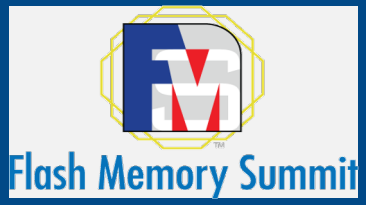


pLUTo

Enabling Massively Parallel Computation in DRAM via Lookup Tables

João Dinis Ferreira Gabriel Falcao **Juan Gómez-Luna** Mohammed Alser
Lois Orosa Mohammad Sadrosadati Jeremie S. Kim Geraldo F. Oliveira
Taha Shahroodi Anant Nori Onur Mutlu

SAFARI

ETH zürich

TU Delft

instituto de telecomunicações

CESGA
GALICIA SUPERCOMPUTING CENTER

intel

Summary

Background: Processing-in-Memory (PiM) alleviates the performance and energy bottlenecks caused by data movement in modern applications

- *Processing-near-Memory (PnM): adds* logic elements near memory arrays
- *Processing-using-Memory (PuM): uses* the analog properties of memory for computation

Problem: Existing Processing-using-DRAM architectures *only support* a limited range of operations (data movement, bitwise logic, bit shifting)

- This *limits* Processing-using-DRAM's applicability to a narrow set of applications

Goal: Extend the applicability of Processing-using-DRAM by designing a PuM substrate with support for complex operations

pLUTo: A Processing-using-DRAM substrate that replaces complex operations with equivalent memory lookups

- *pLUTo LUT Query* operation enables bulk in-memory table lookups
- *Three pLUTo designs* target different performance/energy/area tradeoffs
- *pLUTo API* and *pLUTo Compiler* facilitate programmer adoption

Key Results: Our extensive evaluation shows that pLUTo

- *Greatly* outperforms CPU/GPU/PnM baselines, both in performance and energy
- Incurs *small* DRAM area overheads (between 10.2% and 23.1%)

https://github.com/CMU-SAFARI/pLUTo

Contents

Introduction

pLUTo

- Overview
- pLUTo Designs
- System Integration

Evaluation

Conclusion

Contents

Introduction

pLUTo

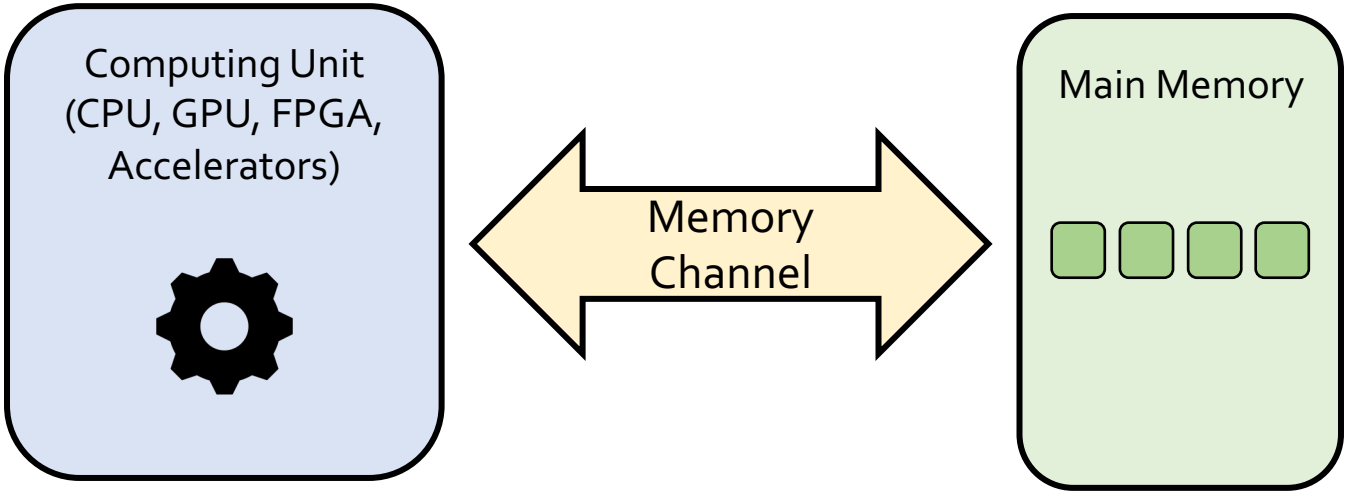
- Overview
- pLUTo Designs
- System Integration

Evaluation

Conclusion

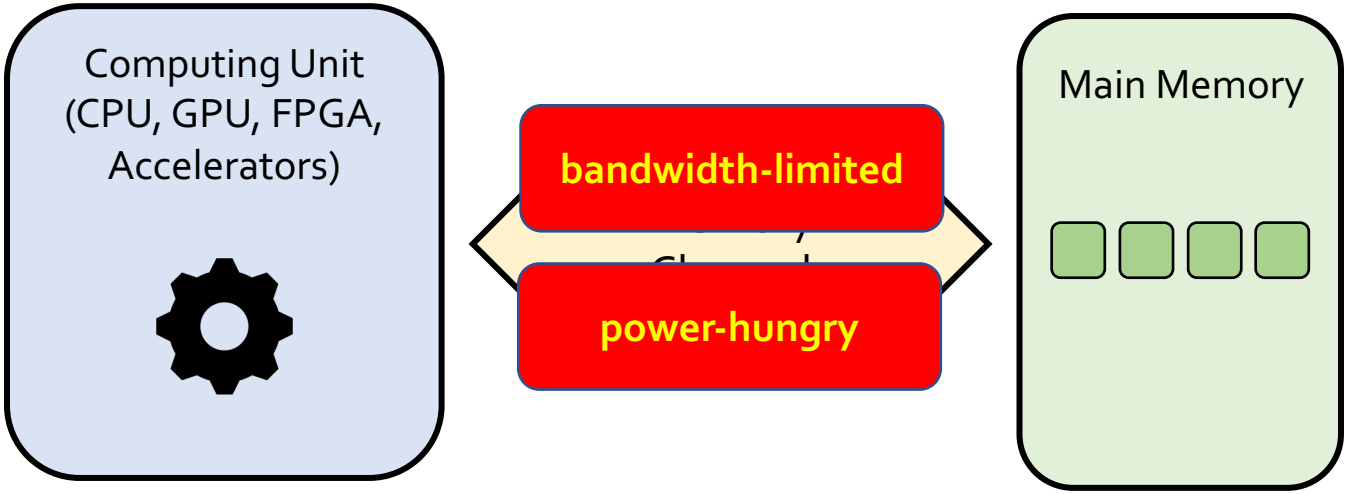
Data Movement Bottleneck

Data movement is a major bottleneck
in modern computer architectures



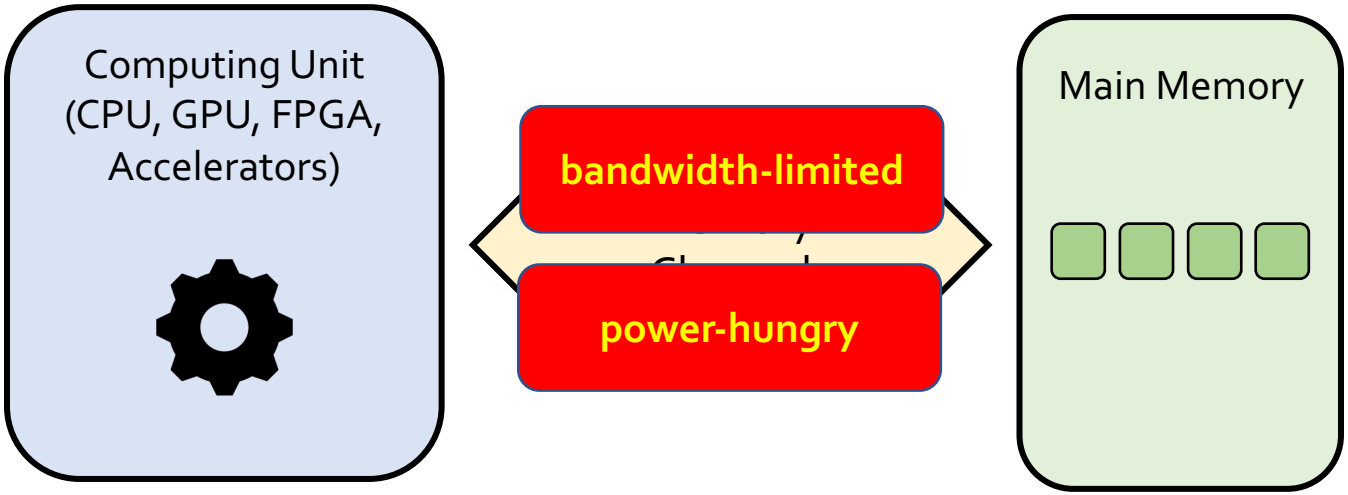
Data Movement Bottleneck

Data movement is a major bottleneck
in modern computer architectures



Data Movement Bottleneck

Data movement is a major bottleneck
in modern computer architectures

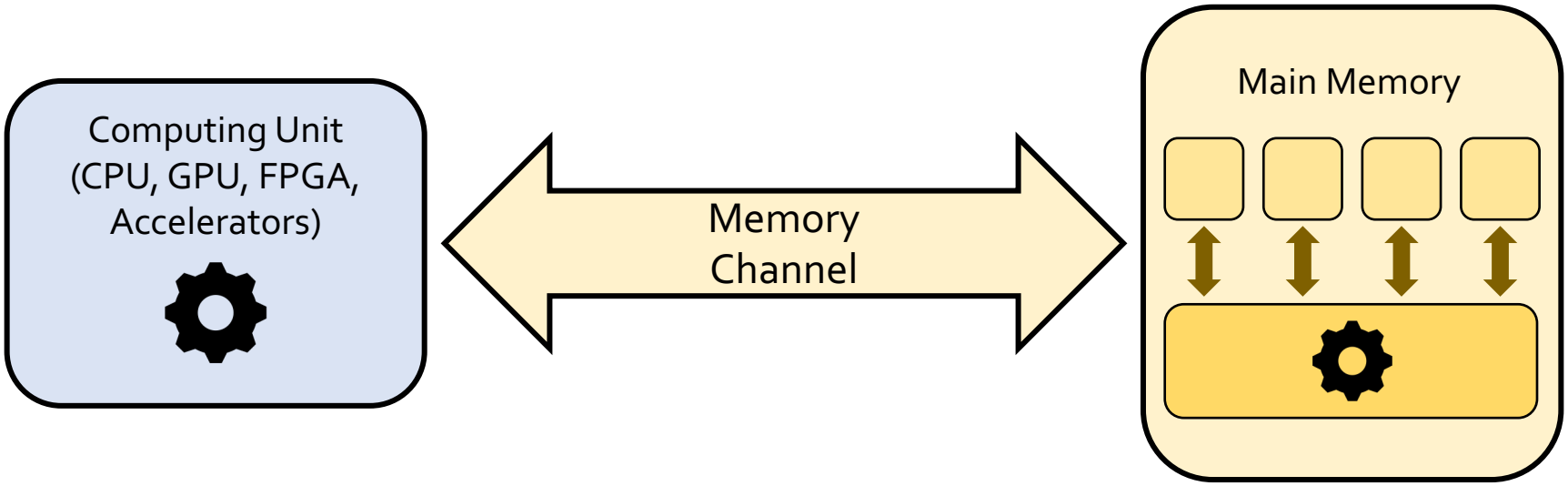


Over 60% of the total system energy is spent on data movement[1]

[1] A. Boroumand et al., "Google Workloads for Consumer Devices: Mitigating Data Movement Bottlenecks," ASPLOS, 2018

Processing-in-Memory

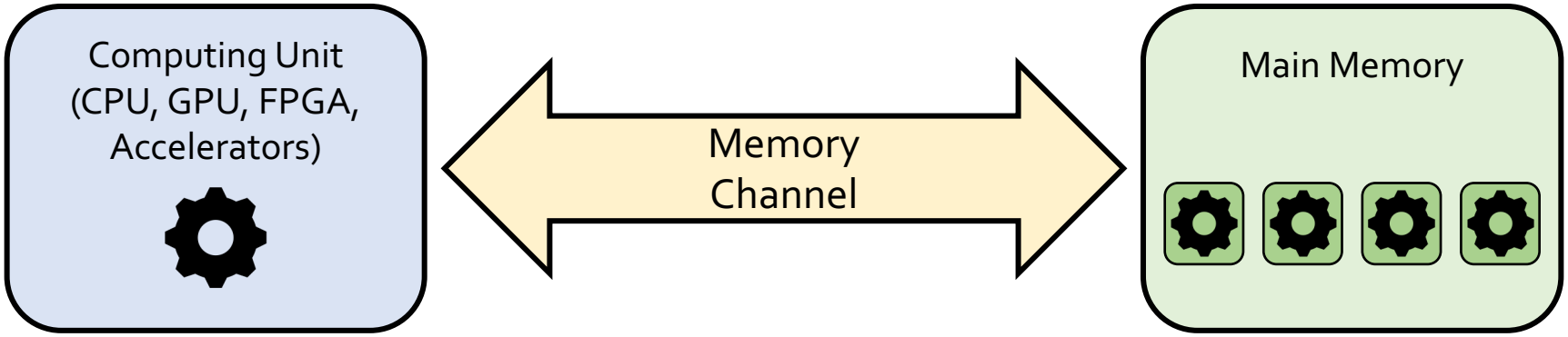
1. Processing-near-Memory



Processing-near-Memory leverages additional logic placed on the same die as memory or on the logic layer of 3D-stacked memory

Processing-in-Memory

2. Processing-using-Memory



Processing-using-Memory leverages the operational principles of memory to perform computation

Limitations of Processing-using-DRAM

Data Movement	*RowClone, Seshadri+ 2013* *LISA, Chang+ 2013*
Bitwise Operations	*Ambit, Seshadri+ 2017*
Bit Shifting	*DRISA, Li+ 2017*
Arithmetic Operations	*SIMDRAM, Hajinazar & Oliveira+ 2021*

Existing Processing-using-DRAM architectures only support a limited range of operations

The Goal of pLUTo

Extend **Processing-using-DRAM to support the execution of** ***arbitrarily complex operations***

Contents

Introduction

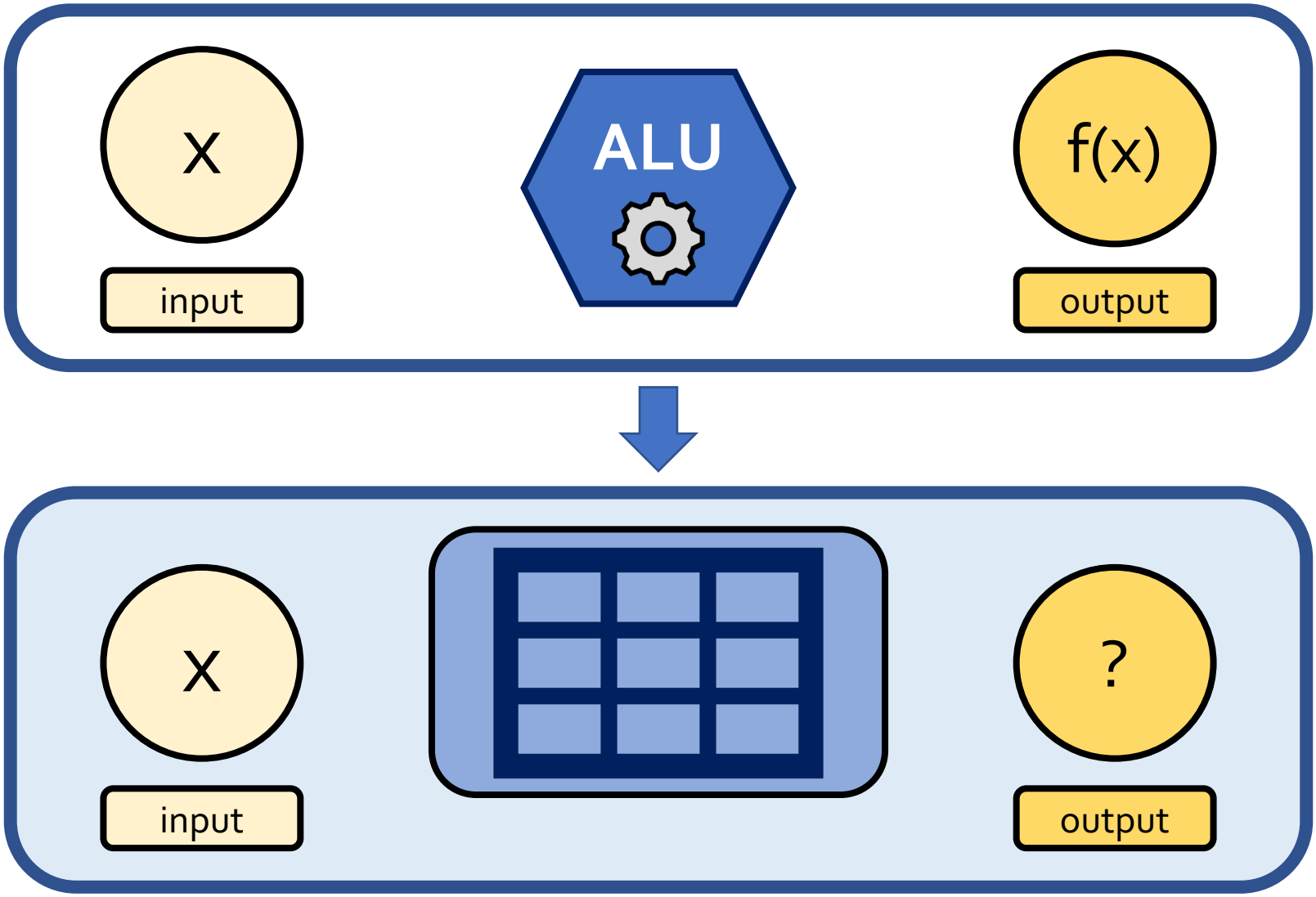
pLUTo

- Overview
- pLUTo Designs
- System Integration

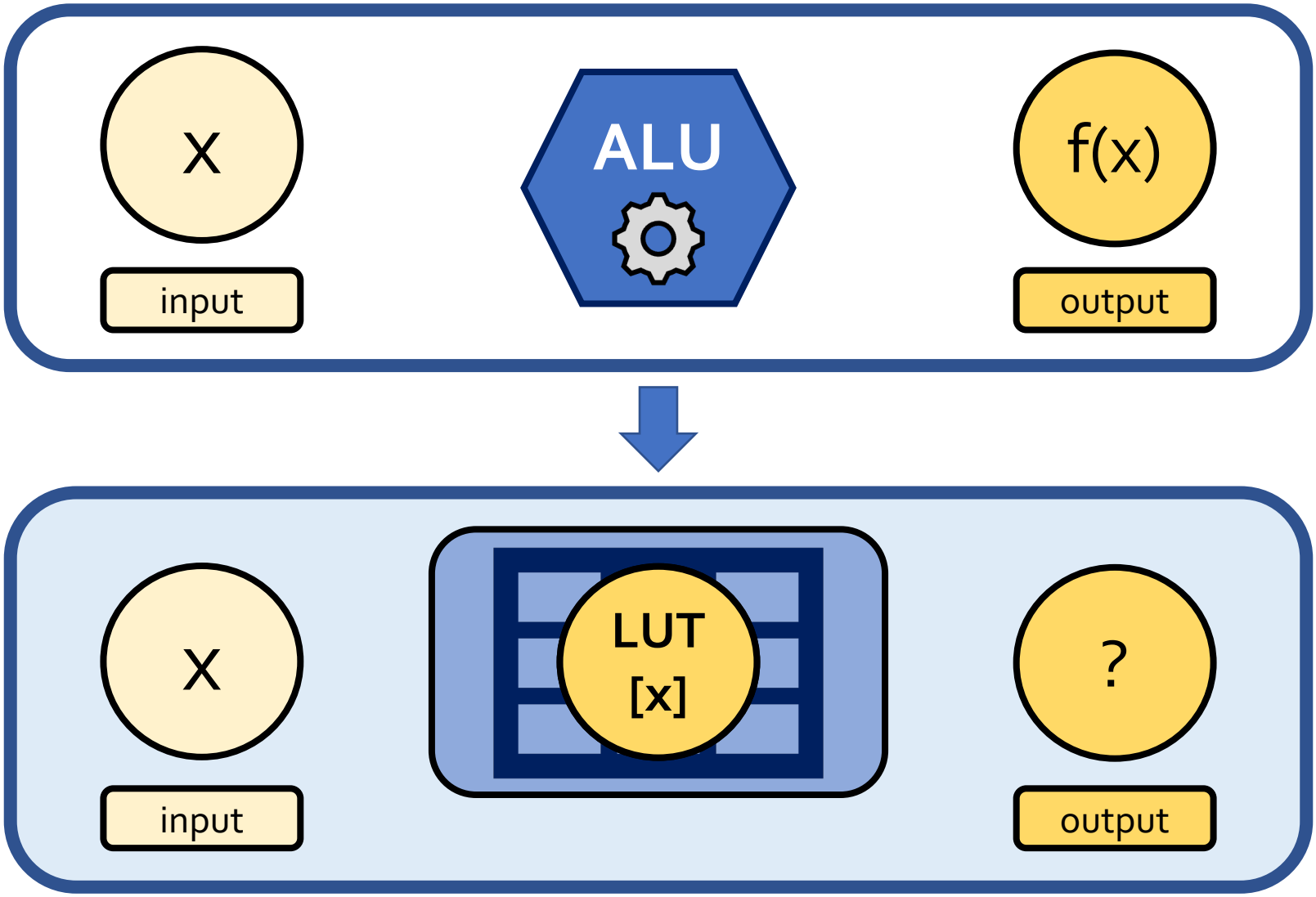
Evaluation

Conclusion

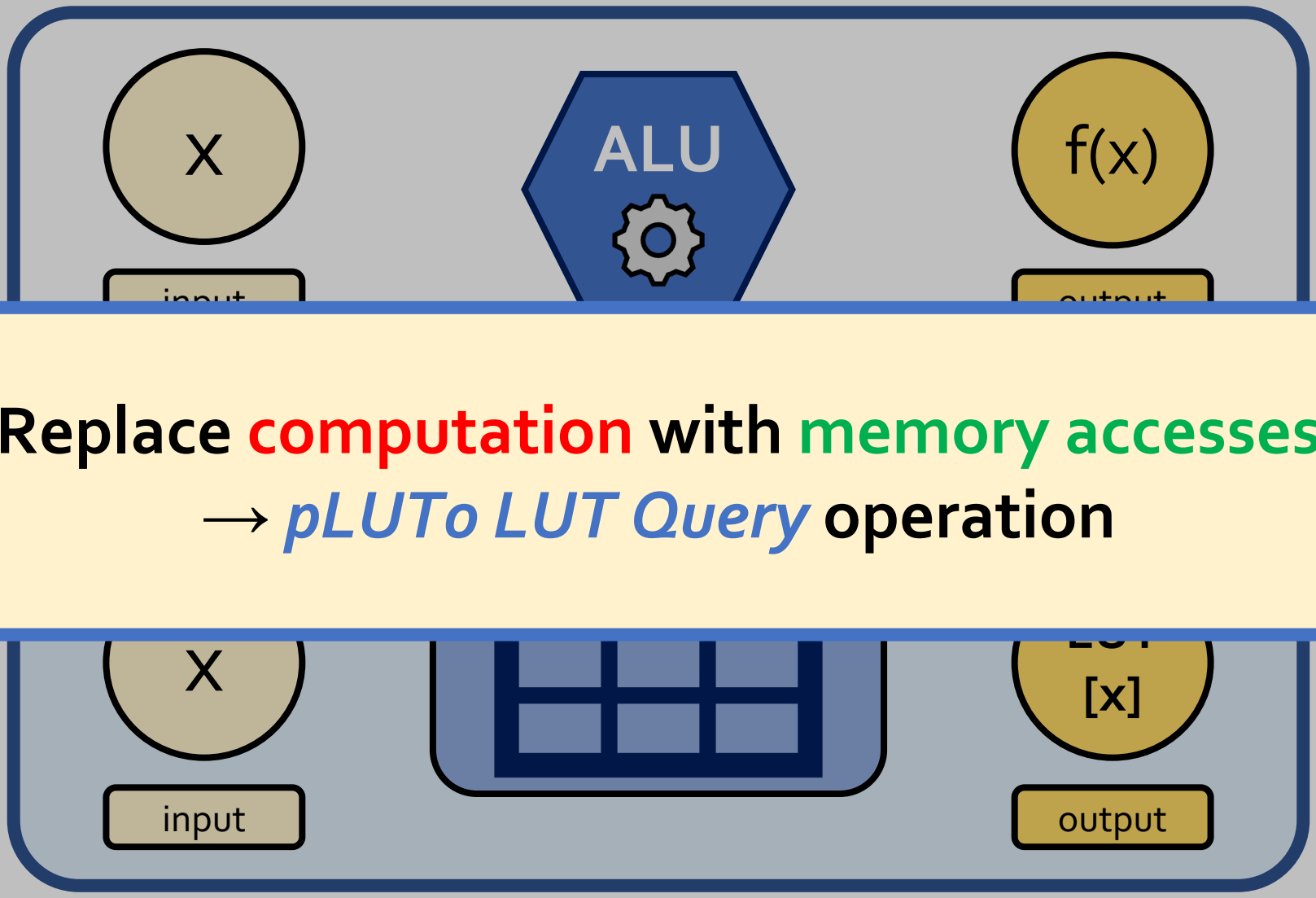
pLUTo: Key Idea



pLUTo: Key Idea

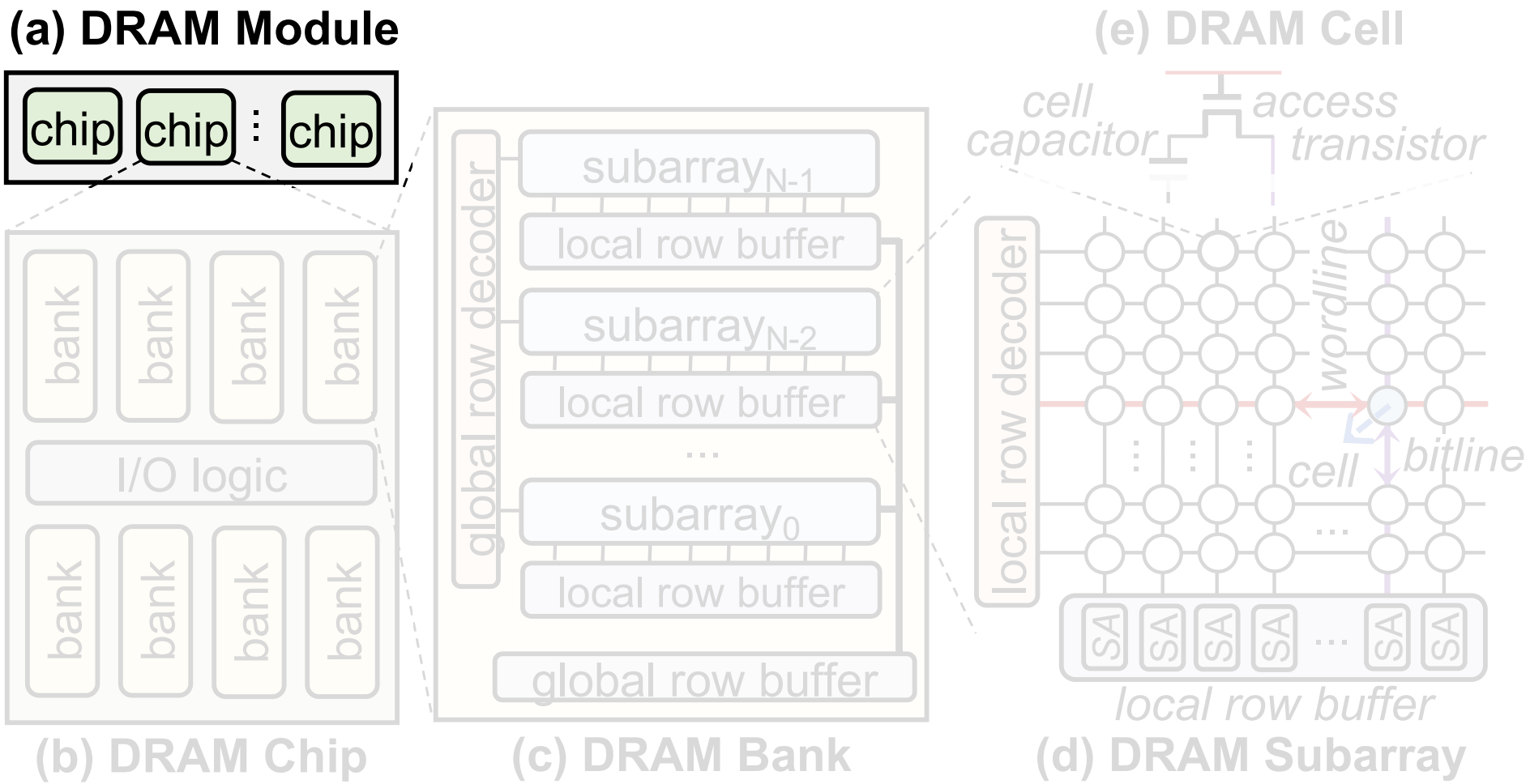


pLUTo: Key Idea

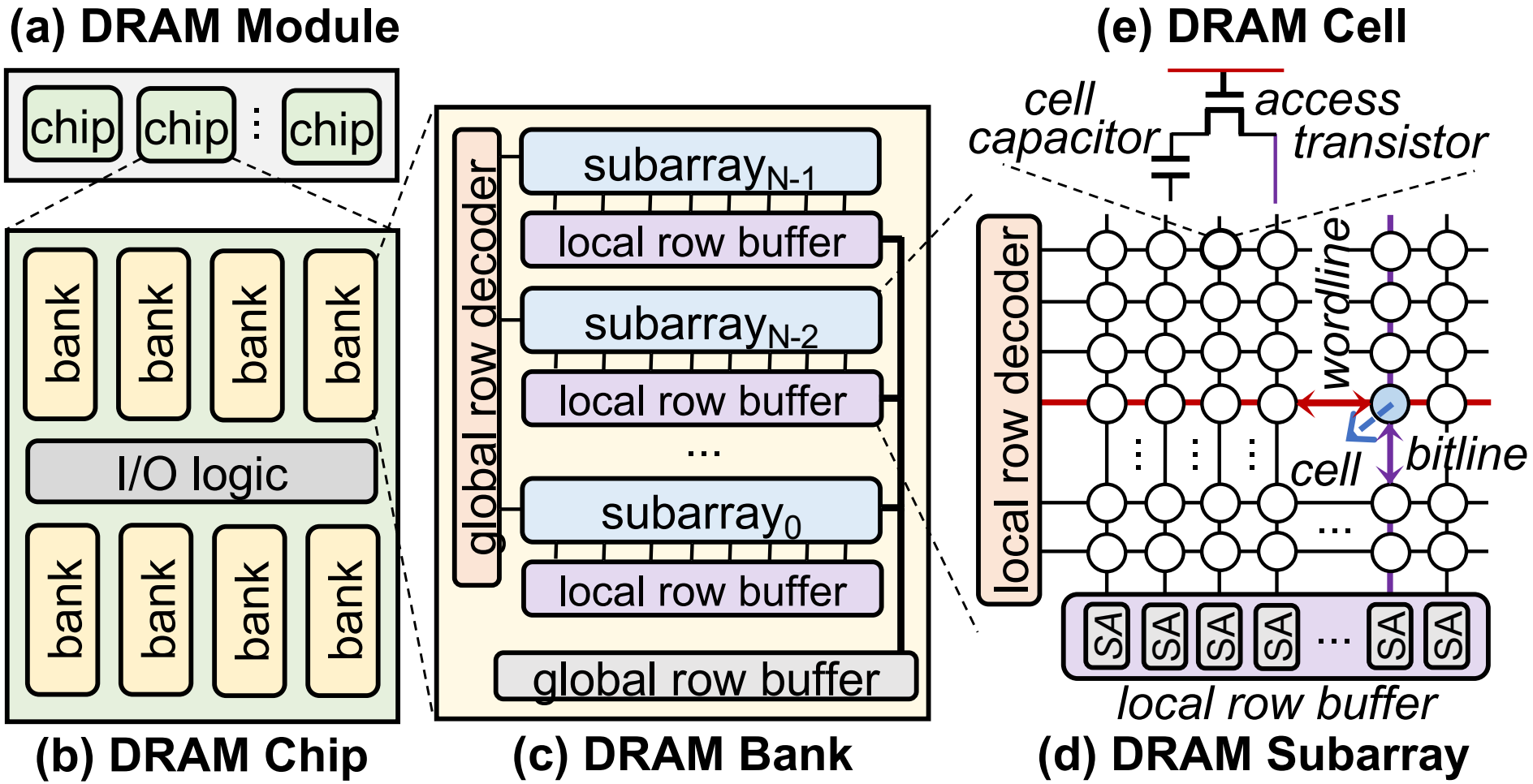


Replace computation with memory accesses
→ *pLUTo LUT Query* operation

DRAM Organization

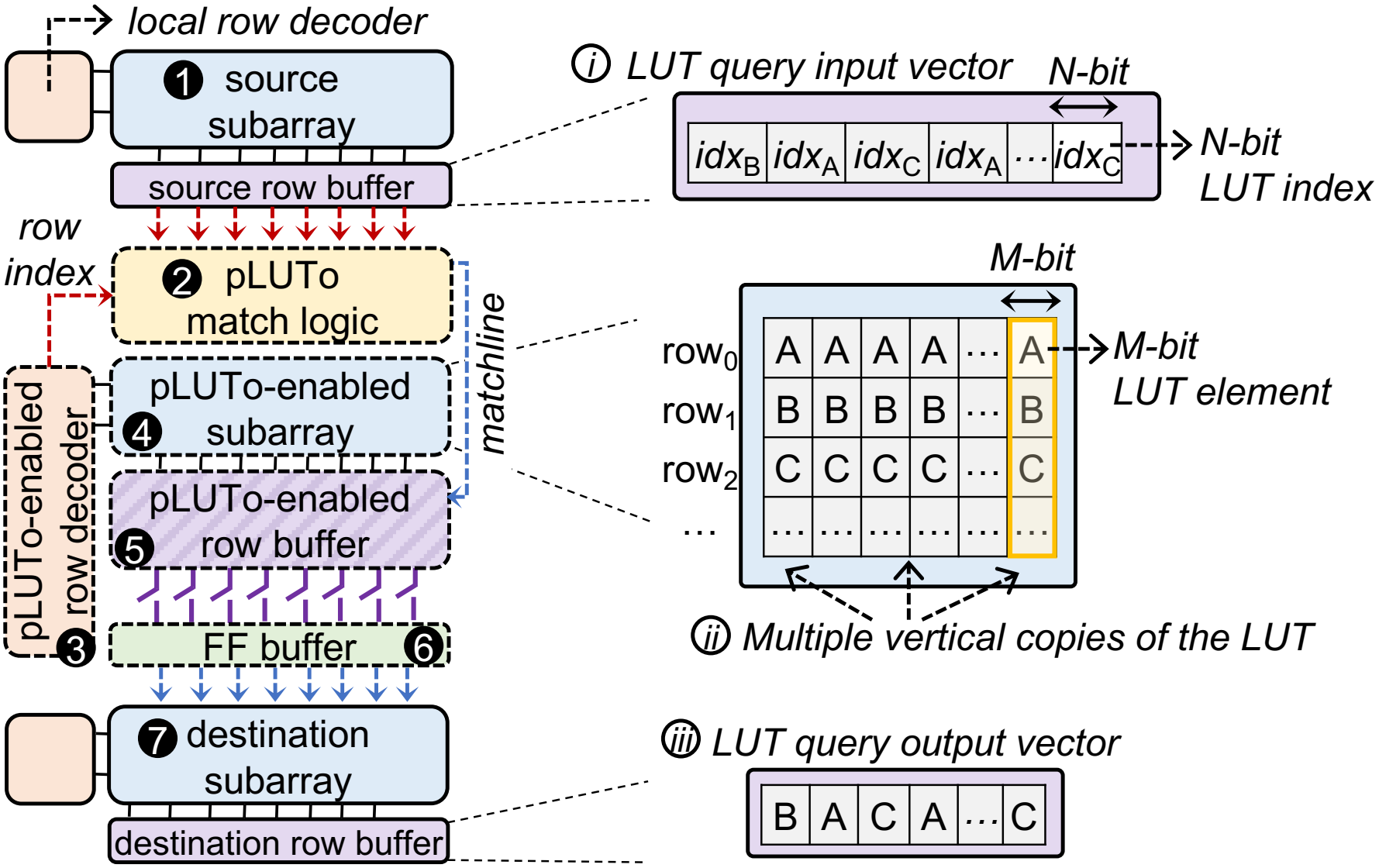


DRAM Organization



pLUTo Components

pLUTo Components



In-DRAM pLUTo LUT Query: Setup

LUT Query:
Return {2nd, 1st, 2nd, 4th}
prime numbers

	Prime numbers	
LUT index	i	f(i)
0	1st	2
1	2nd	3
2	3rd	5
3	4th	7

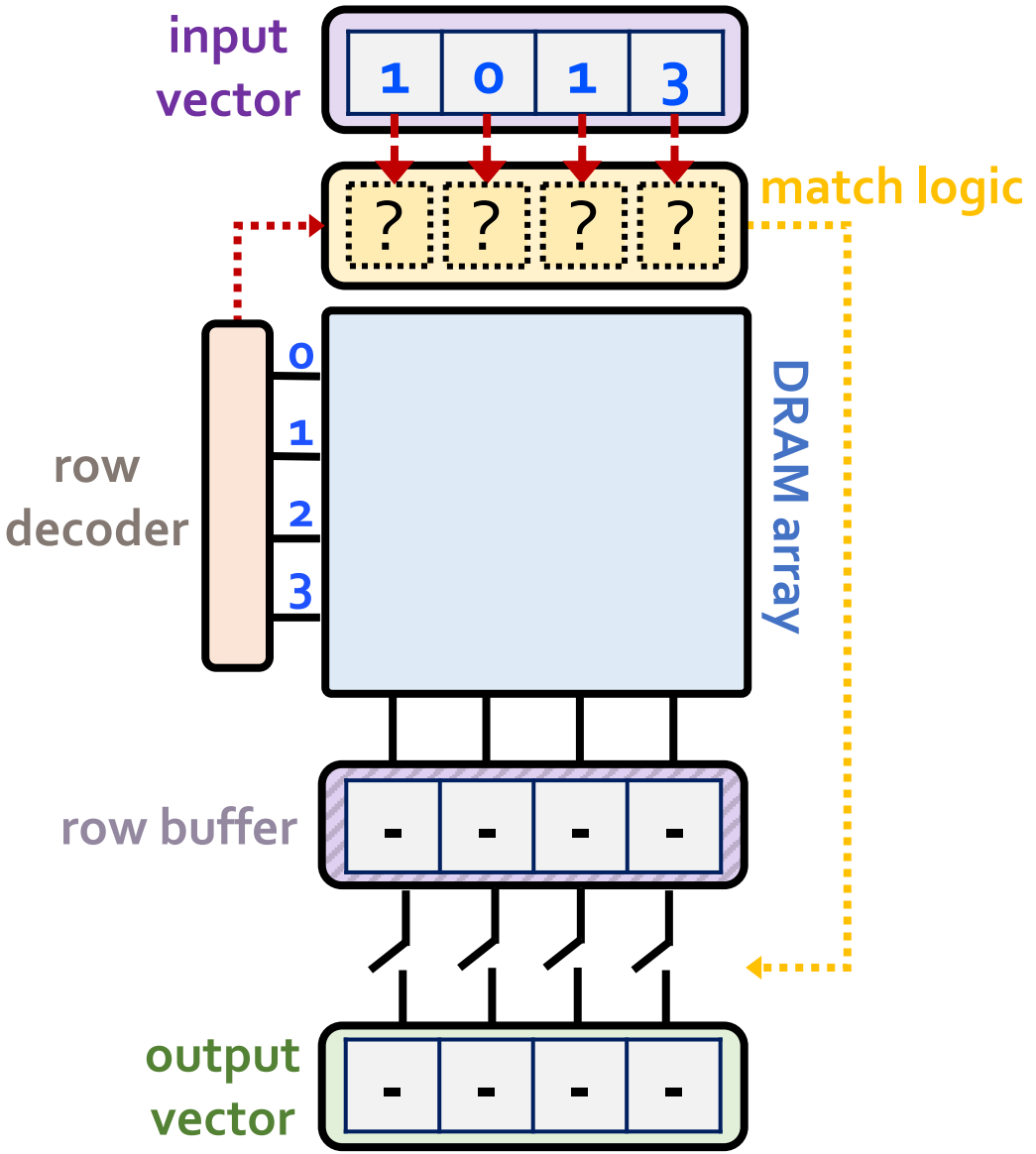
lookup table

1	0	1	3

input vector

3	2	3	7

output vector



In-DRAM pLUTo LUT Query: Setup

LUT Query:
Return {2nd, 1st, 2nd, 4th}
prime numbers

LUT index	Prime numbers: i	Prime numbers: f(i)
0	1st	2
1	2nd	3
2	3rd	5
3	4th	7

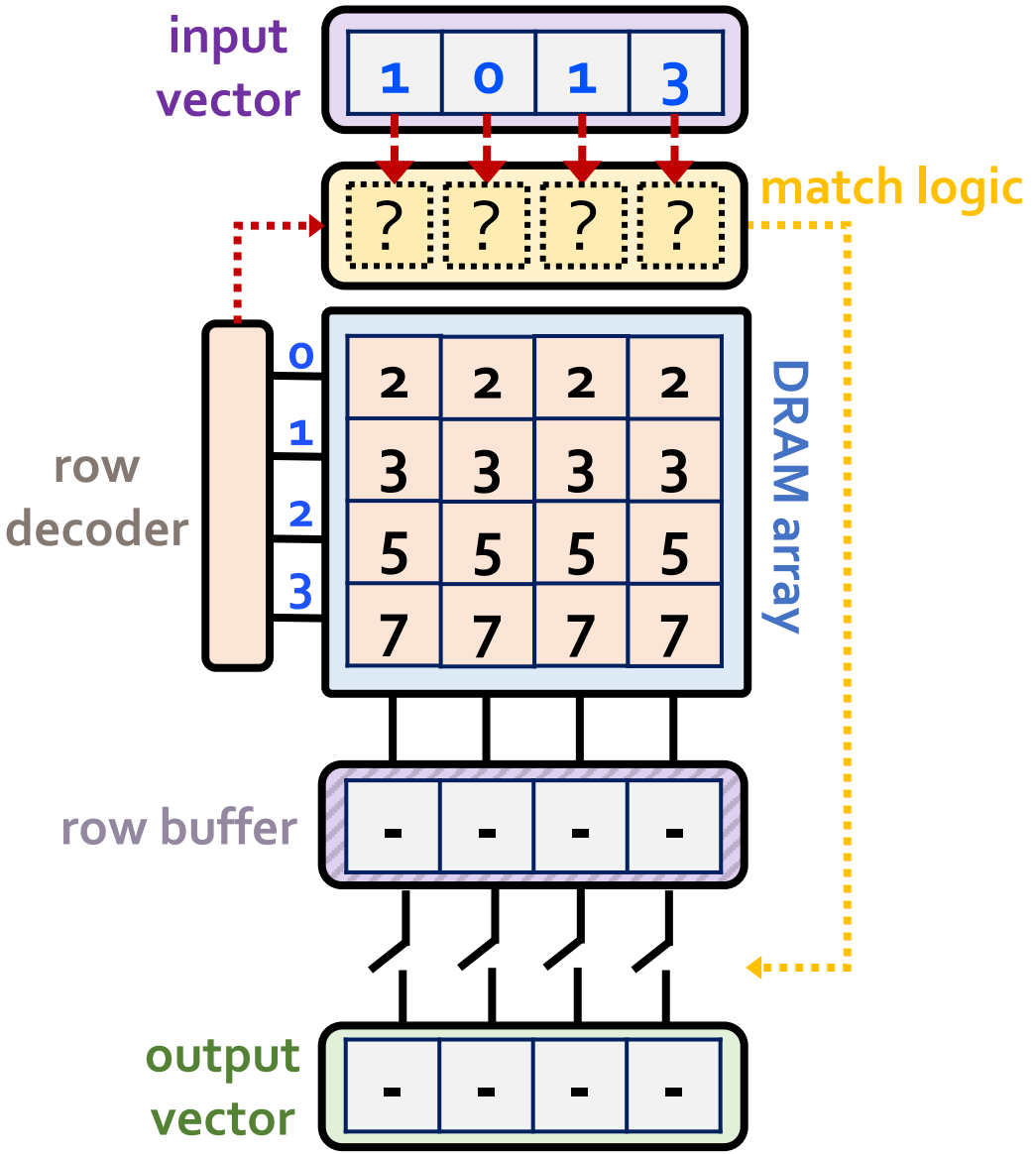
lookup table

1	0	1	3

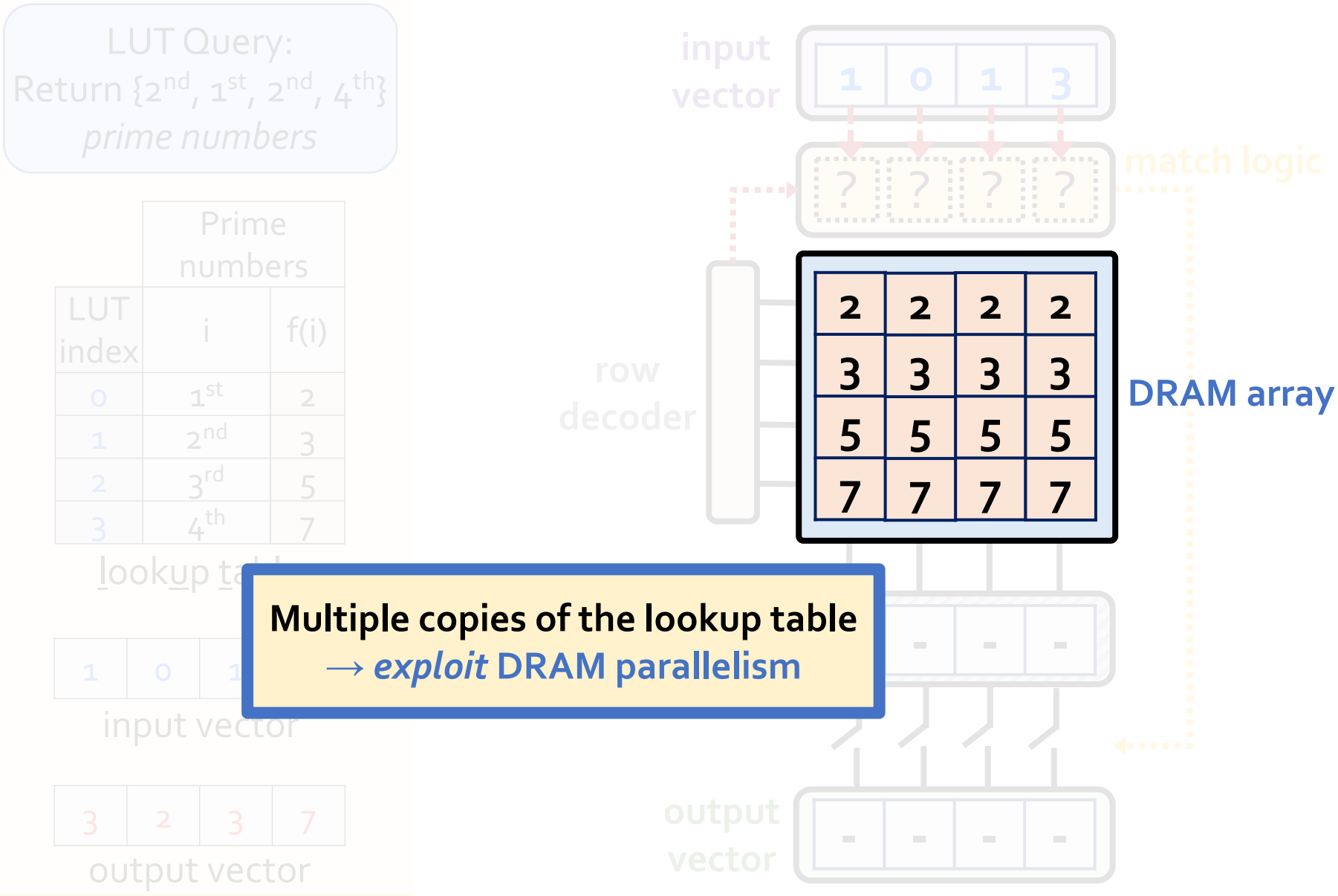
input vector

3	2	3	7

output vector



In-DRAM pLUTo LUT Query: Setup



In-DRAM pLUTo LUT Query: Setup

LUT Query:
Return {2nd, 1st, 2nd, 4th}
prime numbers

LUT index	Prime numbers i	f(i)
0	1st	2
1	2nd	3
2	3rd	5
3	4th	7

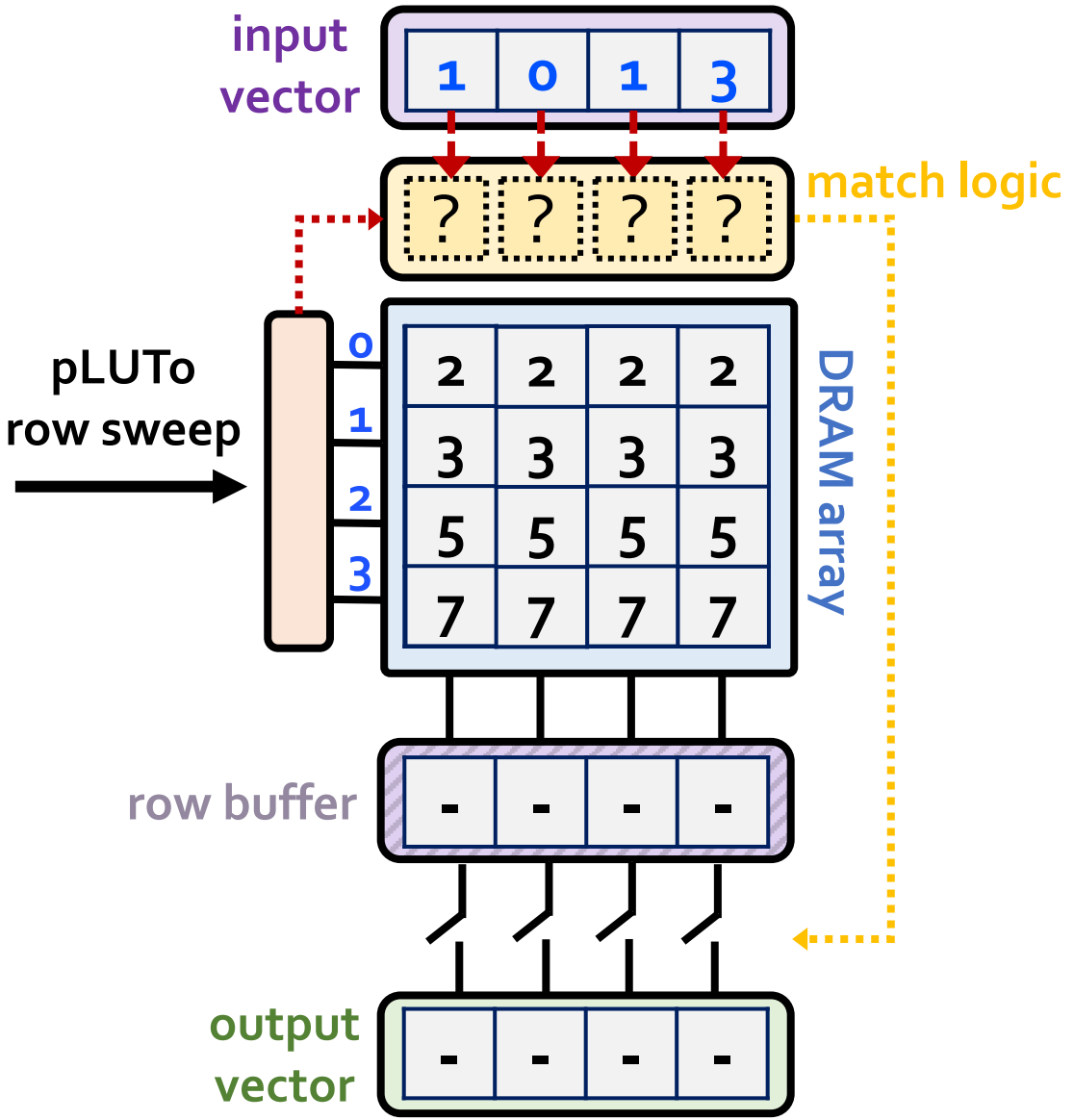
lookup table

1	0	1	3

input vector

3	2	3	7

output vector



In-DRAM pLUTo LUT Query: Step 1

LUT Query:
Return {2nd, 1st, 2nd, 4th}
prime numbers

LUT index	Prime numbers: i	Prime numbers: f(i)
0	1st	2
1	2nd	3
2	3rd	5
3	4th	7

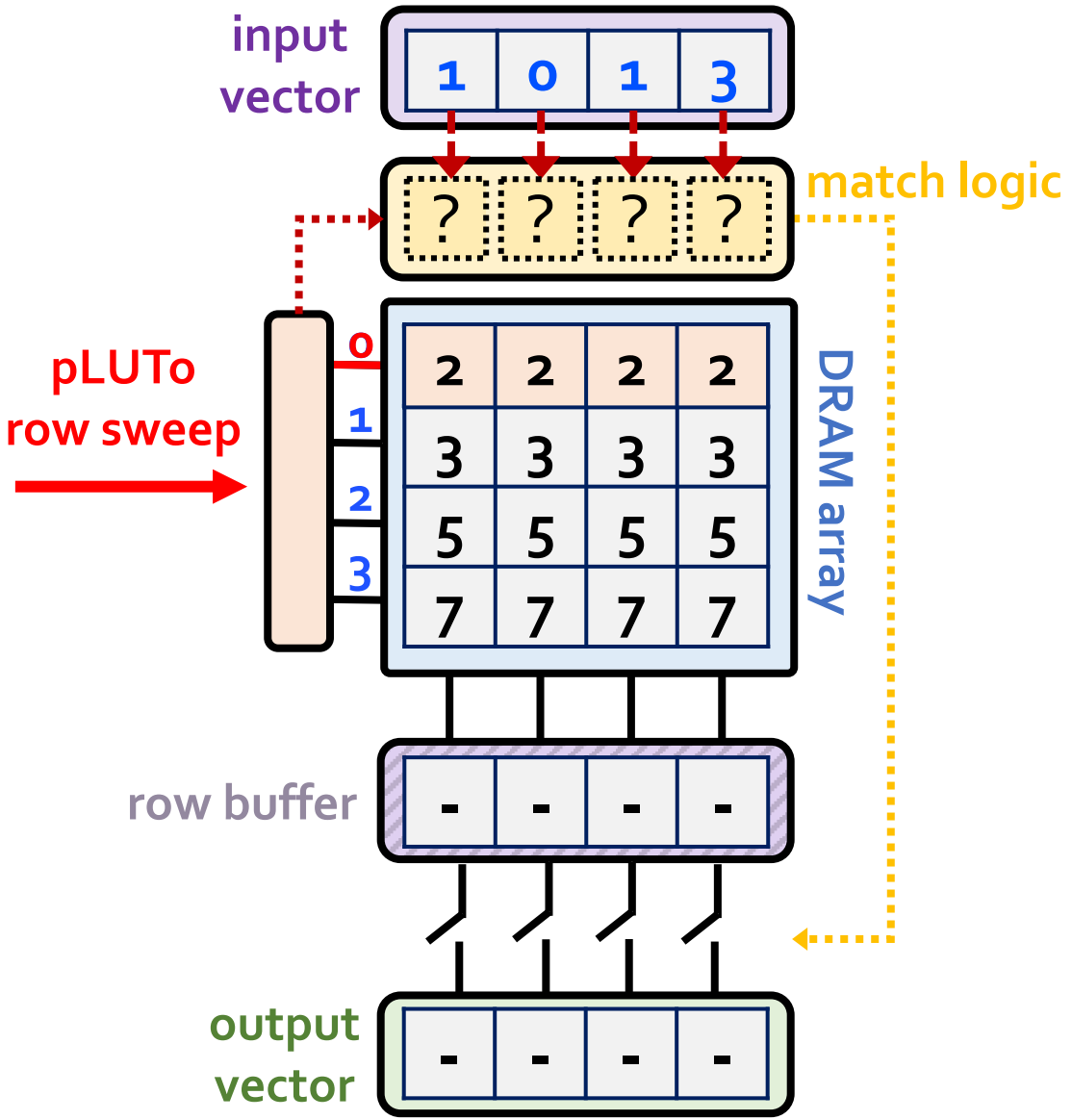
lookup table

1	0	1	3

input vector

3	2	3	7

output vector



In-DRAM pLUTo LUT Query: Step 1

LUT Query:
Return {2nd, 1st, 2nd, 4th}
prime numbers

LUT index	Prime numbers: i	Prime numbers: f(i)
0	1st	2
1	2nd	3
2	3rd	5
3	4th	7

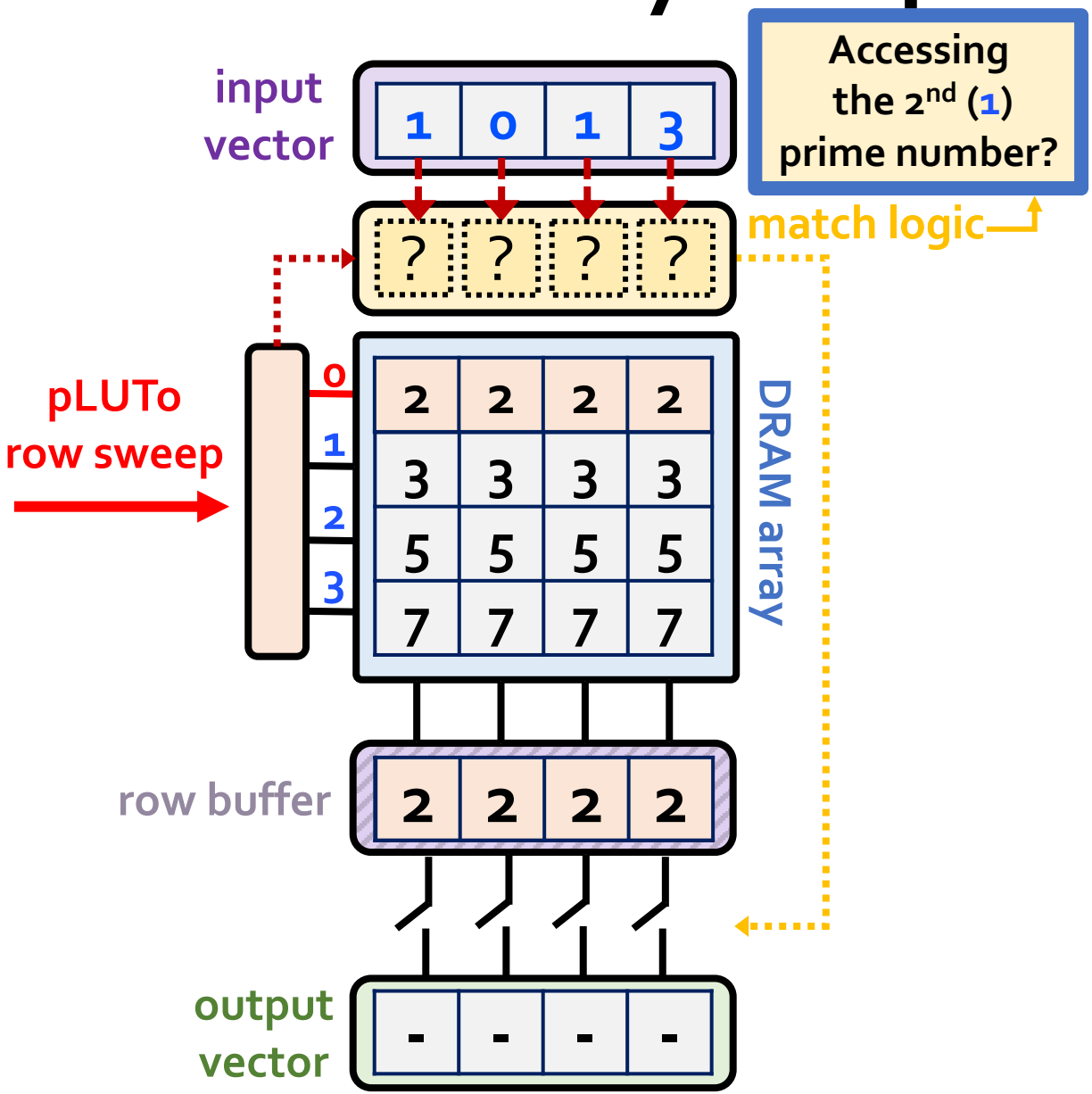
lookup table

1	0	1	3

input vector

3	2	3	7

output vector



In-DRAM pLUTo LUT Query: Step 1

LUT Query:
Return {2nd, 1st, 2nd, 4th} *prime numbers*

LUT index	Prime numbers i	f(i)
0	1st	2
1	2nd	3
2	3rd	5
3	4th	7

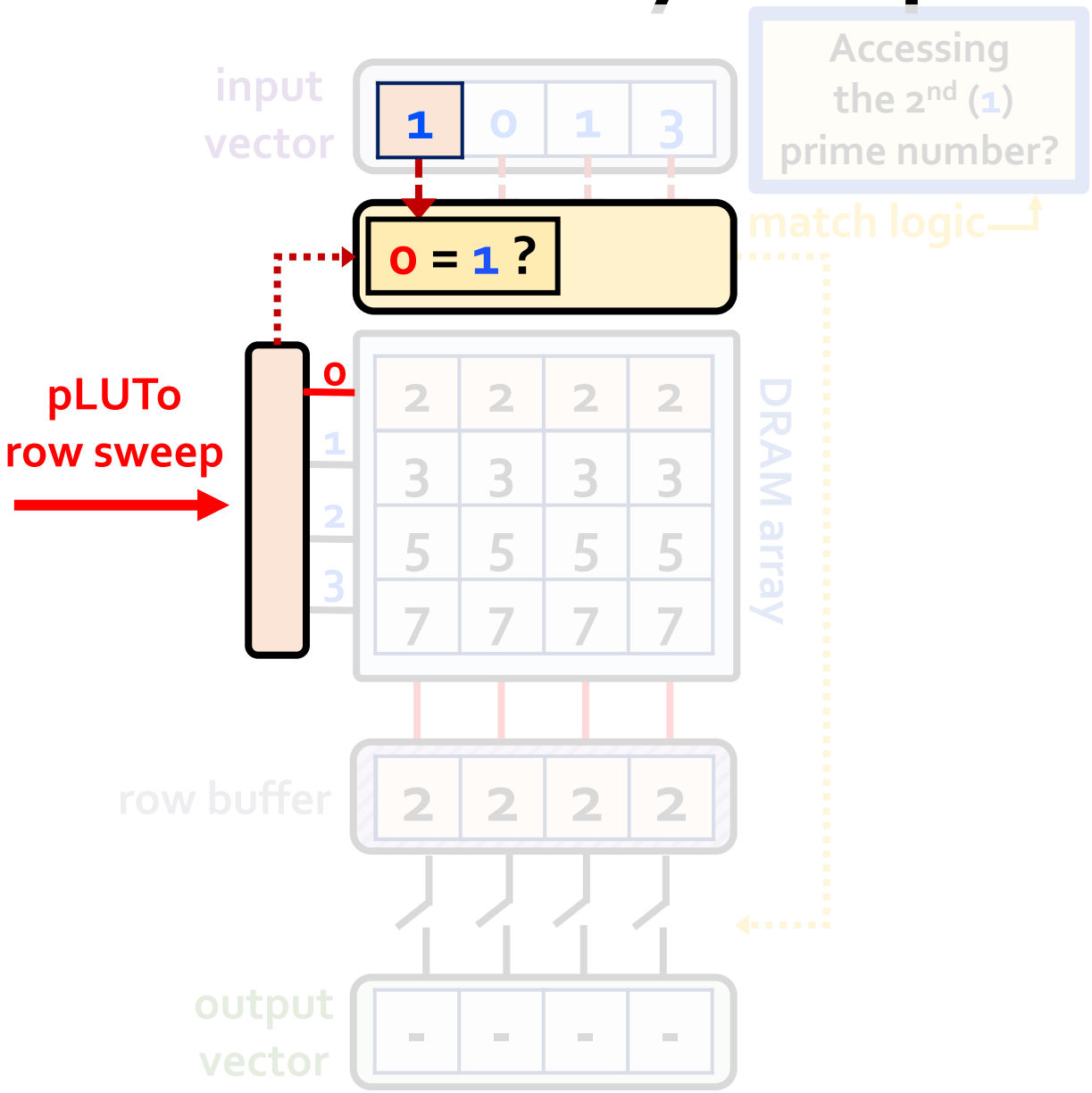
lookup table

1	0	1	3

input vector

3	2	3	7

output vector



In-DRAM pLUTo LUT Query: Step 1

LUT Query:
Return {2nd, 1st, 2nd, 4th}
prime numbers

LUT index	Prime numbers: i	Prime numbers: f(i)
0	1st	2
1	2nd	3
2	3rd	5
3	4th	7

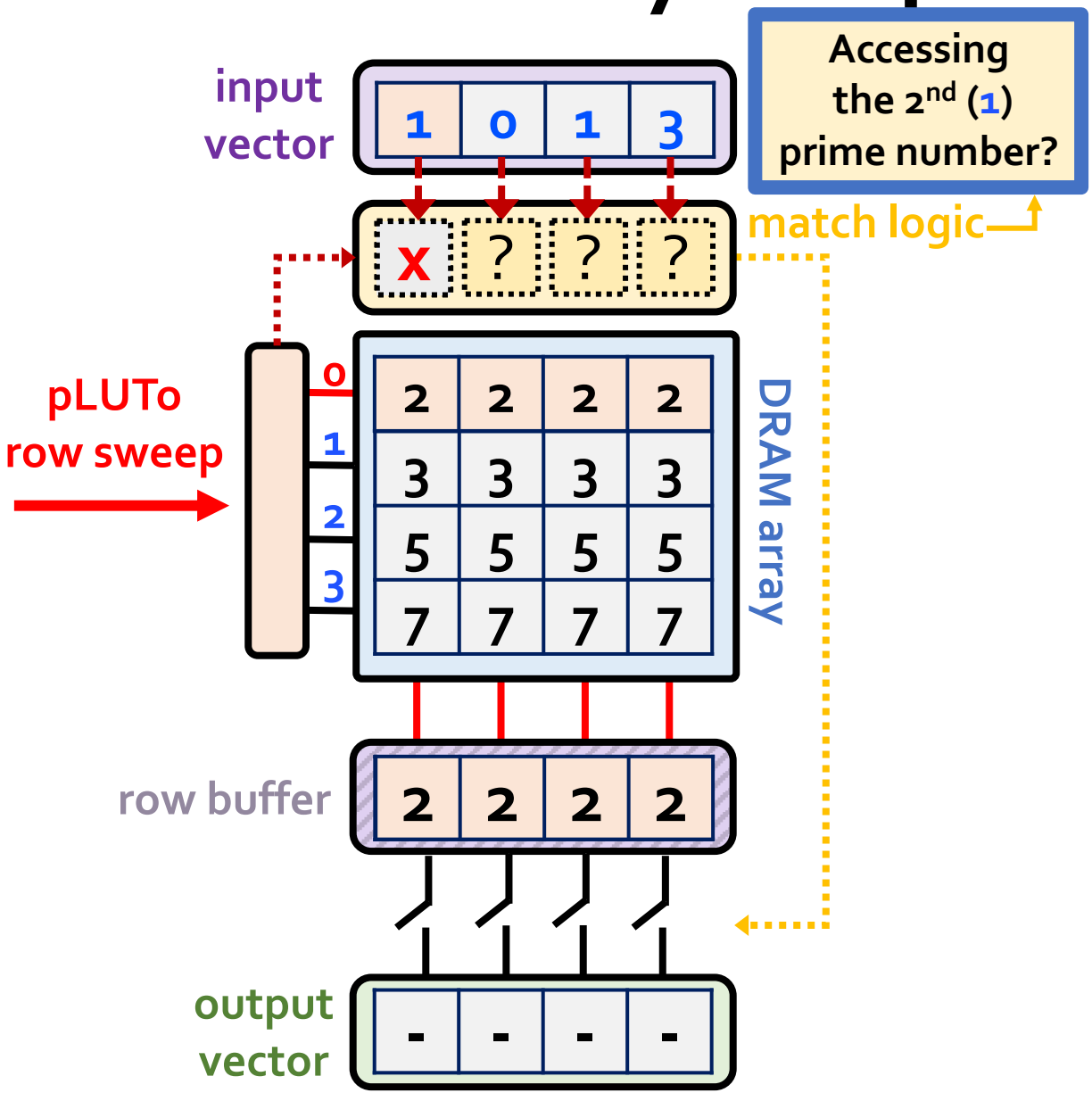
lookup table

1	0	1	3

input vector

3	2	3	7

output vector



In-DRAM pLUTo LUT Query: Step 1

LUT Query:
Return {2nd, 1st, 2nd, 4th}
prime numbers

LUT index	Prime numbers i	f(i)
0	1st	2
1	2nd	3
2	3rd	5
3	4th	7

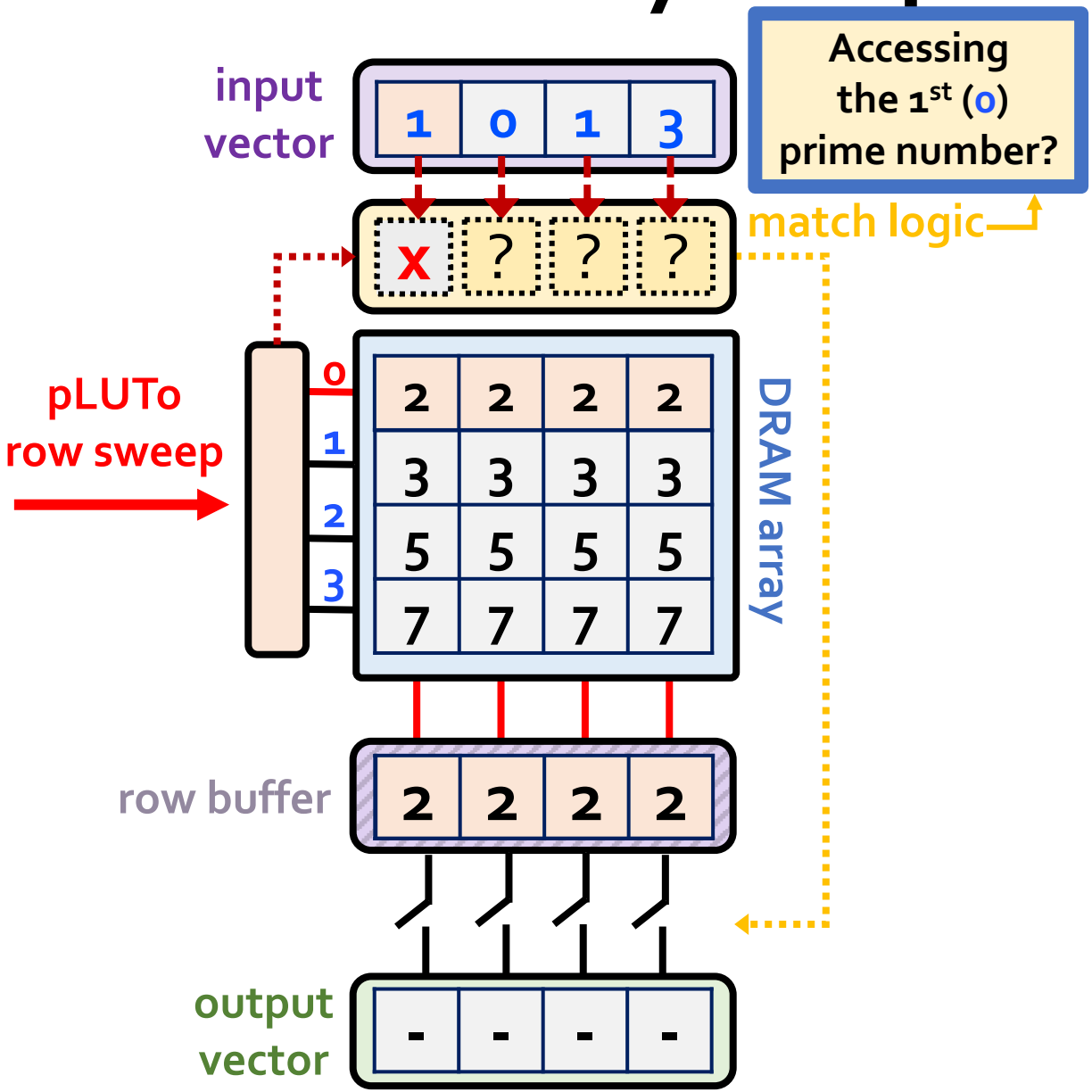
lookup table

1	0	1	3

input vector

3	2	3	7

output vector



In-DRAM pLUTo LUT Query: Step 1

LUT Query:
Return {2nd, 1st, 2nd, 4th} *prime numbers*

LUT index	Prime numbers: i	Prime numbers: f(i)
0	1st	2
1	2nd	3
2	3rd	5
3	4th	7

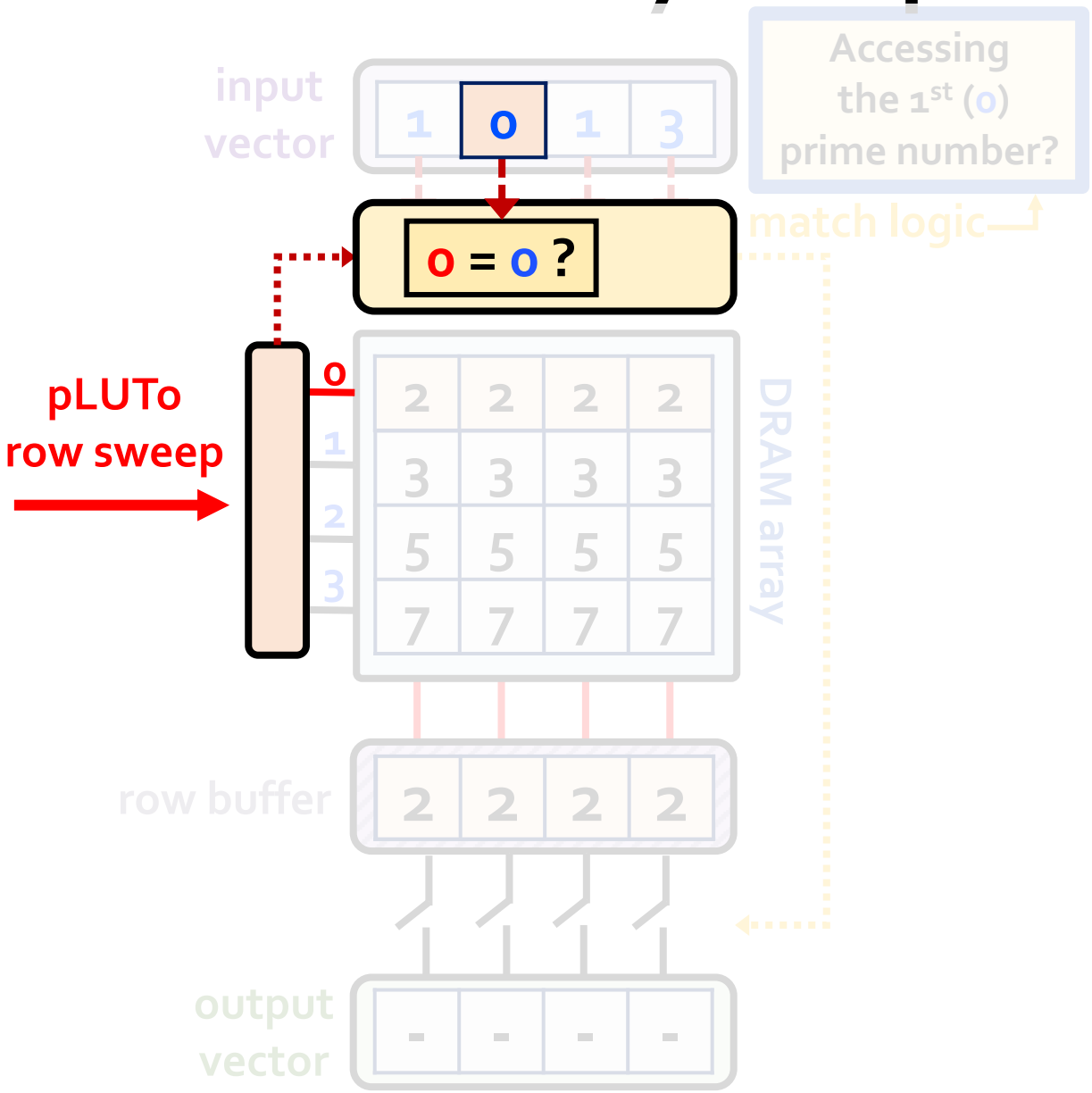
lookup table

1	0	1	3

input vector

3	2	3	7

output vector



In-DRAM pLUTo LUT Query: Step 1

LUT Query:
Return {2nd, 1st, 2nd, 4th}
prime numbers

LUT index	Prime numbers i	f(i)
0	1st	2
1	2nd	3
2	3rd	5
3	4th	7

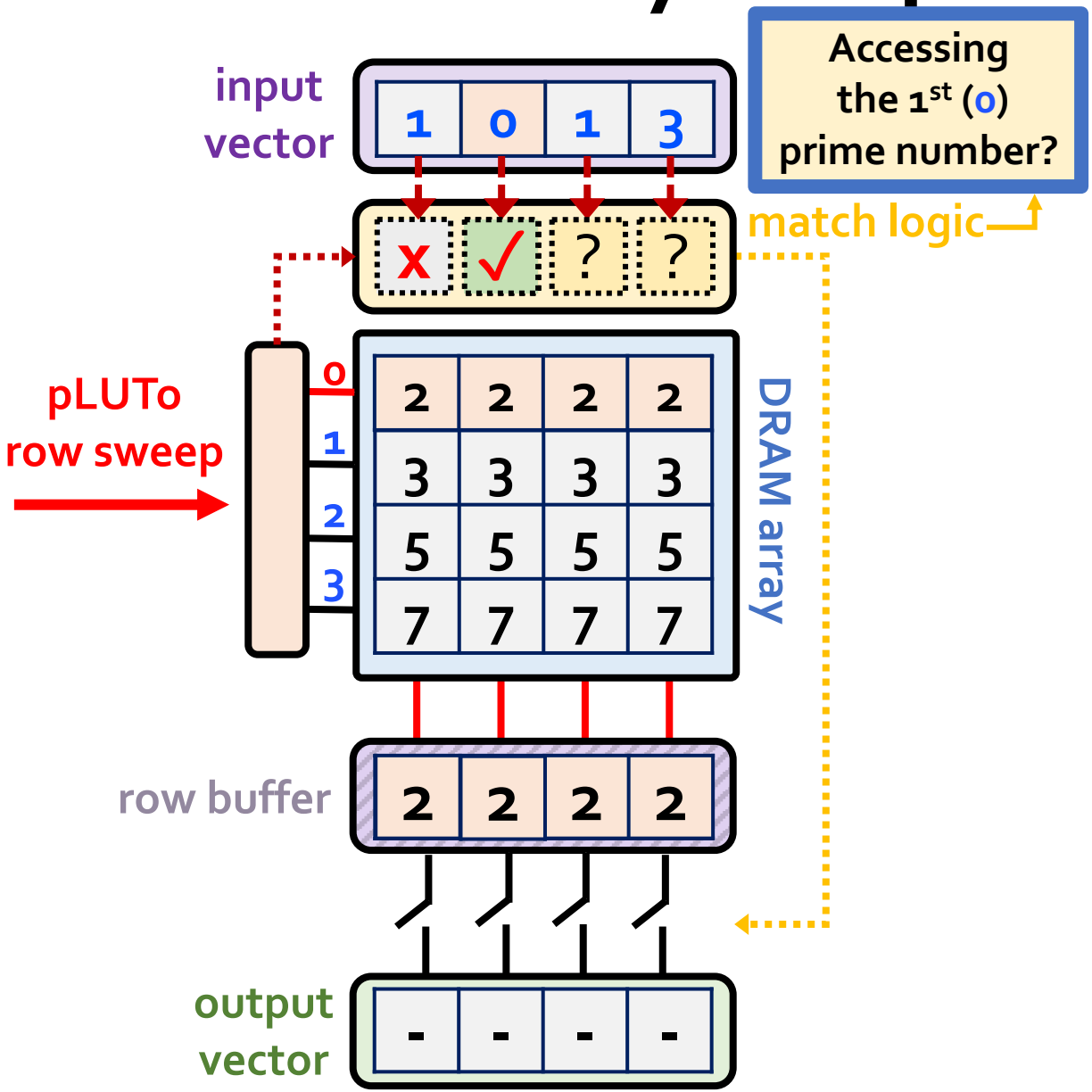
lookup table

1	0	1	3

input vector

3	2	3	7

output vector



In-DRAM pLUTo LUT Query: Step 1

LUT Query:
Return {2nd, 1st, 2nd, 4th}
prime numbers

LUT index	Prime numbers: i	Prime numbers: f(i)
0	1st	2
1	2nd	3
2	3rd	5
3	4th	7

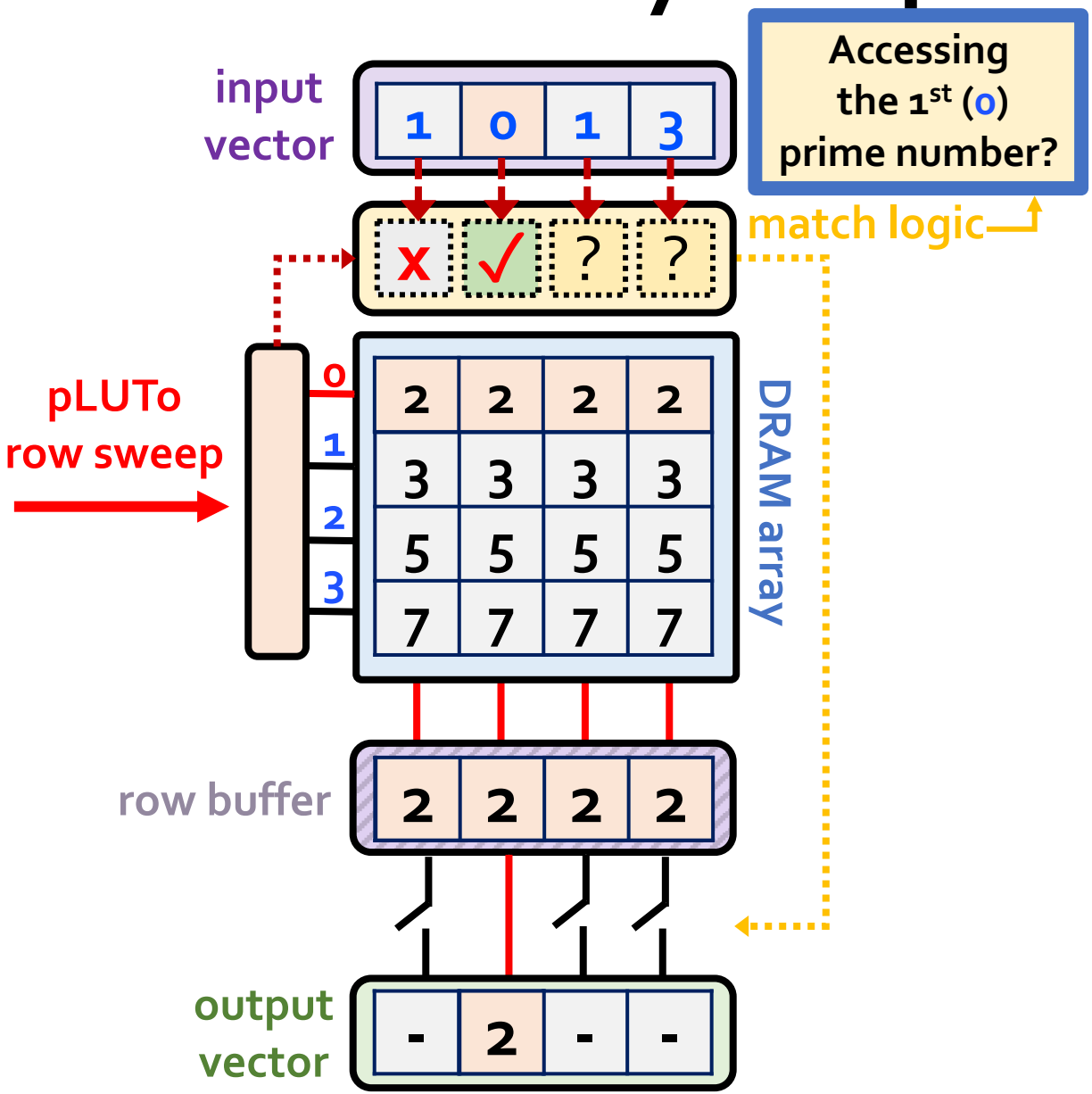
lookup table

1	0	1	3

input vector

3	2	3	7

output vector



In-DRAM pLUTo LUT Query: Step 1

LUT Query:
Return {2nd, 1st, 2nd, 4th}
prime numbers

LUT index	Prime numbers: i	Prime numbers: f(i)
0	1st	2
1	2nd	3
2	3rd	5
3	4th	7

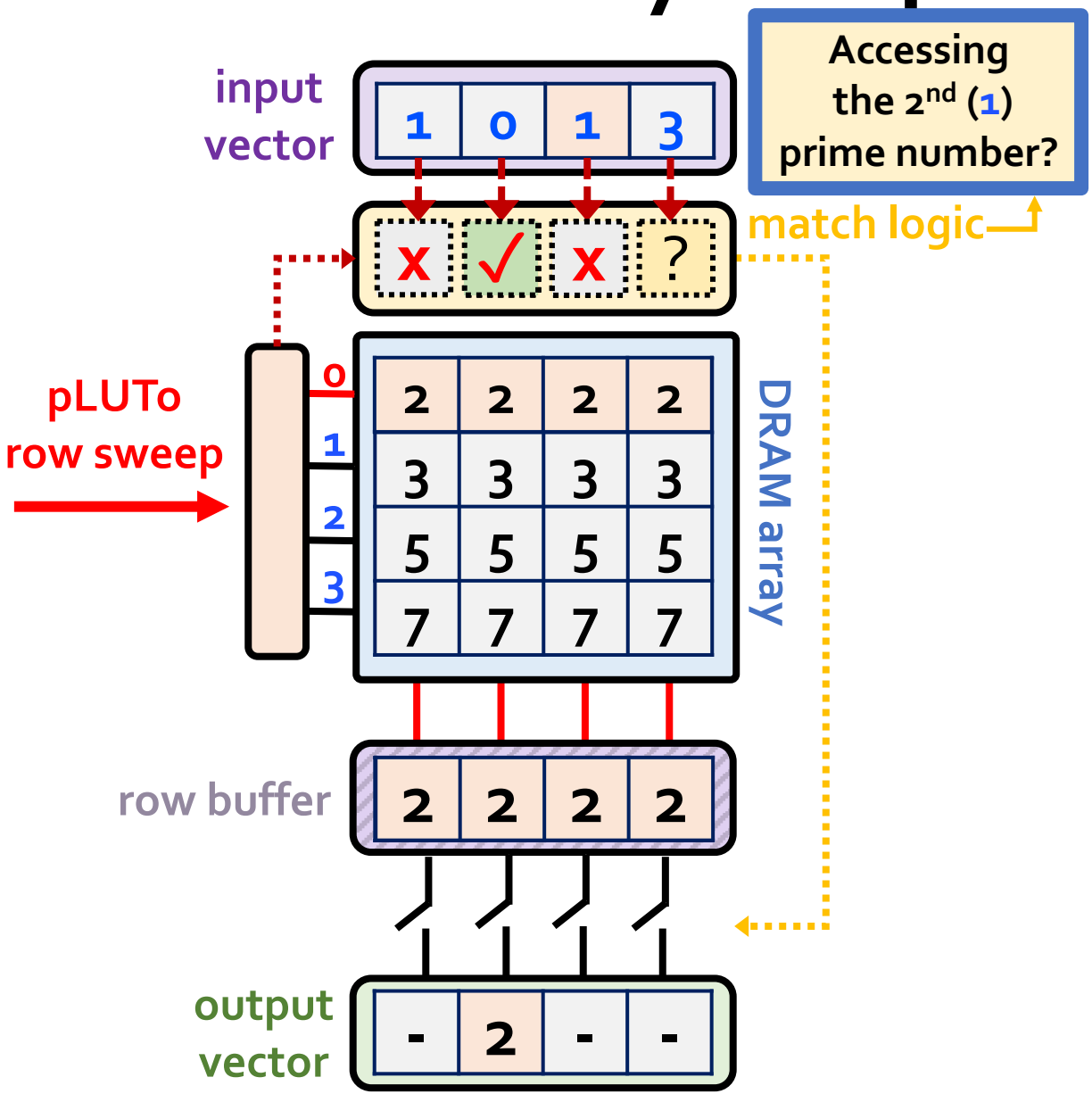
lookup table

1	0	1	3

input vector

3	2	3	7

output vector



SAFARI

In-DRAM pLUTo LUT Query: Step 1

LUT Query:
Return {2nd, 1st, 2nd, 4th}
prime numbers

LUT index	Prime numbers: i	Prime numbers: f(i)
0	1st	2
1	2nd	3
2	3rd	5
3	4th	7

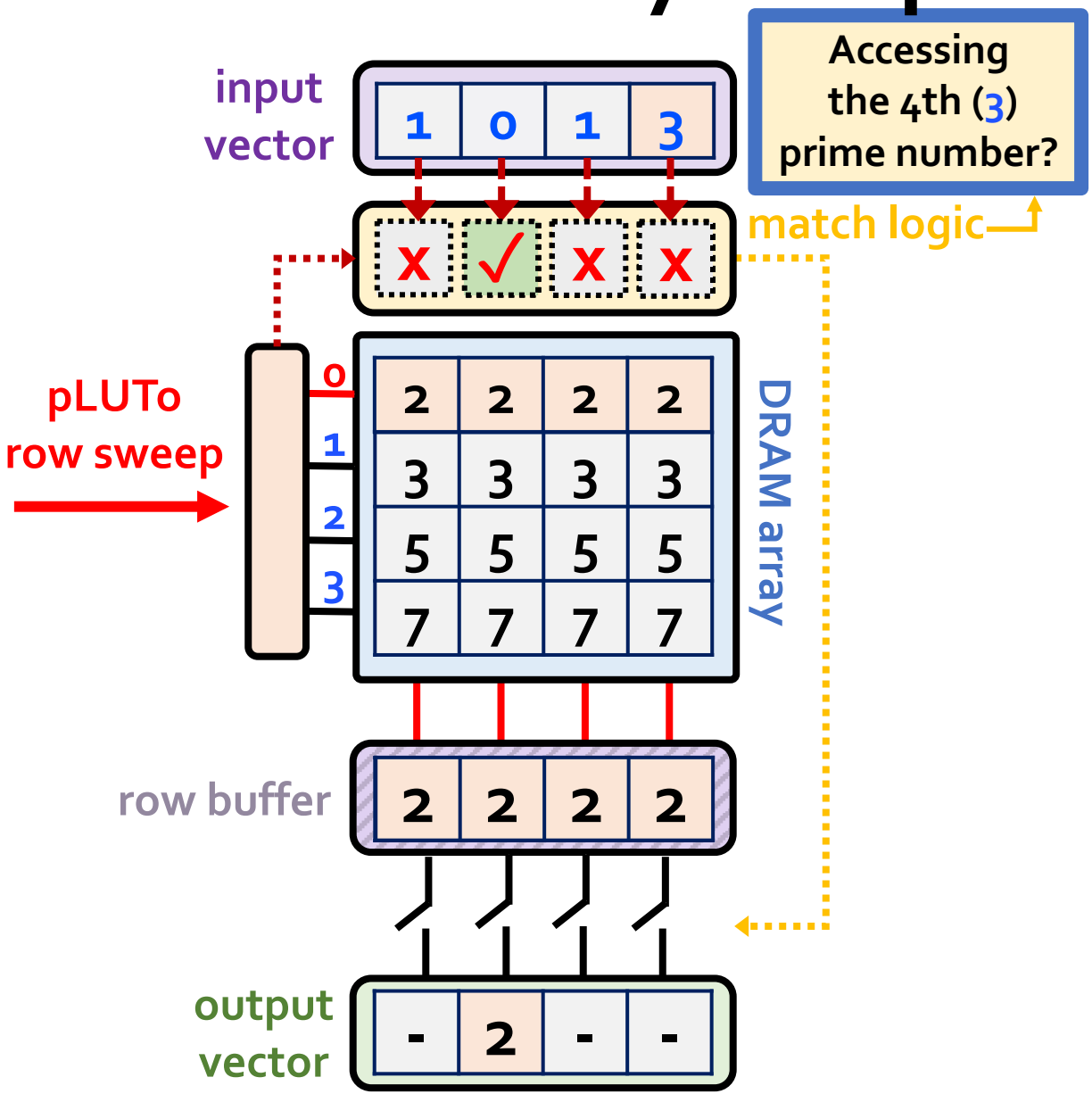
lookup table

1	0	1	3

input vector

3	2	3	7

output vector



In-DRAM pLUTo LUT Query: Step 2

LUT Query:
Return {2nd, 1st, 2nd, 4th}
prime numbers

LUT index	Prime numbers: i	Prime numbers: f(i)
0	1st	2
1	2nd	3
2	3rd	5
3	4th	7

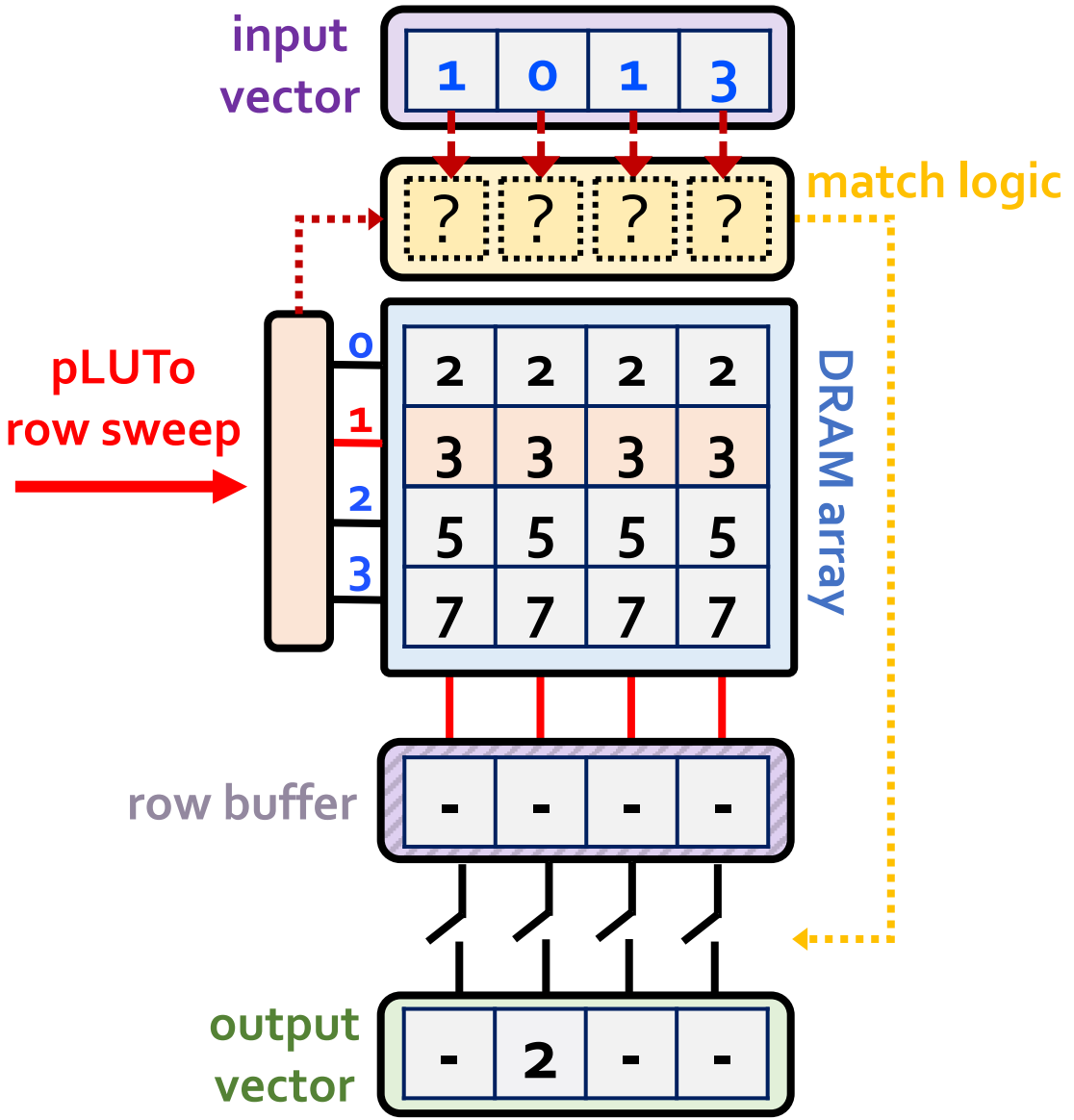
lookup table

1	0	1	3

input vector

3	2	3	7

output vector



In-DRAM pLUTo LUT Query: Step 2

LUT Query:
Return {2nd, 1st, 2nd, 4th}
prime numbers

LUT index	Prime numbers: i	Prime numbers: f(i)
0	1st	2
1	2nd	3
2	3rd	5
3	4th	7

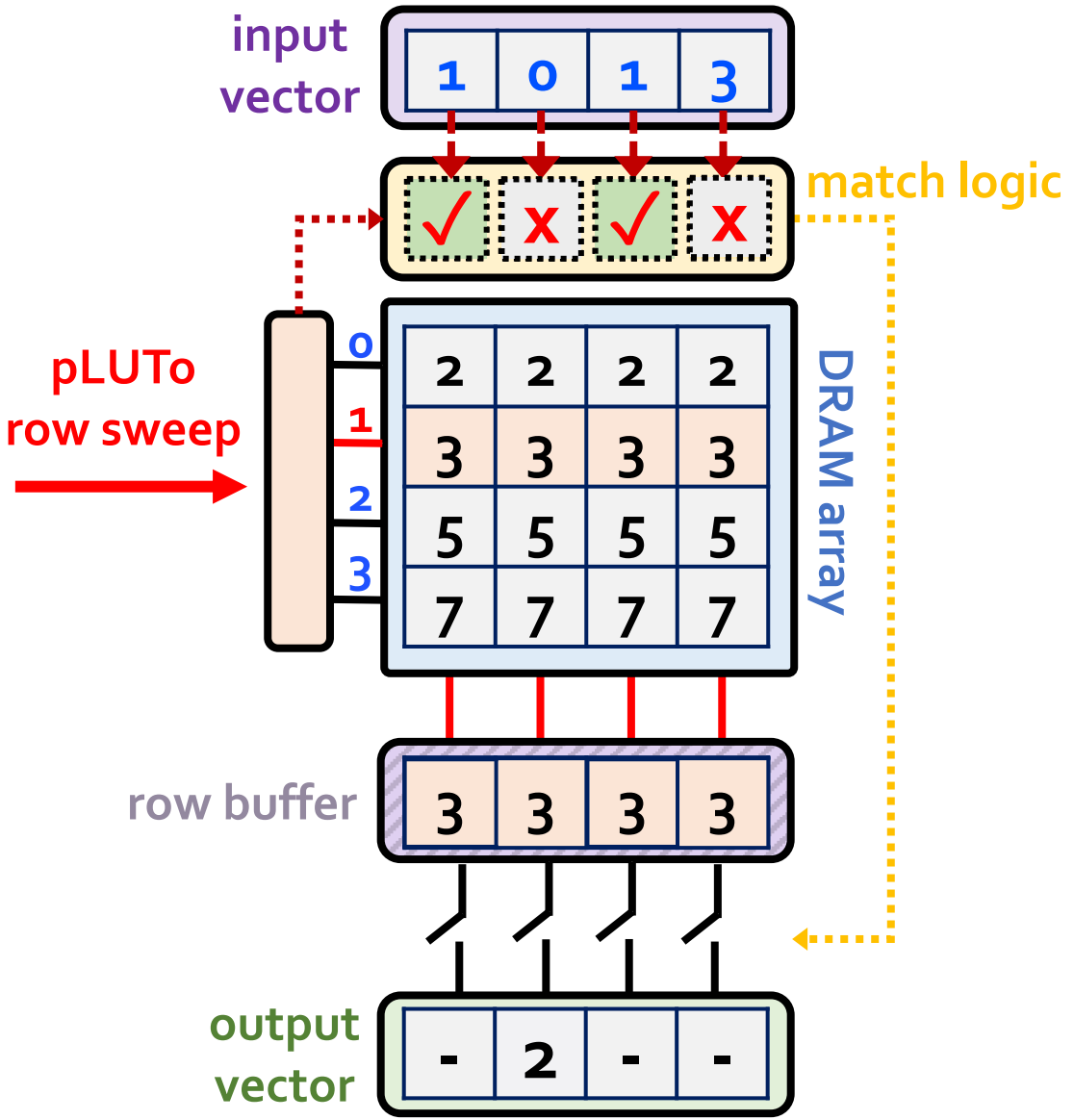
lookup table

1	0	1	3

input vector

3	2	3	7

output vector



In-DRAM pLUTo LUT Query: Step 2

LUT Query:
Return {2nd, 1st, 2nd, 4th}
prime numbers

LUT index	Prime numbers: i	Prime numbers: f(i)
0	1st	2
1	2nd	3
2	3rd	5
3	4th	7

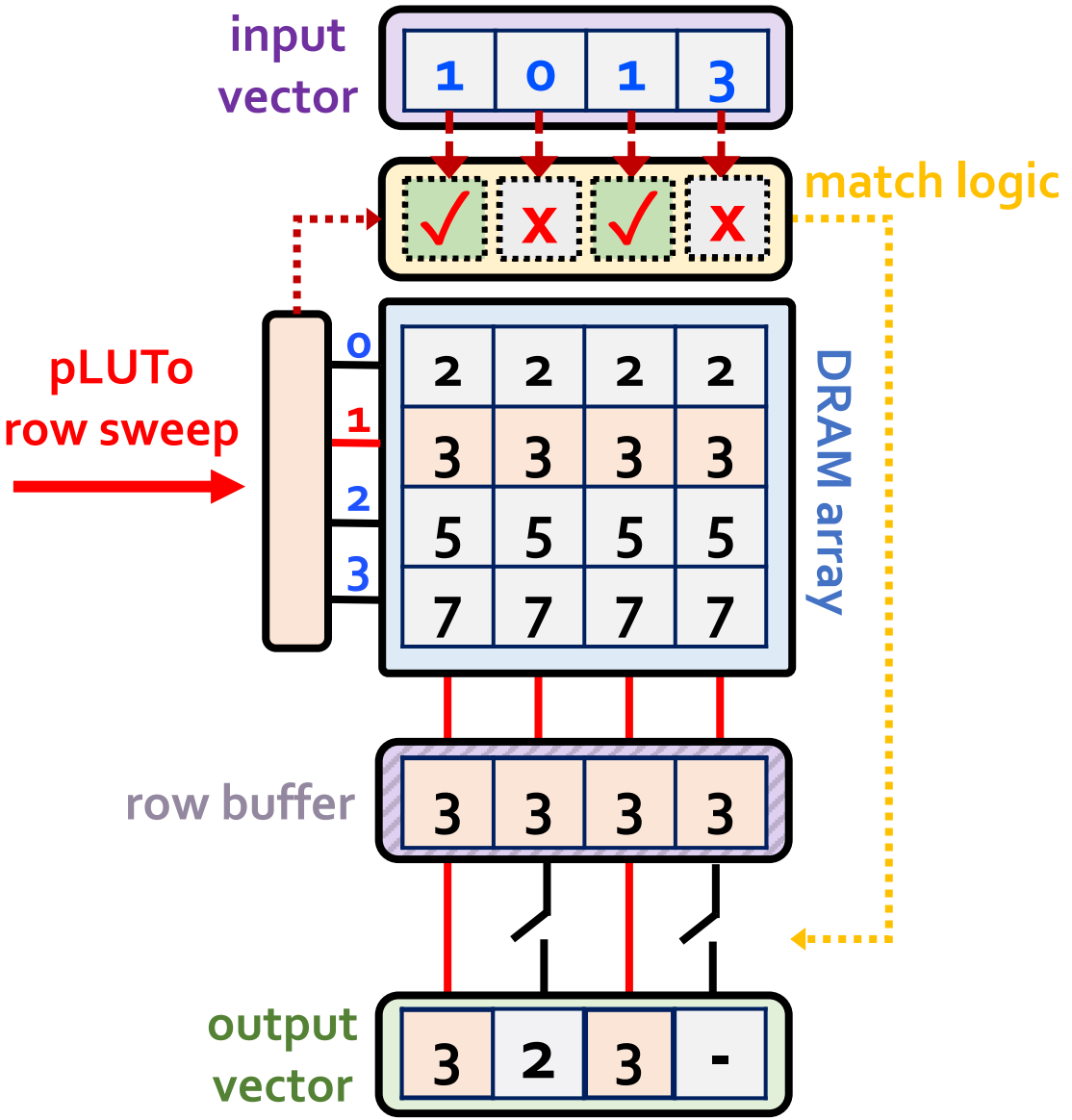
lookup table

1	0	1	3

input vector

3	2	3	7

output vector



In-DRAM pLUTo LUT Query: Step 3

LUT Query:
Return {2nd, 1st, 2nd, 4th}
prime numbers

LUT index	Prime numbers i	f(i)
0	1st	2
1	2nd	3
2	3rd	5
3	4th	7

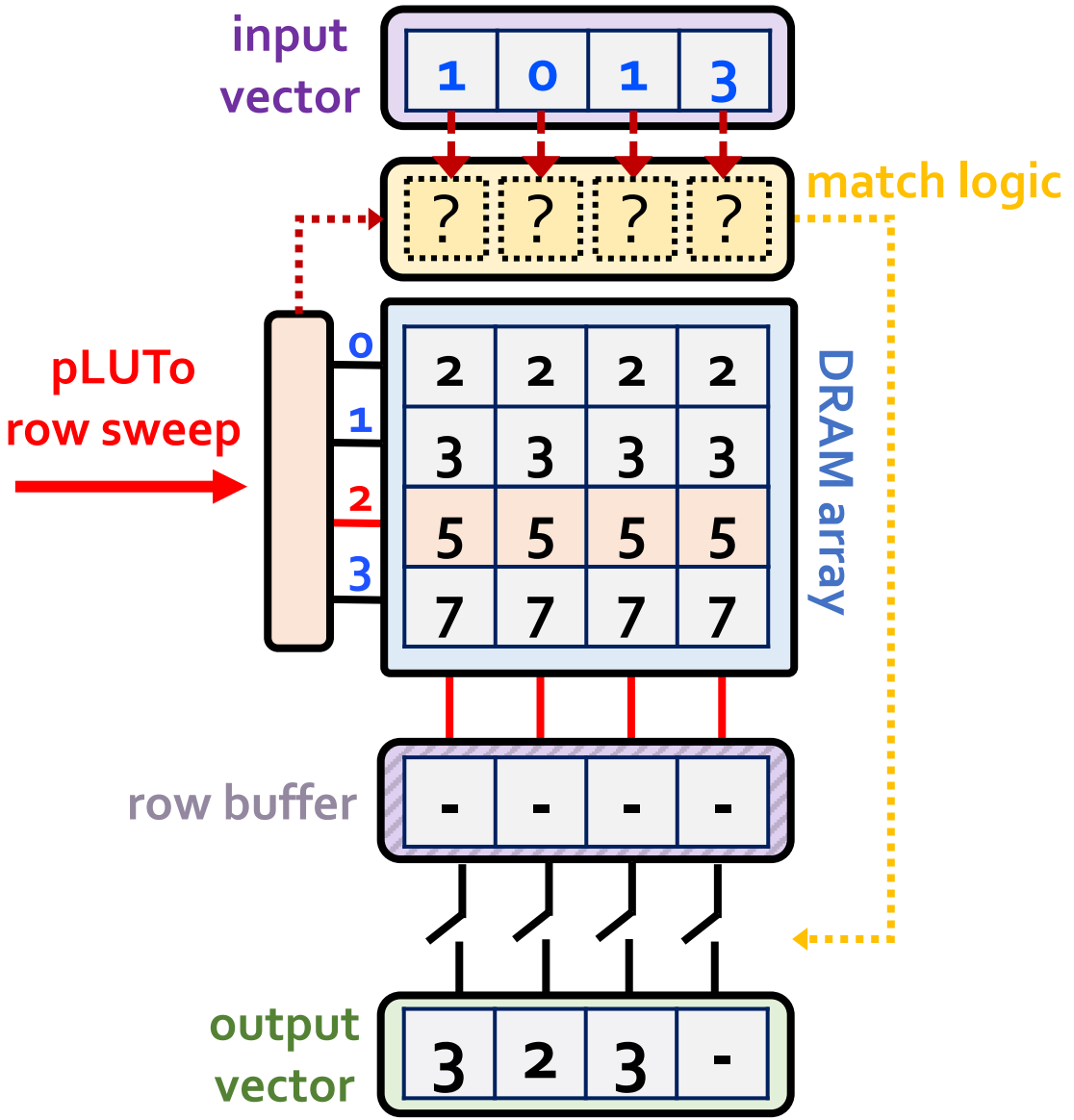
lookup table

1	0	1	3

input vector

3	2	3	7

output vector



In-DRAM pLUTo LUT Query: Step 3

LUT Query:
Return {2nd, 1st, 2nd, 4th}
prime numbers

LUT index	Prime numbers i	f(i)
0	1st	2
1	2nd	3
2	3rd	5
3	4th	7

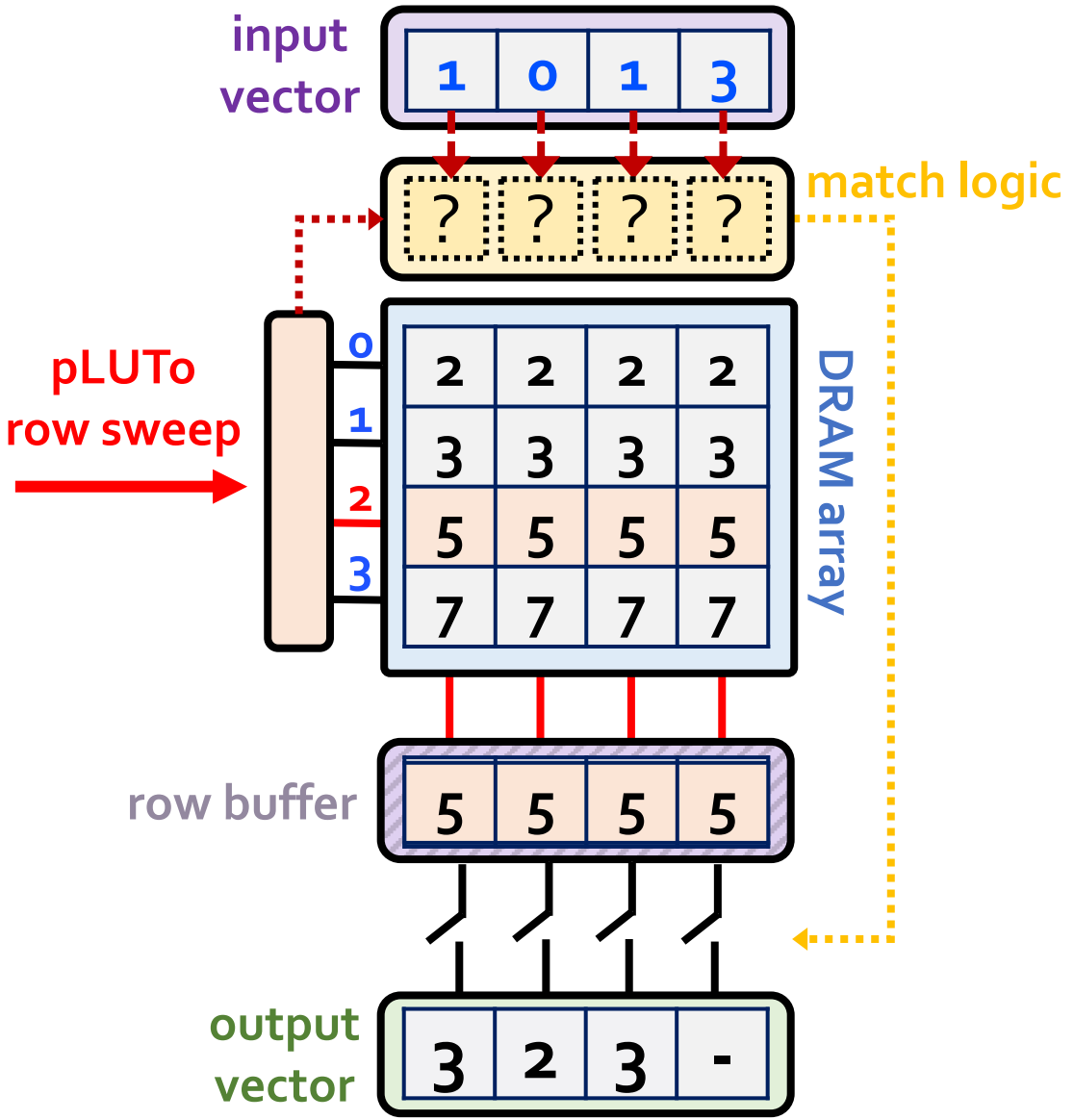
lookup table

1	0	1	3

input vector

3	2	3	7

output vector



In-DRAM pLUTo LUT Query: Step 3

LUT Query:
Return {2nd, 1st, 2nd, 4th}
prime numbers

LUT index	Prime numbers i	f(i)
0	1st	2
1	2nd	3
2	3rd	5
3	4th	7

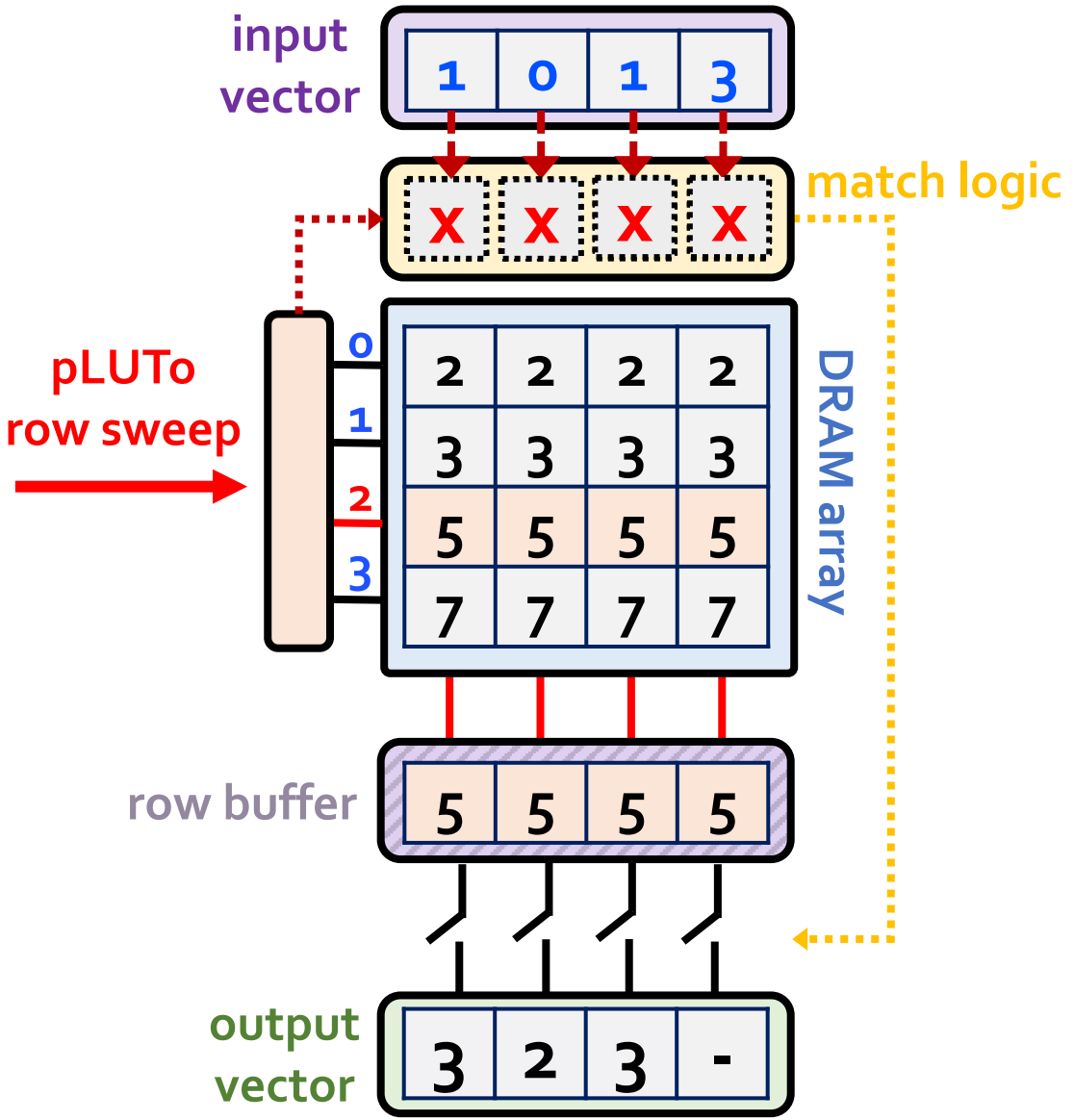
lookup table

1	0	1	3

input vector

3	2	3	7

output vector



In-DRAM pLUTo LUT Query: Step 4

LUT Query:
Return {2nd, 1st, 2nd, 4th}
prime numbers

LUT index	Prime numbers: i	Prime numbers: f(i)
0	1st	2
1	2nd	3
2	3rd	5
3	4th	7

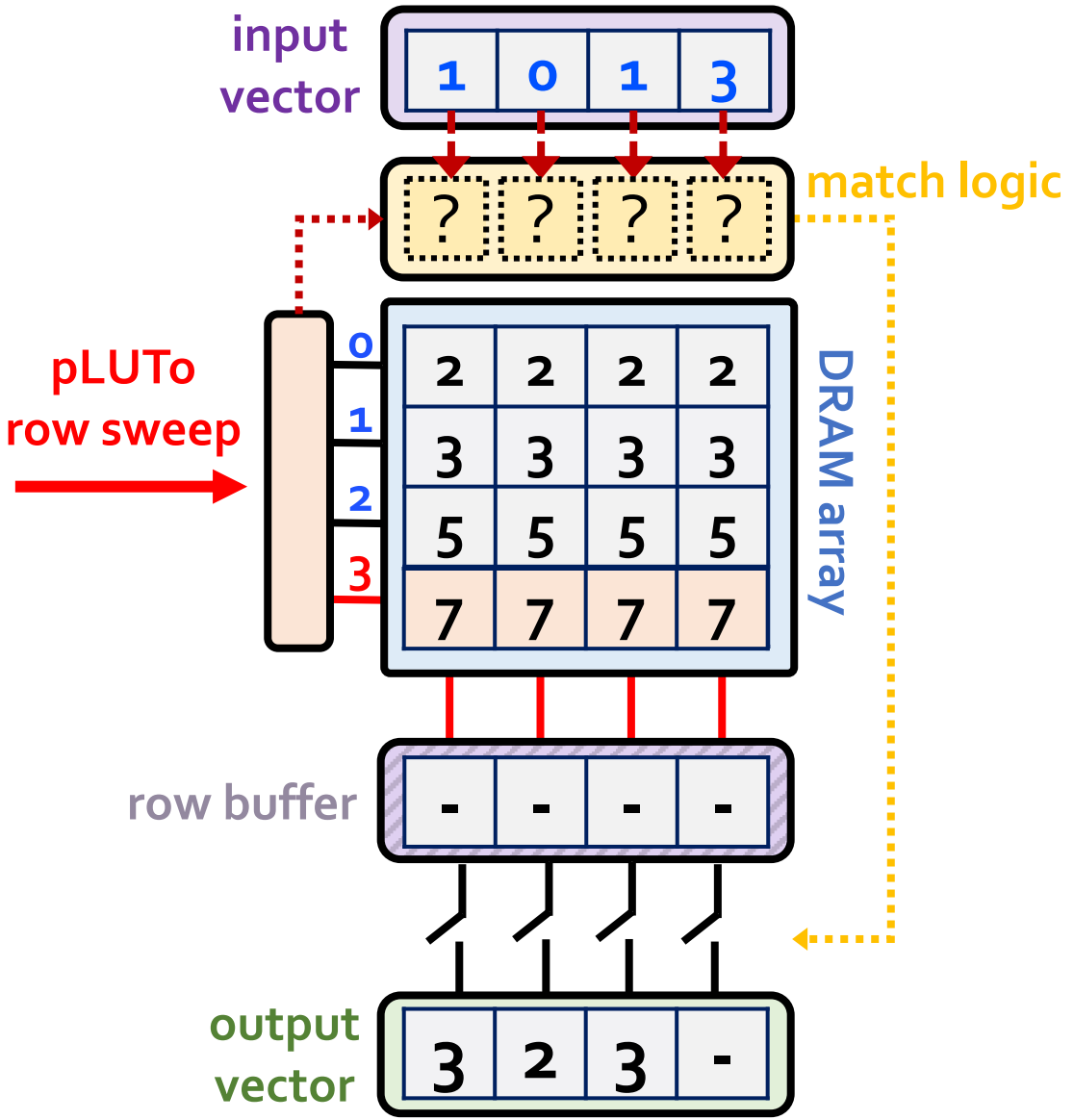
lookup table

1	0	1	3

input vector

3	2	3	7

output vector



In-DRAM pLUTo LUT Query: Step 4

LUT Query:
Return {2nd, 1st, 2nd, 4th}
prime numbers

LUT index	Prime numbers: i	Prime numbers: f(i)
0	1st	2
1	2nd	3
2	3rd	5
3	4th	7

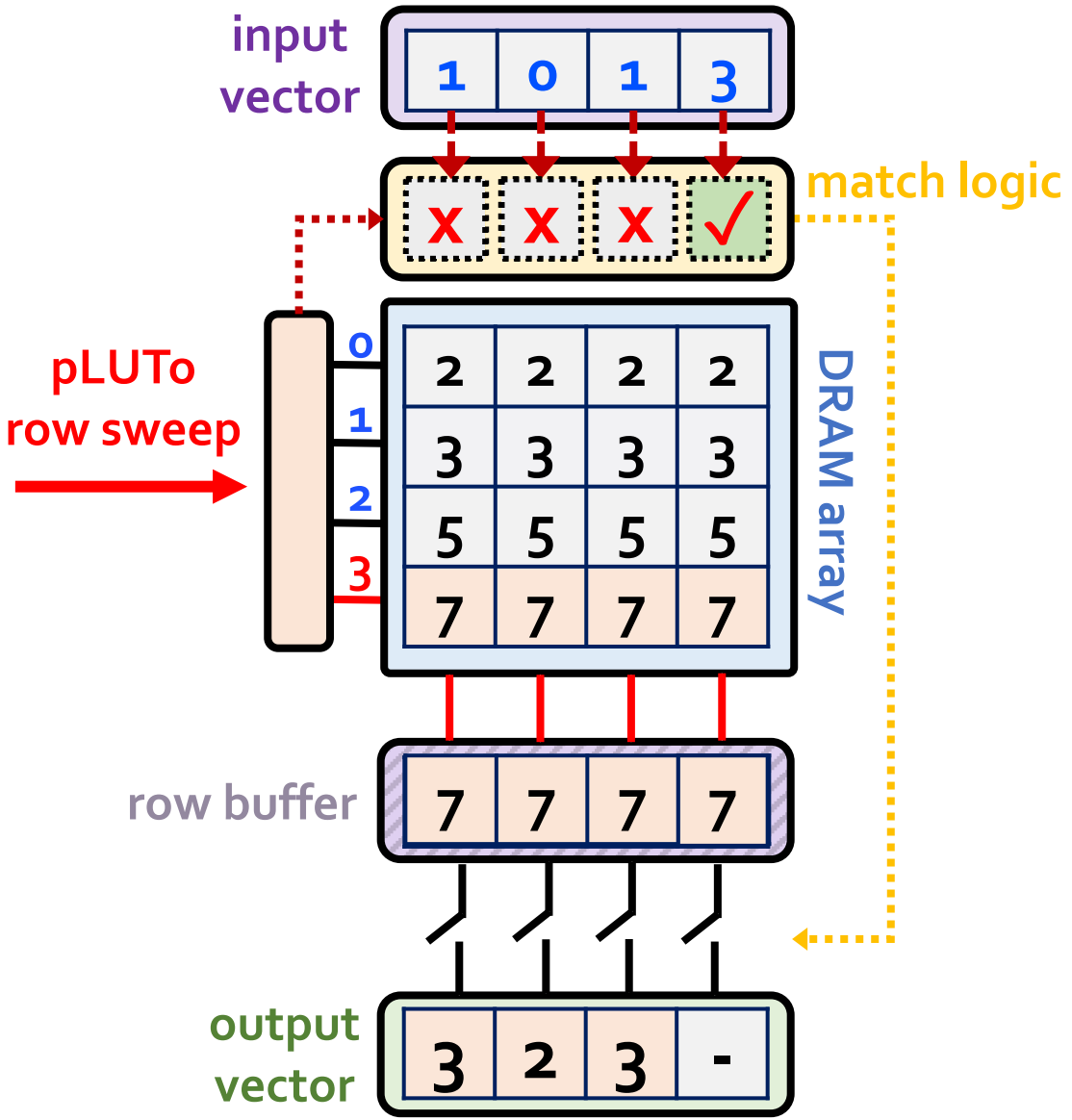
lookup table

1	0	1	3

input vector

3	2	3	7

output vector



In-DRAM pLUTo LUT Query: Step 4

LUT Query:
Return {2nd, 1st, 2nd, 4th}
prime numbers

LUT index	Prime numbers: i	Prime numbers: f(i)
0	1st	2
1	2nd	3
2	3rd	5
3	4th	7

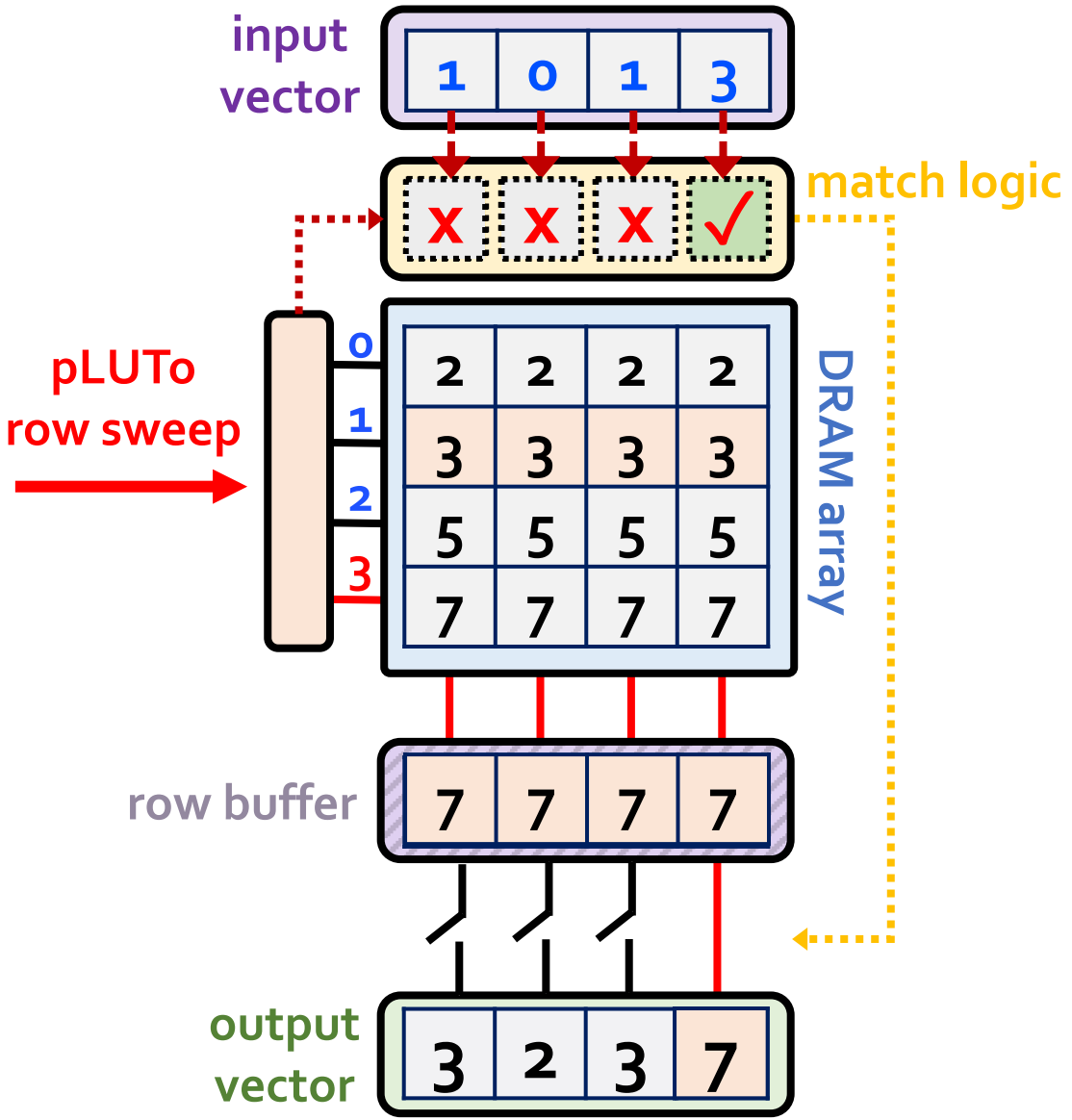
lookup table

1	0	1	3

input vector

3	2	3	7

output vector



Contents

Introduction

pLUTo

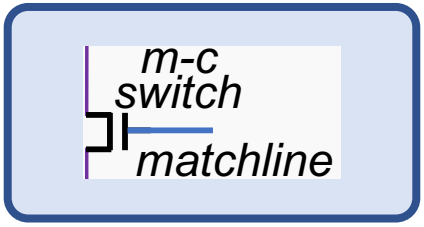
- Overview
- pLUTo Designs
- System Integration

Evaluation

Conclusion

pLUTo Designs

- **Match Logic:** shared by the three pLUTo designs



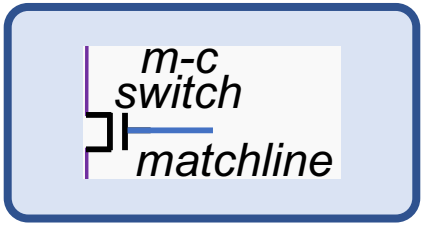
BSA
Buffered Sense Amplifier

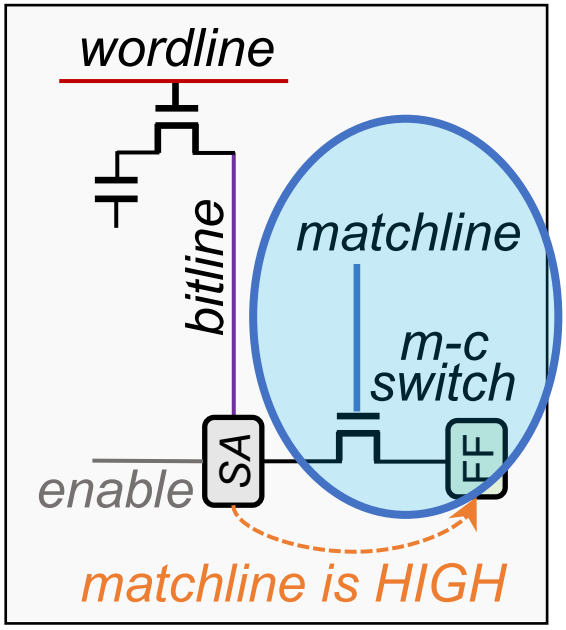
GSA
Gated Sense Amplifier

GMC
Gated Memory Cell

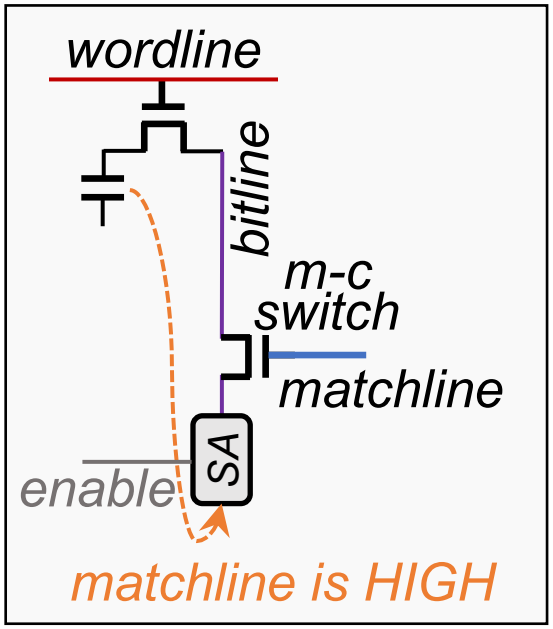
pLUTo Designs

- **Match Logic:** shared by the three pLUTo designs





BSA
Buffered Sense Amplifier

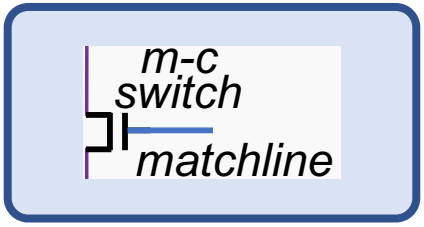


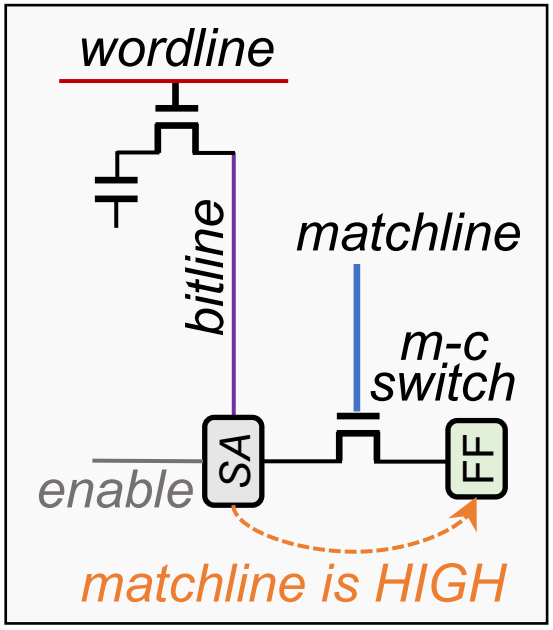
GSA
Gated Sense Amplifier

GMC
Gated Memory Cell

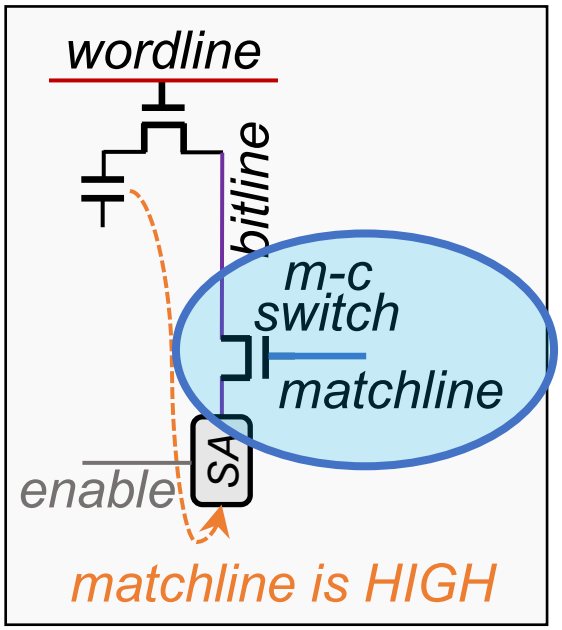
pLUTo Designs

- **Match Logic:** shared by the three pLUTo designs





BSA
Buffered Sense Amplifier

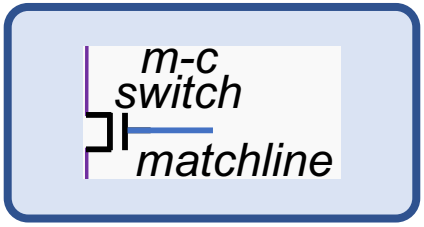


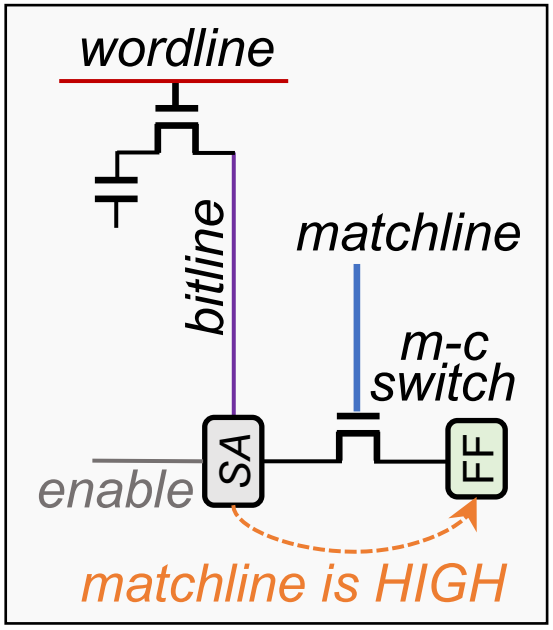
GSA
Gated Sense Amplifier

GMC
Gated Memory Cell

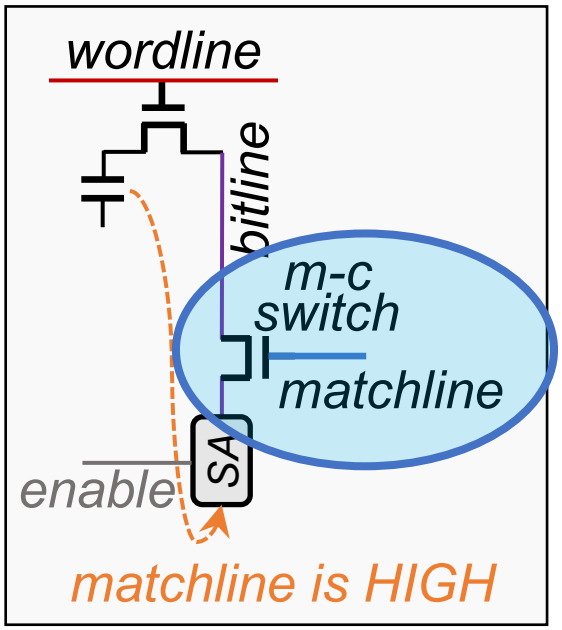
pLUTo Designs

- **Match Logic:** shared by the three pLUTo designs

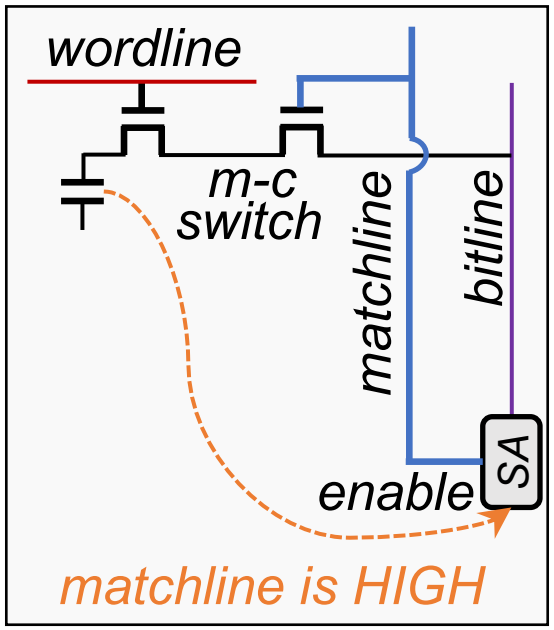




BSA
Buffered Sense Amplifier



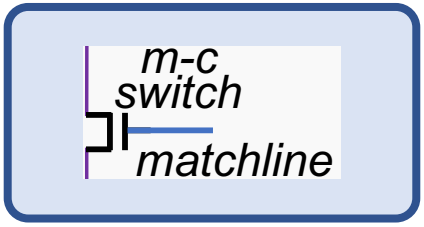
GSA
Gated Sense Amplifier

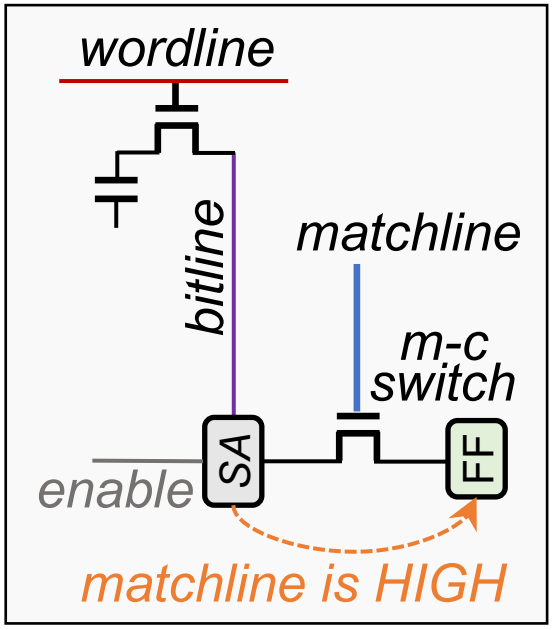


GMC
Gated Memory Cell

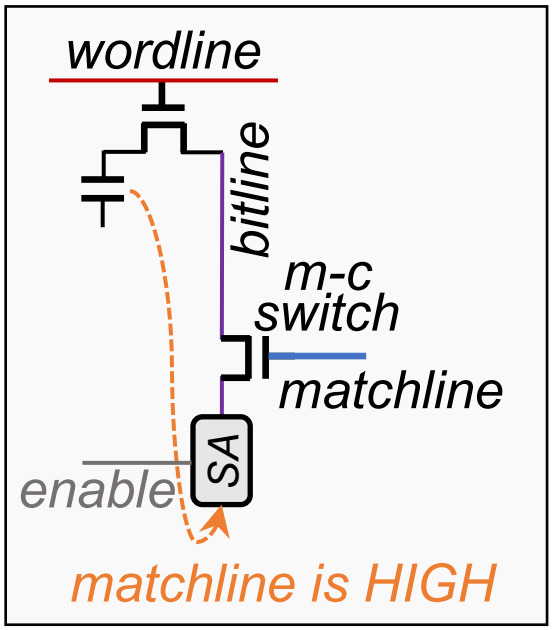
pLUTo Designs

- **Match Logic:** shared by the three pLUTo designs

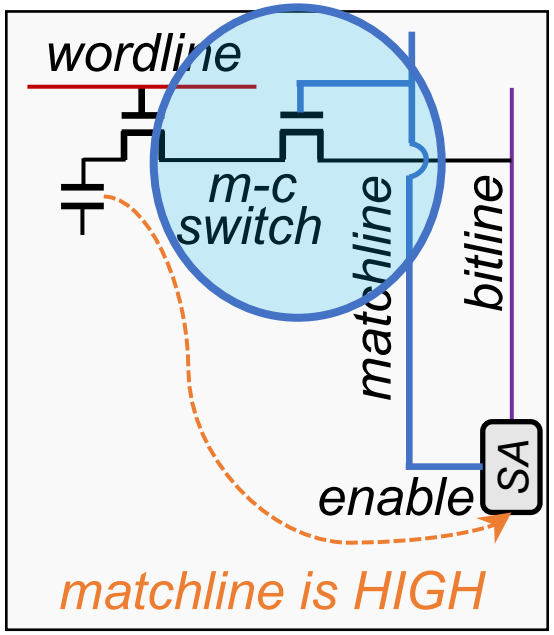




BSA
Buffered Sense Amplifier



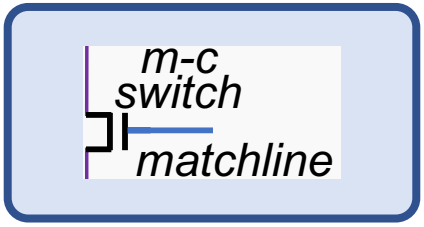
GSA
Gated Sense Amplifier

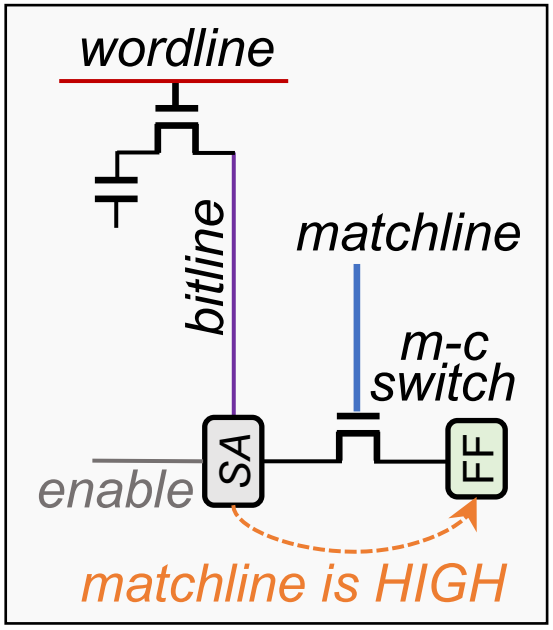


GMC
Gated Memory Cell

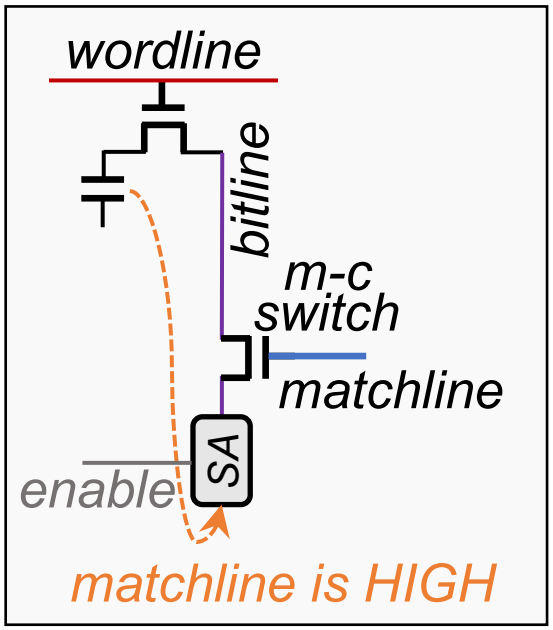
pLUTo Designs

- **Match Logic:** shared by the three pLUTo designs

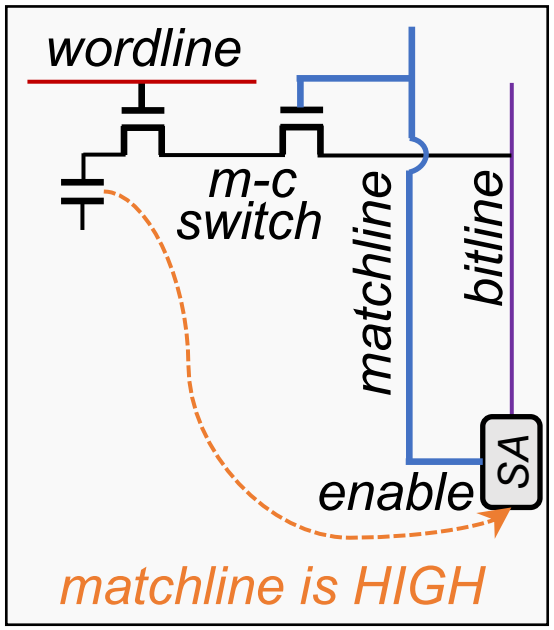




BSA
Buffered Sense Amplifier

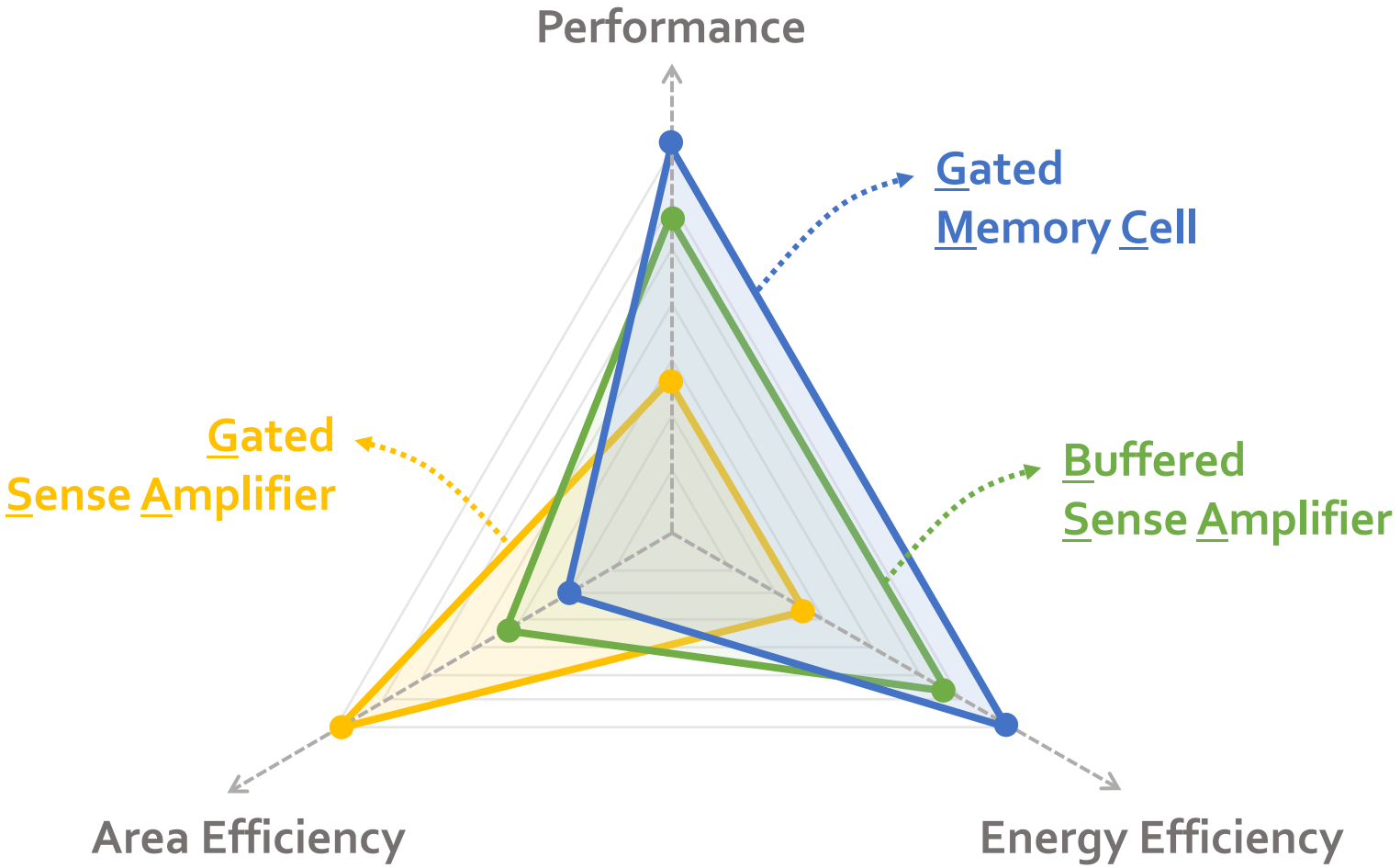


GSA
Gated Sense Amplifier



GMC
Gated Memory Cell

pLUTo Designs: Tradeoff Space



pLUTo designs cover a *broad design space* and provide *different* performance, energy, and area efficiency

Contents

Introduction

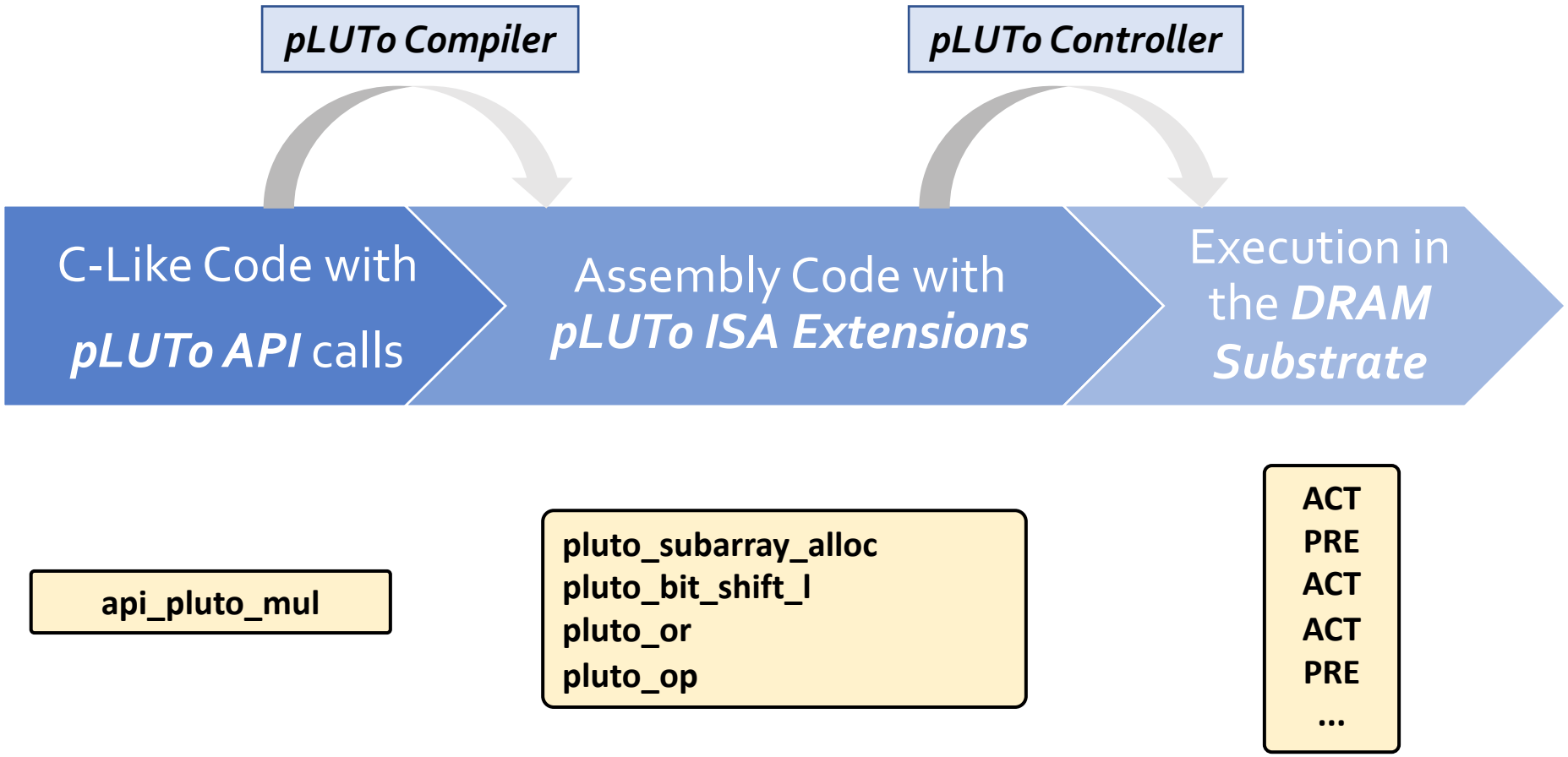
pLUTo

- Overview
- pLUTo Designs
- System Integration

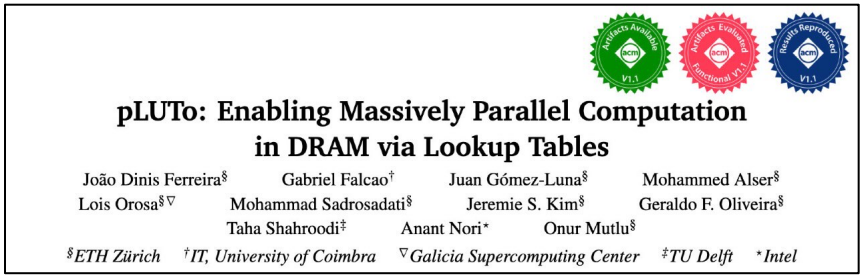
Evaluation

Conclusion

System Integration



More in the Paper



pLUTo: Enabling Massively Parallel Computation in DRAM via Lookup Tables

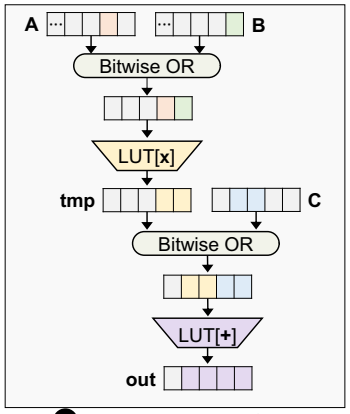
João Dinis Ferreira[§] Gabriel Falcao[†] Juan Gómez-Luna[§] Mohammed Alser[§]
Lois Orosa[§▽] Mohammad Sadrosadati[§] Jeremie S. Kim[§] Geraldo F. Oliveira[§]
Taha Shahroodi[‡] Anant Nori[⋆] Onur Mutlu[§]
[§]ETH Zürich [†]IT, University of Coimbra [▽]Galicia Supercomputing Center [‡]TU Delft [⋆]Intel

```c
uint2_t *A,*B,*C = (uint2_t *)malloc(input_size*2); // Inputs
uint4_t *out = (uint4_t *)malloc(input_size*4); // Output

// Array initialization
// ...
// Multiply-and-add loop
for (int i = 0; i < input_size; i++) {
      out[i] = A[i]*B[i] + C[i];
}
```

(a) Reference C Code

```
// Array allocation
uint2_t *A, *B   = pluto_malloc(size=input_size, bitwidth=2);
uint2_t *C, *tmp = pluto_malloc(size=input_size, bitwidth=4);
uint4_t *out     = pluto_malloc(size=input_size, bitwidth=5);

// Multiply-and-add loop
for (int i = 0; i < input_size/row_size; i++){
    api_pluto_mul(in1 = A, in2 = B,   out = tmp, bitwidth = 2);
    api_pluto_add(in1 = C, in2 = tmp, out = out, bitwidth = 4);
}
```

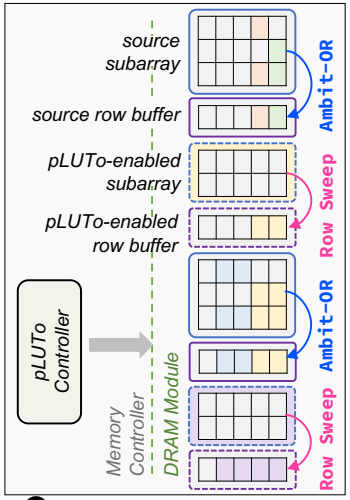
(b) pLUTo API Code



(d) Data Dependency Graph

```
# Array allocation
pluto_row_alloc $prg0, input_size, 2 # Allocate A
pluto_row_alloc $prg1, input_size, 2 # Allocate B
pluto_row_alloc $prg2, input_size, 4 # Allocate C
pluto_row_alloc $prg3, input_size, 4 # Allocate tmp
pluto_row_alloc $prg4, input_size, 5 # Allocate out

# Allocate and load LUTs
pluto_subarray_alloc $lut_rg0, "mul2_lut_file.dat"
pluto_subarray_alloc $lut_rg1, "add4_lut_file.dat"

# Allocate temporary row for OR operation
pluto_row_alloc $prg5, input_size, 8

# Multiply-and-add loop
div $r0, input_size, row_size # Initialize loop counter
LOOP:
    pluto_bit_shift_l $pgr0, 4 # Shift A 4 bits to the left
    pluto_or $prg5, $prg0, $prg1 # $prg5 <- A | B
    pluto_op $prg3, $prg5, $lut_rg0, 256, 4 # tmp <- LUT[A|B]
    pluto_bit_shift_l $pgr3, 4 # Shift tmp 4 bits to the left
    pluto_or $prg5, $prg3, $prg2 # $pgr5 <- tmp | C
    pluto_op $prg4, $prg5, $lut_rg1, 256, 8 # out <- LUT[tmp|C]
    # Update input addresses
    subi $r0, 0     # decrement loop counter
    bne $r0, LOOP   # next loop iteration
```

(c) pLUTo ISA Instructions



(e) pLUTo Controller & Execution

https://arxiv.org/pdf/2104.07699.pdf

Contents

Introduction

pLUTo

- Overview
- pLUTo Designs
- System Integration

Evaluation

Conclusion

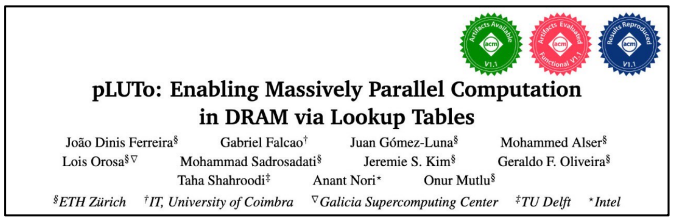
Methodology: Experimental Setup

- **Baselines**
  - **CPU** (Intel® Xeon Gold 5118)
  - **GPU** (NVIDIA® GeForce RTX 3080 Ti)
  - **Processing-near-Memory (PnM)**
- **pLUTo**
  - In-house simulator, open-sourced
  - https://github.com/CMU-SAFARI/pLUTo
- **We evaluate:**
  - Performance
  - Energy consumption
  - Area overhead
  - Circuit-level reliability and correctness
  - …

Methodology: Experimental Setup

- **Baselines**
  - **CPU** (Intel® Xeon Gold 5118)
  - **GPU** (NVIDIA® GeForce RTX 3080 Ti)
  - **Processing-near-Memory (PnM)**
- **pLUTo**
  - In-house simulator, open-sourced
  - https://github.com/CMU-SAFARI/pLUTo
- **We evaluate:**
  - **Performance**
  - **Energy consumption**
  - Area overhead
  - Circuit-level reliability and correctness
  - …

pLUTo SAFARI


pLUTo: Enabling Massively Parallel Computation in DRAM via Lookup Tables

João Dinis Ferreira[§] Gabriel Falcao[†] Juan Gómez-Luna[§] Mohammed Alser[§]
Lois Orosa[§▽] Mohammad Sadrosadati[§] Jeremie S. Kim[§] Geraldo F. Oliveira[§]
Taha Shahroodi[‡] Anant Nori[⋆] Onur Mutlu[§]

[§]ETH Zürich [†]IT, University of Coimbra [▽]Galicia Supercomputing Center [‡]TU Delft [⋆]Intel

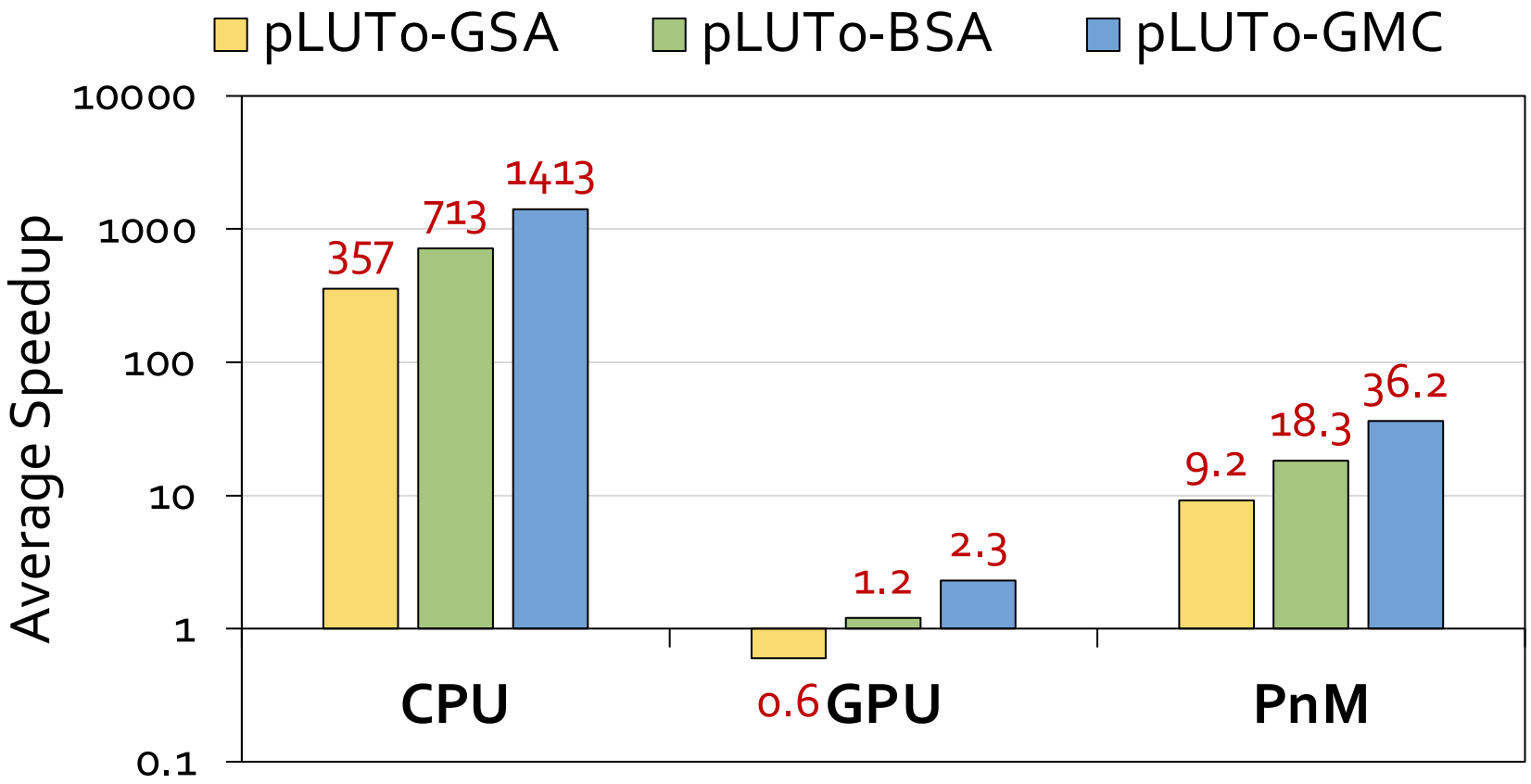
Methodology: Workloads

- **7 real-world workloads (not well supported by prior work)**
  - CRC-8/16/32
  - Salsa20
  - VMPC
  - Image Binarization
  - Color Grading

- **4 synthetic workloads (supported by prior work)**
  - Vector Addition
  - Vector Point-Wise Multiplication
  - Row-Level Bitwise Logic Operations
  - Bit Counting

Performance

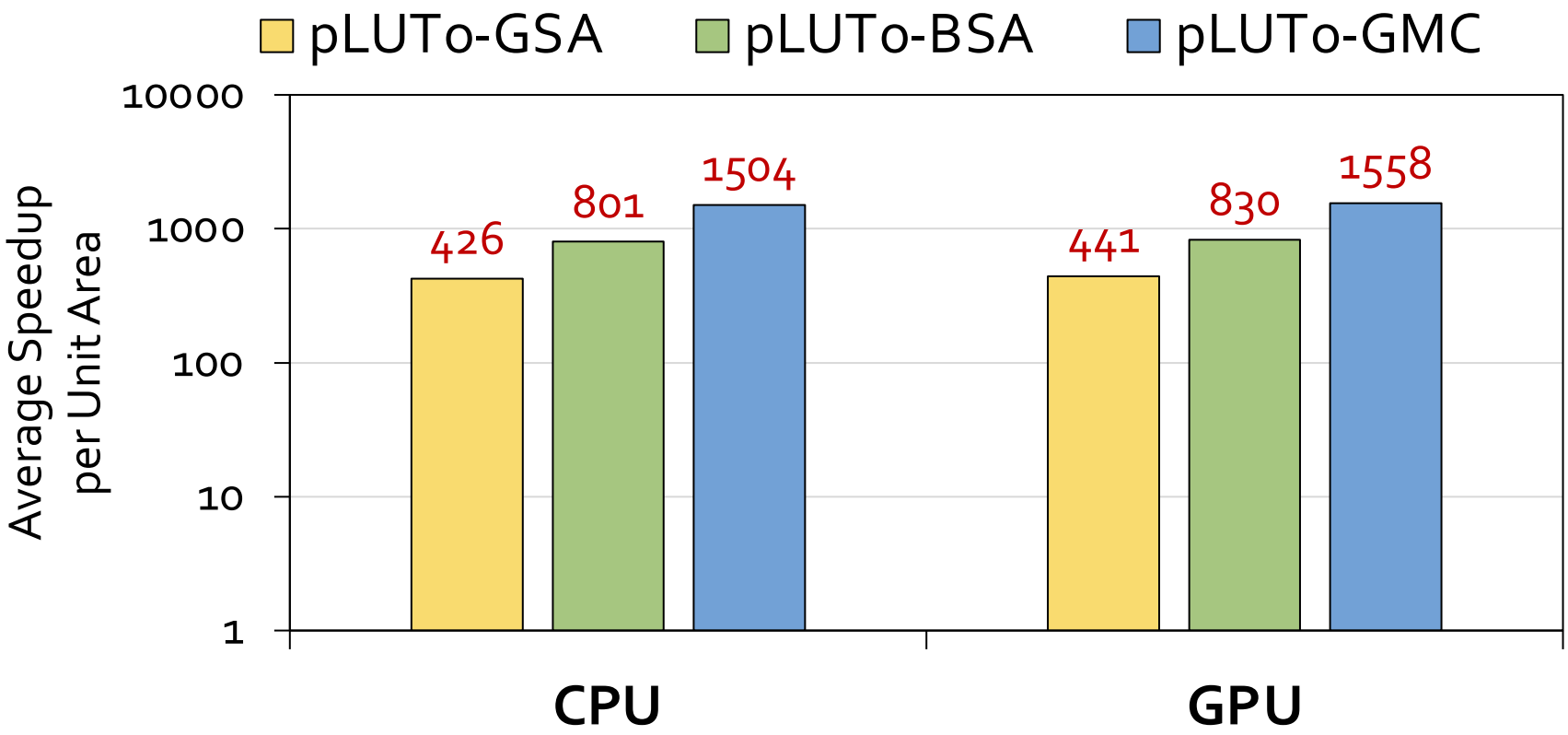
Average speedup across 7 real-world workloads



pLUTo *significantly outperforms* CPU, GPU and PnM baselines

Performance (normalized to area)

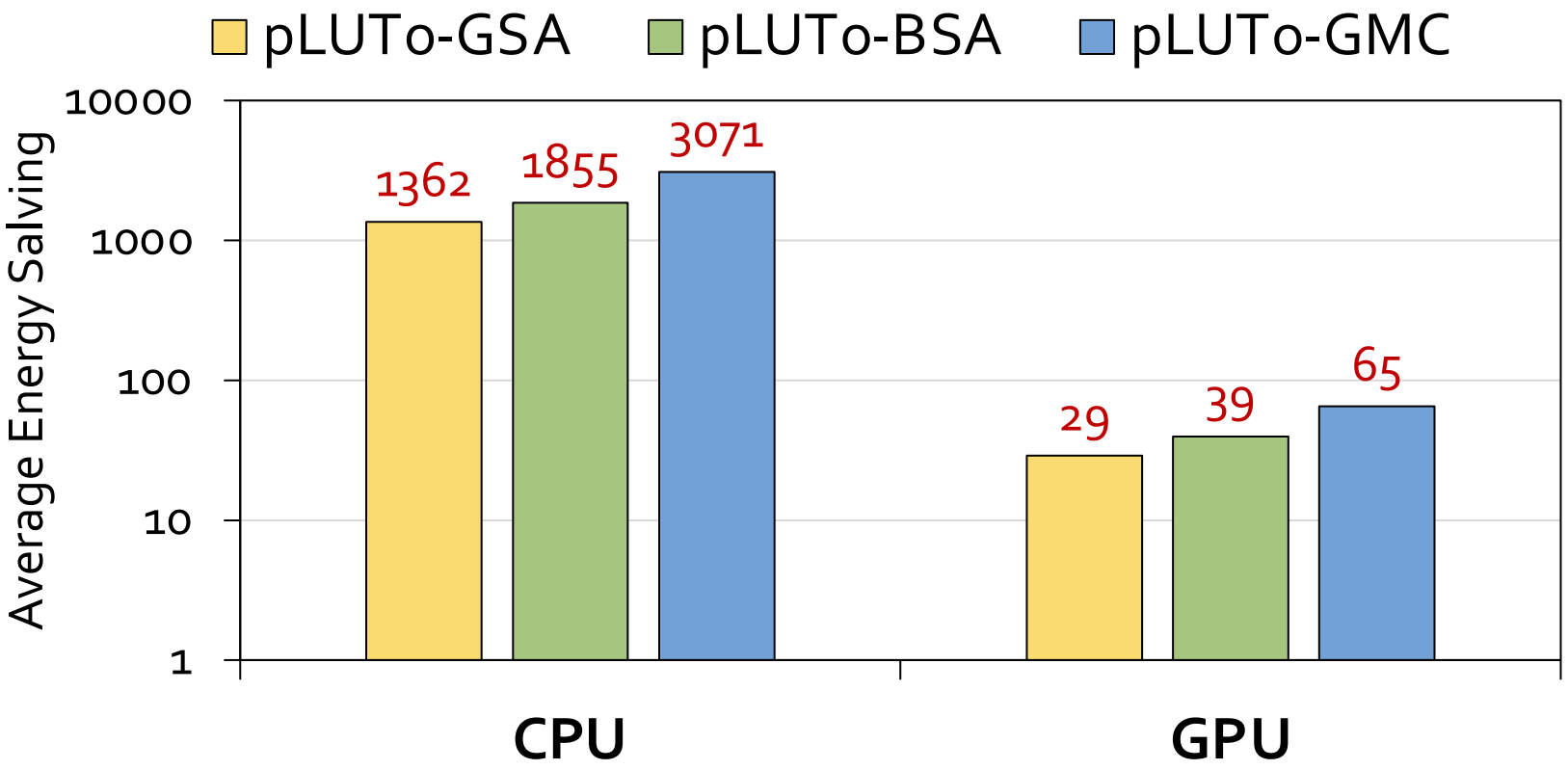
Average speedup normalized to area across 7 real-world workloads



pLUTo provides *substantially higher* performance per unit area than *both* the CPU and the GPU

Energy Consumption

Average energy consumption across 7 real-world workloads



pLUTo *significantly reduces energy consumption* compared to processor-centric architectures for various workloads

More Results in the Paper

- Comparison with FPGA
- Area Overhead Analysis
- Circuit-Level Reliability & Correctness
- Subarray-Level Parallelism
- LUT Loading Overhead
- Range of Supported Operations

pLUTo: Enabling Massively Parallel Computation in DRAM via Lookup Tables

João Dinis Ferreira[§] Gabriel Falcao[†] Juan Gómez-Luna[§] Mohammed Alser[§]
Lois Orosa[§▽] Mohammad Sadrosadati[§] Jeremie S. Kim[§] Geraldo F. Oliveira[§]
Taha Shahroodi[‡] Anant Nori[⋆] Onur Mutlu[§]

[§]*ETH Zürich* [†]*IT, University of Coimbra* [▽]*Galicia Supercomputing Center* [‡]*TU Delft* [⋆]*Intel*

Contents

Introduction

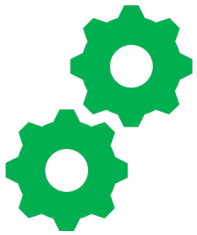
pLUTo

- Overview
- pLUTo Designs
- System Integration

Evaluation

Conclusion

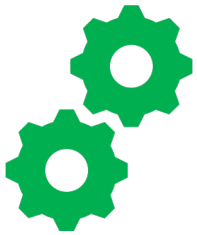
pLUTo Summary



The ***pLUTo Lookup Table Query*** introduces support for

the execution of arbitrarily complex operations in DRAM

pLUTo Summary

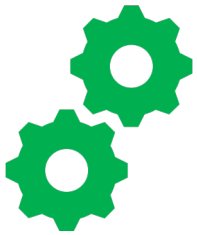


The *pLUTo Lookup Table Query* introduces support for **the execution of arbitrarily complex operations** in DRAM

Buffered Sense Amplifier, *Gated Sense Amplifier* & *Gated Memory Cell*

target different **performance/energy/area tradeoffs**

pLUTo Summary

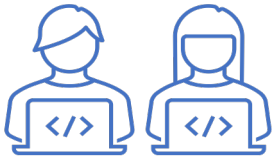


The *pLUTo Lookup Table Query* introduces support for

the execution of arbitrarily complex operations in DRAM

Buffered Sense Amplifier, *Gated Sense Amplifier* & *Gated Memory Cell*

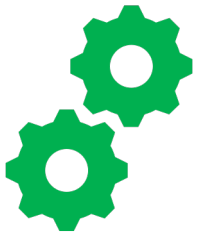
target different **performance/energy/area tradeoffs**



pLUTo API, *pLUTo Compiler* & *pLUTo Controller*

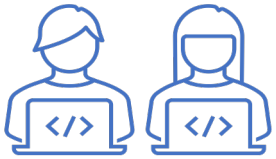
facilitate programmer adoption of **pLUTo**

pLUTo Summary



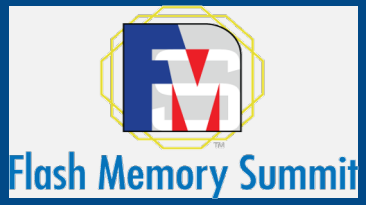
The *pLUTo Lookup Table Query* introduces support for **the execution of arbitrarily complex operations** in DRAM

Buffered Sense Amplifier, *Gated Sense Amplifier* & *Gated Memory Cell* target different **performance/energy/area tradeoffs**



pLUTo API, *pLUTo Compiler* & *pLUTo Controller* **facilitate programmer adoption** of **pLUTo**

pLUTo greatly outperforms the CPU/GPU/PnM baselines in **performance and energy** while incurring **small area overheads**



pLUTo

Enabling Massively Parallel Computation in DRAM via Lookup Tables

João Dinis Ferreira Gabriel Falcao **Juan Gómez-Luna** Mohammed Alser

Lois Orosa Mohammad Sadrosadati Jeremie S. Kim Geraldo F. Oliveira

Taha Shahroodi Anant Nori Onur Mutlu

pLUTo

Enabling Massively Parallel Computation
in DRAM via Lookup Tables

pLUTo Summary

The *pLUTo Lookup Table Query* introduces support for the execution of arbitrarily complex operations in DRAM

Buffered Sense Amplifier, Gated Sense Amplifier & Gated Memory Cell target different performance/energy/area tradeoffs

pLUTo API, pLUTo Compiler & pLUTo Controller facilitate programmer adoption of pLUTo

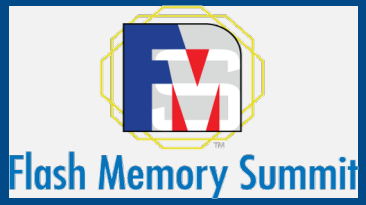
pLUTo greatly outperforms the CPU/GPU/PnM baselines in performance and energy while incurring small area overheads

SAFARI

85

arXiv

GitHub



pLUTo

Enabling Massively Parallel Computation in DRAM via Lookup Tables

João Dinis Ferreira Gabriel Falcao **Juan Gómez-Luna** Mohammed Alser

Lois Orosa Mohammad Sadrosadati Jeremie S. Kim Geraldo F. Oliveira

Taha Shahroodi Anant Nori Onur Mutlu