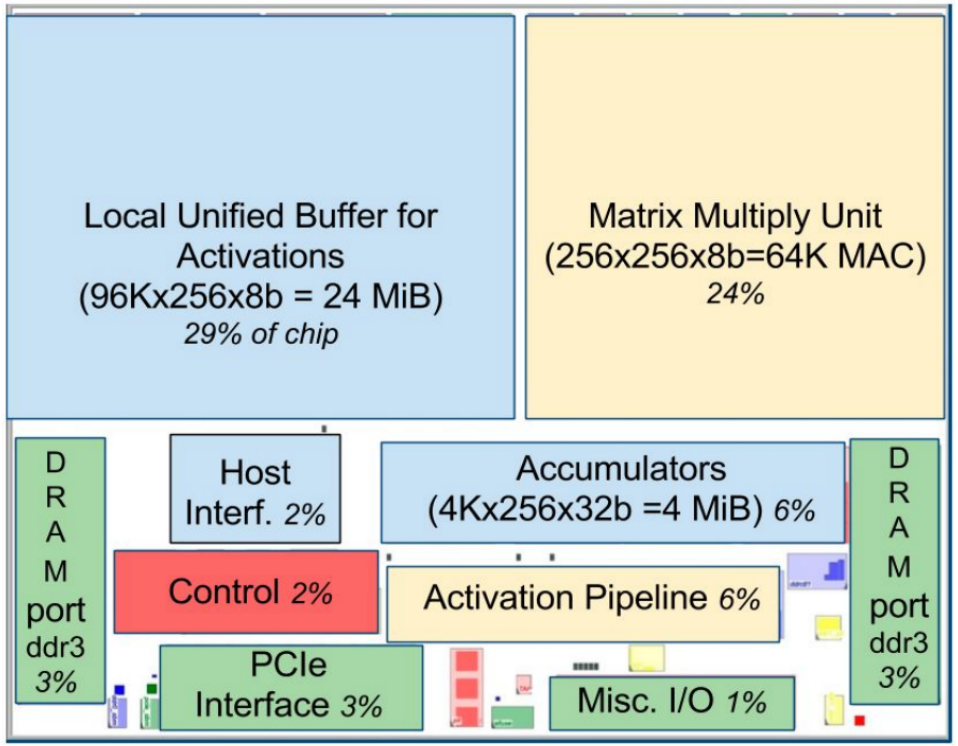


"In-Datacenter Performance Analysis of a Tensor Processing Unit", ISCA 2017

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN-MADISON

Low-power GEMM

- Challenges on the edge
  - Binary computing is not efficient
    - Mandatory on-chip SRAM
      - systolic TPU, 2-D mesh ShiDianNao, etc.
    - Superquadratic area overhead
      - wire congestion

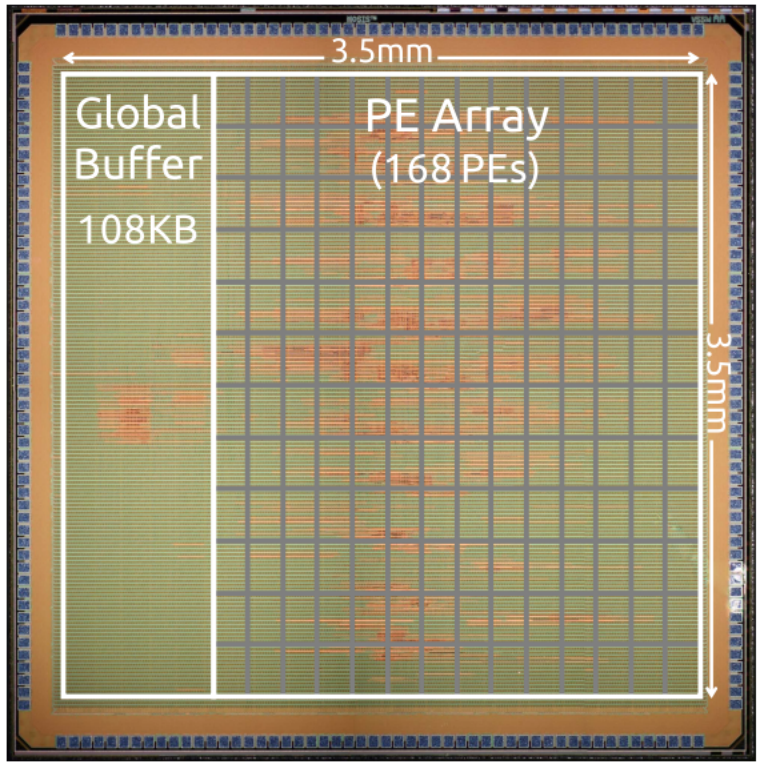


"In-Datacenter Performance Analysis of a Tensor Processing Unit", ISCA 2017

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN-MADISON

Low-power GEMM

- Challenges on the edge
  - Binary computing is not efficient
    - Mandatory on-chip SRAM
      - systolic TPU, 2-D mesh ShiDianNao, etc.
    - Superquadratic area overhead
      - wire congestion

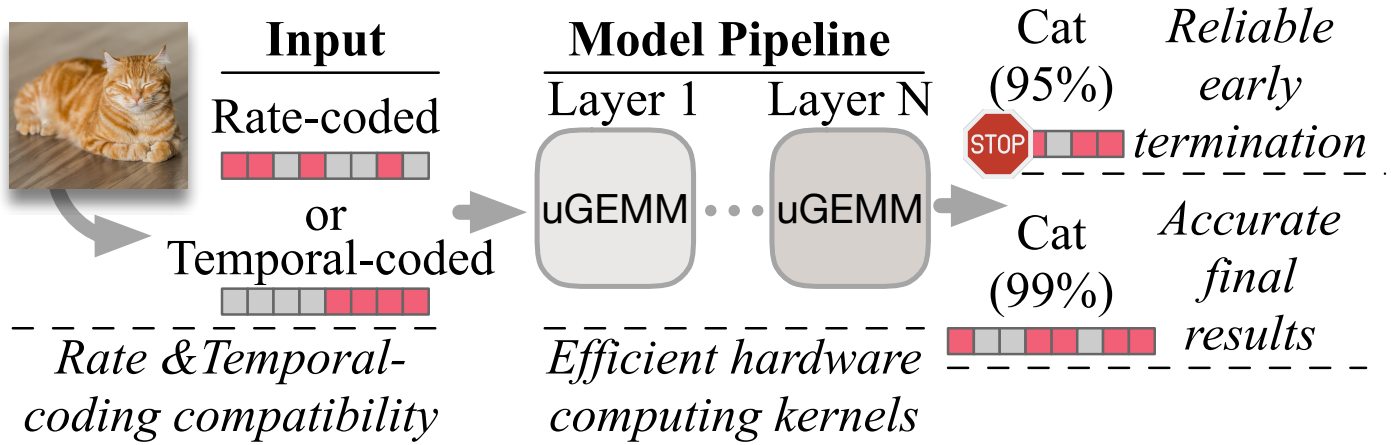


"Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks", JSSC 2017

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN-MADISON

Low-power GEMM

- Challenges



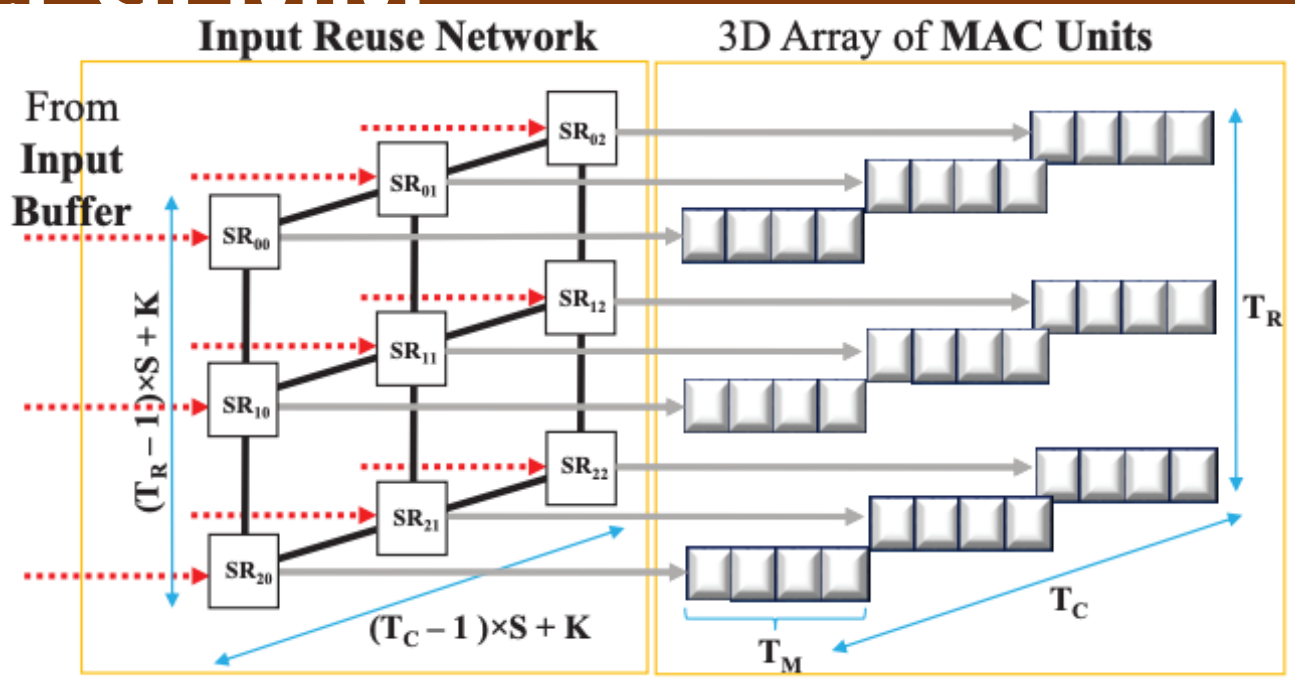
- Unary computing is not generalizable
  - Fully streaming unary (FSU) architectures
    - full customization

"uGEMM: Unary Computing Architecture for GEMM Applications", ISCA 2020

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN-MADISON

Low-power GEMM

- Challenges



  - Unary computing is not generalizable
    - Fully streaming unary (FSU) architectures
      - full customization
    - Hybrid unary binary (HUB) architectures
      - low efficiency for matrix multiplication

"A New Stochastic Computing Multiplier with Application to Deep Convolutional Neural Networks", DAC 2017

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN-MADISON

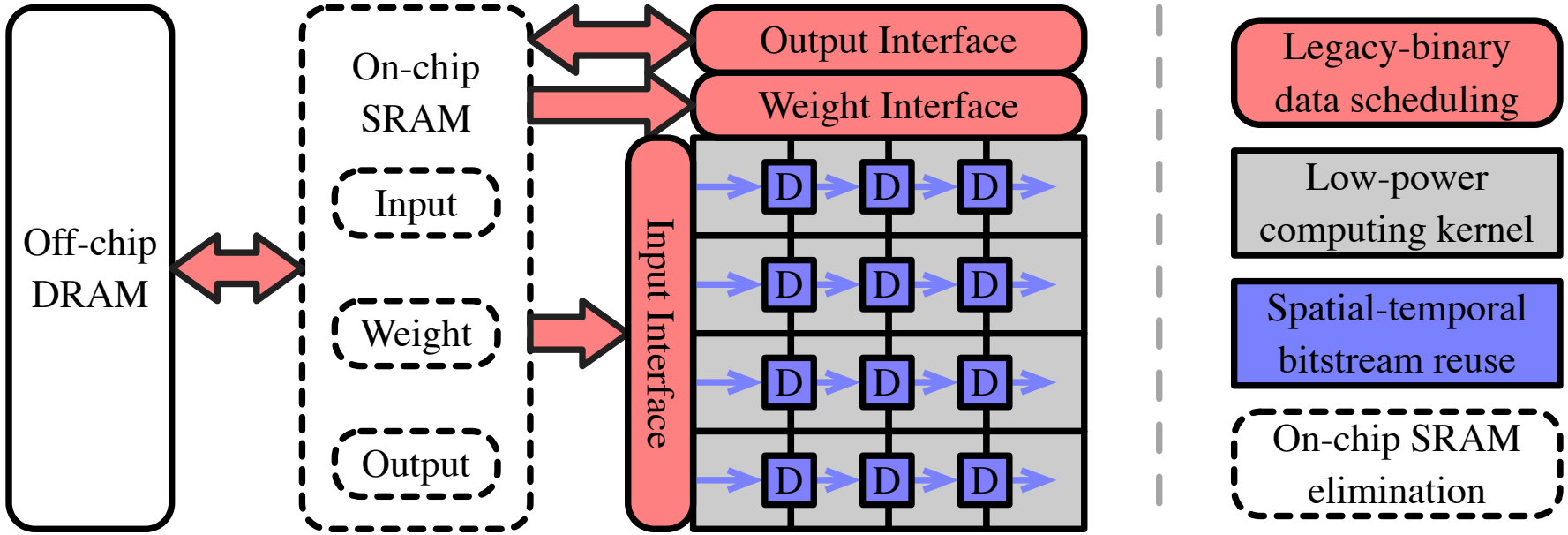
Low-power GEMM

- Compare GEMM architectures

Architecture	Accuracy	Power Efficiency	Scalability	Generalizability
B-Systolic	Precise	Low	High	High
B-Mesh	Precise	Low	Low	High
FSU	Low-High	High	Low	Low
HUB	High	High	Low	Medium
Our goal	High	High	High	High

Low-power GEMM

- Our solution
  - uSystolic: A systolic array with hybrid unary binary computing kernels



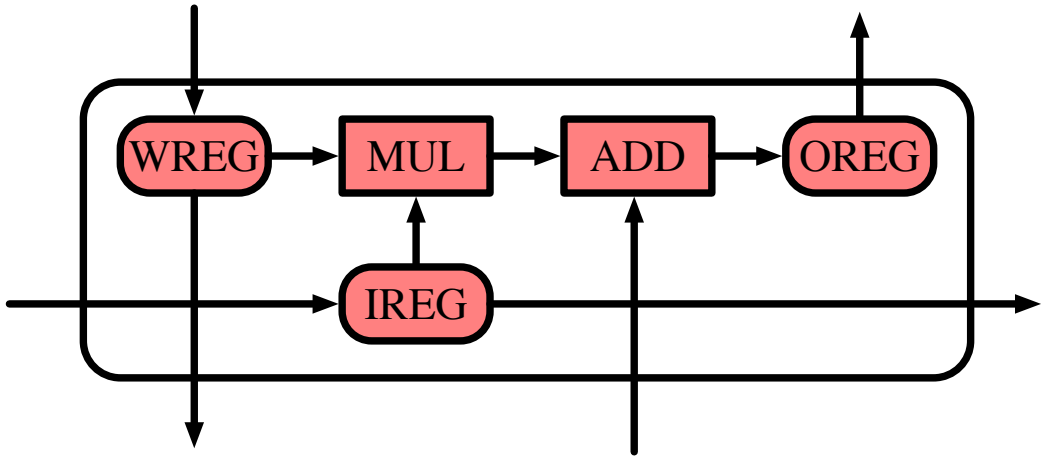
Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Outline

- ❑ Motivation
- ❑ Background
- ❑ Architecture
- ❑ Evaluation
- ❑ Conclusion

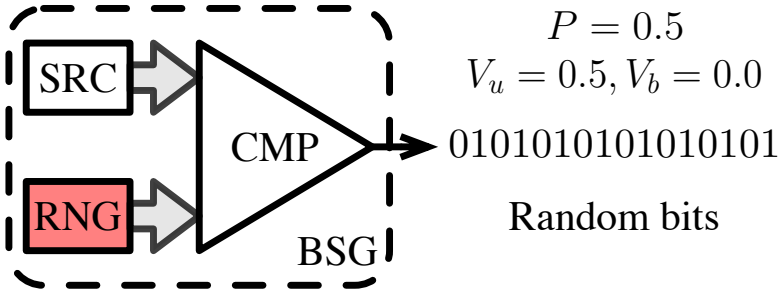
Binary systolic array

- Processing element
  - MAC (MUL+ADD)
  - Weight (WREG)
    - Preloaded from top to bottom
  - Input feature map (IREG)
    - Streamed in from left to right
  - Output feature map (OREG)
    - Streamed out from bottom to top



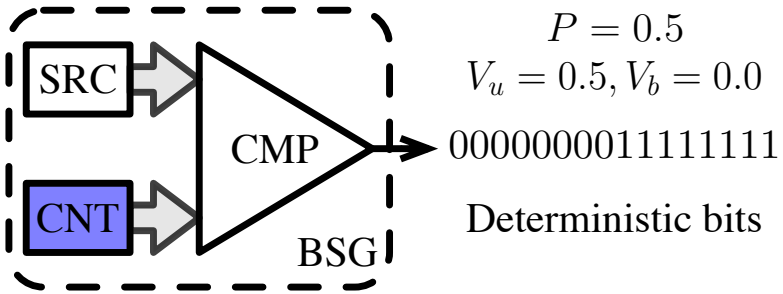
Unary data

- Bitstream
  - Rate coding



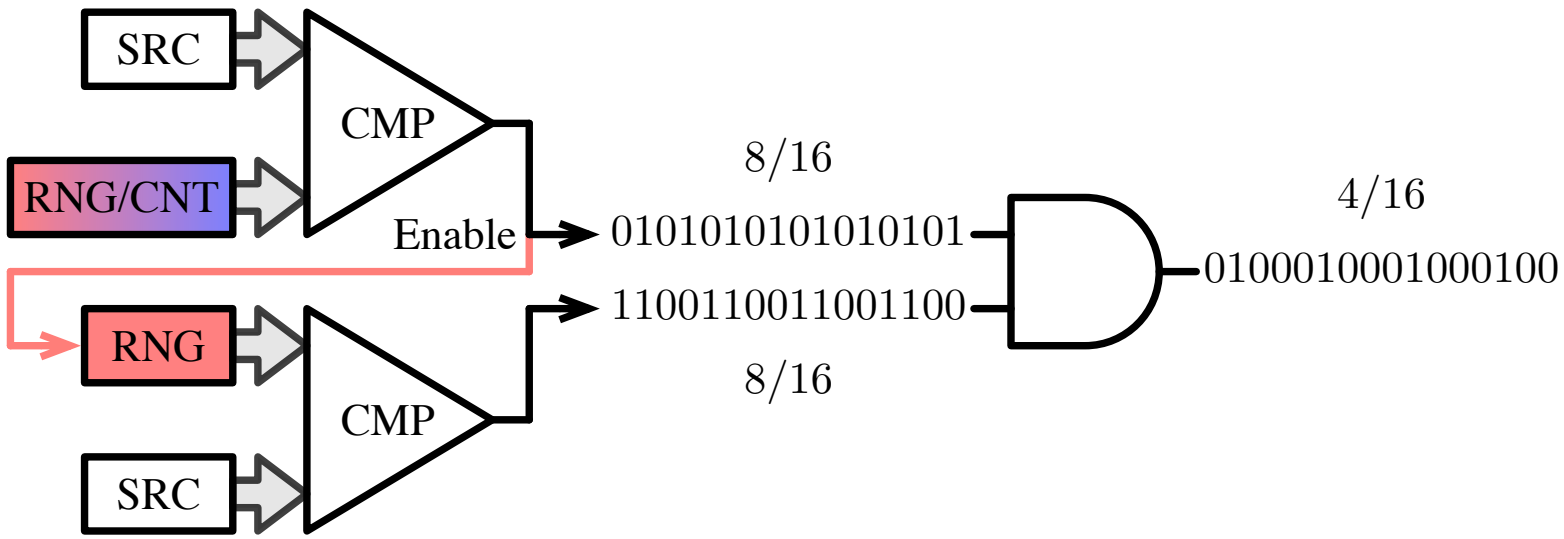
SRC: source binary data
CMP: comparator
RNG: random number generator
CNT: counter
BSG: bitstream generation

  - Temporal coding



Unary multiplier

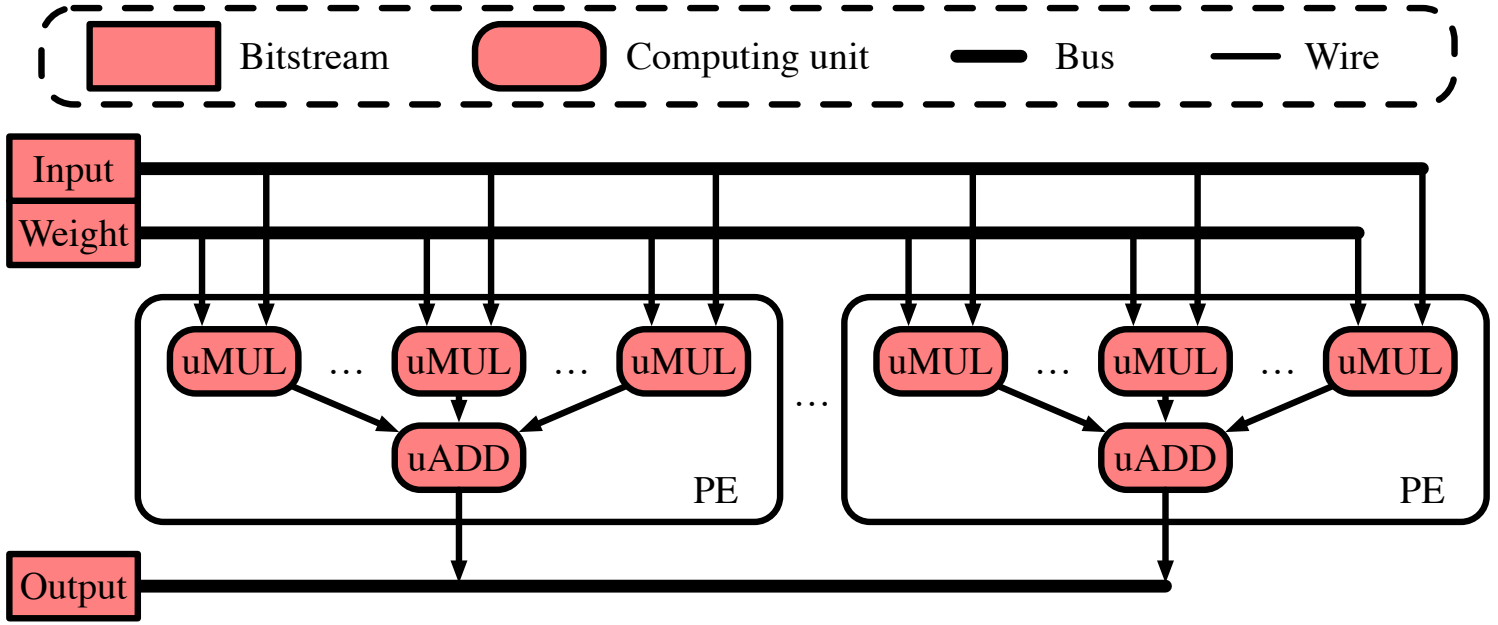
- uMUL
  - Conditional bitstream generation
    - Control the RNG update of one input with another input
    - Rate coding for controllee; Rate or temporal coding for controller



"uGEMM: unary computing architecture for GEMM applications", ISCA 2020

Unary GEMM architecture

- uGEMM
  - SIMD FSU architecture
  - Spatial reuse with low scalability



"uGEMM: unary computing architecture for GEMM applications", ISCA 2020

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN-MADISON

Outline

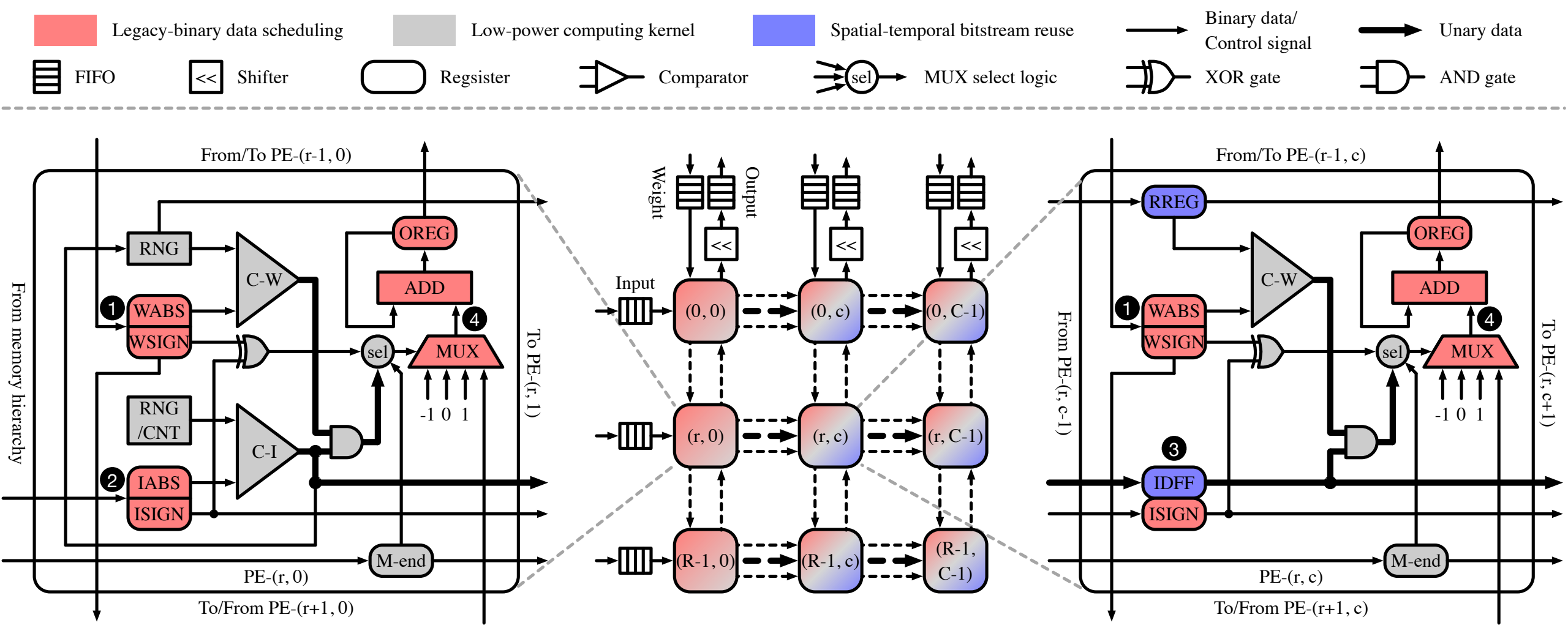
- ❑ Motivation
- ❑ Background
- ❑ Architecture
- ❑ Evaluation
- ❑ Conclusion

uSystolic highlight

- uSystolic vs uGEMM
  - Binary accumulation
    - High accuracy
  - Unipolar multiplication on sign-magnitude formatted data
    - Half multiplier area and latency: over 2X improvement in energy efficiency
  - Systolic array with local interconnections
    - Spatial-temporal bitstream reuse: high scalability with no accuracy drop and minimized area overhead
  - Mature data schedule from binary systolic arrays

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

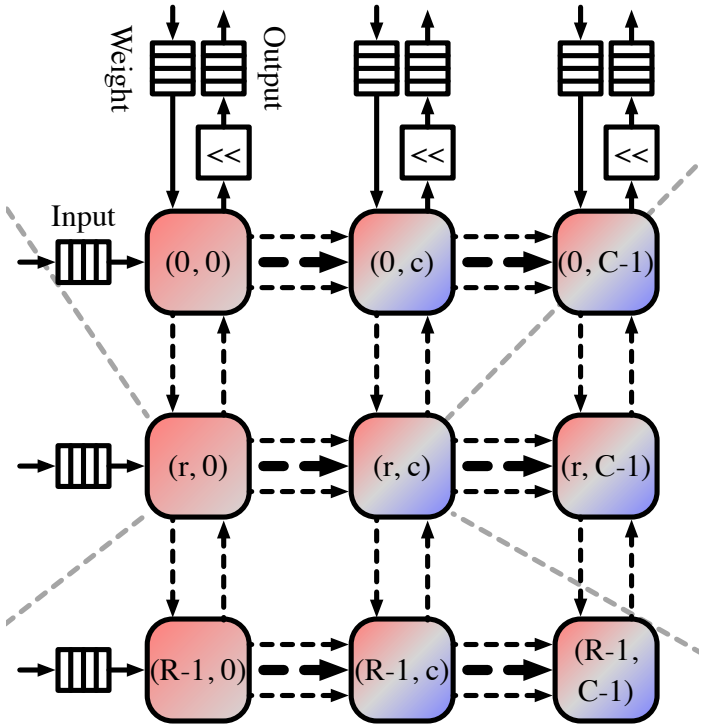
uSystolic architecture



Systolic array

Legacy-binary data scheduling
Low-power computing kernel
Spatial-temporal bitstream reuse
Binary data/ Control signal
Unary data
FIFO
Shifter
Regsister
Comparator
MUX select logic
XOR gate
AND gate

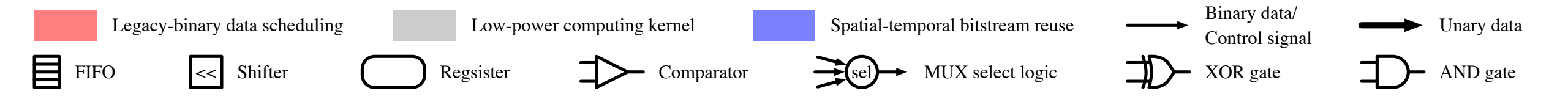
- Weight: top -> bottom
- Input: left -> right
- Output: bottom -> top

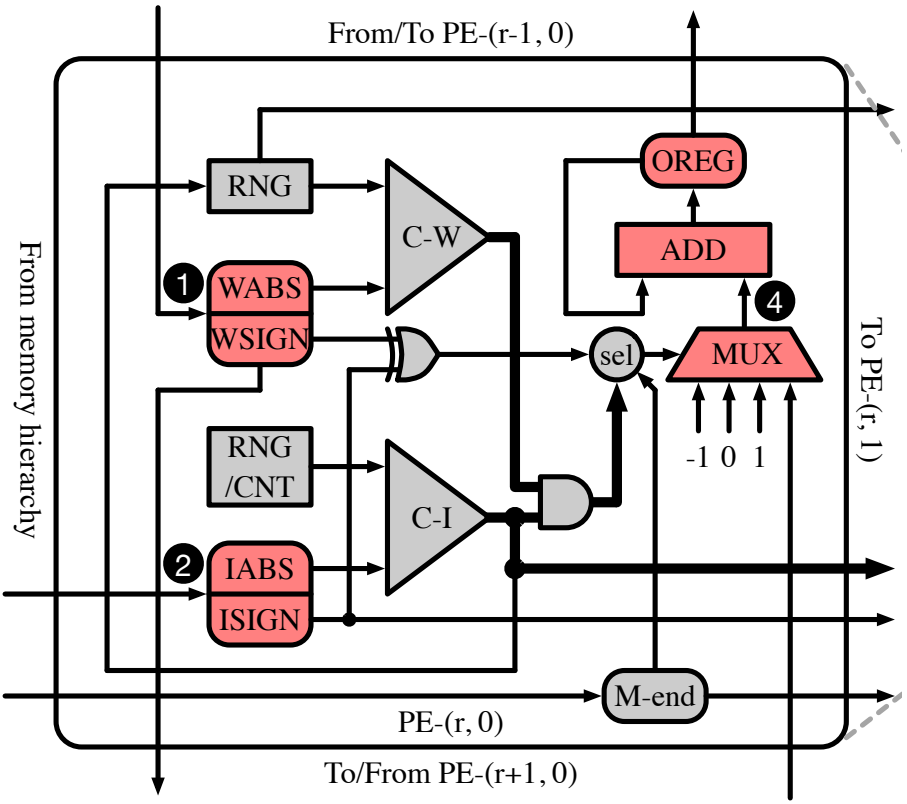


- Leftmost PE
- Non-leftmost PE

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Leftmost PE





1: Weight preload
2: Input streaming
4: Output accumulation

Gray: unipolar uMUL, halving the area/latency
Red: binary interface, inheriting data schedule

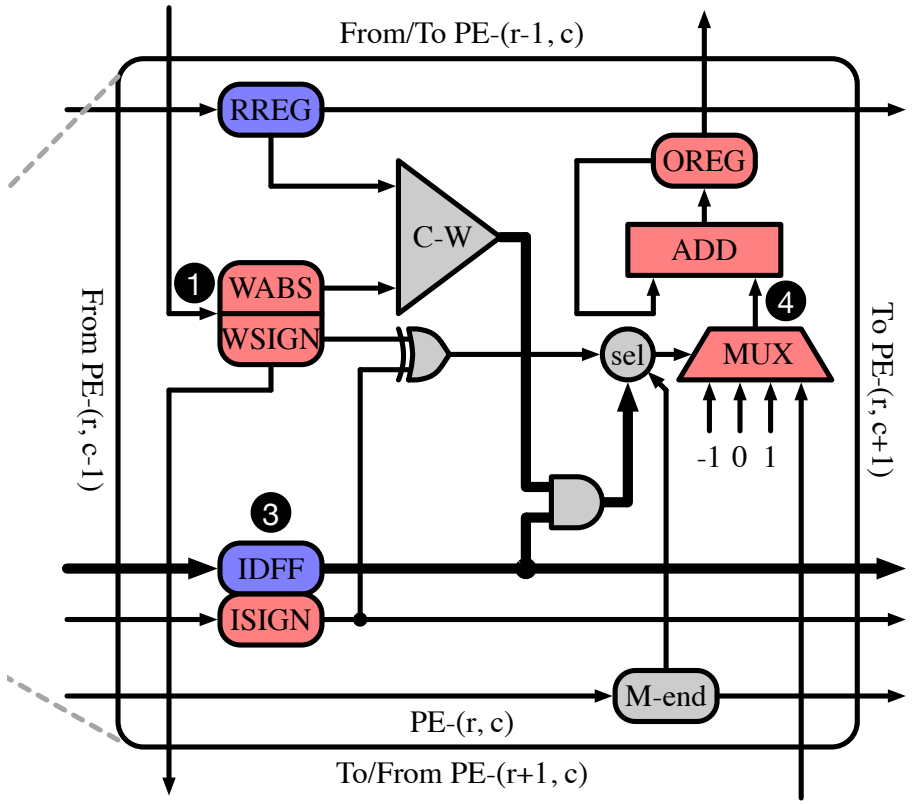
Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Non-leftmost PE

Legacy-binary data scheduling
Low-power computing kernel
Spatial-temporal bitstream reuse
Binary data/ Control signal
Unary data
FIFO
Shifter
Regsister
Comparator
MUX select logic
XOR gate
AND gate

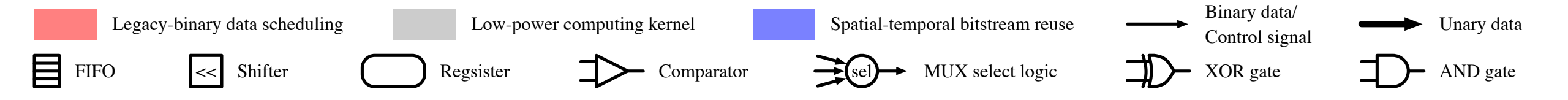
1: Weight preload
3: Spatial-temporal bitstream reuse
4: Output accumulation

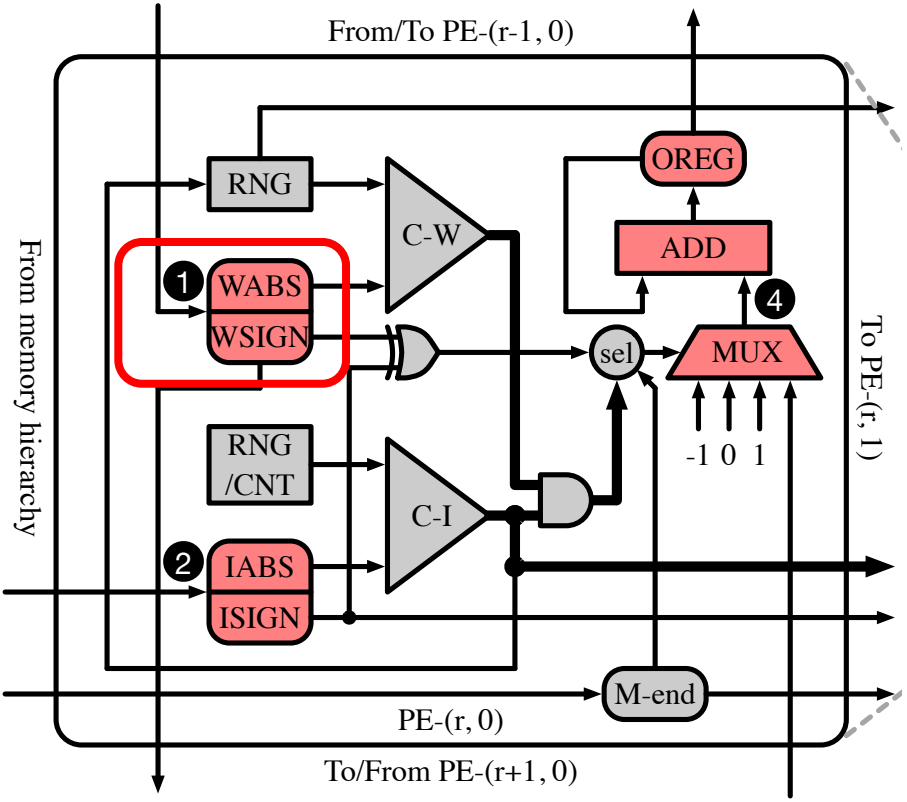
Gray: unipolar uMUL, halving the area/latency
Red: binary interface, inheriting data schedule
Blue: spatial-temporal bitstream reuse, reducing cost



Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

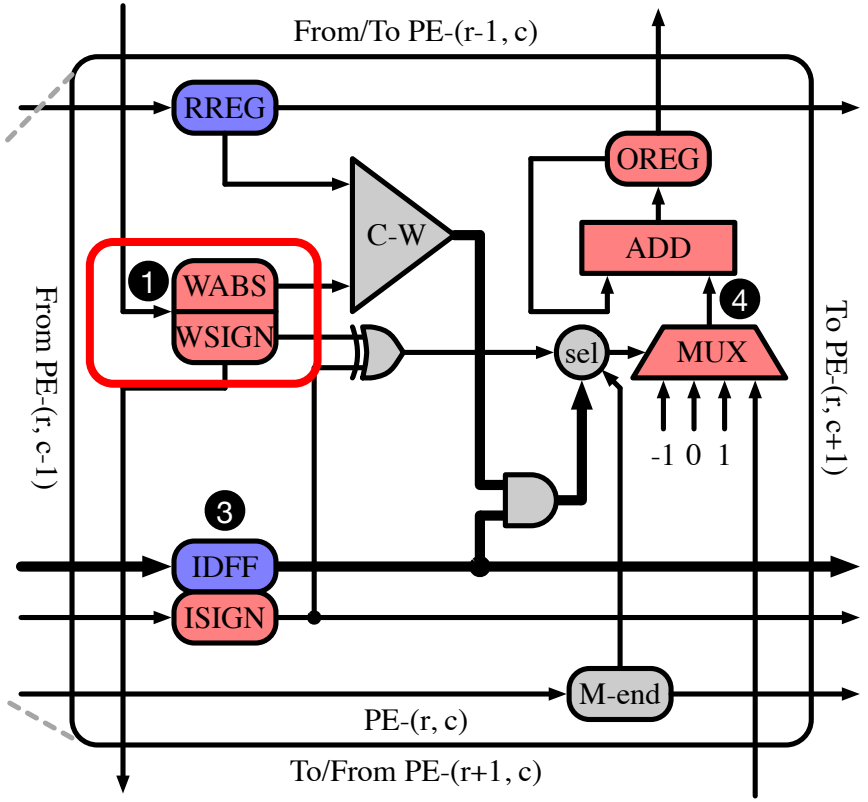
Walkthrough example





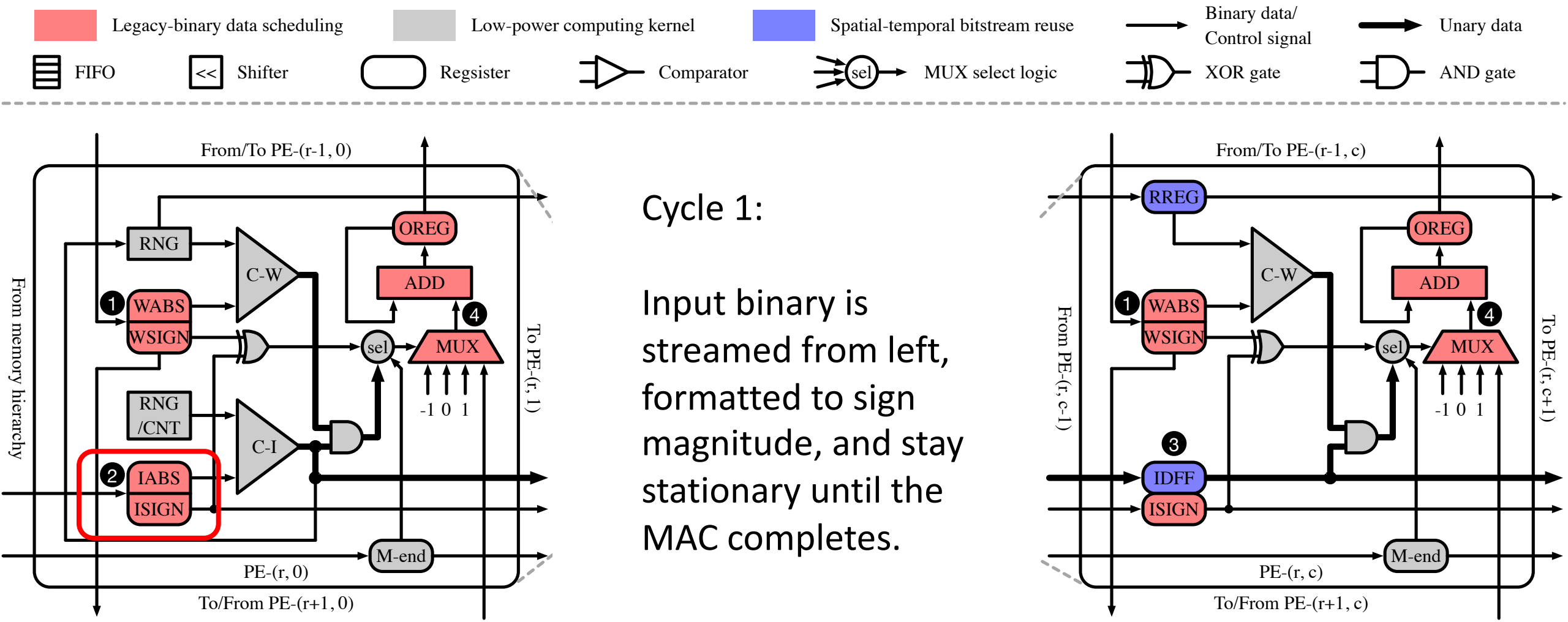
Cycle 0:

Weight binary is preloaded from top, formatted to sign magnitude, and stays stationary until the GEMM completes.



Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Walkthrough example

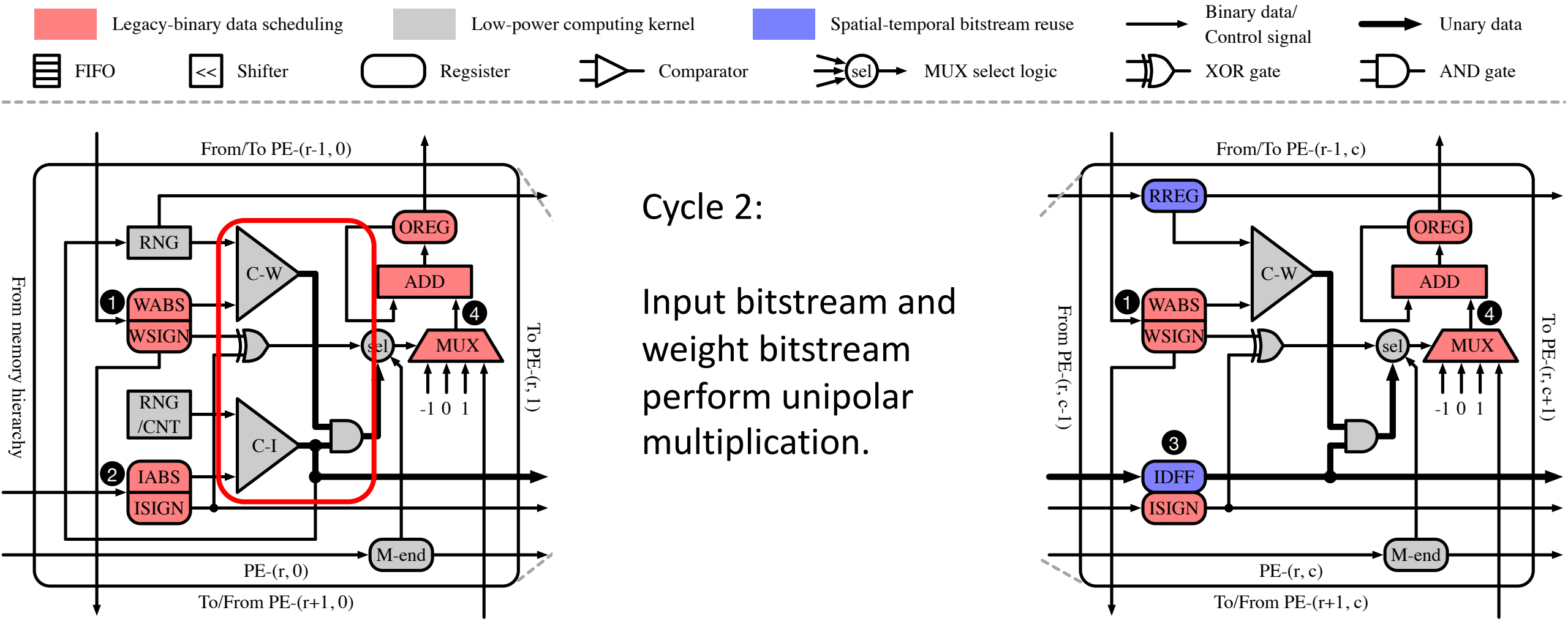


Cycle 1:

Input binary is streamed from left, formatted to sign magnitude, and stay stationary until the MAC completes.

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Walkthrough example

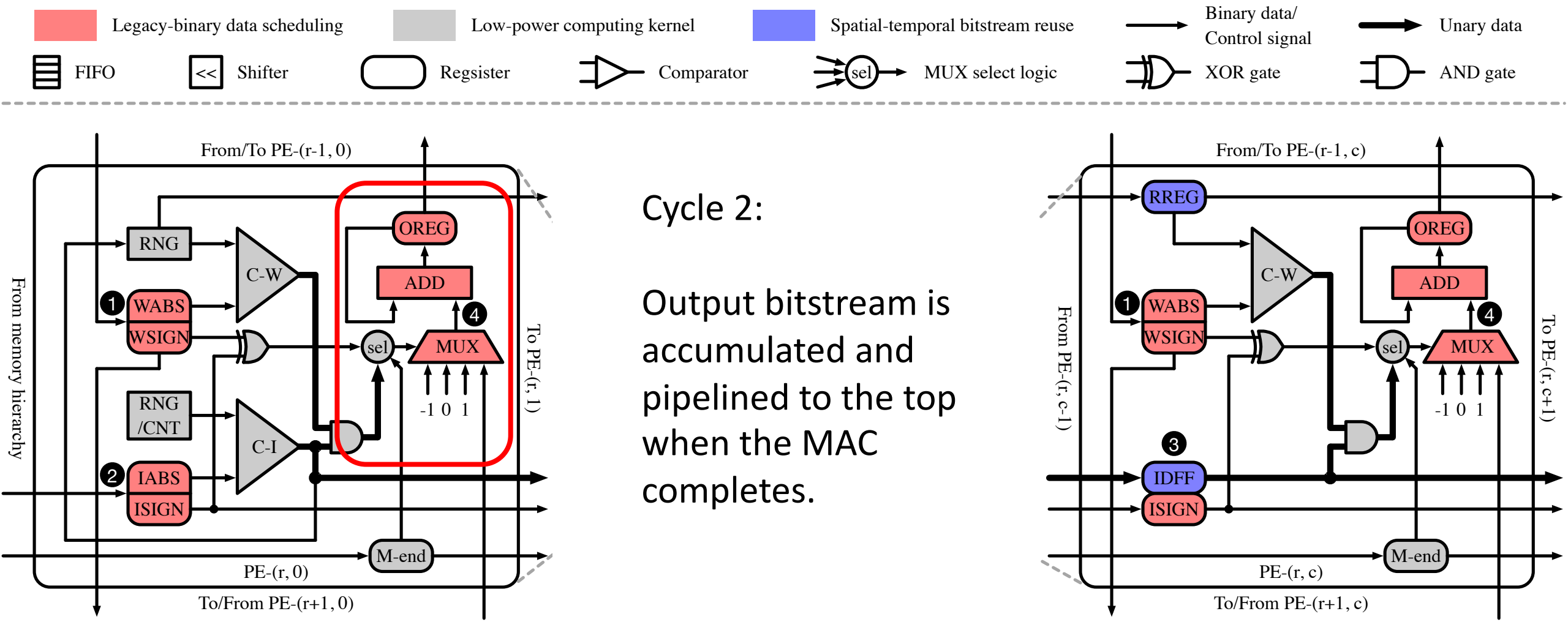


Cycle 2:

Input bitstream and weight bitstream perform unipolar multiplication.

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

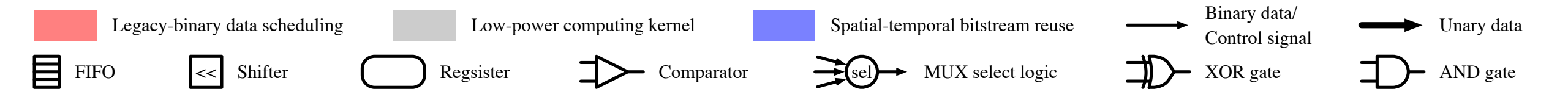
Walkthrough example

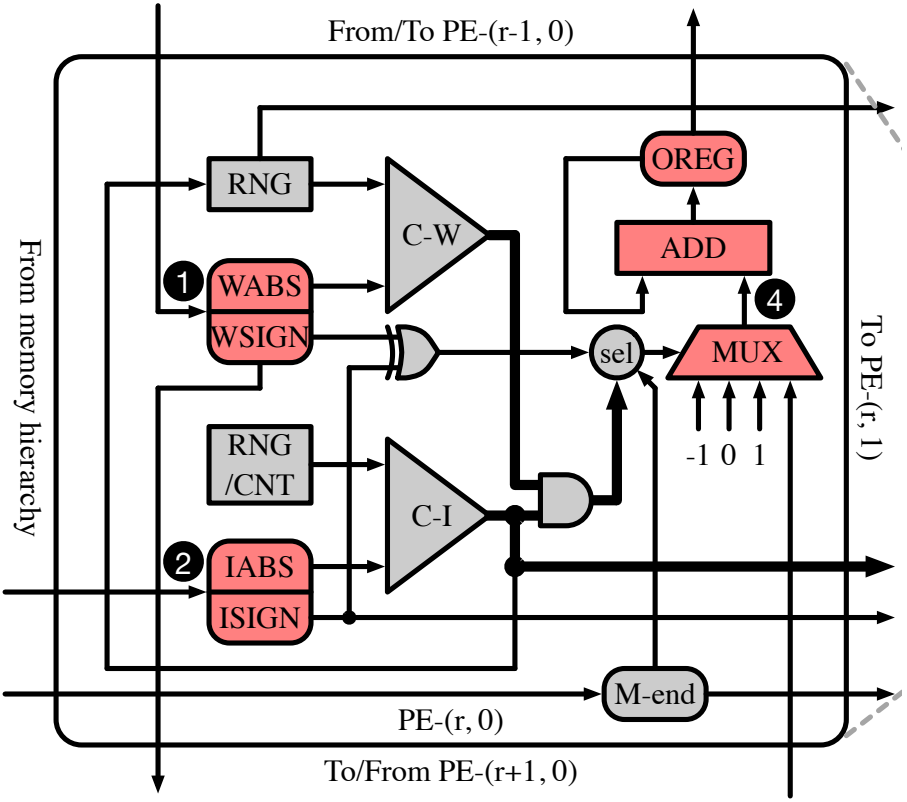


Cycle 2:

Output bitstream is accumulated and pipelined to the top when the MAC completes.

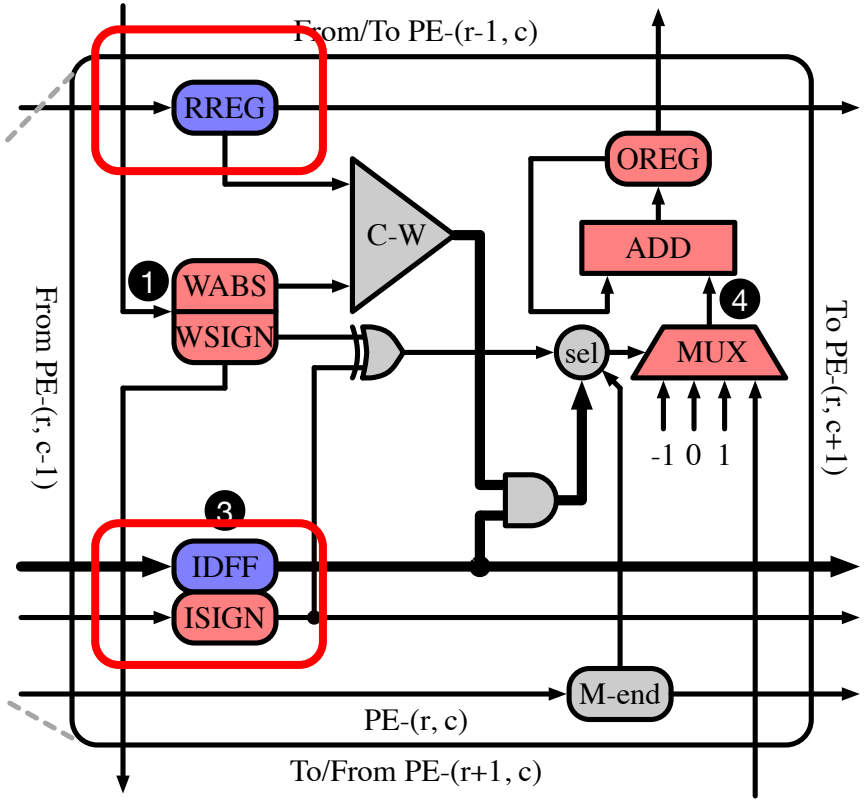
Walkthrough example





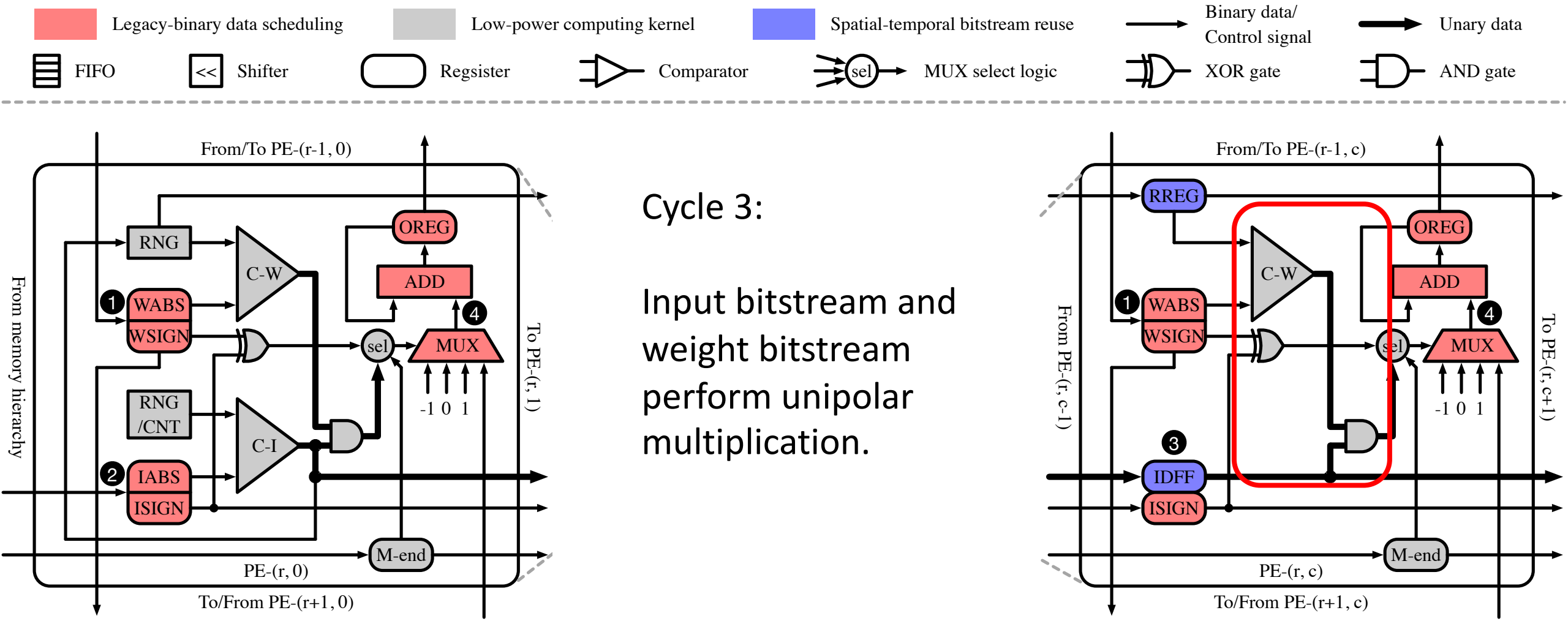
Cycle 2:

Input bitstream is pipelined to right, together with weight random number.



Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Walkthrough example

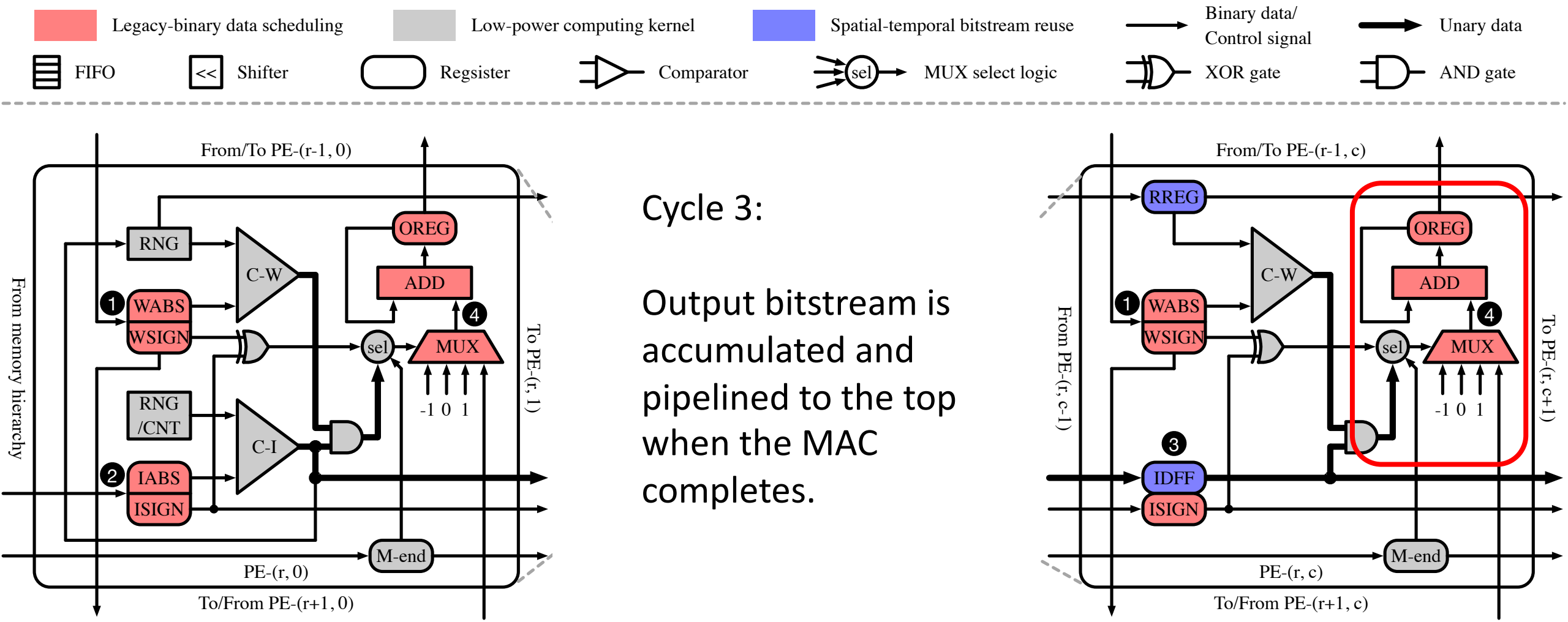


Cycle 3:

Input bitstream and weight bitstream perform unipolar multiplication.

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Walkthrough example

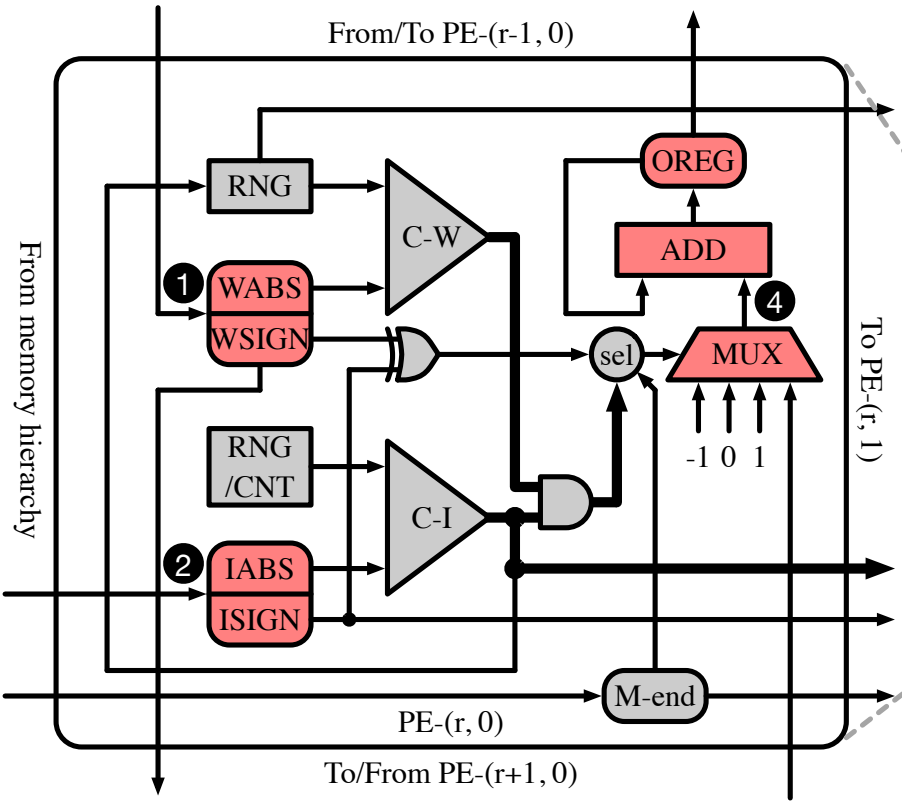


Cycle 3:

Output bitstream is accumulated and pipelined to the top when the MAC completes.

Spatial-temporal bitstream reuse

Legacy-binary data scheduling
Low-power computing kernel
Spatial-temporal bitstream reuse
Binary data/ Control signal
Unary data
FIFO
Shifter
Regsister
Comparator
MUX select logic
XOR gate
AND gate



Proved high accuracy with conditional BSG

$$SCC = 0 \cong C\text{-}BSG_r(B_0, R_0)$$

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN-MADISON

Spatial-temporal bitstream reuse

Legacy-binary data scheduling | Low-power computing kernel | Spatial-temporal bitstream reuse | Binary data/ Control signal | Unary data

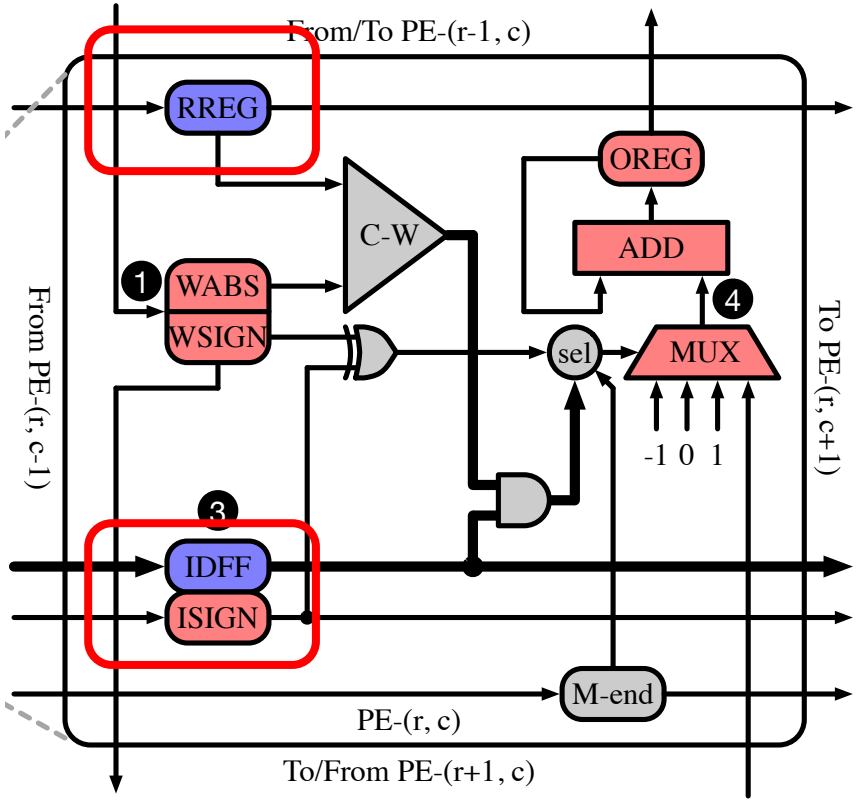
FIFO | Shifter | Regsister | Comparator | MUX select logic | XOR gate | AND gate

Proved high accuracy with conditional BSG

$$SCC = 0 \cong C\text{-}BSG_r(B_0, R_0)$$

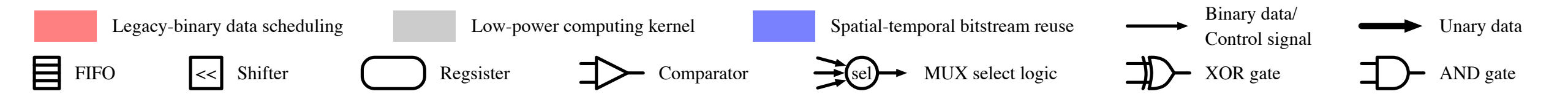
Duplicate B and R with delay

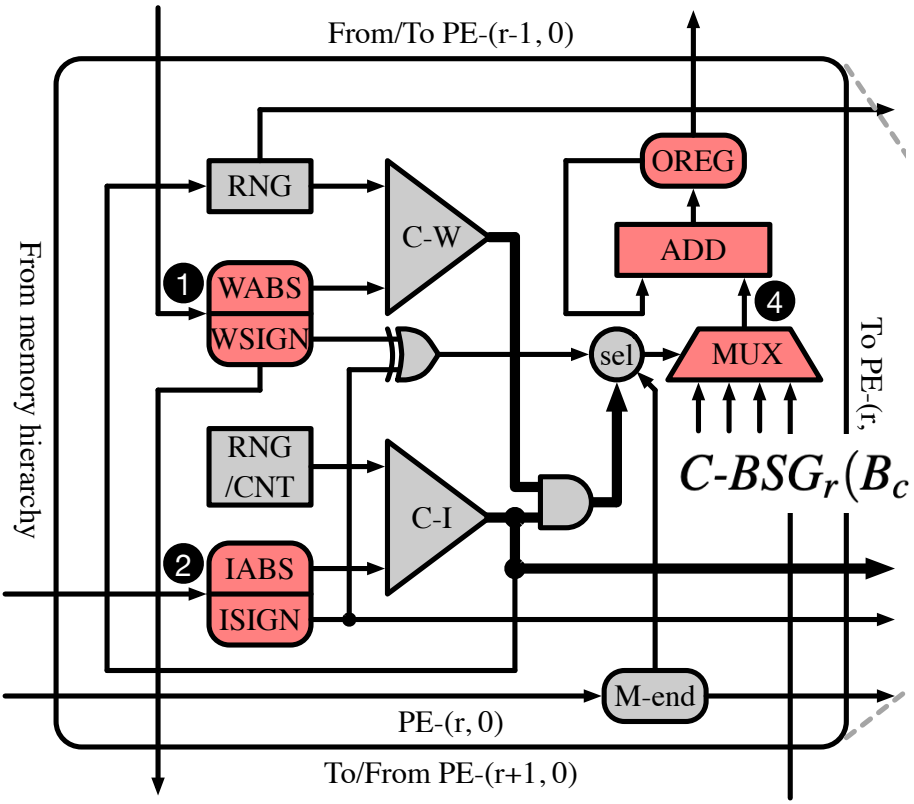
$$C\text{-}BSG_r(B_c, R_c) \Longrightarrow C\text{-}BSG_r(B_{c+1}, R_{c+1})$$



Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Spatial-temporal bitstream reuse





Proved high accuracy with conditional BSG

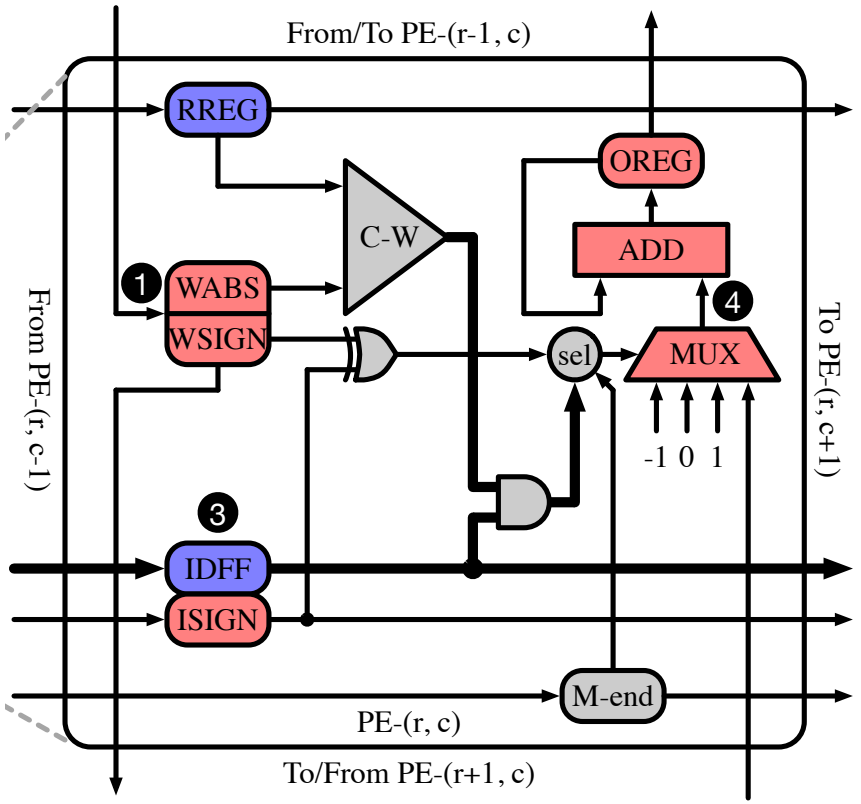
$$SCC = 0 \cong C\text{-}BSG_r(B_0, R_0)$$

Duplicate B and R with delay

$$C\text{-}BSG_r(B_c, R_c) \Longrightarrow C\text{-}BSG_r(B_{c+1}, R_{c+1})$$

For row PEs, accuracy is consistently high

$$SCC = 0 \cong C\text{-}BSG_r(B_c, R_c)$$



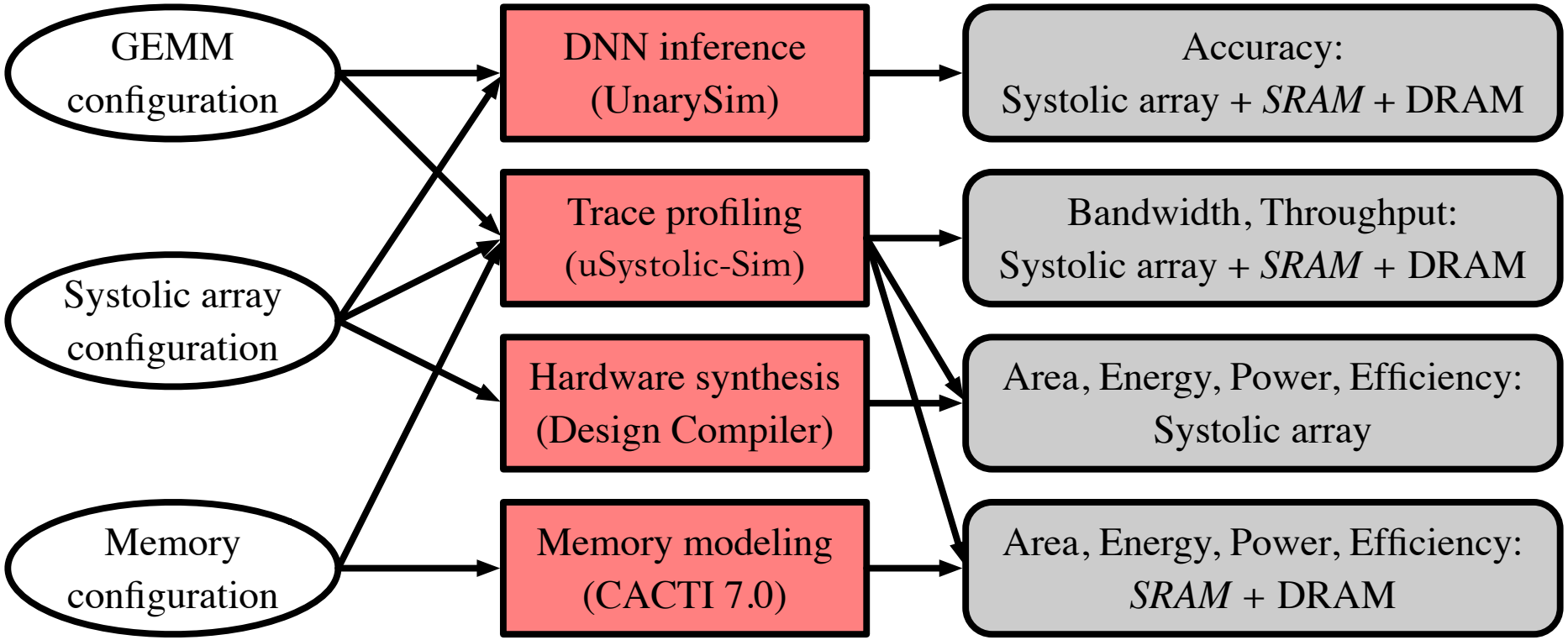
Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Outline

- ❑ Motivation
- ❑ Background
- ❑ Architecture
- ❑ Evaluation
- ❑ Conclusion

Evaluation framework

- Consider both application and hardware



Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Application accuracy evaluation setup

- UnarySim (ISCA 2020, IEEE Micro Top Pick for 2020)
  - DNN

DNN Model	Dataset	# of weight (M)
4-layer CNN	MNIST	1.2
ResNet18	Cifar10	11.7
AlexNet	ImageNet	61.1

  - Design

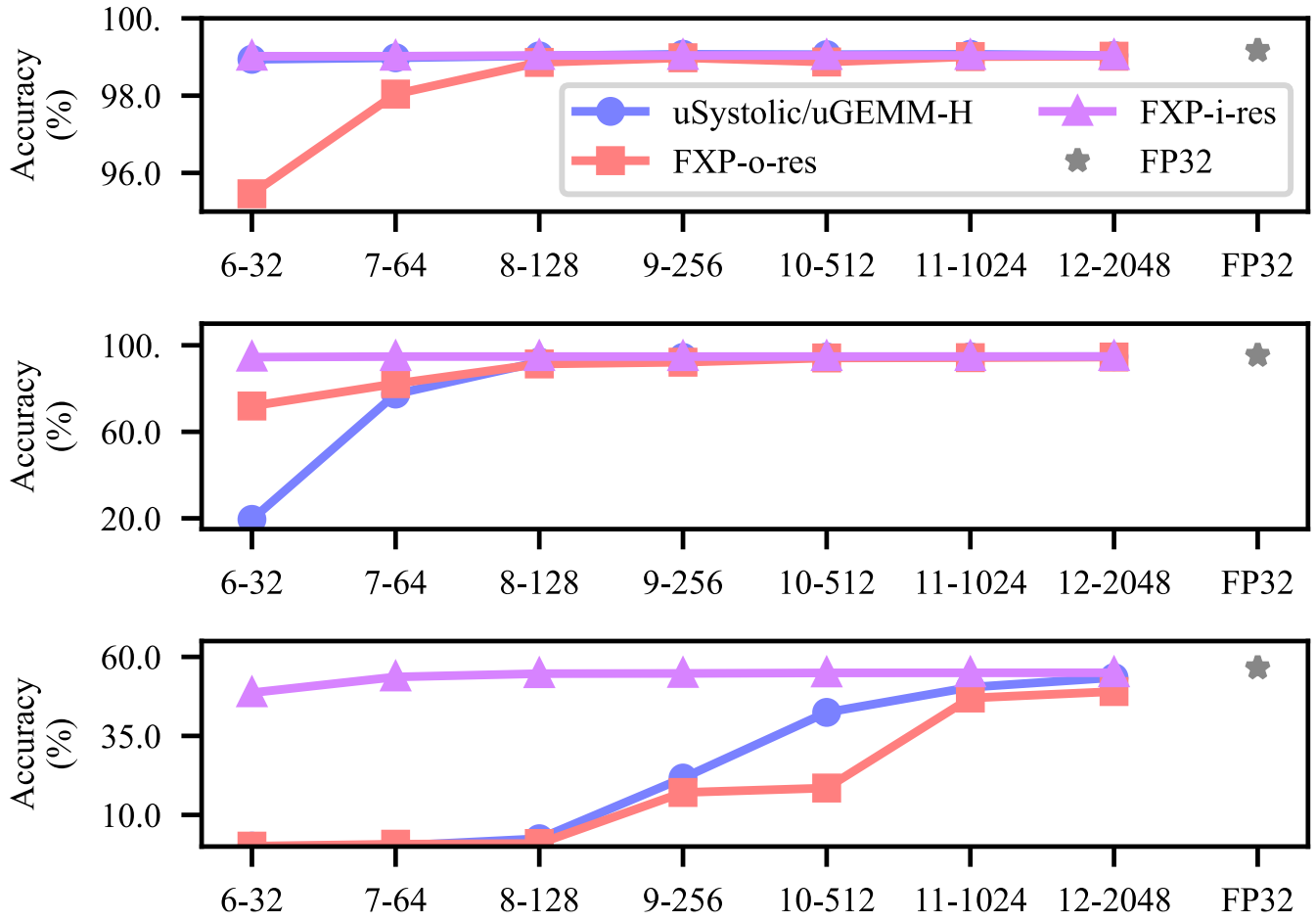
Design	Input resolution	Output resolution	MUL latency
uSystolic	N	N	2^(N-1)
uGEMM-H	N	N	2^N
FXP-i-res	N	2N	1
FXP-o-res	N/2	N	1

Accuracy of DNN inference

4-layer CNN on MNIST (1.2M)

ResNet18 on Cifar10 (11.7M)

AlexNet on ImageNet (61.1M)



X axis is (binary resolution)-(cycle count) for uSystolic multiplication

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

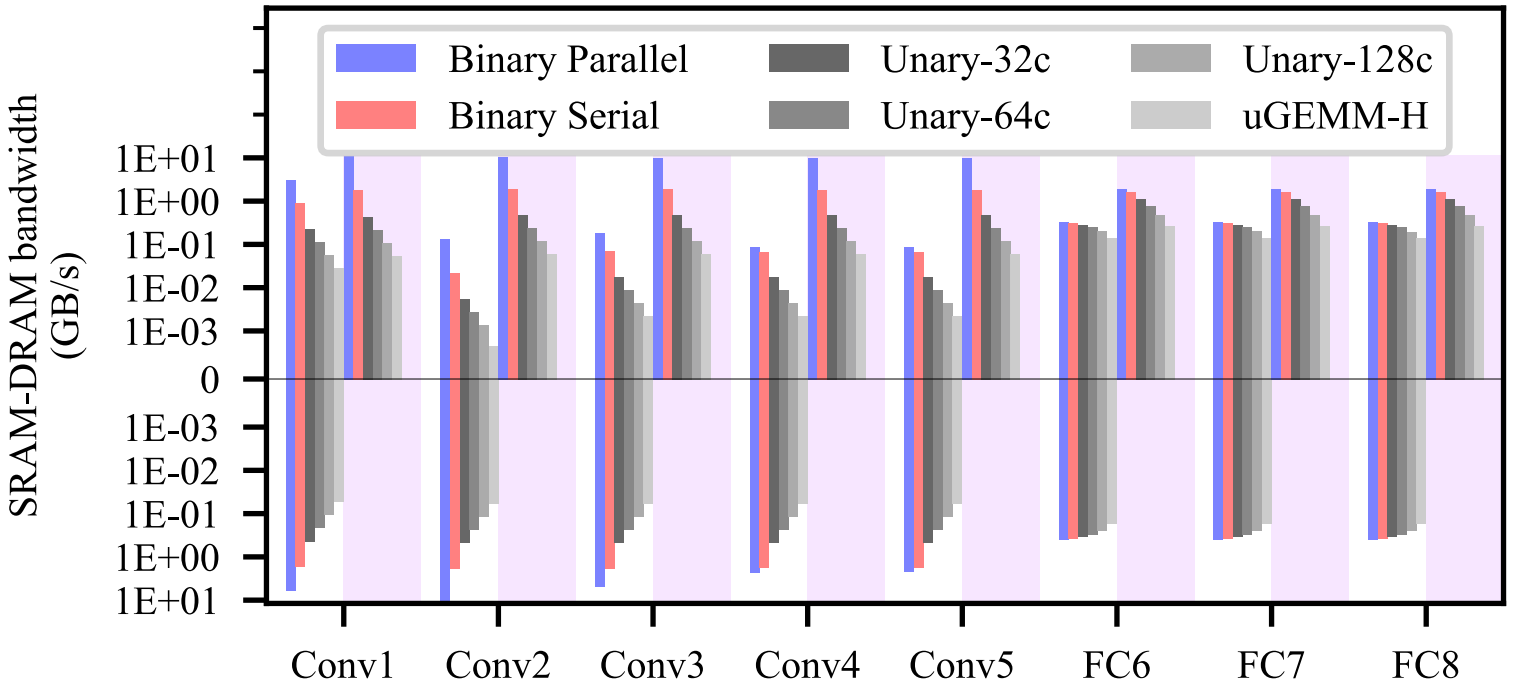
Hardware performance evaluation setup

- uSystolic-Sim
  - DNN: AlexNet and MLPerf
  - Design: systolic array taken from Eyeriss for edge computing
    - 12-by-14 PEs, 192KB SRAM

Design	Input resolution	Output resolution	MUL latency	ACC resolution
uSystolic	8	8	32, 64, 128	24
uGEMM-H	8	8	256	24
Bit parallel	8	16	1	32
Bit serial	8	16	8	32

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Bandwidth

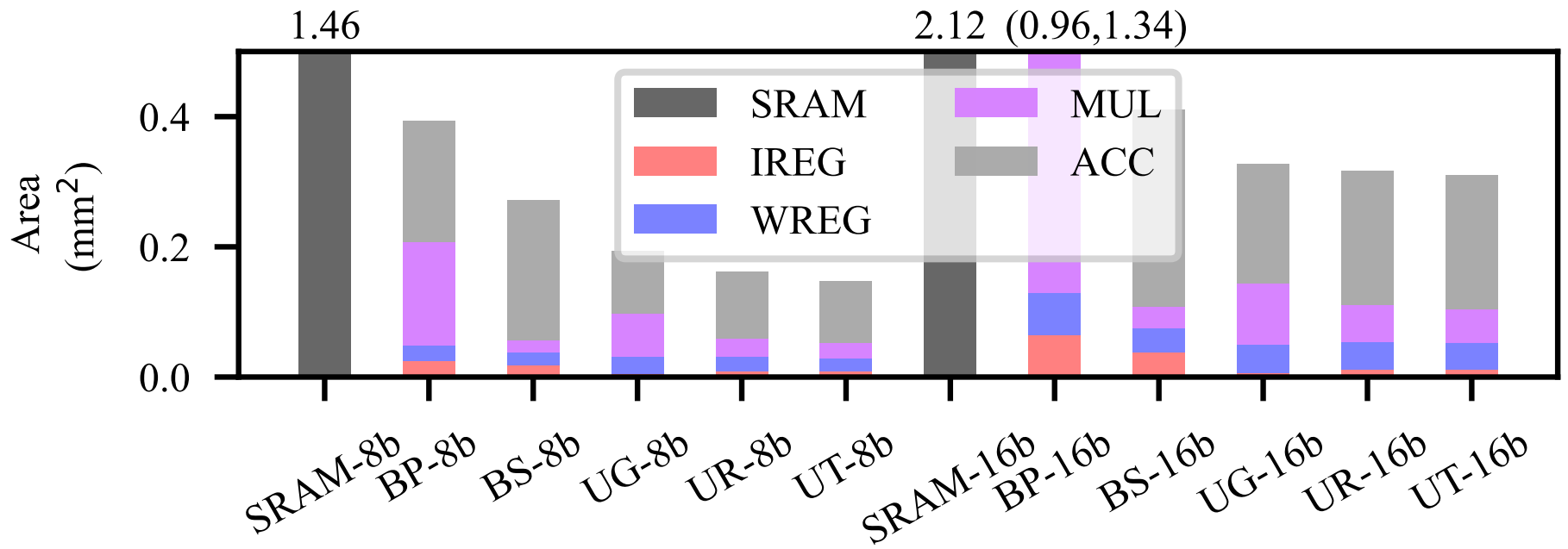


White and purple background colors are for designs with/without on-chip SRAM

- uSystolic requires 0.1~1GB/s DRAM bandwidth, even when SRAM is absent
- Eliminating SRAM from uSystolic without idling computing kernel is possible

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Area

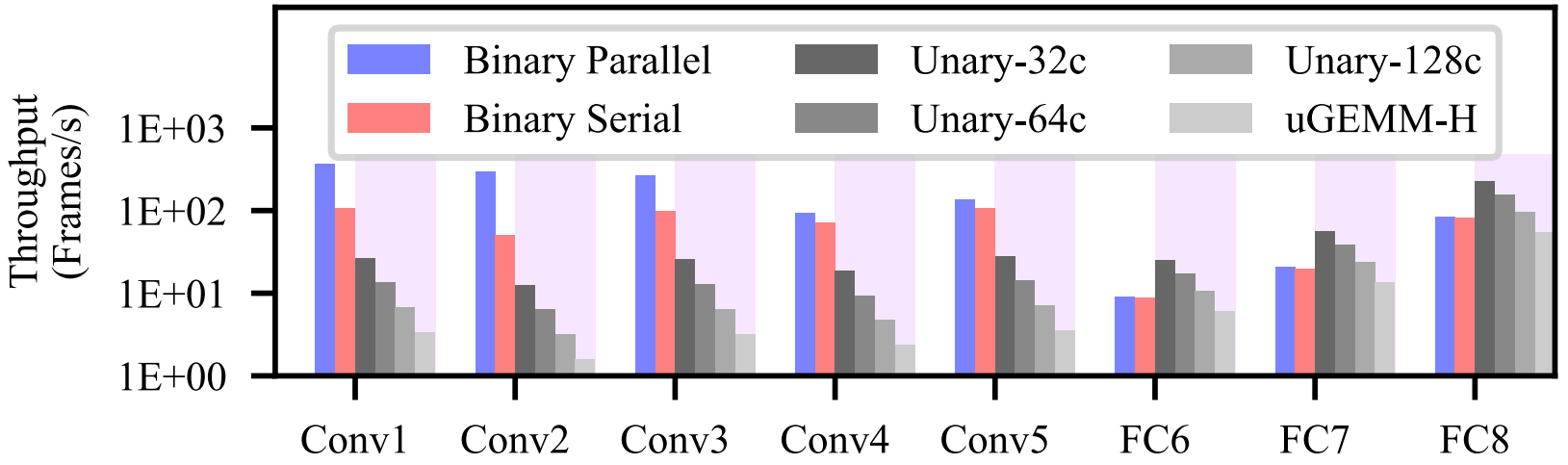


BP: bit parallel; BS: bit serial; UG: uGEMM-H; UR: uSystolic with rate coding; UT: uSystolic with temporal coding

- For 8-bit designs, rate-coded uSystolic without SRAM exhibits 91.3% and 90.7% area reduction compared to binary parallel and binary serial designs with SRAM.

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

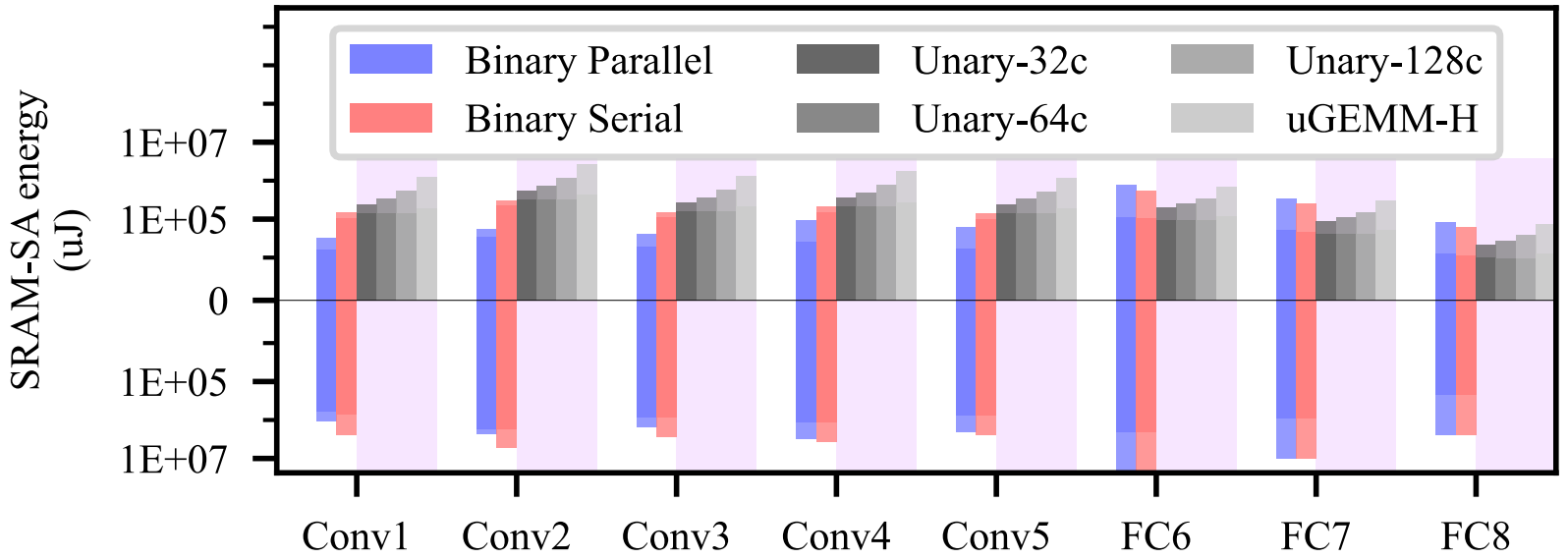
Throughput



- Runtime overhead: 161.8% for bit parallel vs 13.4% for 128 cycle uSystolic
- uSystolic exhibits good throughput scaling, due to low memory contention

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Energy



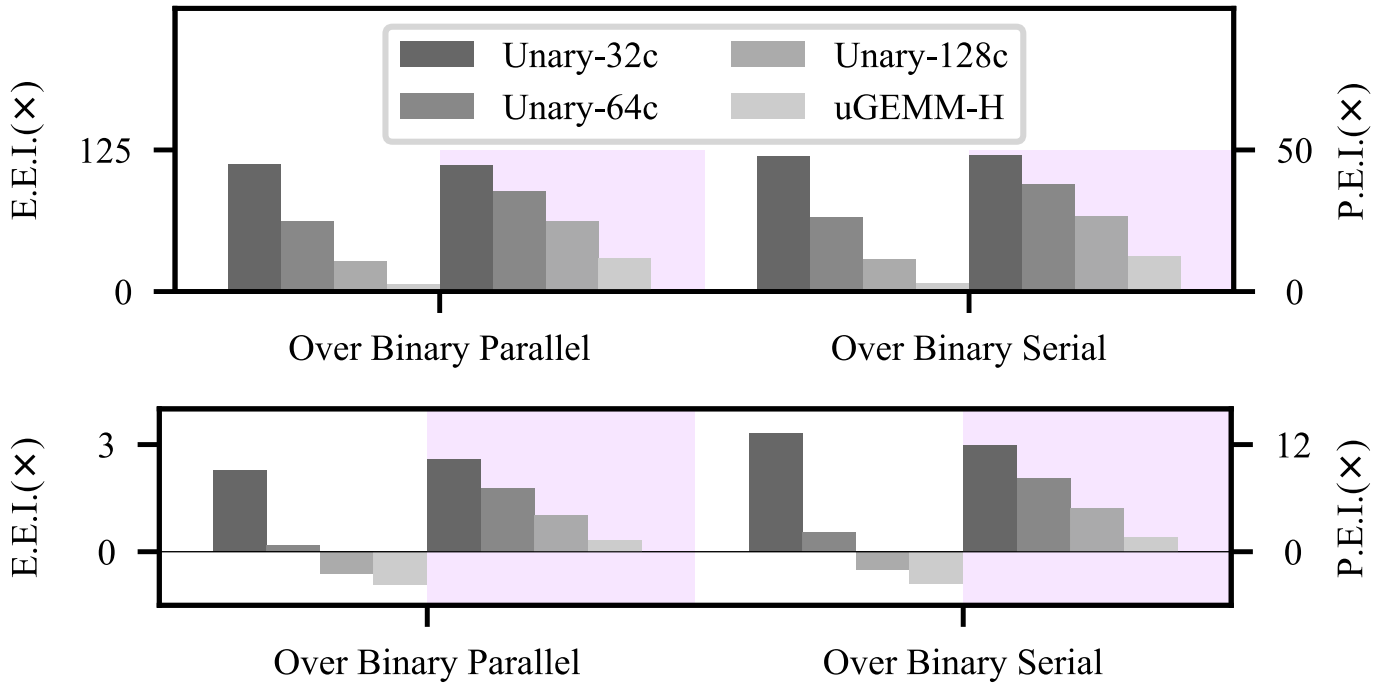
- Early termination in uSystolic reduces the on-chip energy and achieves dynamic accuracy-energy scaling
- Early termination cannot significantly reduce the total energy dominated by DRAM

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Efficiency

AlexNet

MLPerf



White and purple background colors are for energy efficiency improvement (E.E.I.) and power efficiency improvement (P.E.I)

- Early termination in uSystolic always increases the on-chip energy efficiency and power efficiency, thanks to the linearly increased throughput.

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Outline

- ❑ Motivation
- ❑ Background
- ❑ Architecture
- ❑ Evaluation
- ❑ Conclusion

Conclusion

- uSystolic provides
  - High accuracy
    - Binary accumulation
    - Spatial-temporal bitstream reuse
  - High efficiency
    - Unipolar multiplication
    - Spatial-temporal bitstream reuse
    - On-chip SRAM elimination
  - High scalability and generalizability
    - Systolic array

Related resource

- Unary computing website
  - https://unarycomputing.github.io

- Publicly available simulator
  - https://github.com/diwu1990/UnarySim
  - https://github.com/diwu1990/uSystolic-Sim

Thank you!
Q & A

Di Wu and Joshua San Miguel
di.wu@ece.wisc.edu and jsanmiguel@wisc.edu

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON