http://synergy.ece.gatech.edu

MAERI: Enabling Flexible Dataflow Mappi over DNN Accelerator via Reconfigurable Interconnects

http://synergy.ece.gatech.edu/tools/maeri/maeri-tutorial-hpca-2019

Tushar Krishna
Georgia Tech

Tutorial @ HPCA 2019
Feb 16 2019

Schedule: Afternoon

MAERI

Time	Topic	Presenter
2:00 - 2:20	**Overview of MAERI**	Tushar
2:20 – 3:00	**MAERI Mapper**	Zhongyuan
3:00 – 3:20	**MAERI RTL**	Hyoukjun
3:20 – 3:40	**MAERI Demo + Hands-on Exercises**	Hyoukjun
3:40 – 4:00	**Coffee Break**	
4:00 – 4:30	**Extensions**	Michael
4:30 – 4:40	**Wrap-Up**	Tushar

Dataflows within DNN Accelerators

- **DNN Topologies**
  - Layer size / shape
  - Layer types: Convolution / Pool / FC / LSTM
  - New sub-structure: e.g., Inception in Googlenet
- **Compiler/Mapper (e.g., MAESTRO)**
  - Loop reordering
  - Loop tiling size
  - Cross-layer mapping
- **Algorithmic Optimization**
  - Weight pruning: Sparse workload

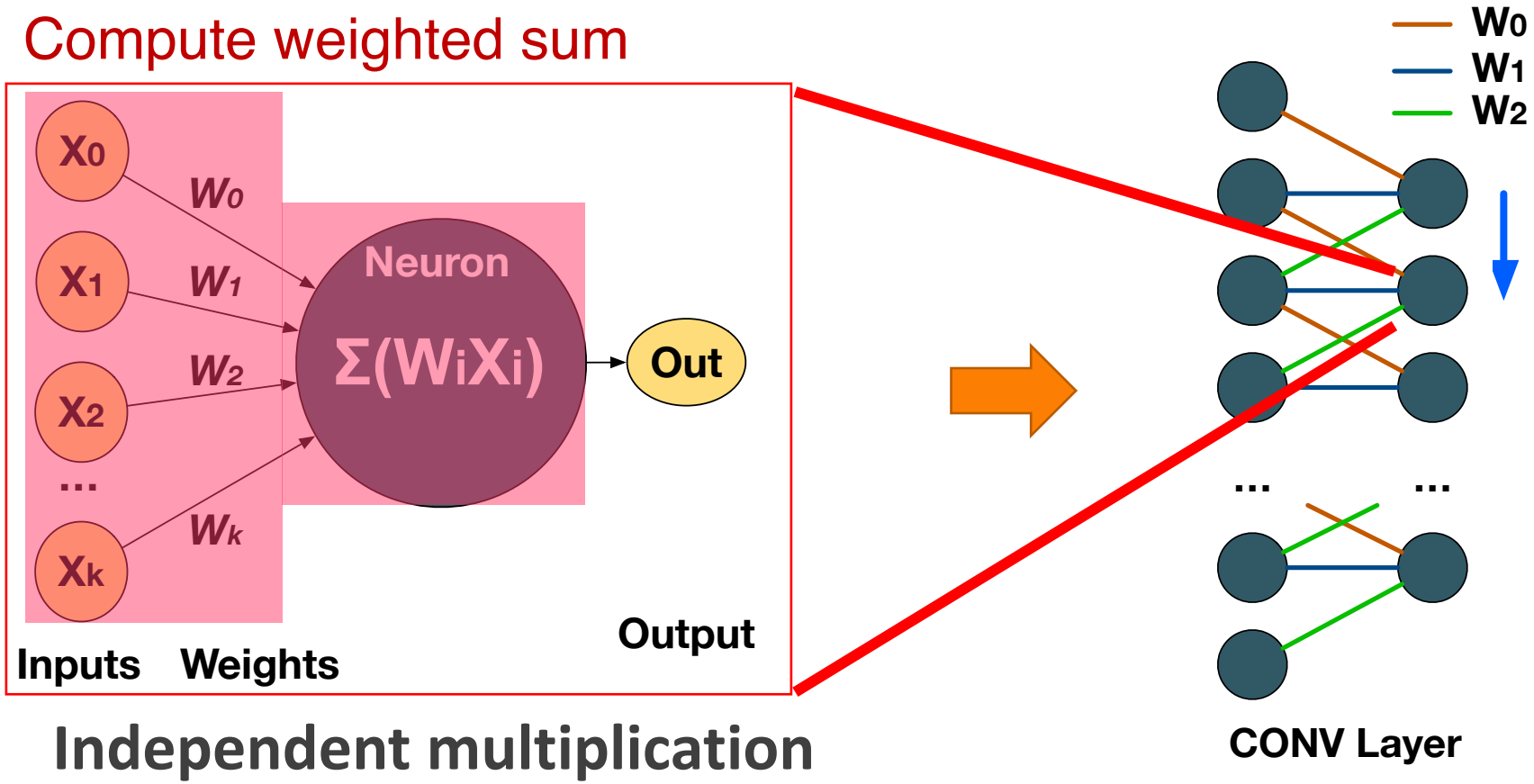
The current trend for supporting multiple dataflows

- New Dataflow → New Accelerator
  - **Data reuse:** FlexFlow (2017), Eyeriss (2016), ...
  - **Cross-layer:** Fused CNN (2016)
  - **Sparse CNN:** SCNN (2017), EIE(2016), ...
  - **LSTM:** ESE (2017), ...

Can we have one architectural solution that can handle arbitrary dataflows and provides ~100% utilization?

What is the computation in a DNN?

Compute weighted sum

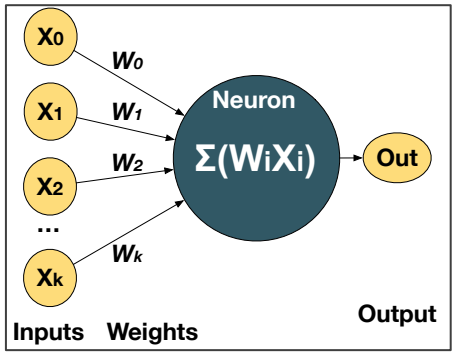


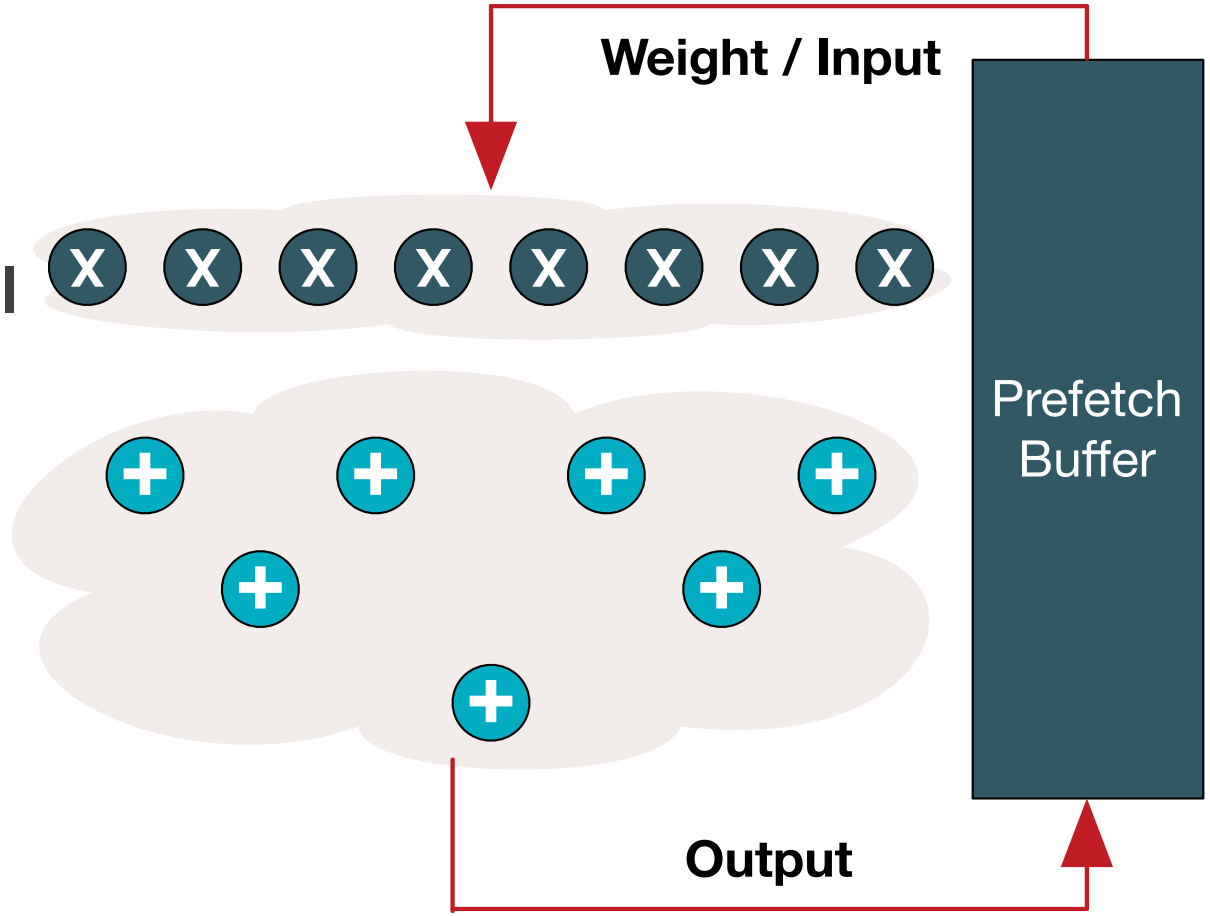
Independent multiplication

Accumulation of partial products

Our Key insight: Each dataflow translates into neurons of different sizes

The MAERI Abstraction

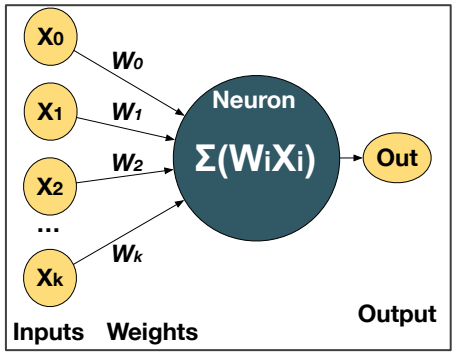




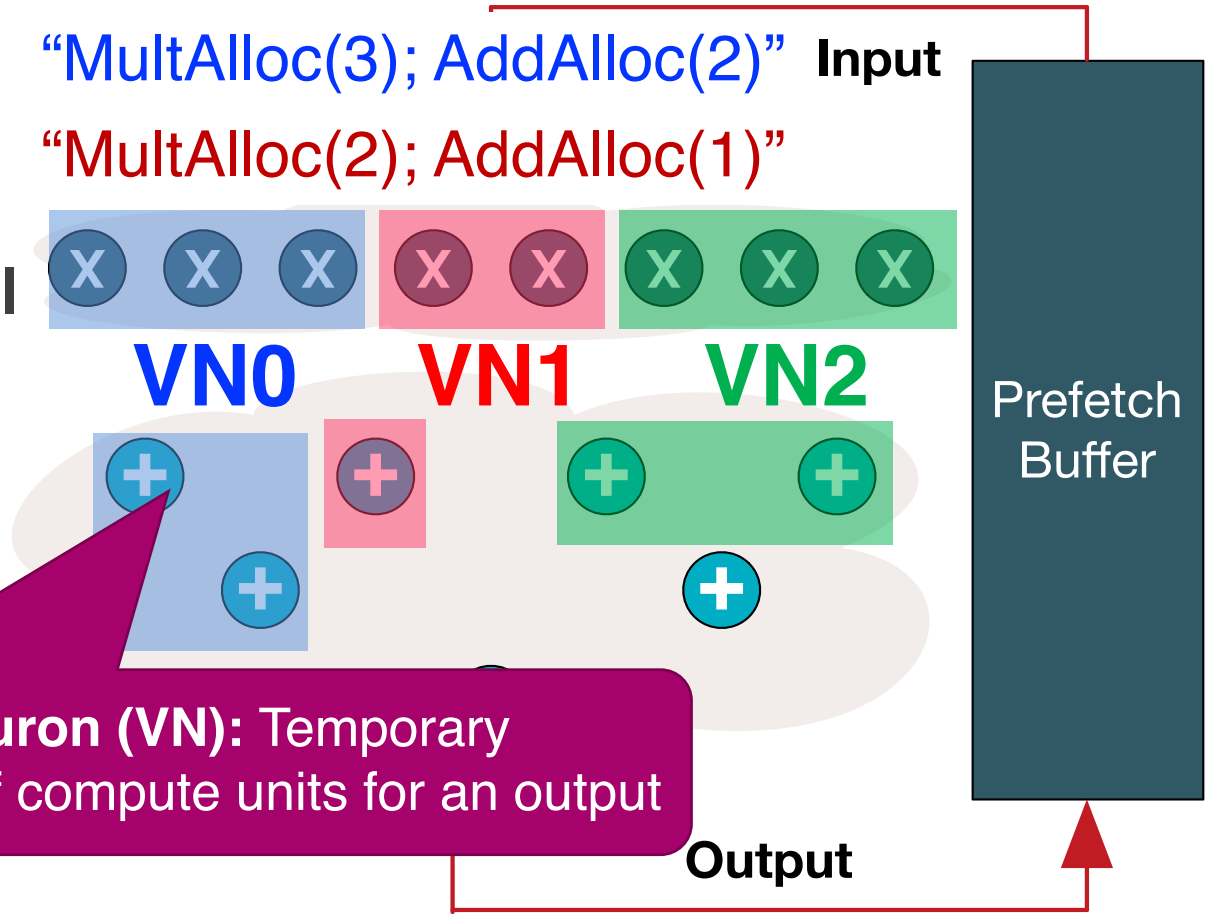
How to enable flexible grouping?

Reconfigurable Interconnects!

The MAERI Abstraction



Multiplier Pool



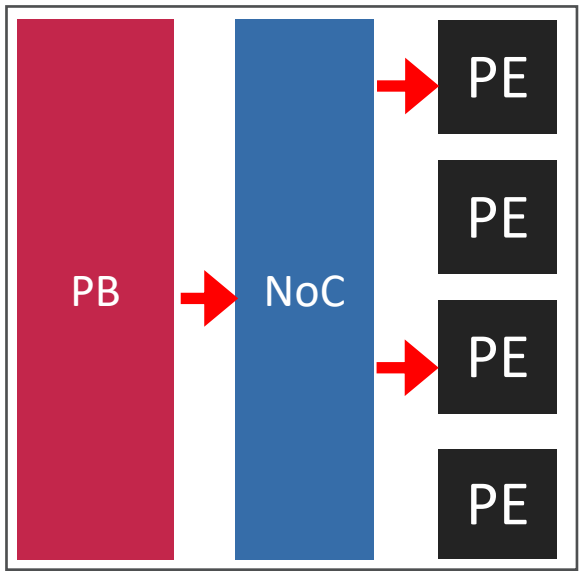
Adder Pool

Virtual Neuron (VN): Temporary grouping of compute units for an output

How to enable flexible grouping?

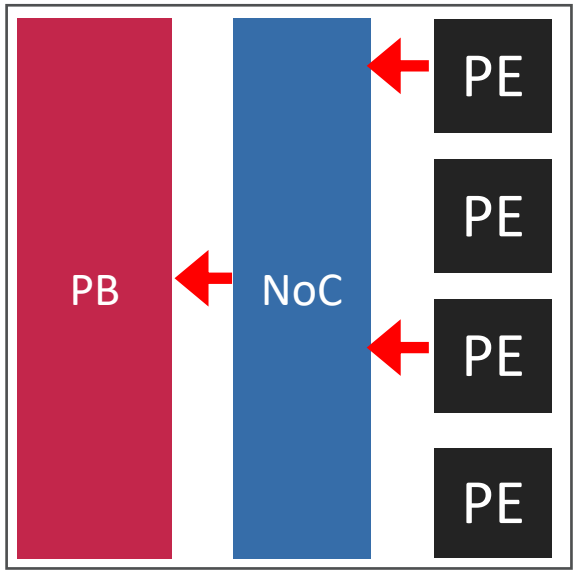
Reconfigurable Interconnects!

Traffic Patterns in DNN Accelerators*



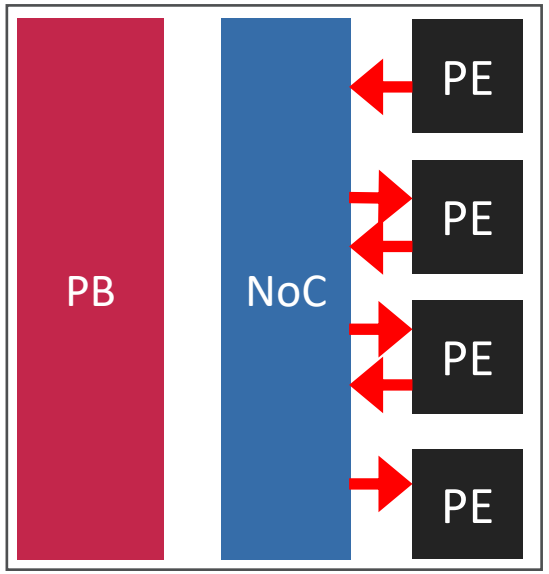
Distribution

e.g. weight distribution to PEs in weight-stationary accelerator



Collection

e.g. output collection in output-stationary accelerator



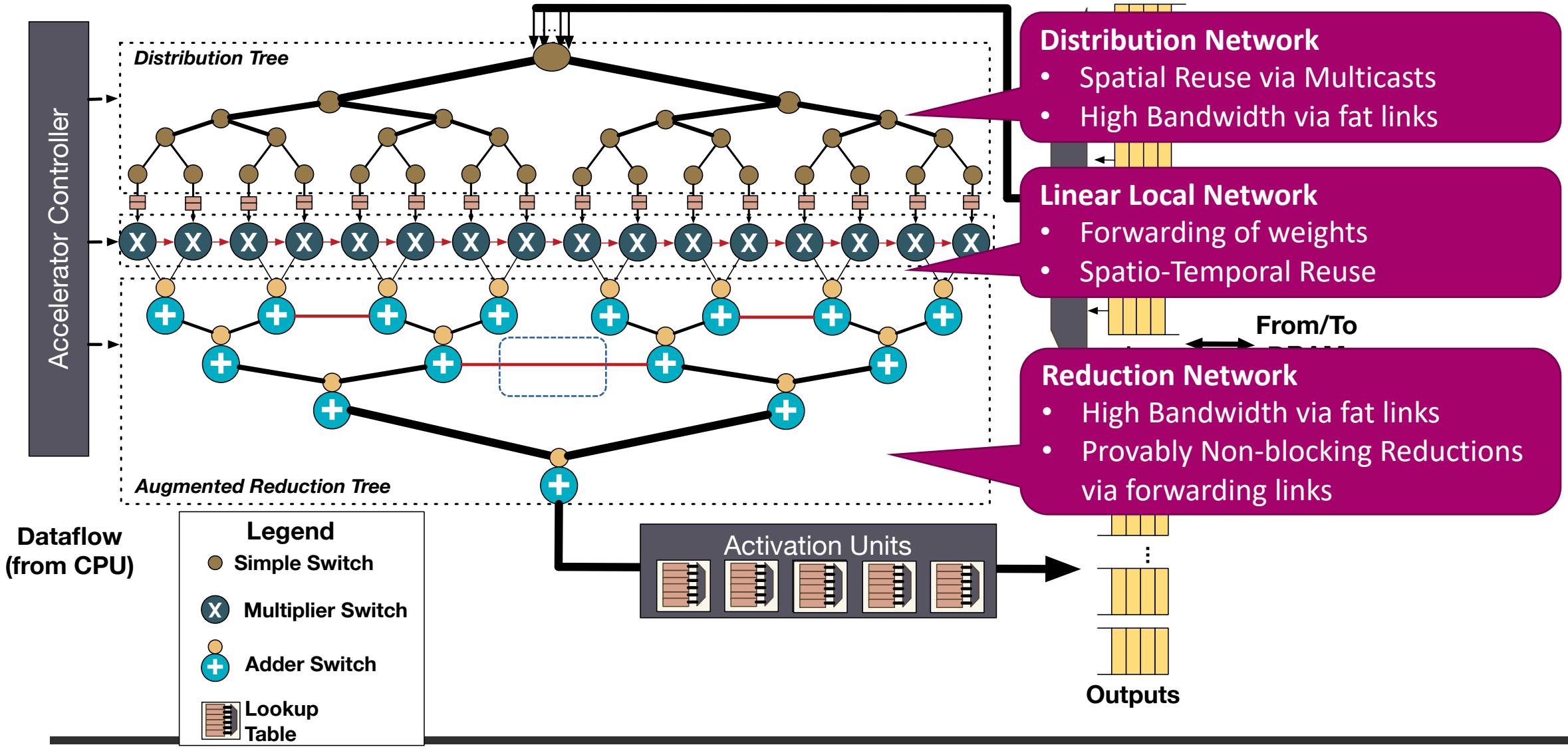
Local Forwarding

e.g. Input forwarding in row-stationary accelerator

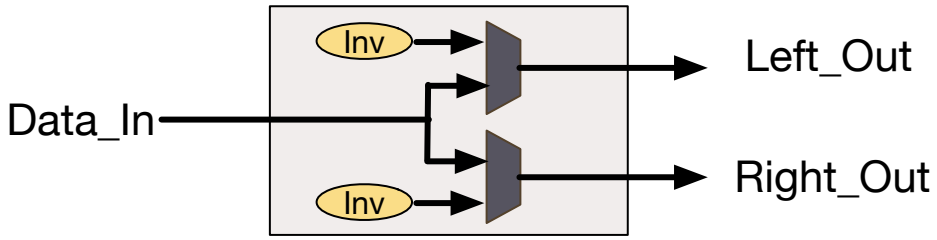
* **PB:** Prefetch buffer (Global buffer)
* **NoC:** Network-on-Chip (Interconnection network)
* **PE:** Processing element (Compute units)

* H. Kwon et al., Rethinking NoCs for Spatial DNN Accelerators, NOCS 2017

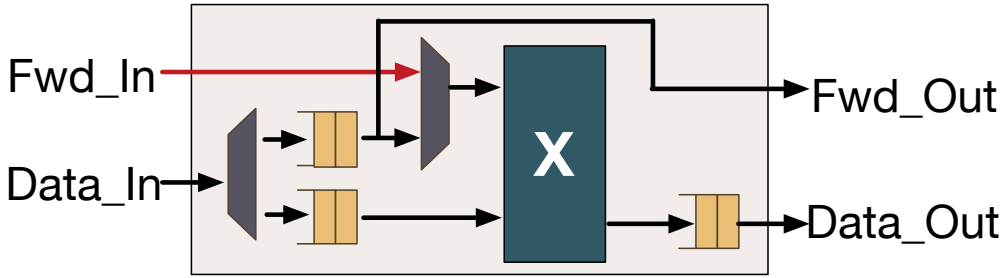
The MAERI Implementation



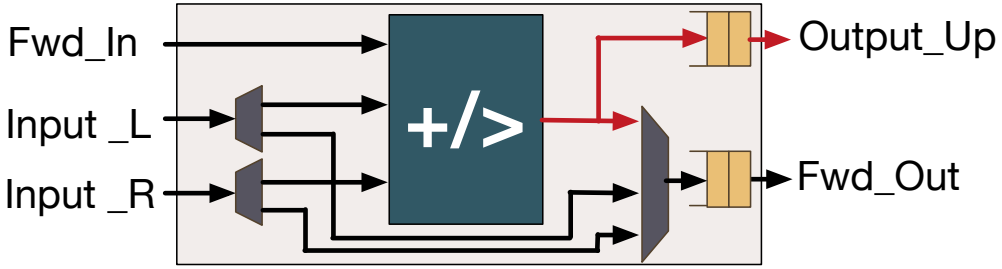
ntation



Distribute Switch (1x2 Switch)



Multiplier Switch (multiplier+ 2x2 switch)



Adder Switch (adder+ 3x2 switch)

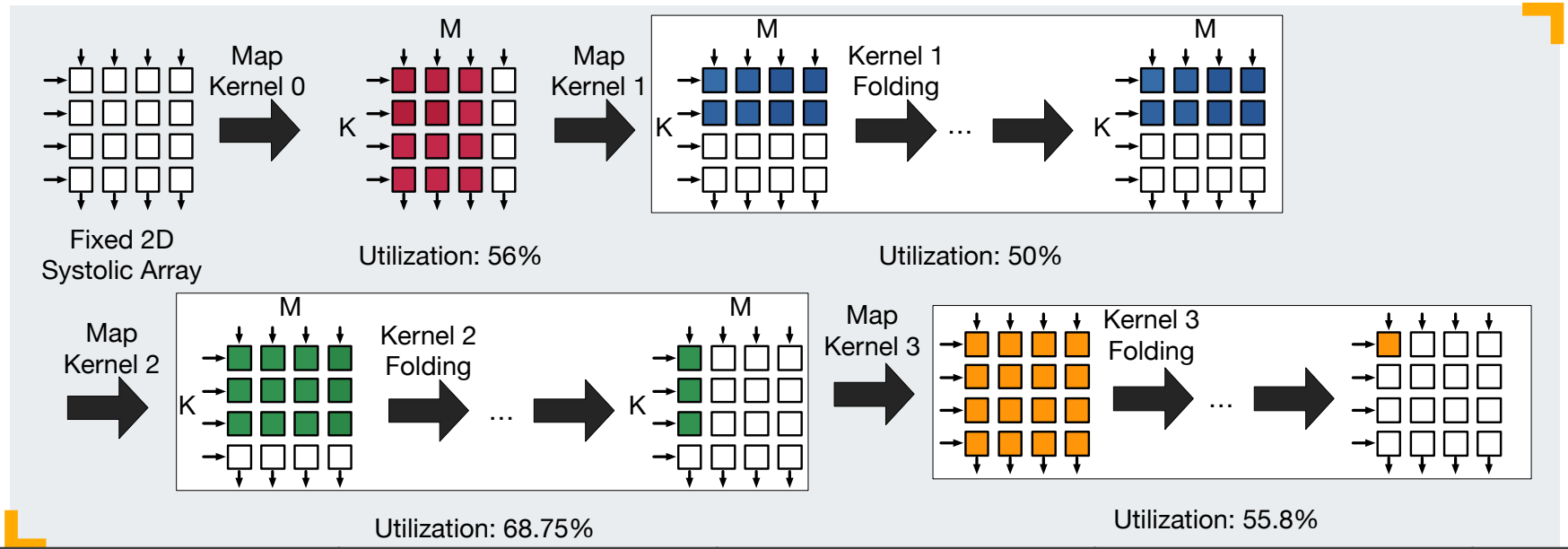
Distribution Network
- Spatial Reuse via Multicasts
- High Bandwidth via fat links

Linear Local Network
- Forwarding of weights
- Spatio-Temporal Reuse

Reduction Network
- High Bandwidth via fat links
- Provably Non-blocking Reductions via forwarding links

Why 1D Configurable Design?

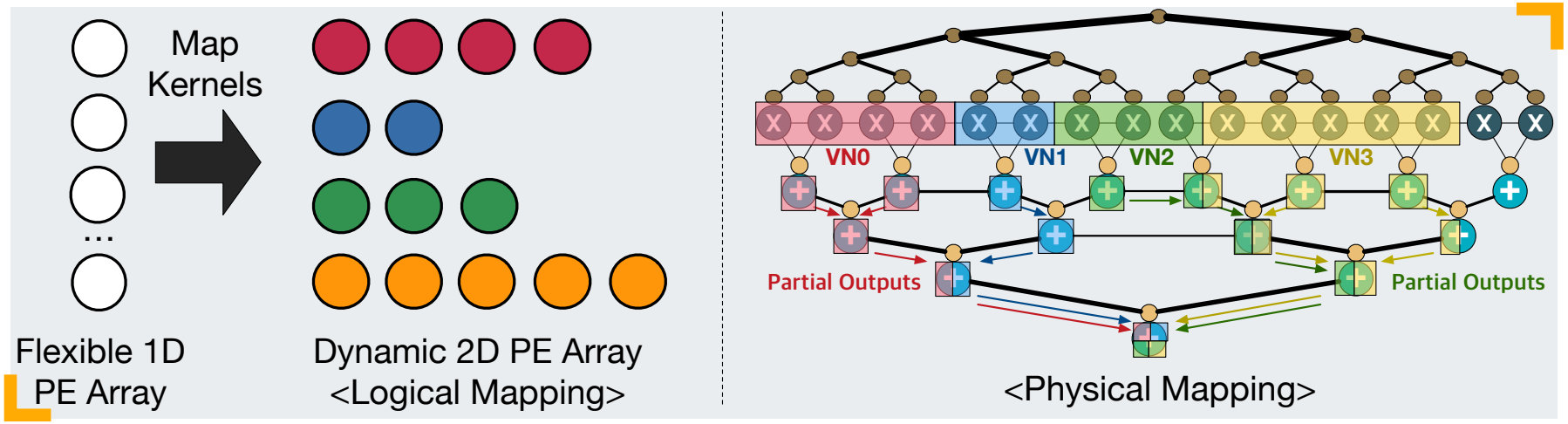
Systolic Array (TPU): Map kernels on fixed 2D array



Dimension	GEMM1	GEMM2	GEMM3	GEMM4
M	3	128	25	25
N	128	256	64	12
K	4	2	3	5

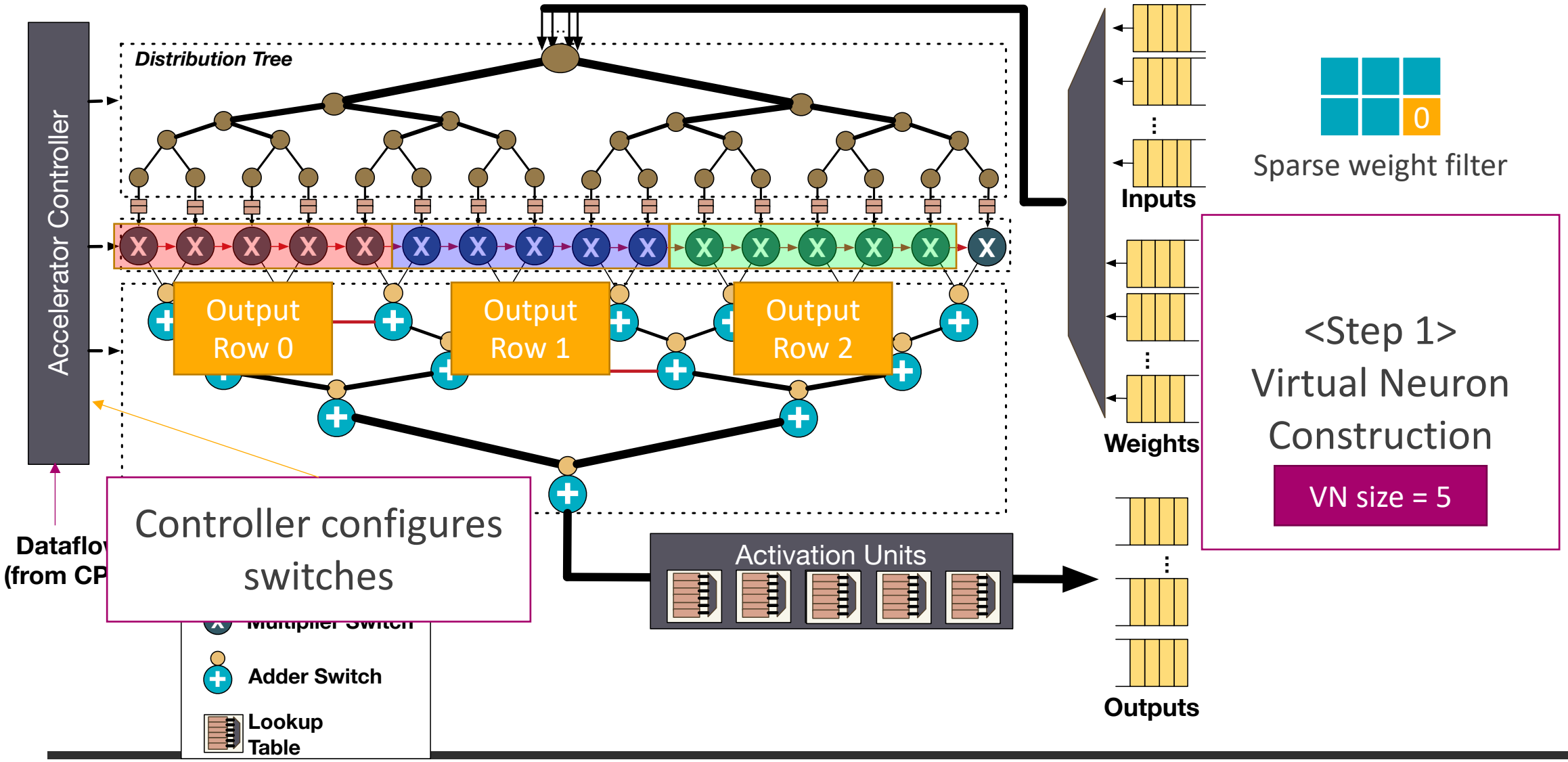
Why 1D Configurable Design?

MAERI: Constructs 2D array at run time

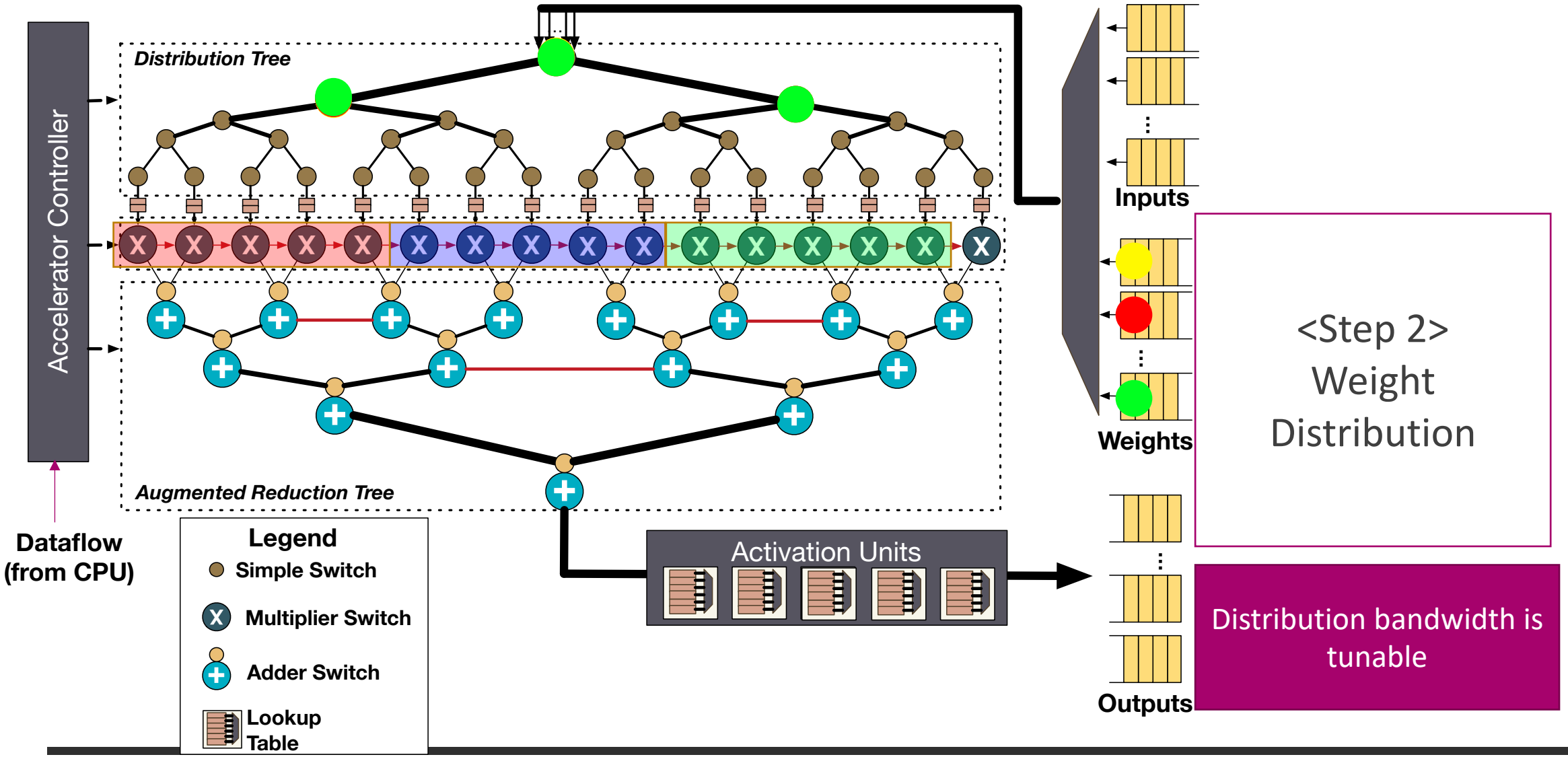


Dimension	GEMM1	GEMM2	GEMM3	GEMM4
M	3	128	25	25
N	128	256	64	12
K	4	2	3	5

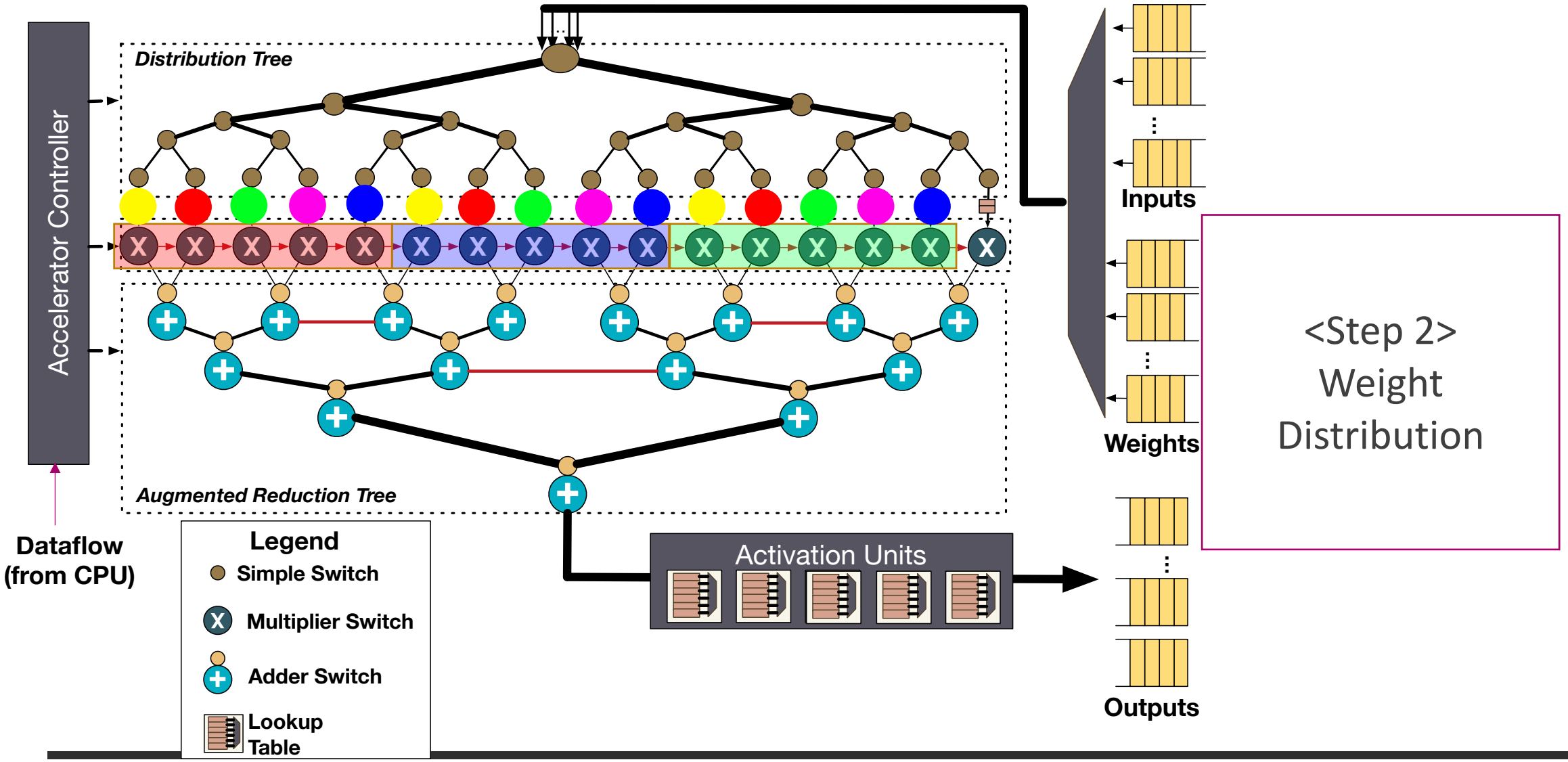
MAERI Operation Example



MAERI Operation Example

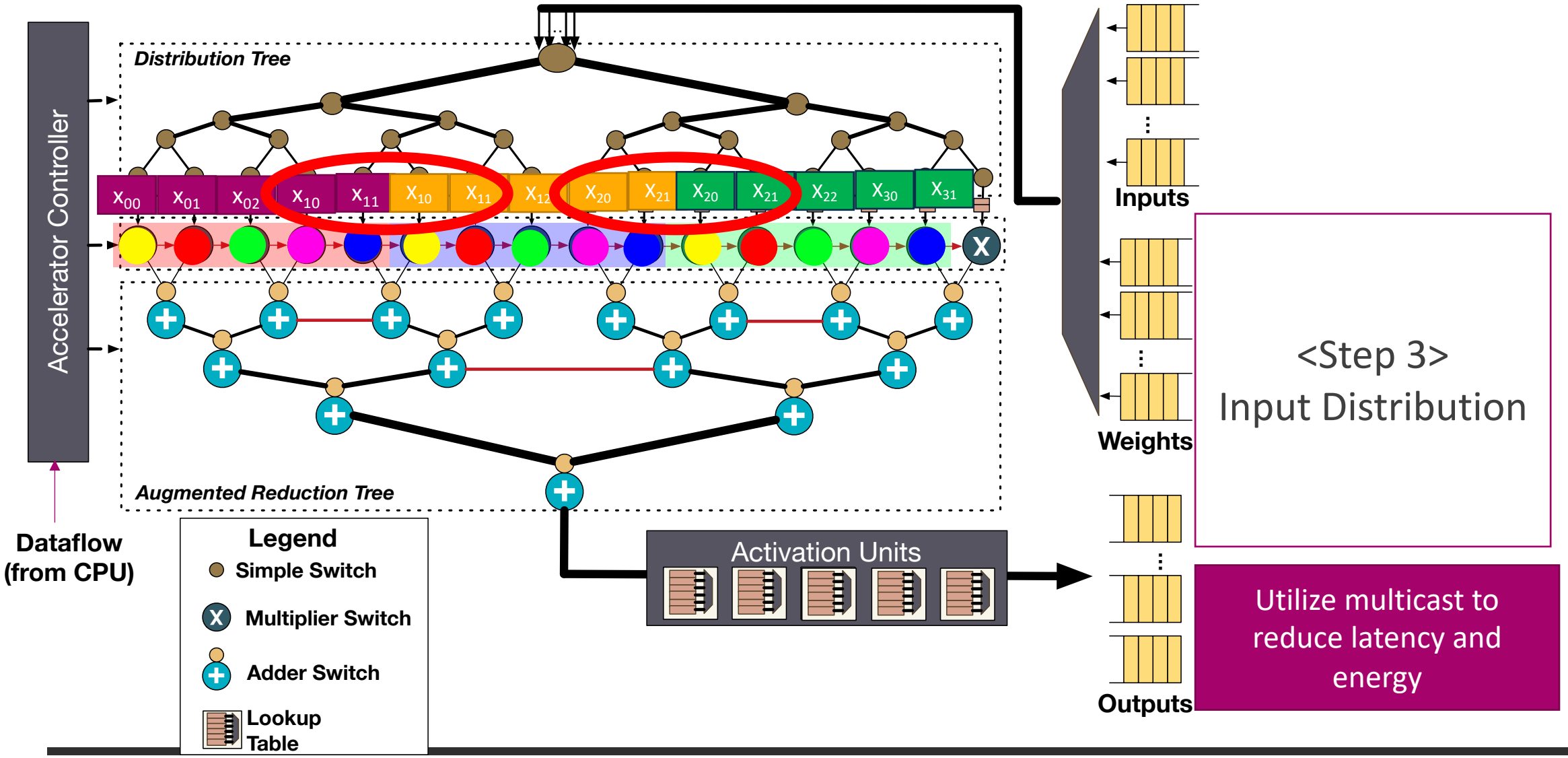


MAERI Operation Example



<Step 2>
Weight
Distribution

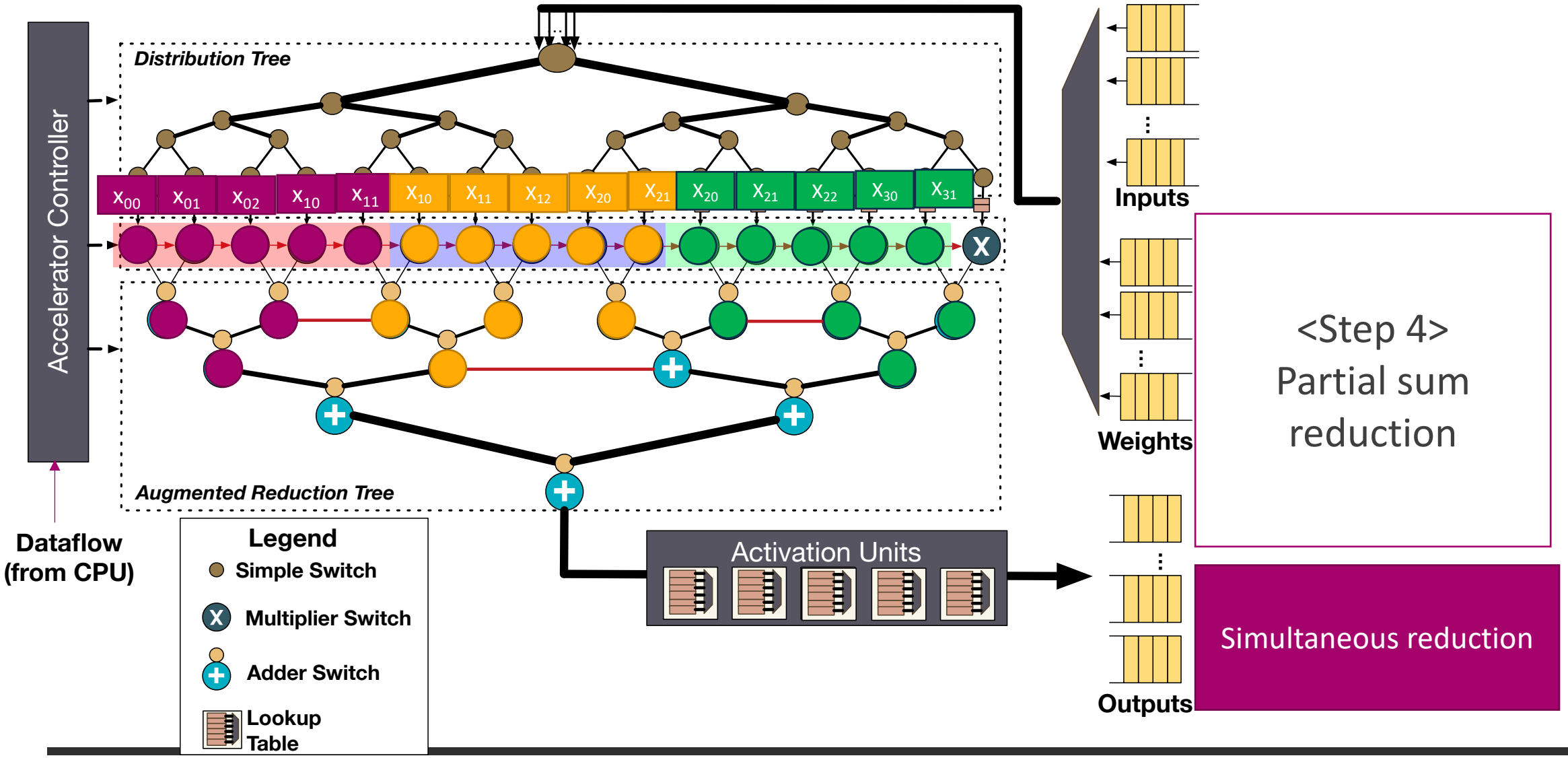
MAERI Operation Example



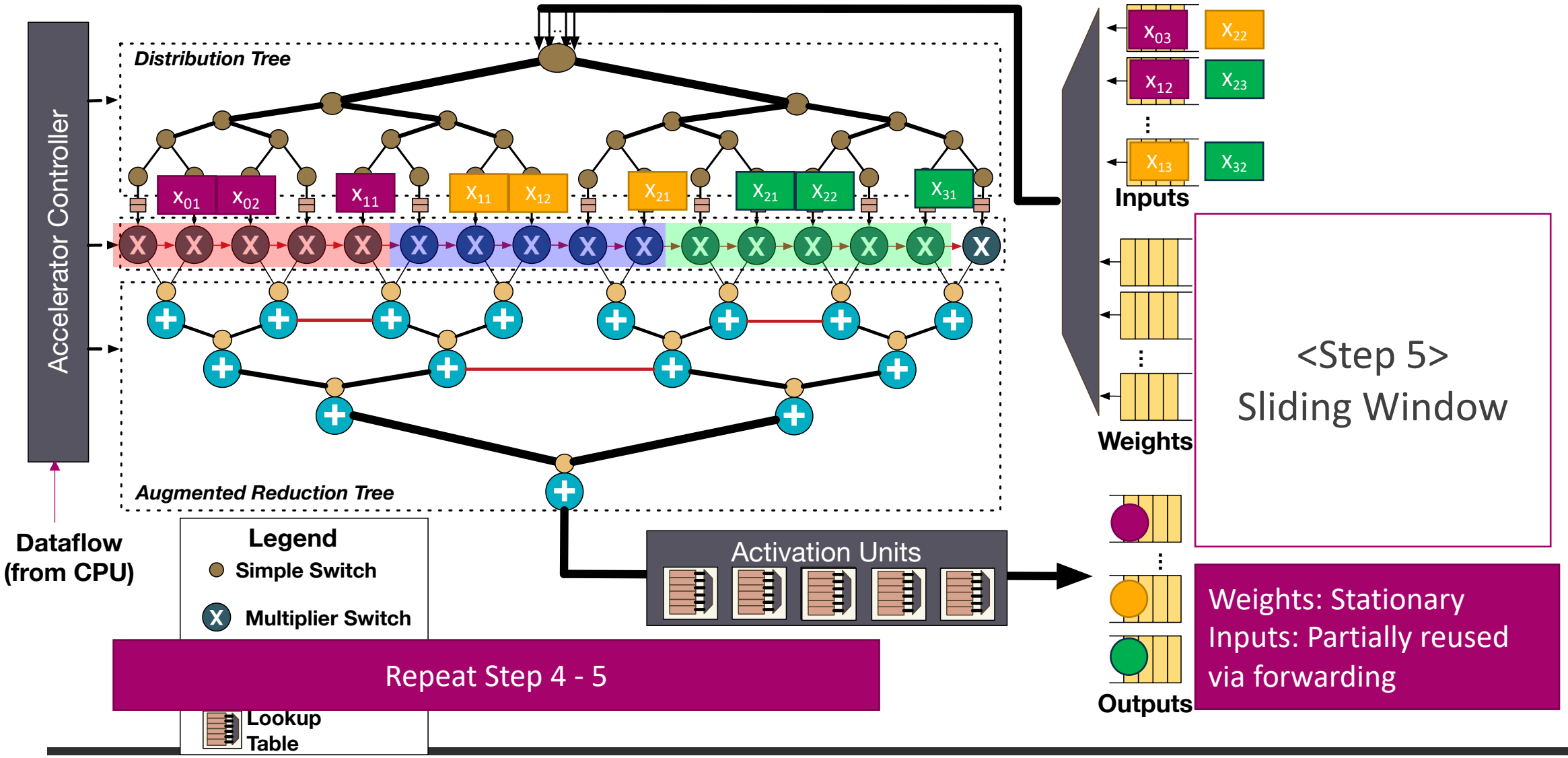
<Step 3>
Input Distribution

Utilize multicast to reduce latency and energy

MAERI Operation Example

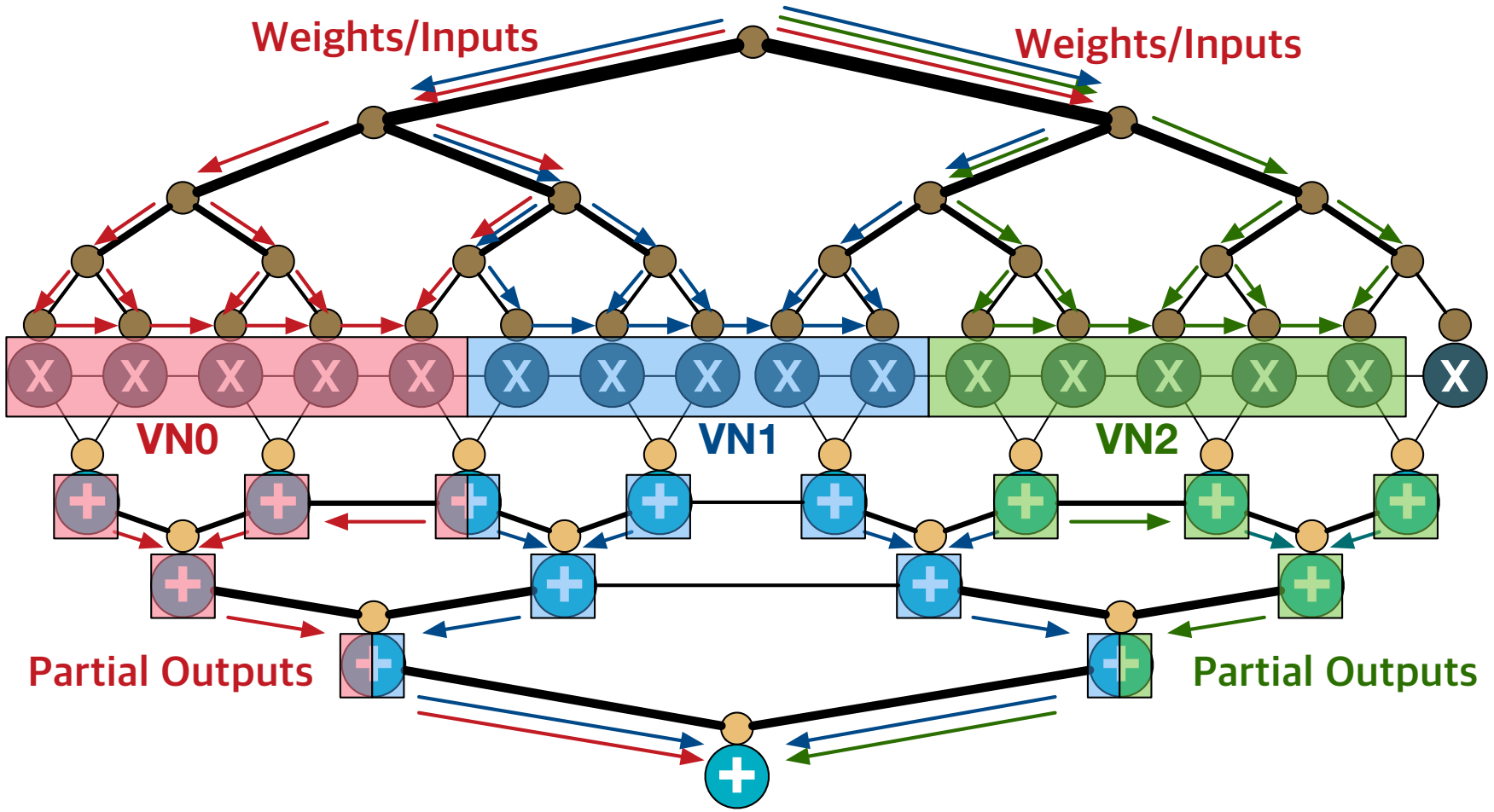


MAERI Operation Example



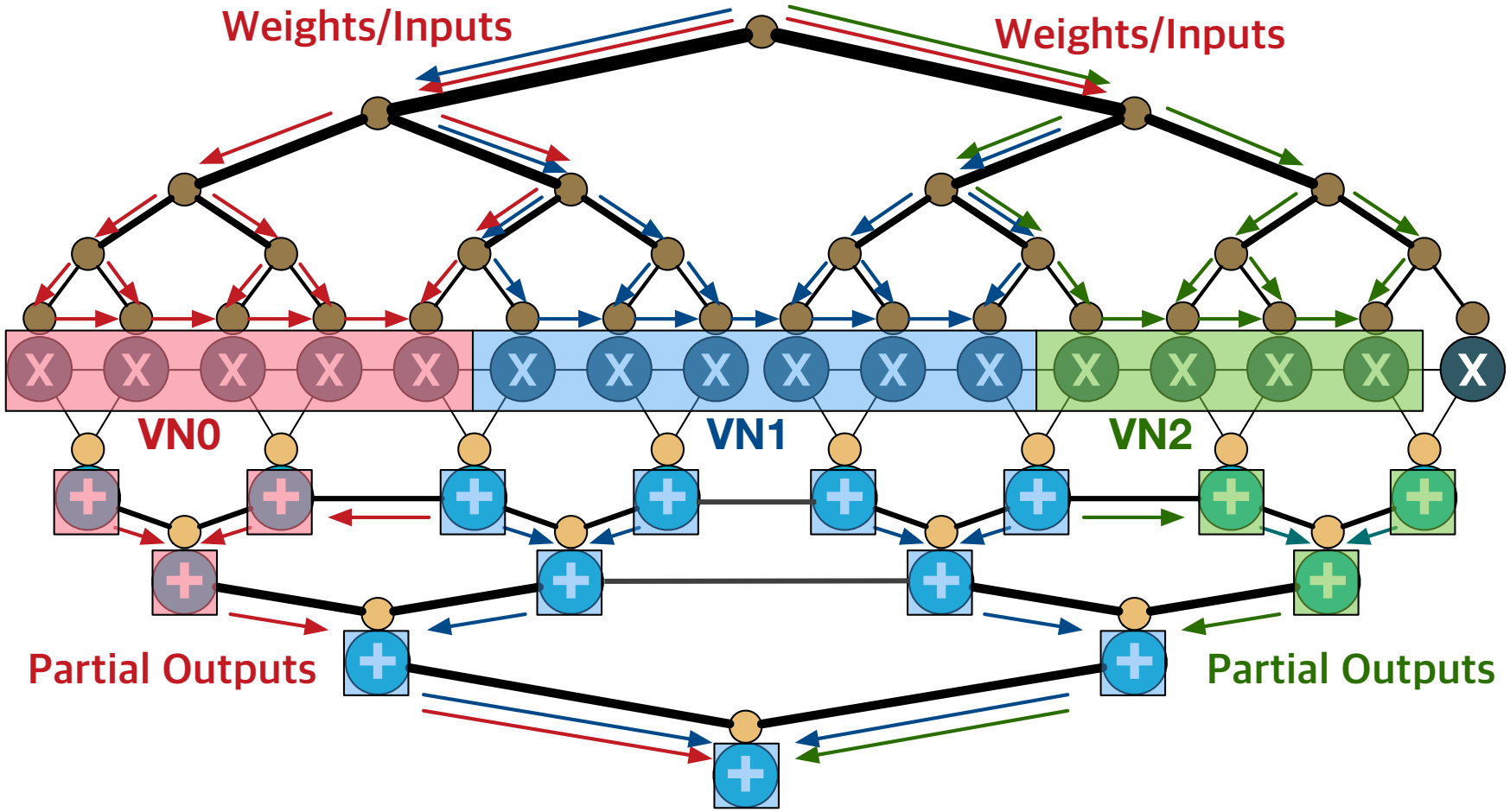
Example Mapping – Dense CNN

Our Key insight: Each dataflow translates into neurons of different sizes



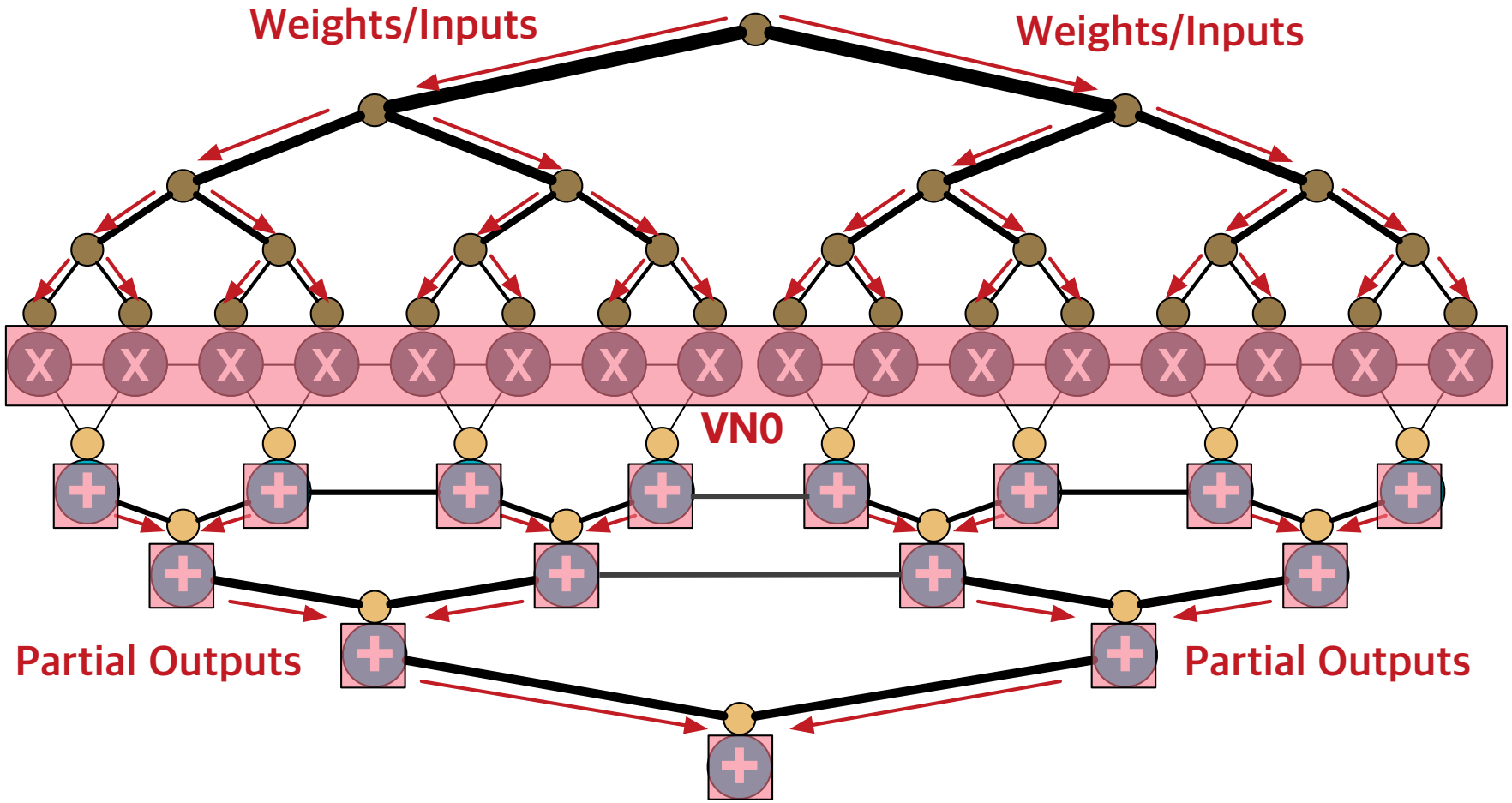
Example Mapping – Sparse CNN

Our Key insight: Each dataflow translates into neurons of different sizes



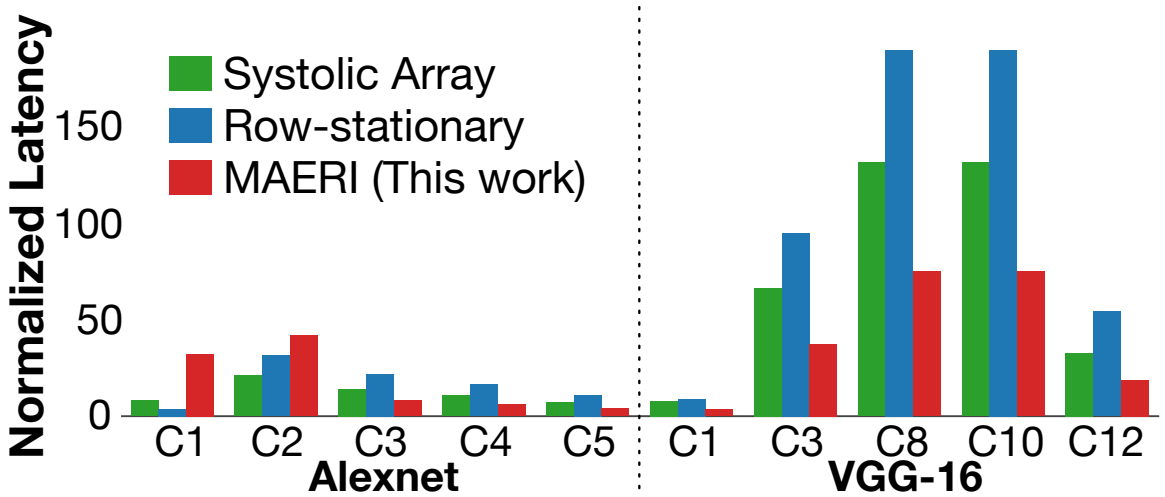
Example Mapping – LSTM / FC

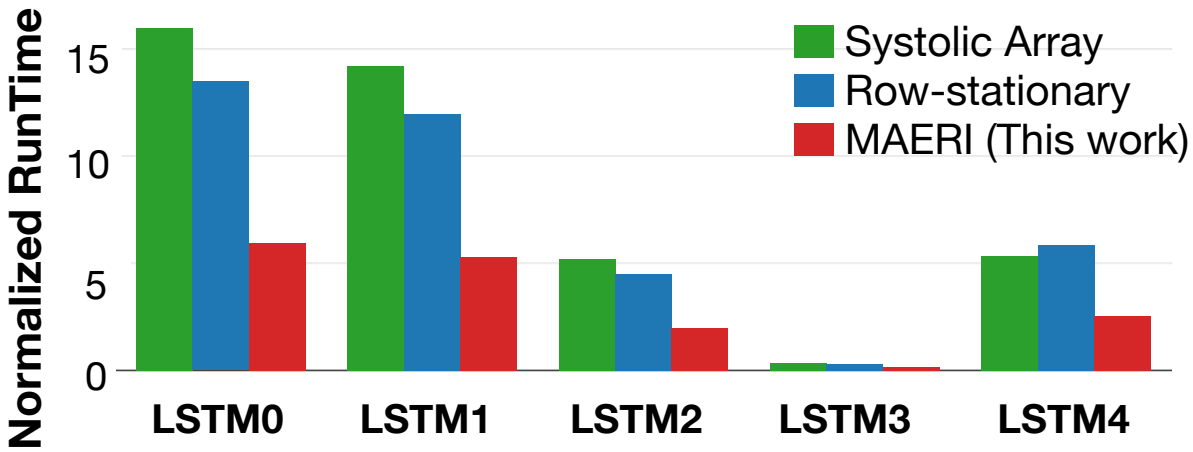
Our Key insight: Each dataflow translates into neurons of different sizes



Performance with Dense Workload

- **Total Latency (Runtime) for Convolution**





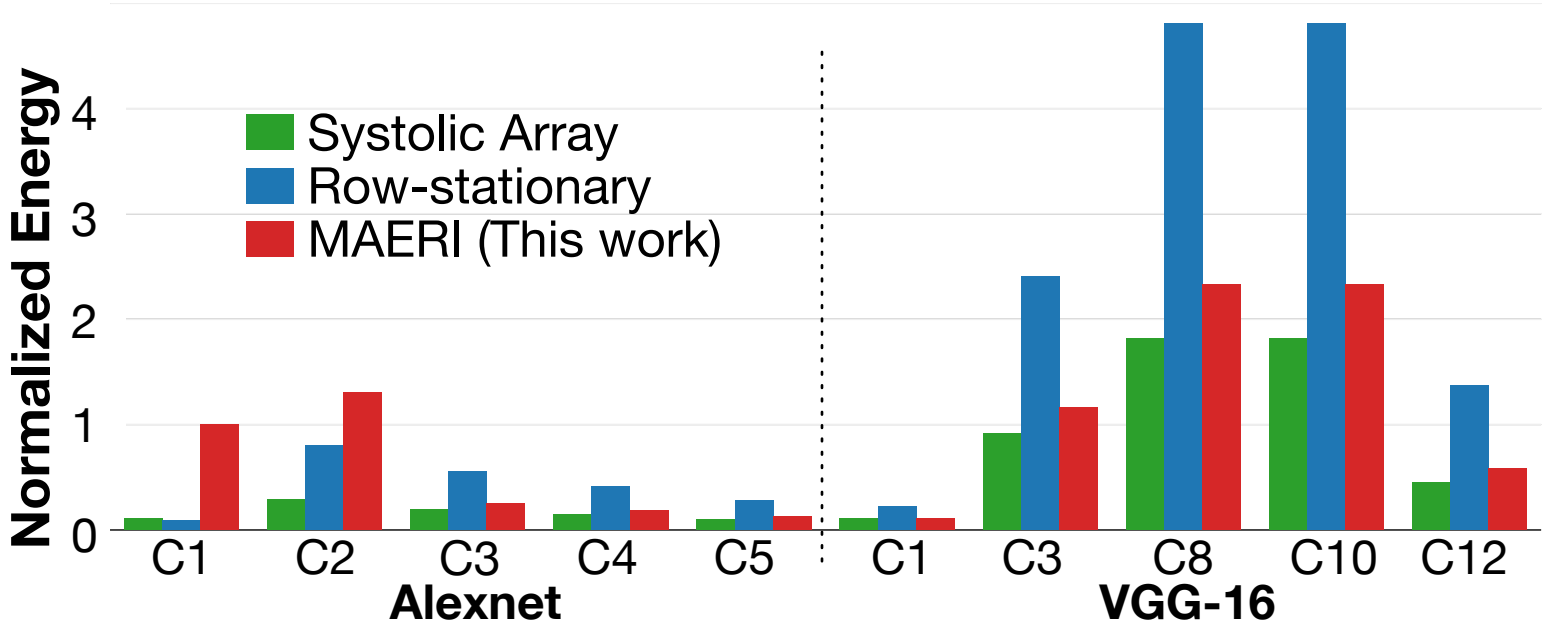
* Normalized to ideal case (100% utilization, Infinite bandwidth)

MAERI reduces runtime upto 65%, 42% in avg.

MAERI reduces runtime upto 63%, 57% in avg.

LSTMs from Yonghui We, et. al., "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation.", Arxiv Preprint, 2016

Energy with Convolution Layers



* Normalized to MAERI energy with Alexnet C1

MAERI reduces energy upto 57% and 28% in average compared to Row-Stationary (dense dataflow) and 7.1% in average compared to Systolic Array (sparse dataflow)

References

- **MAERI: Enabling Flexible Dataflow Mapping over DNN Accelerators via Reconfigurable Interconnects**
  Hyoukjun Kwon, Ananda Samajdar, and Tushar Krishna
  *In Proc of the 23rd ACM International Conference on Architectural Support for Programming Languages and Operating Systems **(ASPLOS)**, Mar 2018*
  - *Selected as Honorable Mention in IEEE Micro Top Picks in Computer Architecture 2019*

- **A Communication-Centric Approach for Designing Flexible DNN Accelerators**
  Hyoukjun Kwon, Ananda Samajdar, and Tushar Krishna
  In IEEE Micro Special Issue on Hardware Acceleration, Nov/Dec 2018

- **Enabling Efficient Mapping Space Exploration on a Reconfigurable Neural Accelerator**
  Zhongyuan Zhao, Hyoukjun Kwon, Sachit Kuhar, Weiguang Sheng, Zhigang Mao, and Tushar Krishna
  *In Proc of the IEEE Inter. Symp. on Performance Analysis of Systems and Software **(ISPASS)**, Mar 2019*

- **Rethinking NoCs for Spatial Neural Network Accelerators**
  Hyoukjun Kwon, Ananda Samajdar, and Tushar Krishna
  *In Proc of 11th International Symposium on Networks-on-Chip **(NOCS)**, Oct 2017*

Schedule: Afternoon

MAERI

Time	Topic	Presenter
2:00 - 2:20	**Overview of MAERI**	Tushar
2:20 – 3:00	**MAERI Mapper**	Zhongyuan
3:00 – 3:20	**MAERI RTL**	Hyoukjun
3:20 – 3:40	**MAERI Demo + Hands-on Exercises**	Hyoukjun
3:40 – 4:00	**Coffee Break**	
4:00 – 4:30	**Extensions**	Michael
4:30 – 4:40	**Wrap-Up**	Tushar