ECE 590: Neuromorphic Computing

Lecture 9: Memristor-based Neuromorphic – Architecture

Yiran Chen
Duke University

Outline

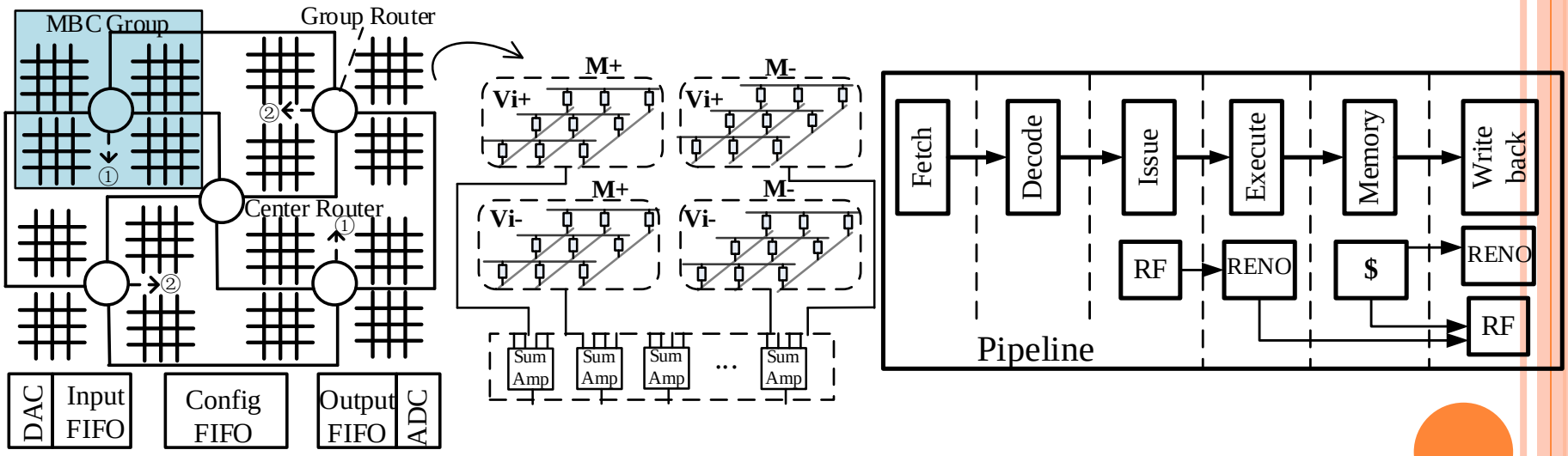
- Paper review session agenda
- Memristor-based arrays with NoC
- Larger Implementation: PipeLayer
- Memristor-based Graph-computing

Paper Review Session Agenda

Feb. 20, 2018 (Tue.)		Feb. 22, 2018 (Thu.)	
1	Zhiyao Xie	1	Arjun Chaudhuri
2	Huanrui Yang	2	Xiong Cao
3	Wei Wen	3	Rana Elnaggar
4	Jiachen Mao	4	Naman Jain
5	Fan Chen	5	Mengyun Liu
6	Hsin-Pai Cheng	6	Xin Liu
7	Nate Inkawhich	7	Atefeh Mehrabi
		8	Jingchi Zhang

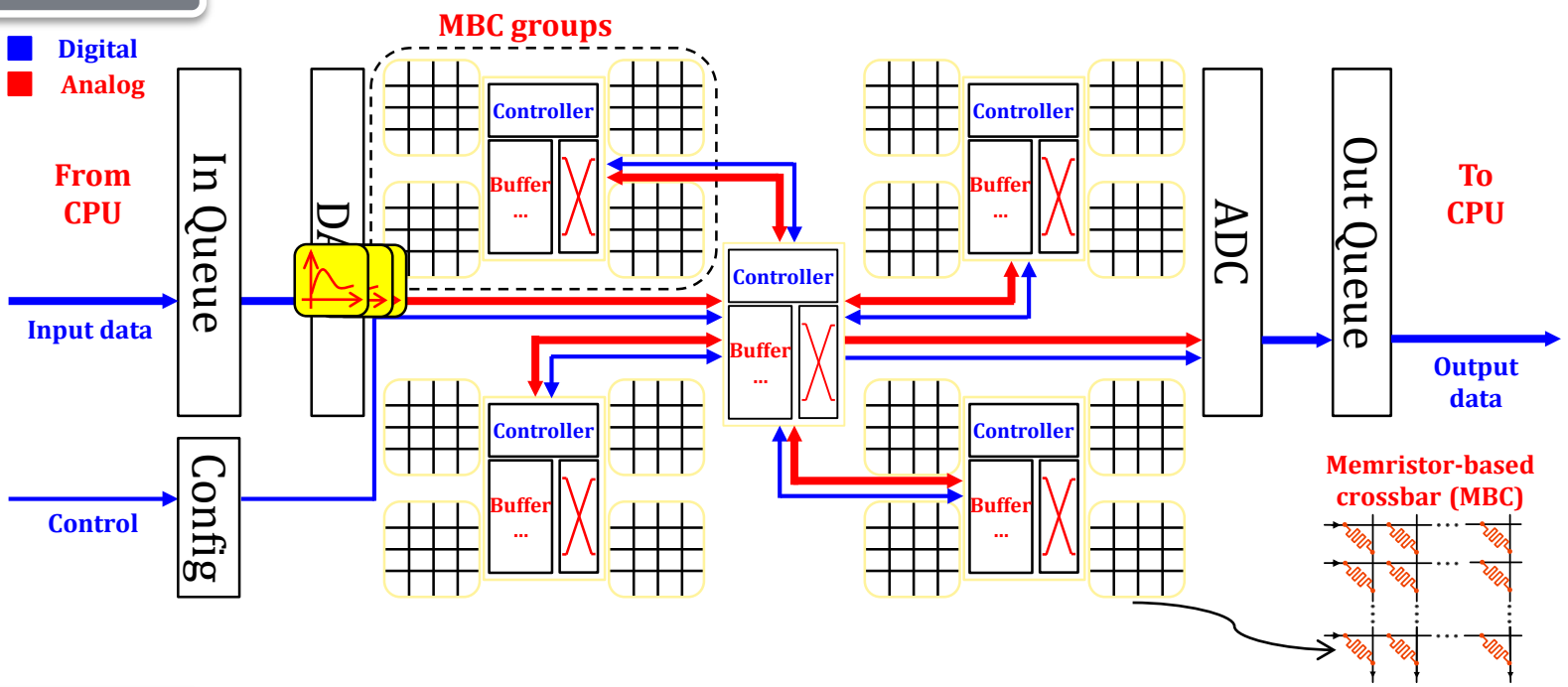
RENO Overview

- ***An efficient memristor-based mixed-signal accelerator*** is designed to speed up neuromorphic computing and support the implementations of a variety of neural network topologies;
- ***A mixed-signal interconnection network (M-Net)*** is proposed to assist the communication of computational signals among the MBCs;
- ***An optimized configuration*** is discussed and established by analyzing the impact of various design parameters on the system performance/accuracy.



Neuromorphic Computing Acceleration (NCA)

RENO Hardware



Software Support

```
bool Recall(float *vec, float *wm)
{   /* simulate the synapse network*/
   for(i=0;i<BsbSize;++i) wx[i] +=
wm[i*BsbSize+j] * vec[j];
  ……
```

Find the candidate codes

Source-to-source translation

```
bool Recall(float *vec)
{
   Send(RENO.id, vec);
   return Receive(RENO.id)
   ……
```

The neural topology

RENO-aware compilation

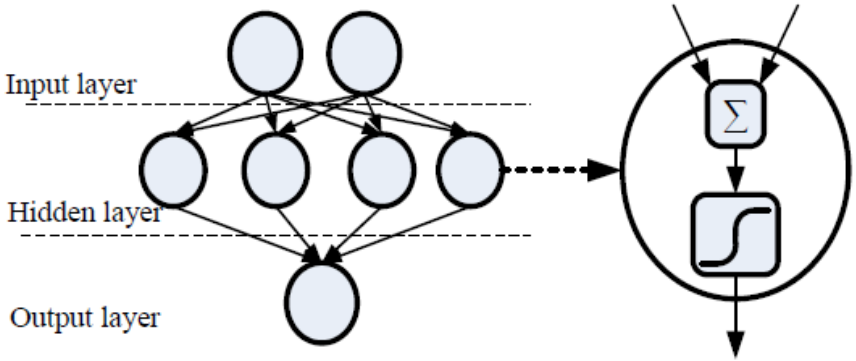
```
MOVD RENO.id, R1
……
SET RENO.id, #VAL
LAUNCH
DEQ R1, RENO.id
```

The NCA-aware executable

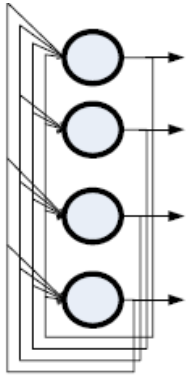
System Level Evaluation

- Two implementations representing tradeoffs between computation performance and accuracy

Multi-layer perception (MLP)



Auto-associative memory (AAM)



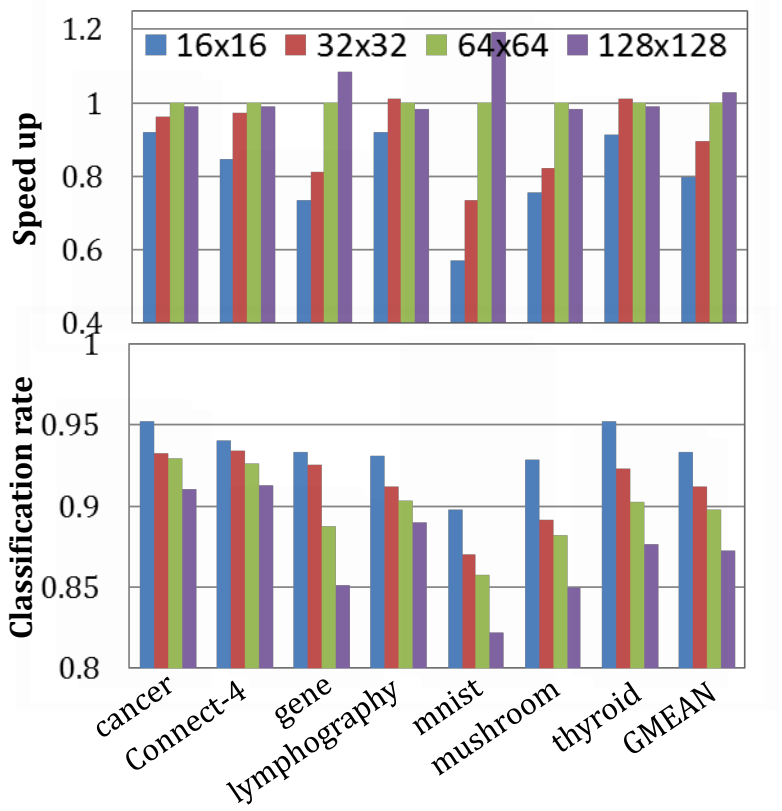
- 7 classification benchmarks
- Classification rate is used as reliability metric

Benchmark	Description
cancer	breast cancer diagnose
connect-4	connect-4 game
gene	nucleotide sequences detection
lymphography	lymph diagnose
MNIST	digit recognition
mushroom	poisonous mushroom discrimination
thyroid	thyroid diagnose

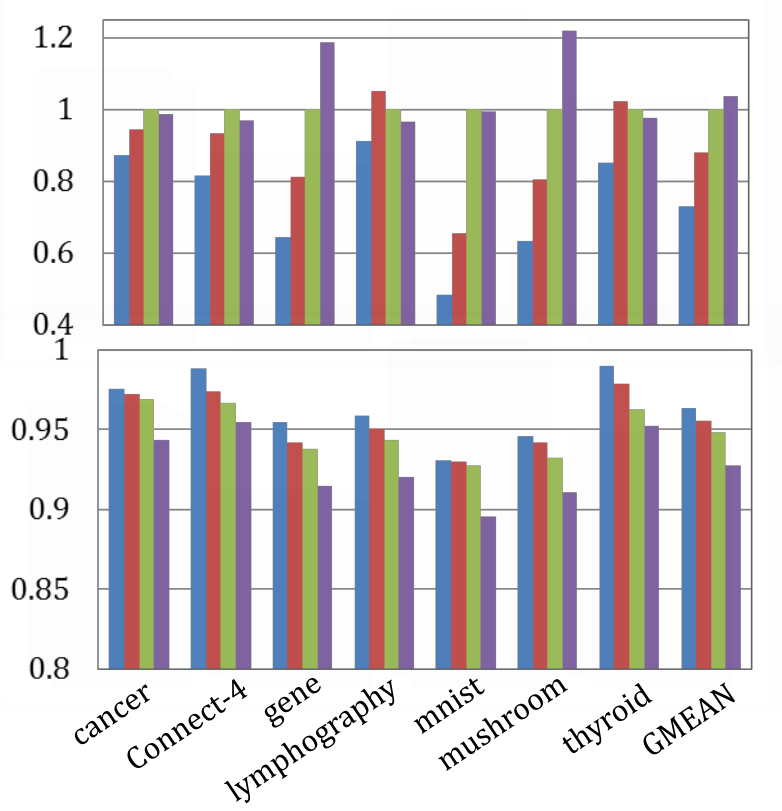
Optimize Configuration

- Exploration of MBC size
  - Performance: large size is preferable
  - But…with decreasing classification rate
  - Because the aggravated variations at a large MBC size
  - 64x64 is the best tradeoff

MLP



AAM

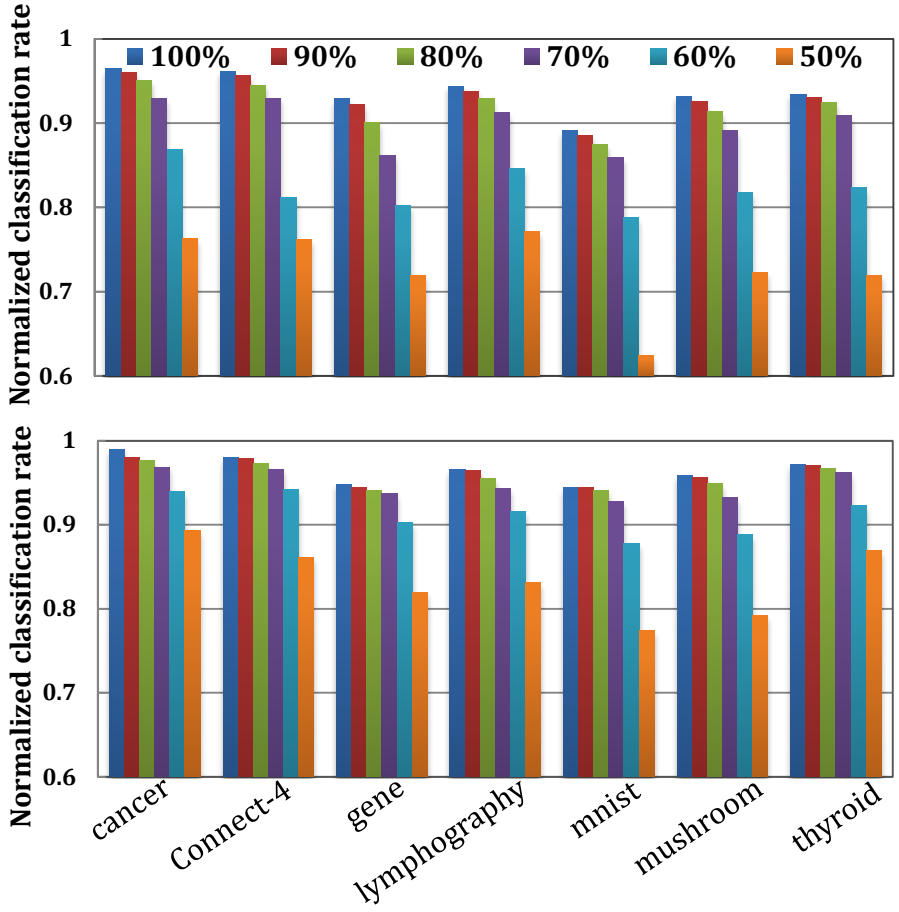


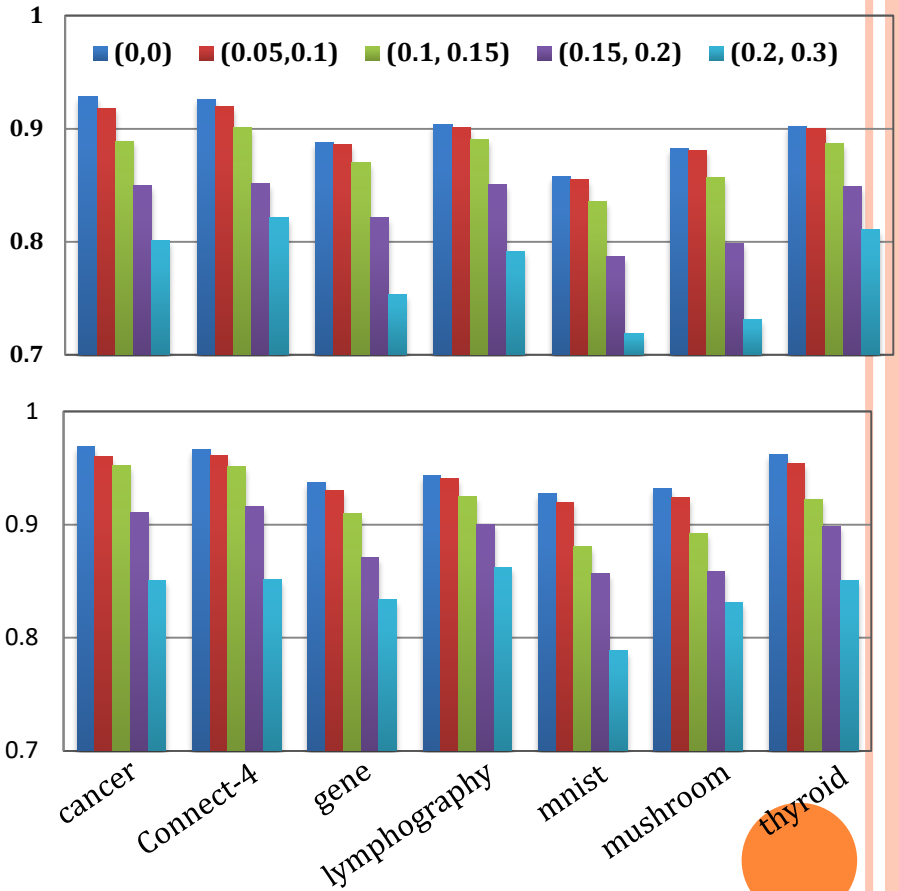
Deficient Training & Hardware

- MBC training effort: Limited accuracy of training
  - MLP (a) is more sensitive than AAM (b)

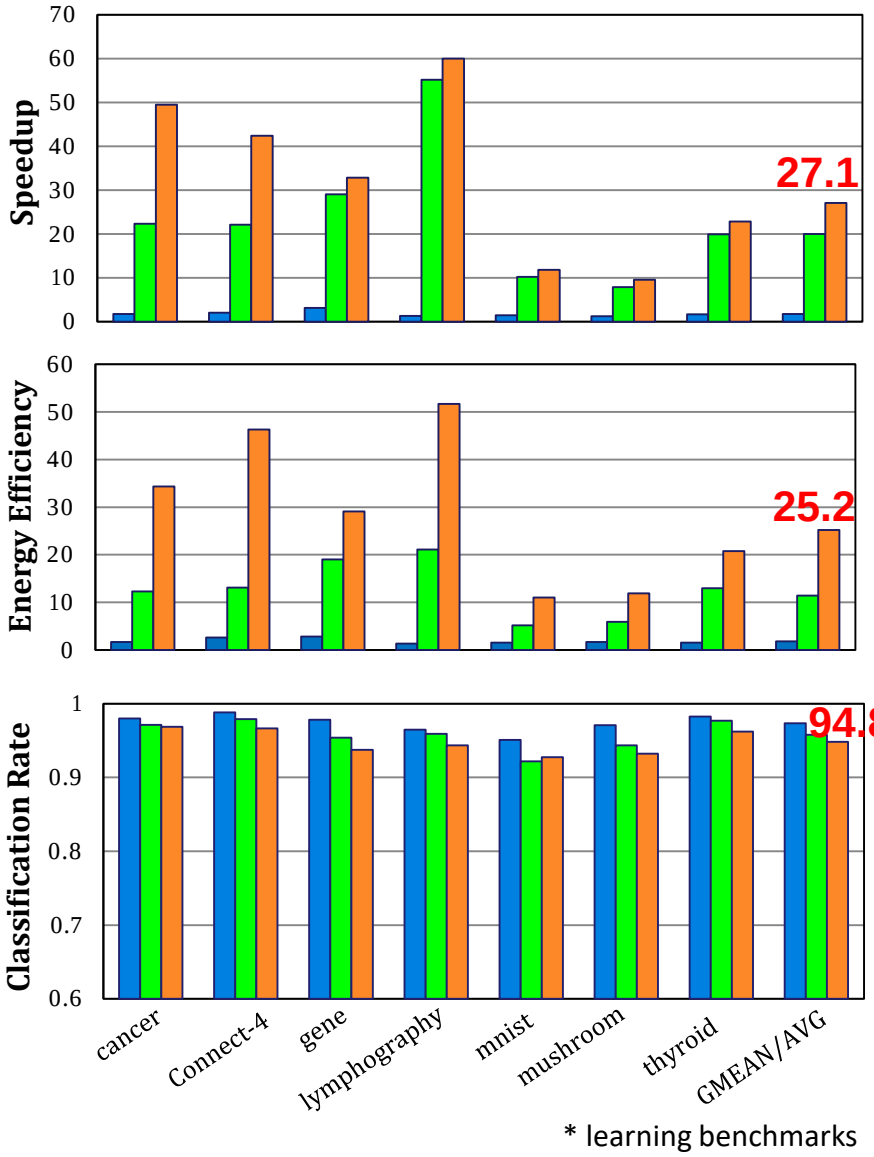
- Variation: Device & Signal
  - AAM (b) is much more robust compared to MLP (a)

(device variation, signal fluctuation)





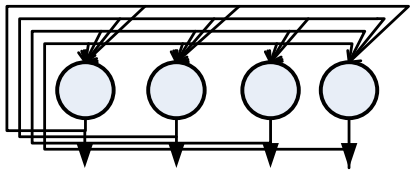
Comparison w/ Other Designs



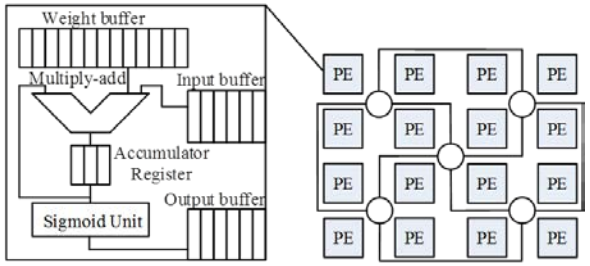
* learning benchmarks

Examples:

Auto-associative memory (AAM)



***Digital NPU + Digital NOC*[1]**



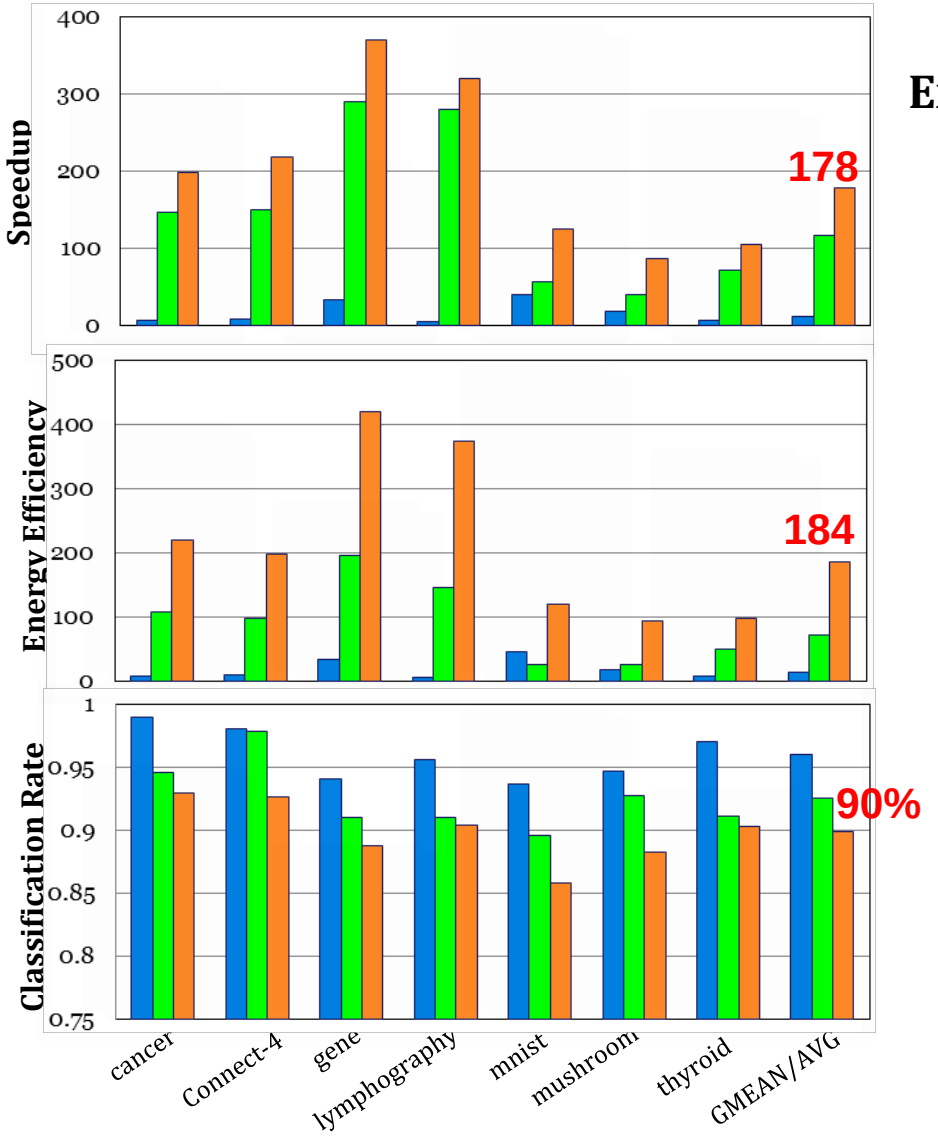
MBC + Digital NoC

RENO (MBC + Mixed-signal NoC)

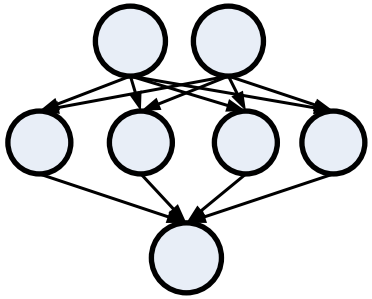
All the results are normalized to the baseline CPU.

[1] H. Esmaeilzadeh *et al.*, MICRO'12

Comparison w/ Other Designs



Example: Multilayer Perception (MLP)

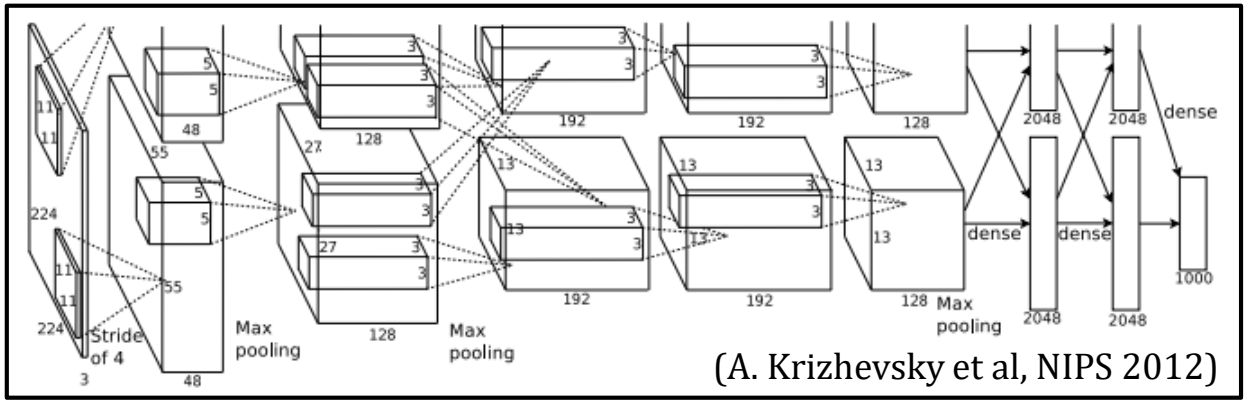


- ***Digital NPU + Digital NOC***[1]
- ***MBC + Digital NoC***
- ***RENO (MBC + Mixed-signal NoC)***

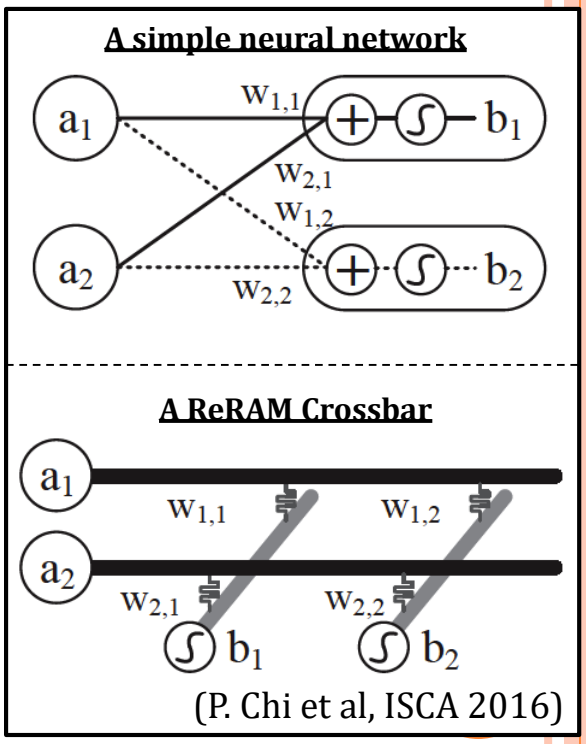
[1] H. Esmaeilzadeh *et al.*, MICRO'12

PipeLayer: Motivation

- **Convolutional Neural Networks (CNNs)**
  - Heart of deep learning
  - Computation and memory intensive
- **Resistive Random Access Memory (ReRAM) Based Acceleration**
  - Capability of combined computation and storage
  - Processing in memory to reduce data movement



(A. Krizhevsky et al, NIPS 2012)



(P. Chi et al, ISCA 2016)

PipeLayer: Motivation

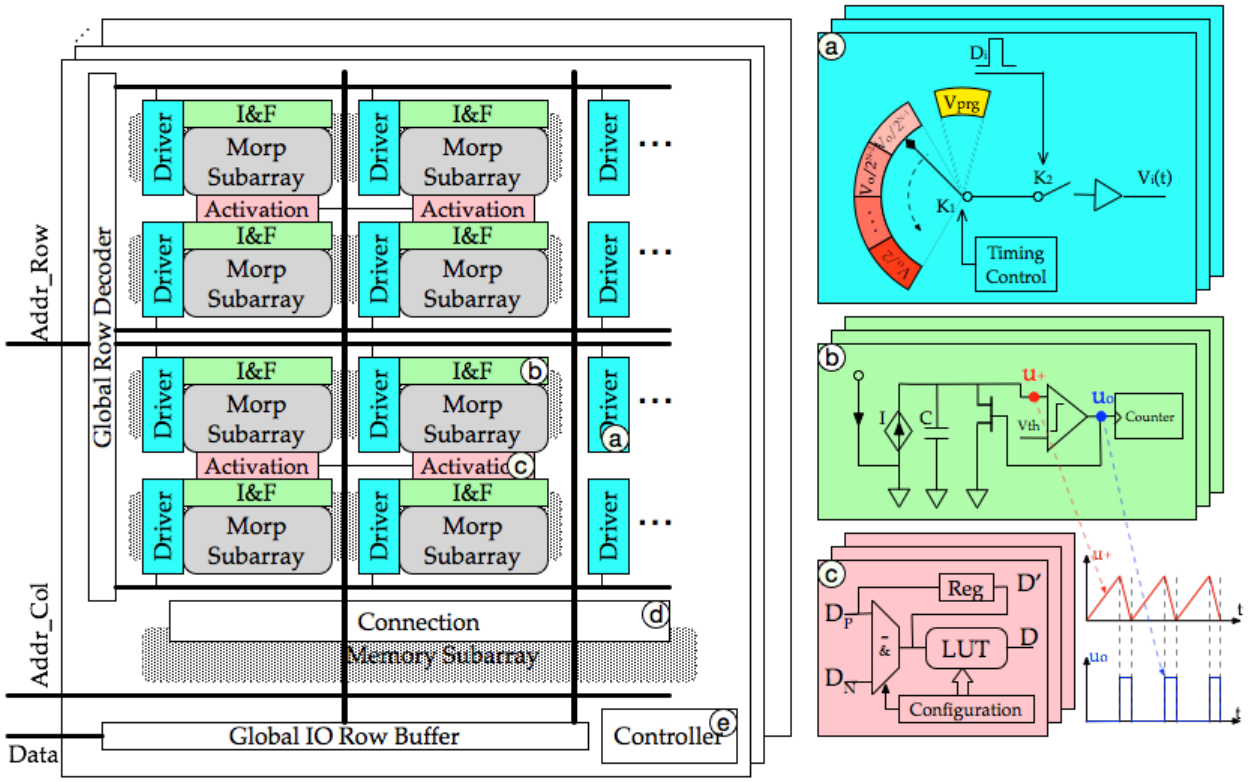
- **Limitations on Current ReRAM Based Approaches**
  - Do not support neural network training *#
  - Deep pipeline may introduce bubbles #
  - Kernel mapping was not clear *
  - Analog/digital converters (ADCs & DACs) overhead *

*P. Chi et al, "PRIME: A novel processing-in-memory architecture for neural network computation in ReRAM-based main memory," ISCA 2016

#A. Shafiee et al, "ISAAC: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars," ISCA 2016.

PipeLayer: Contributions of This Work

- Accelerating Both **Training** and **Testing**
- Intra- and Inter-layer **Pipeline** Design
- **Spike-based** Data Input and Output



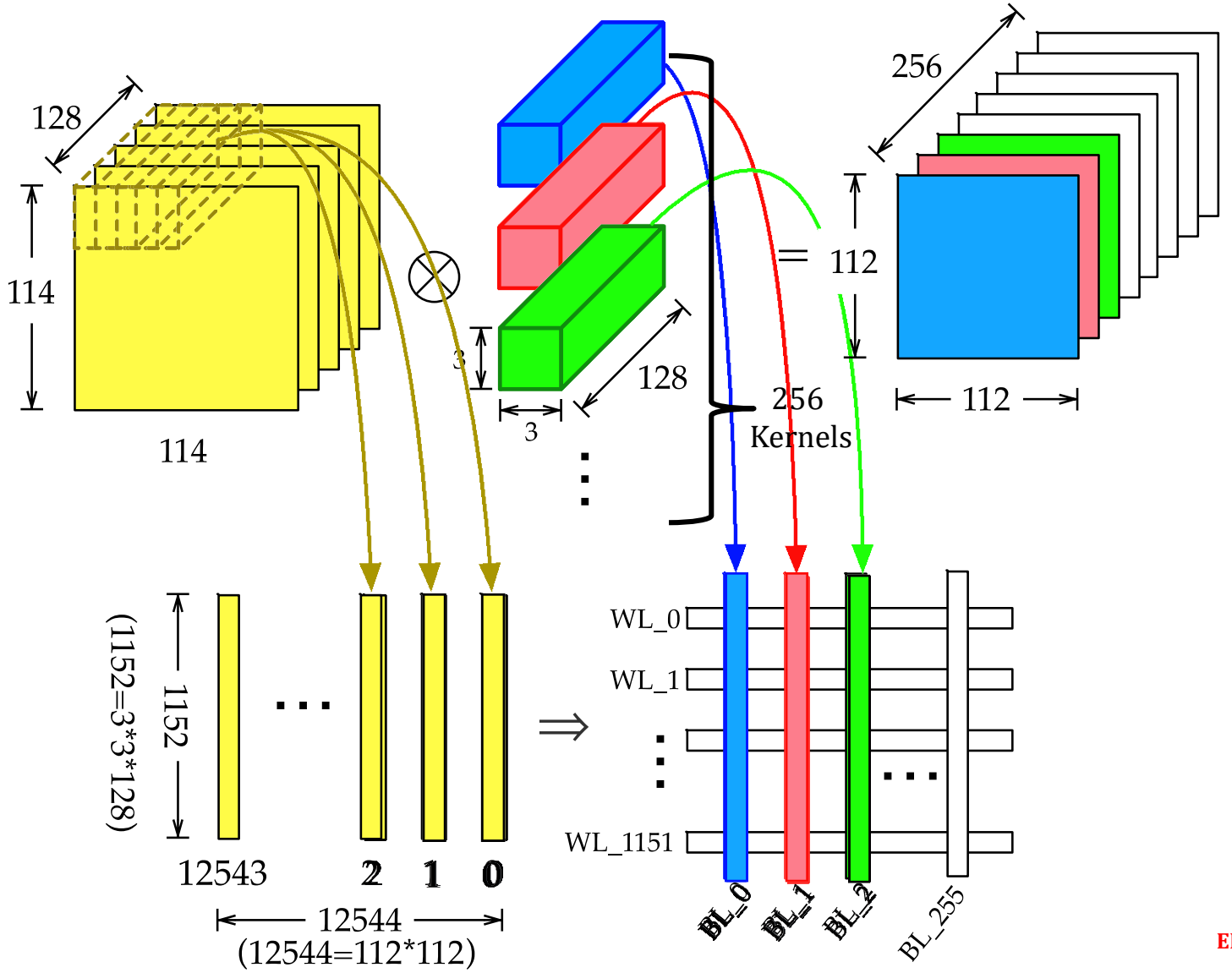
PipeLayer [HPCA, 2017]

A Pipelined ReRAM-Based Accelerator for Deep Learning

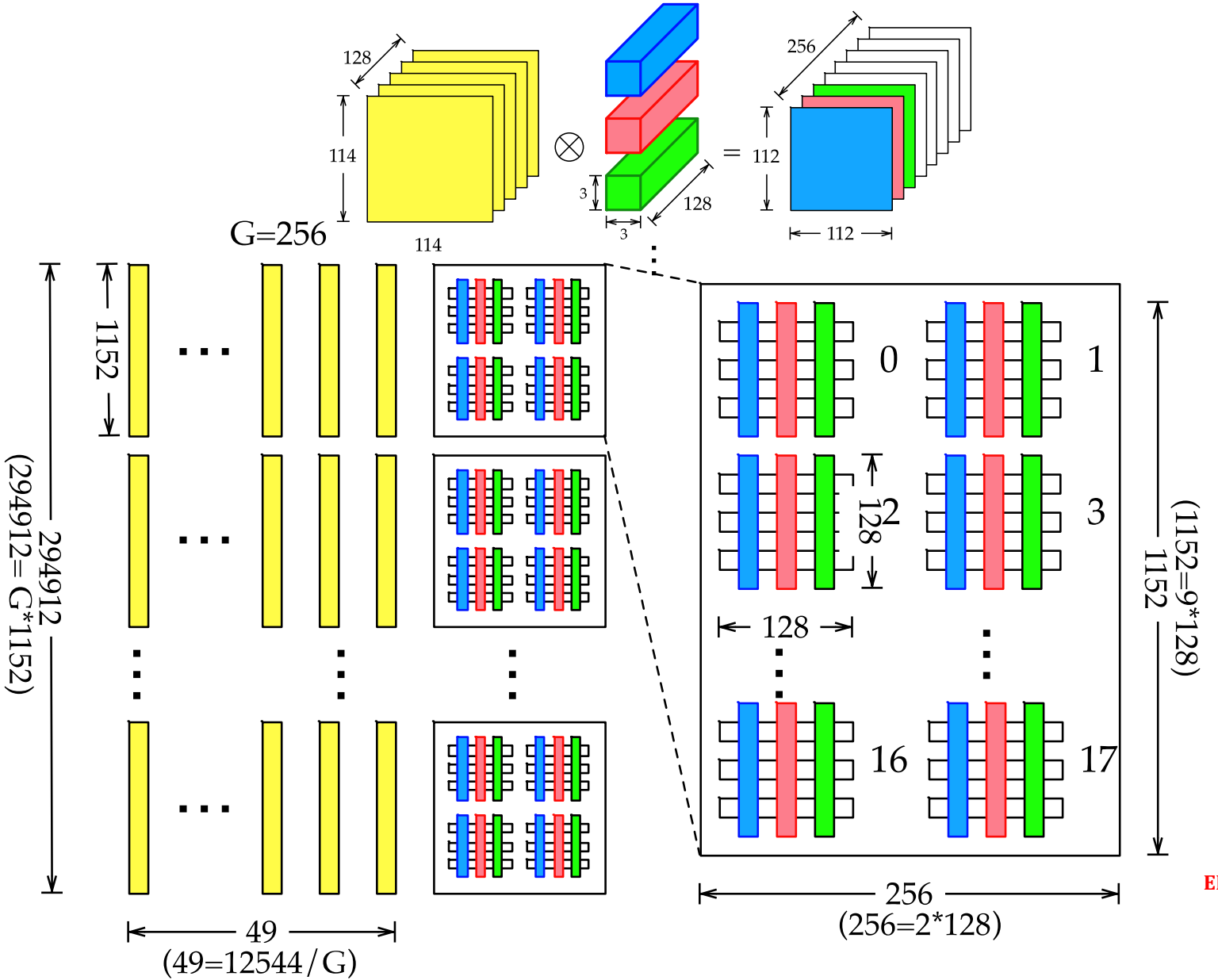
42.5x speedup

7.17x energy saving over GPU

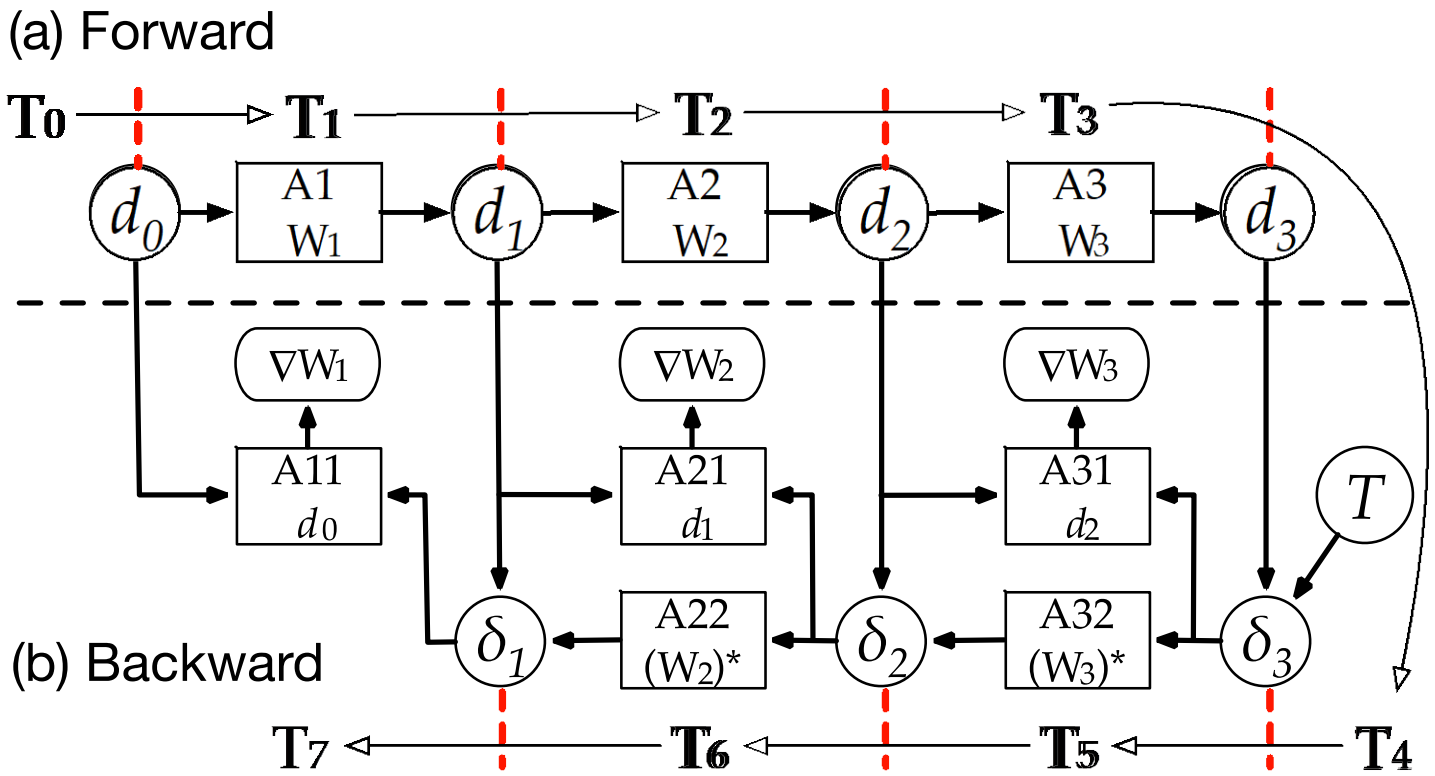
PipeLayer: Kernel Mapping



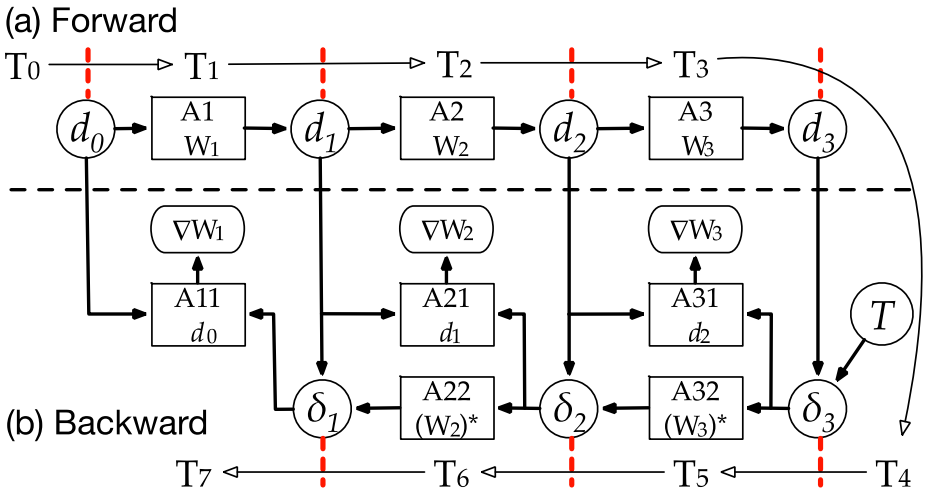
PipeLayer: Intra Layer Parallelism

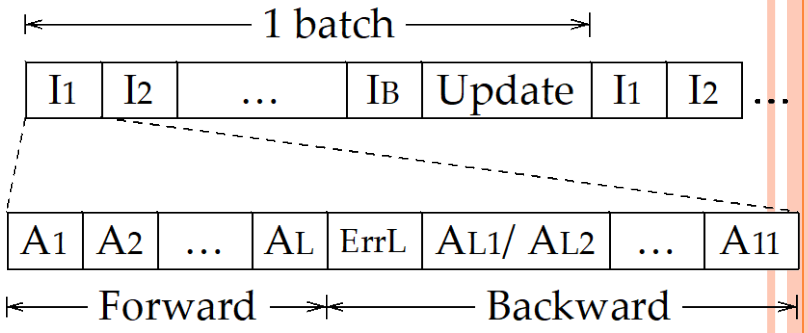


PipeLayer: Training Support

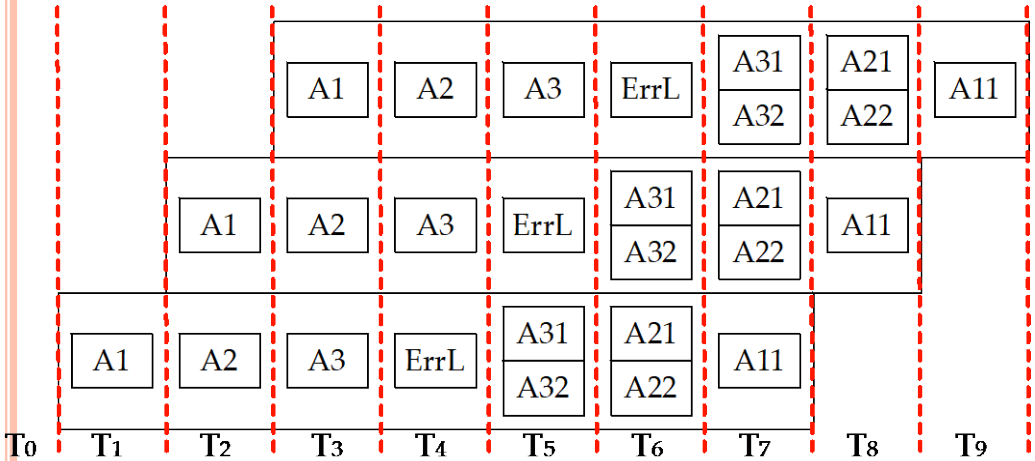


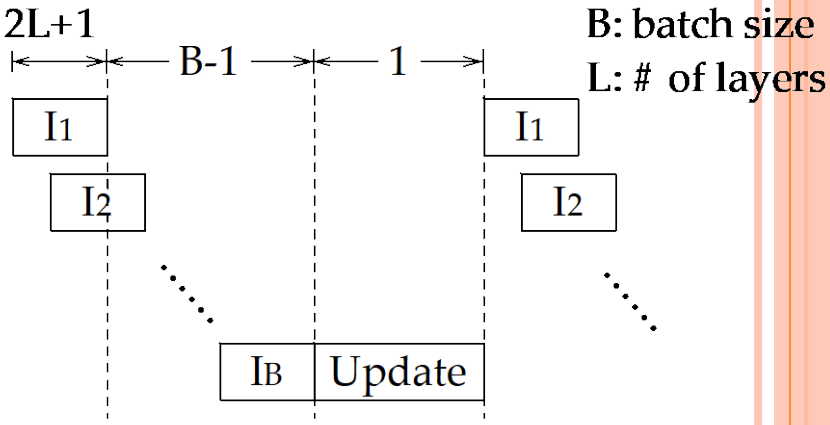
PipeLayer: Inter Layer Parallelism





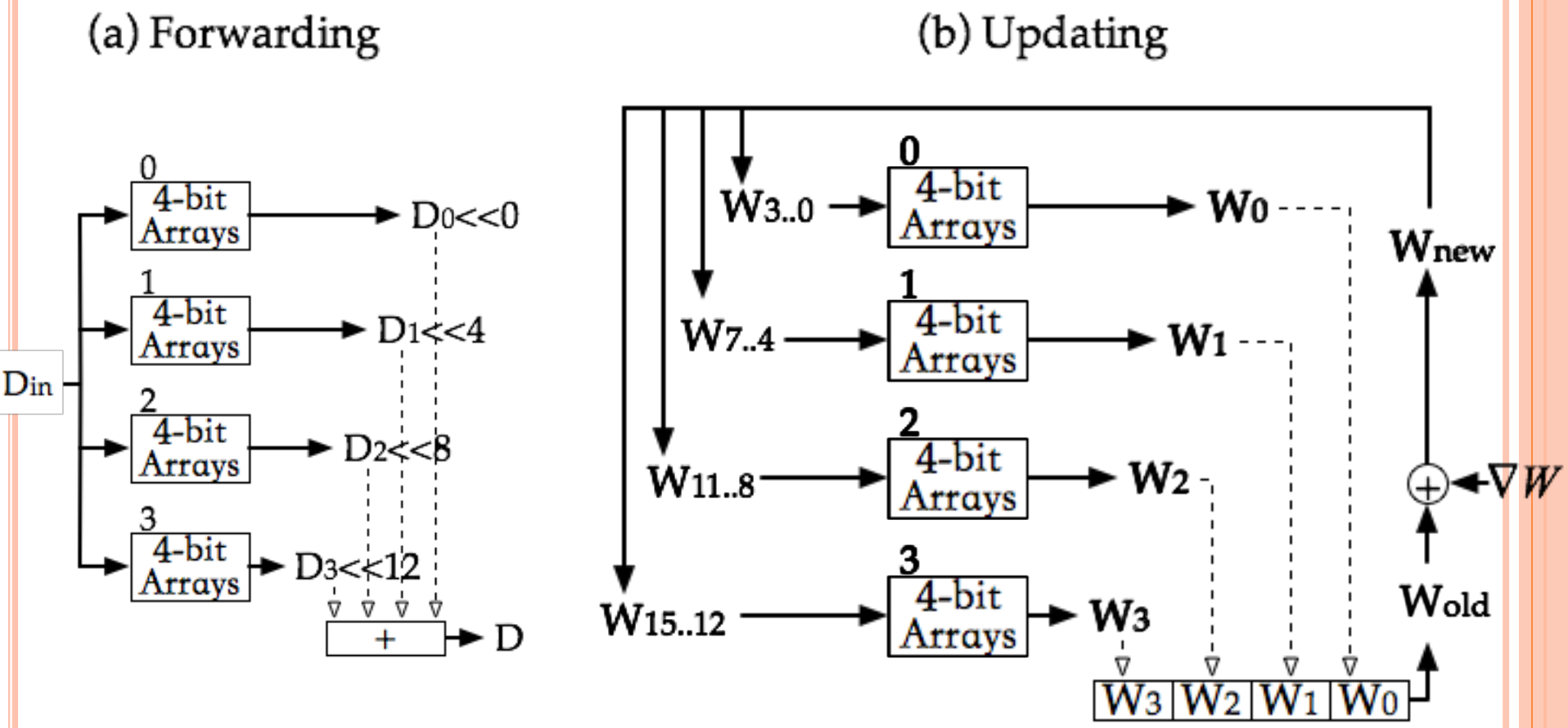
(a) Latency of PipeLayer without pipeline



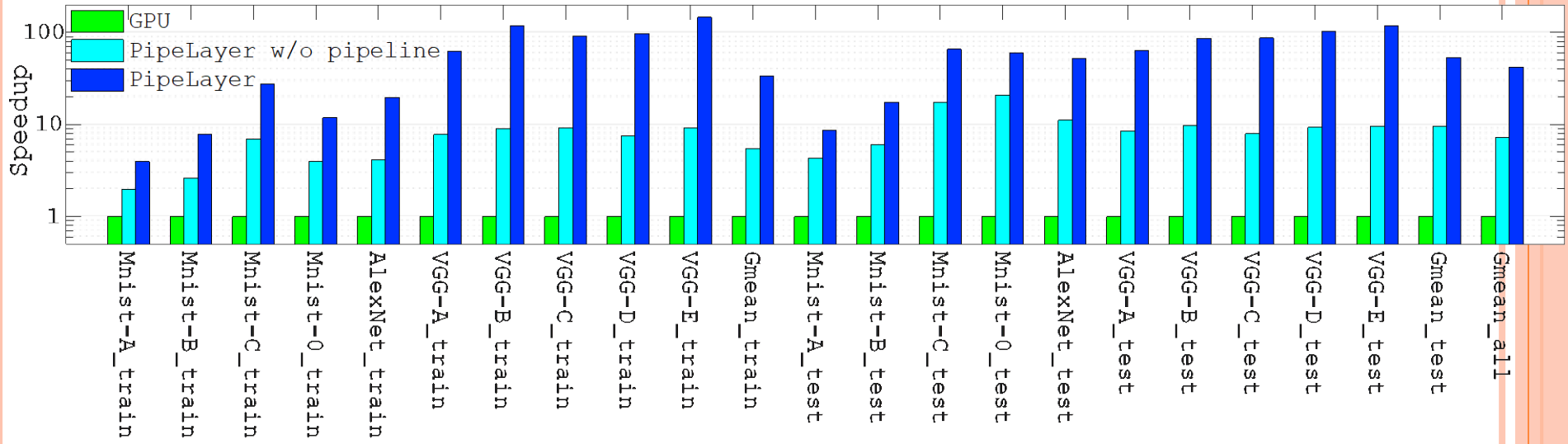


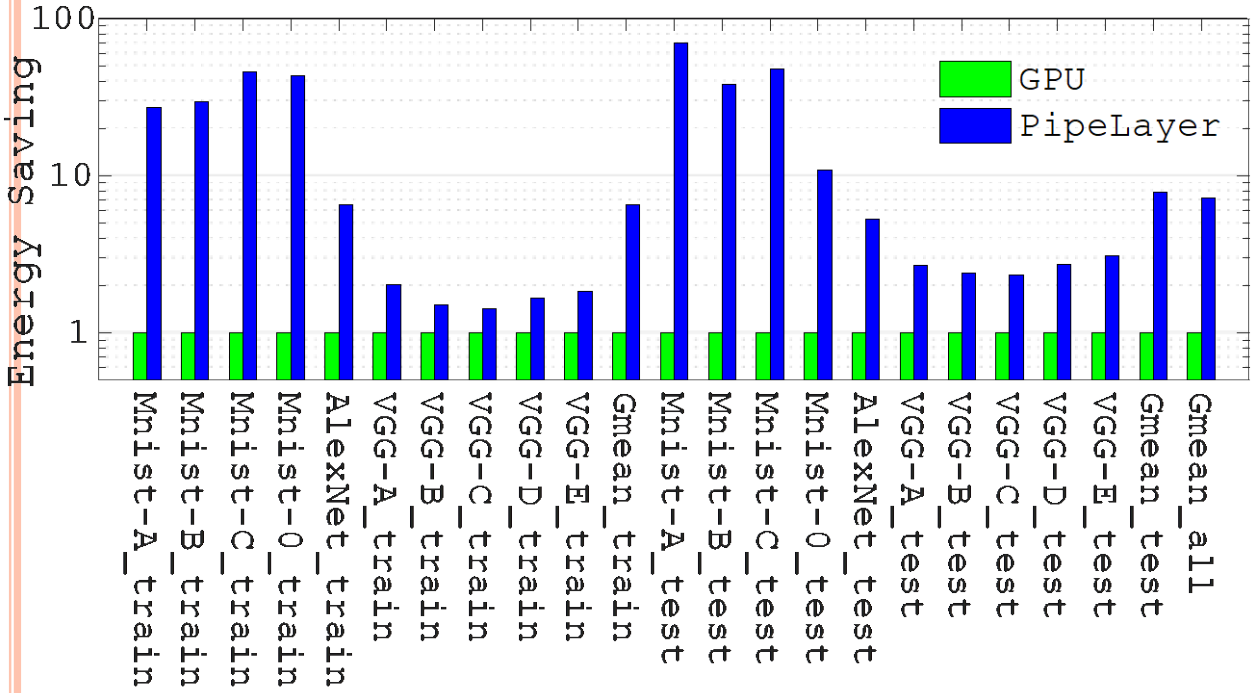
(b) Latency of PipeLayer with pipeline

PipeLayer: Resolution Compensation



PipeLayer: Speedup & Energy saving (vs GTX 1080)

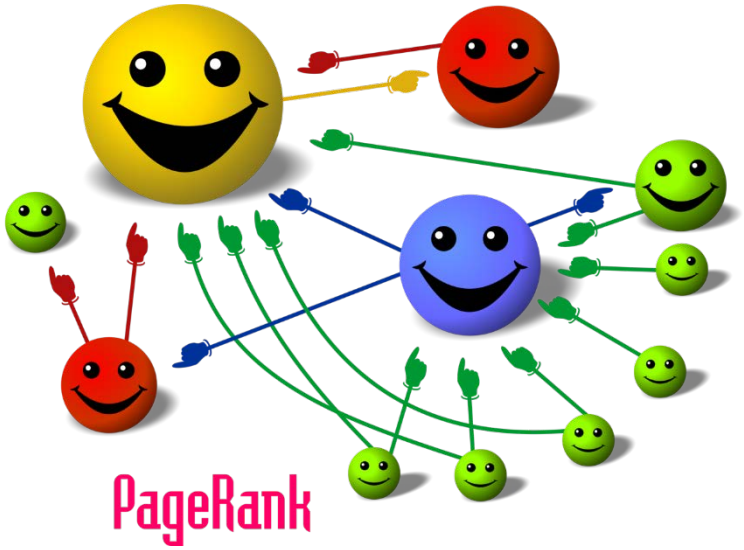




- **Speedup:**
  -Gmean: 42.5x
  -Max: 146.6x
- **Energy Saving:**
  -Gmean: 7.2x
  -Max: 70.0x

Graph Processing

- Applications: PageRank, Breadth First Search, etc.

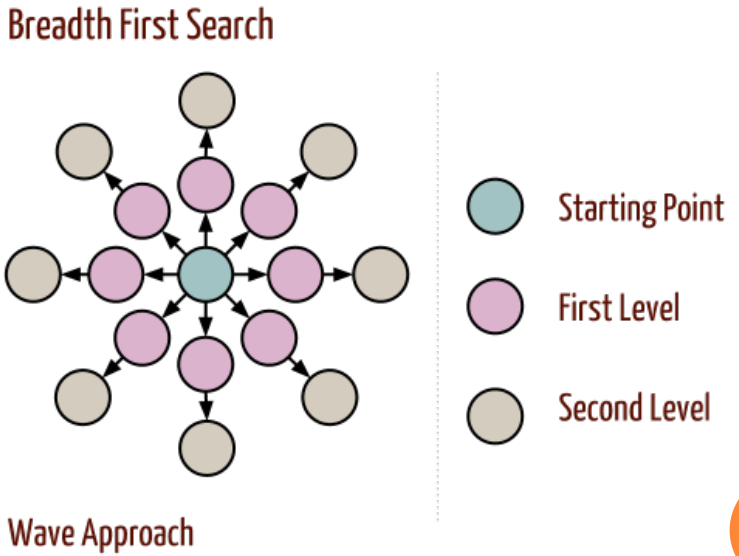


PageRank

website page importance measuring

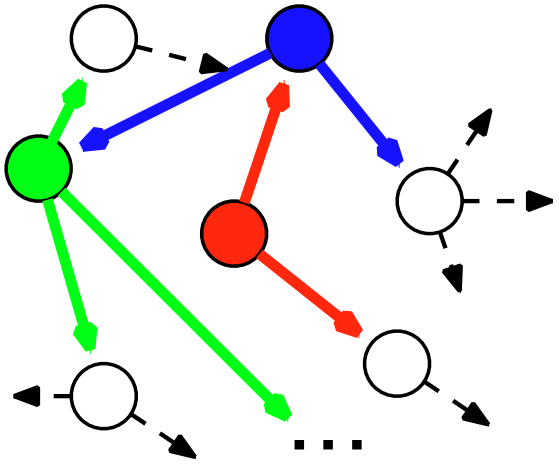
BFS

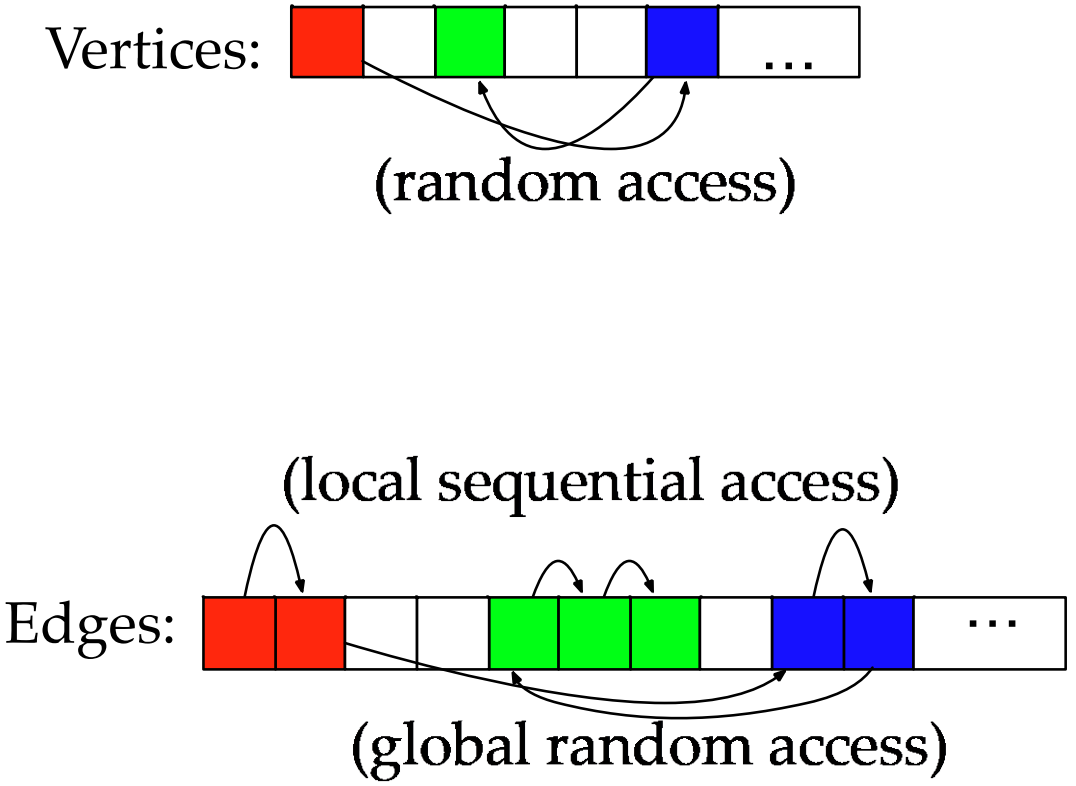
maximum flow problem, bipartiteness testing



Graph Programming Models

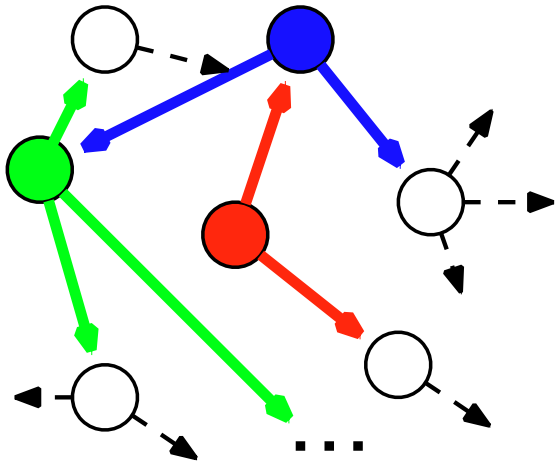
- Vertex-centric programing

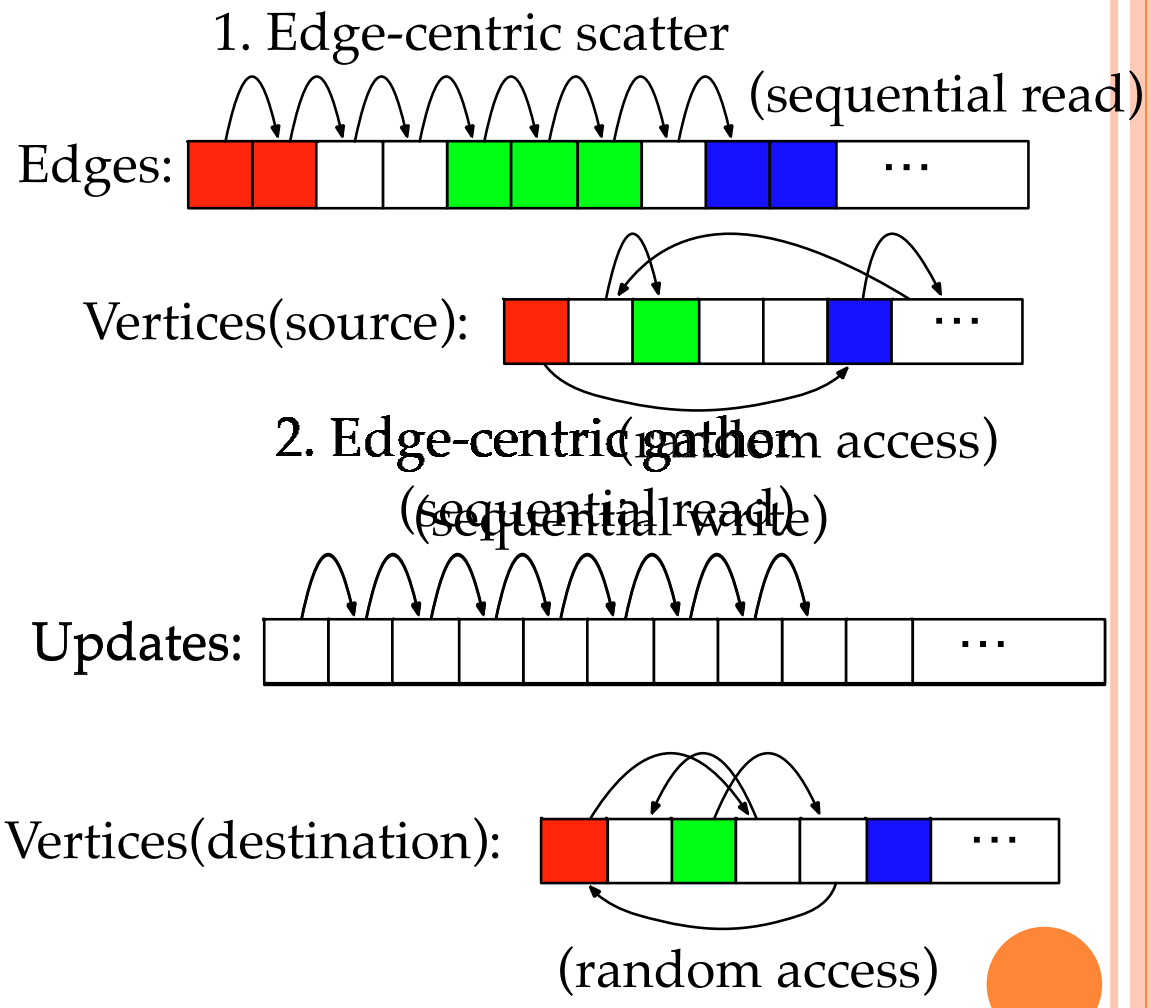




Graph Programming Models

- Edge-centric programing





GraphR: Graph Representation

	0	1	2	3
0	0	0	3	8
1	0	0	7	0
2	1	0	0	0
3	0	4	0	2

- Compressed Sparse Column (CSC)
- Compressed Sparse Row (CSR)
- Coordinate List(COO)

CSC

(row,val)		colptr	
(2,1)	0	0	0
(3,4)	1	1	1
(0,3)	2	2	2
(1,7)	3	4	3
(0,8)	4	6	
(3,2)	5		

CSR

(col,val)		rowptr	
(2,3)	0	0	0
(3,8)	1	2	1
(2,7)	2	3	2
(0,1)	3	4	3
(1,4)	4	6	
(3,2)	5		

COO

(row,col,val)
(0,2,3)
(0,3,8)
(1,2,7)
(2,0,1)
(3,1,4)
(3,3,2)

GraphR: Where is it?

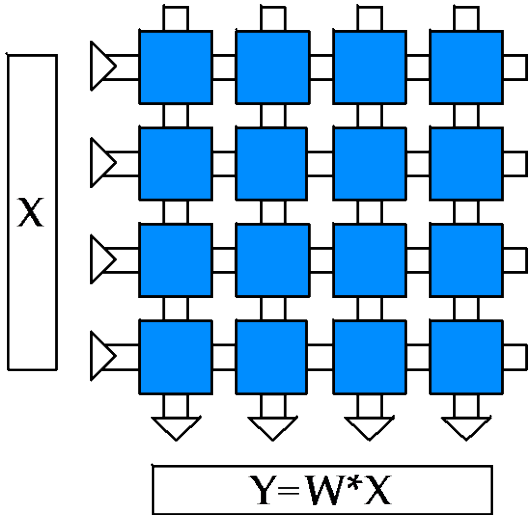
Compressed Form

(row,col,val)

(0,2,3)
(0,3,8)
(1,2,7)
(2,0,1)
(3,1,4)
(3,3,2)

GraphR: I'm Here!

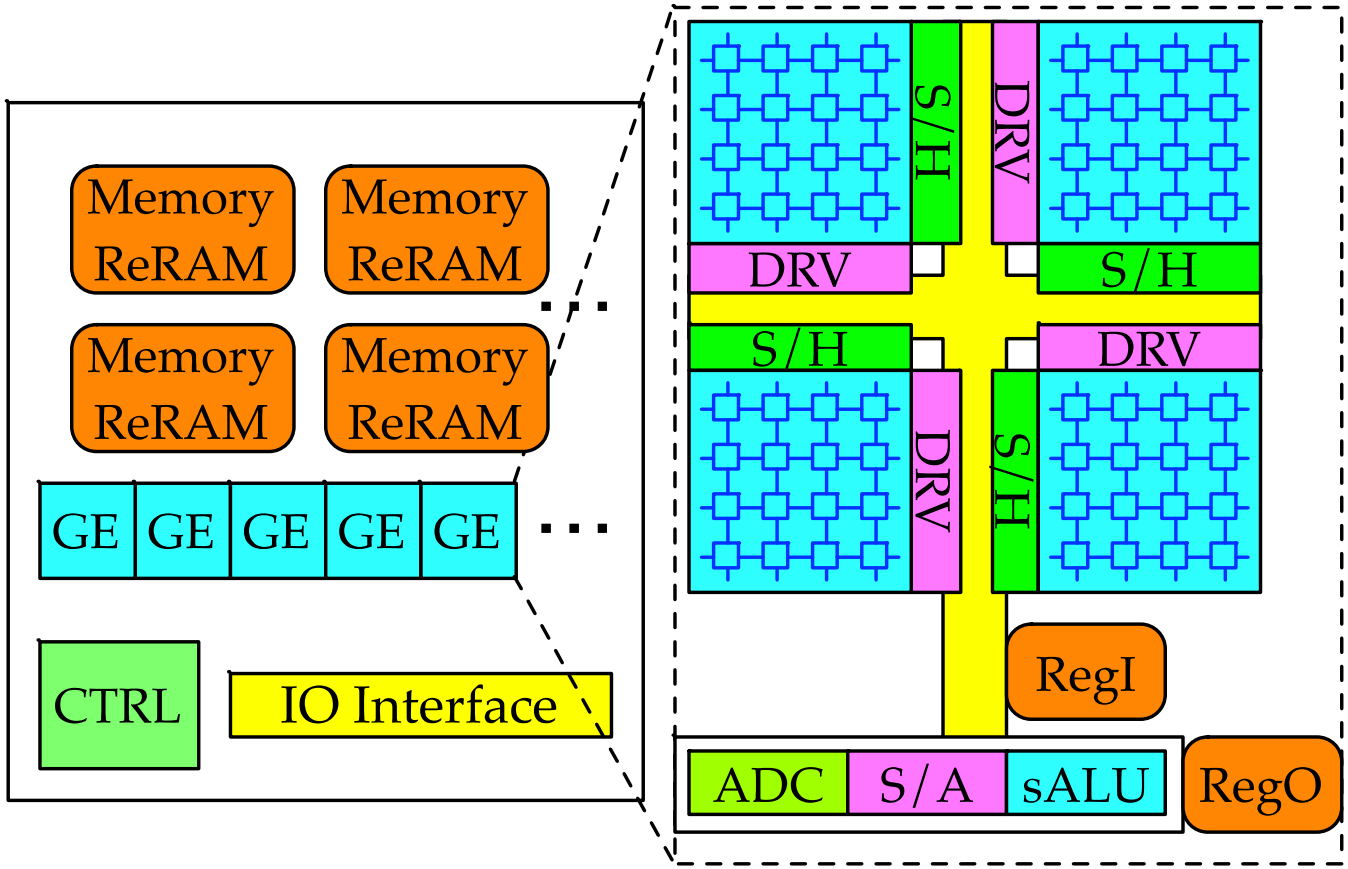
Matrix-vector Multiplication (MVM) in ReRAM



Storage Efficiency

Computation Efficiency

GraphR Architecture



CTRL Controller

S/H Sample & hold

S/A Shift & add unit

DRV Driver

GE Graph Engine

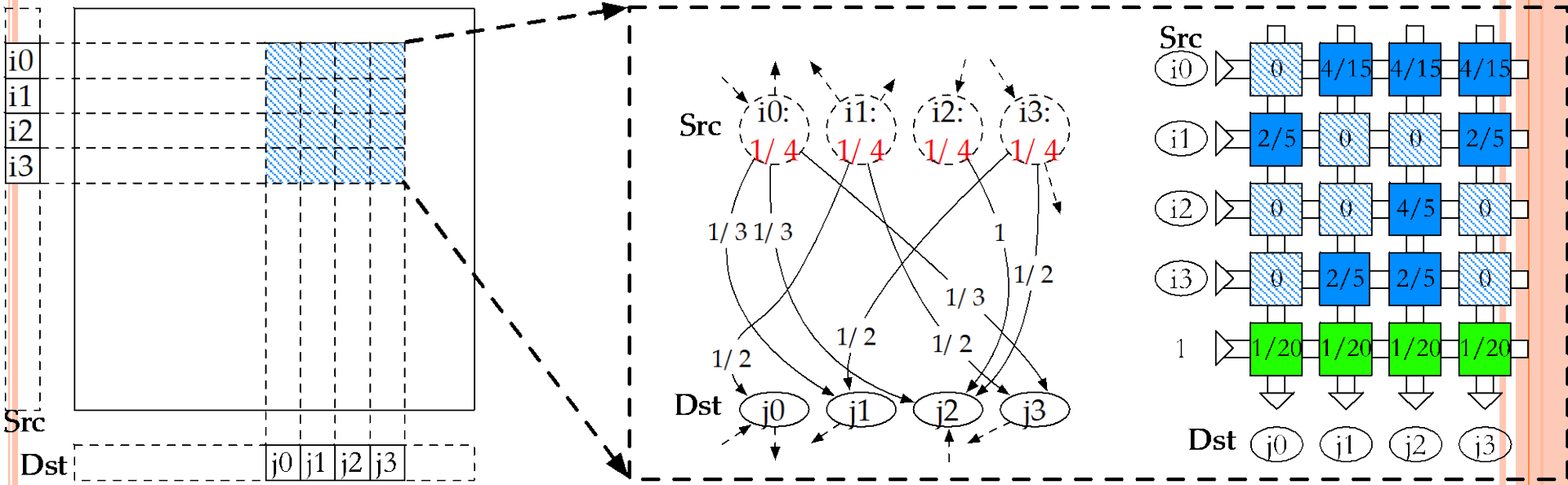
ADC Analog to digital

sALU Simple ALU

RegI Input register

RegO Output register

GraphR: Parallel Processing (PageRank)

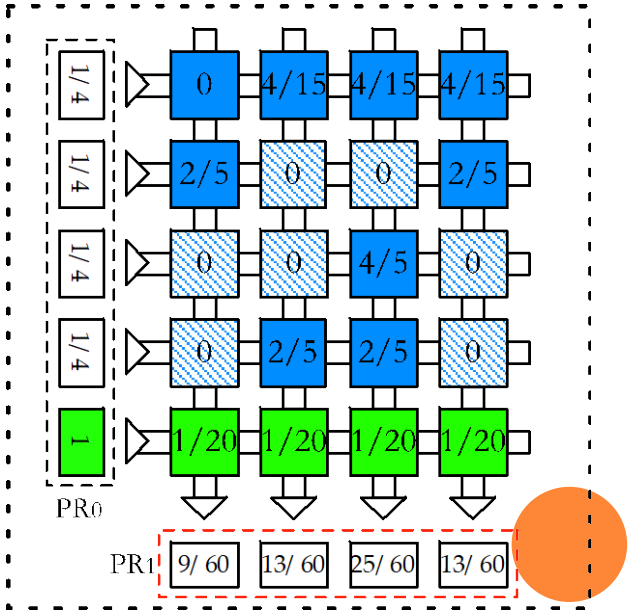


$$PR_1 = W\ PR_0 + e$$

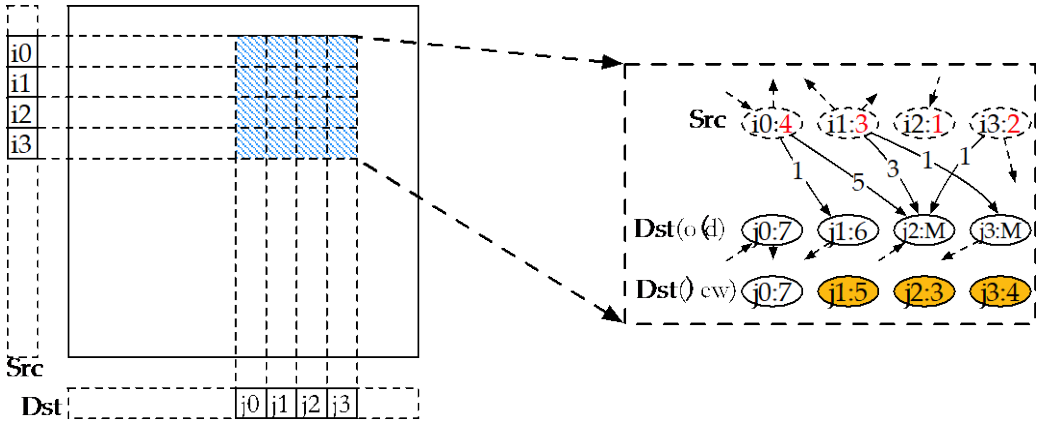
$$W = \begin{bmatrix} 0 & 4/15 & 4/15 & 4/15 \\ 2/5 & 0 & 0 & 2/5 \\ 0 & 0 & 4/5 & 0 \\ 0 & 2/5 & 2/5 & 0 \end{bmatrix}^T$$

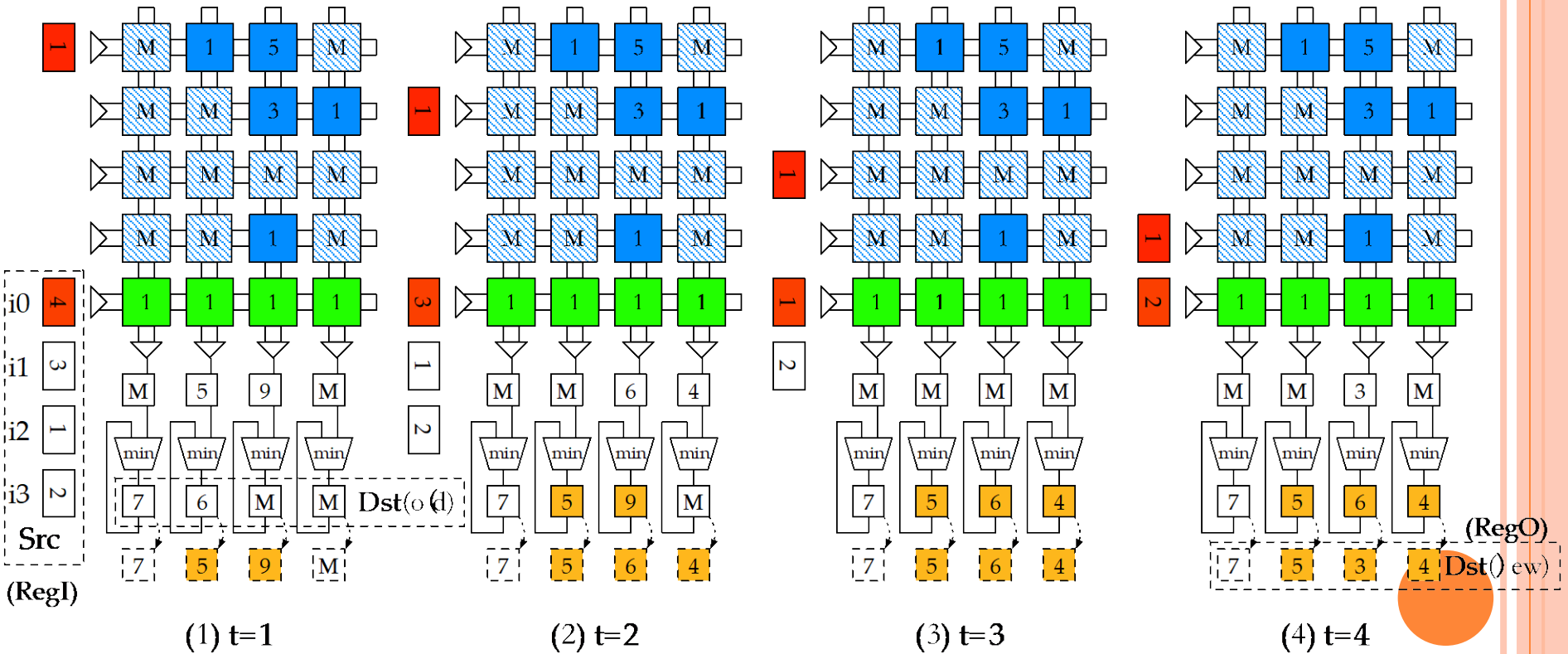
$$PR_0 = [1/4\ \ 1/4\ \ 1/4\ \ 1/4]^T$$

$$e = [1/20\ \ 1/20\ \ 1/20\ \ 1/20]^T$$

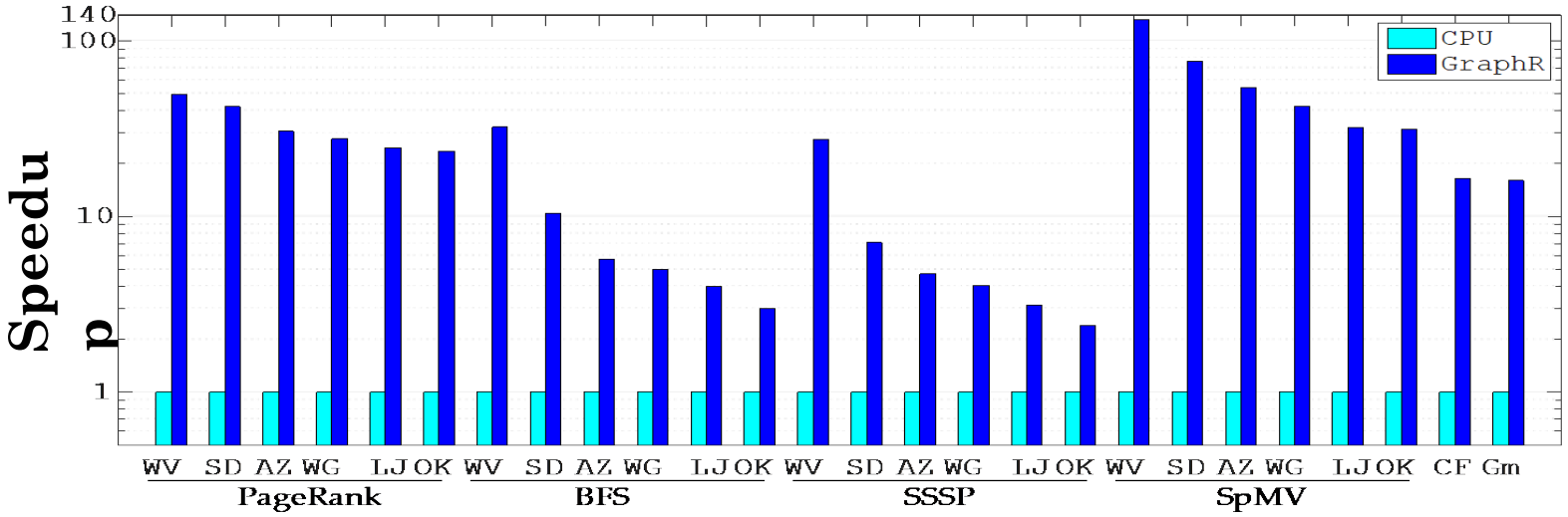


GraphR: Serial Processing (SSSP)

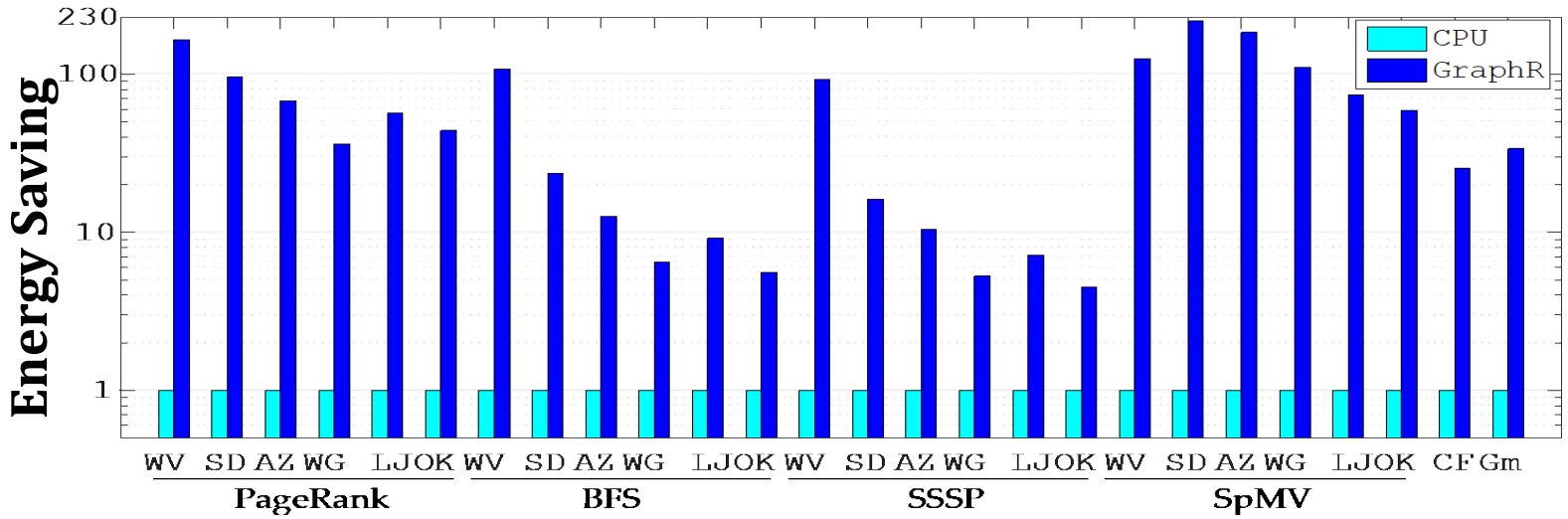




GraphR: Speedup and Energy Saving



16.01x speedup over CPU



33.82x energy saving over CPU