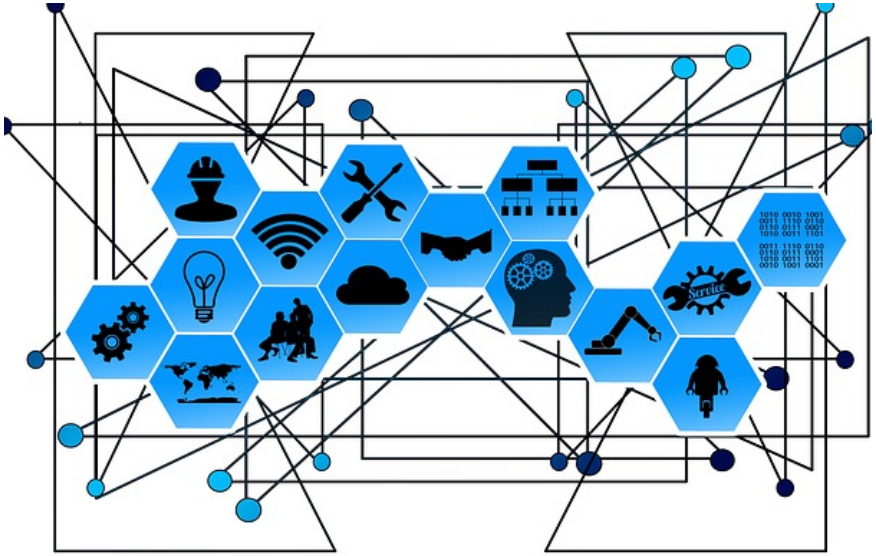
Refinements in Data Manipulation Method for Coarse Grained Reconfigurable Architectures

Takuya Kojima and Hideharu Amano
Keio University, Japan

14th International Symposium on Reconfigurable Communication-centric Systems-on-Chip (ReCoSoC 2019)

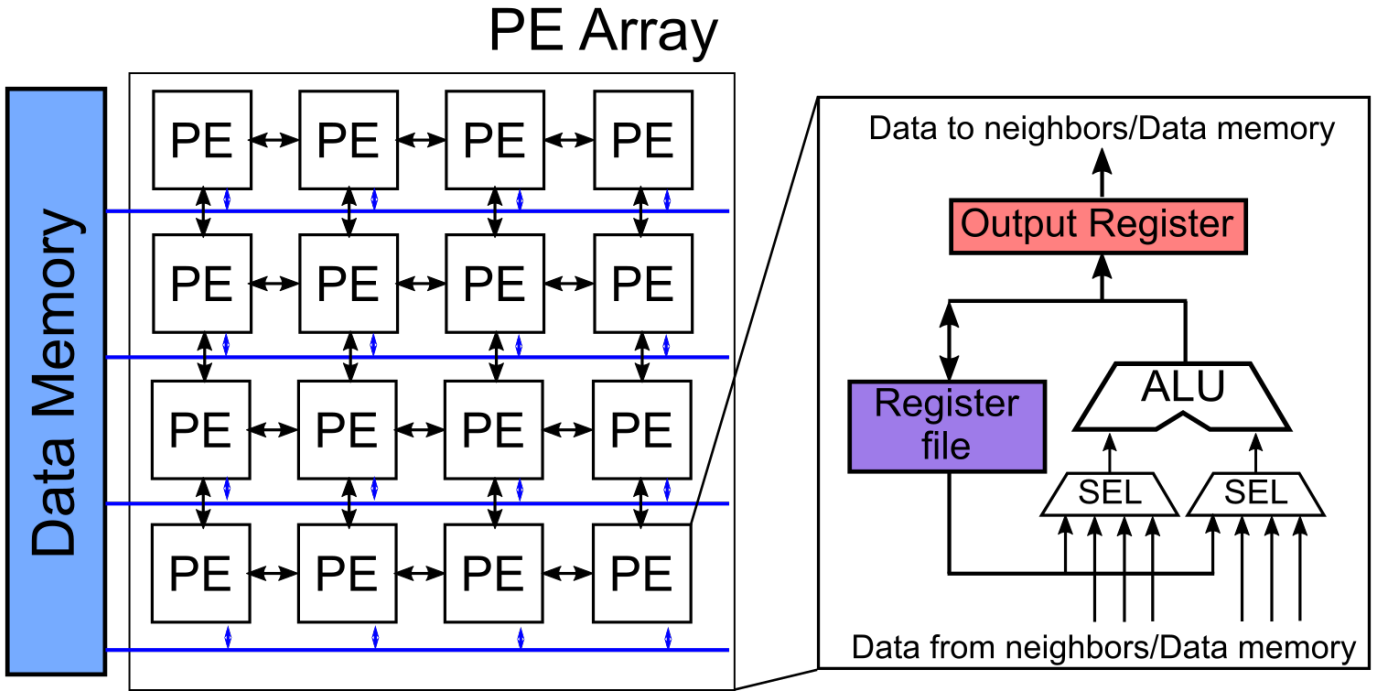
Importance of Programmability and High Energy Efficiency

- Forthcoming
  - IoT devices
  - Wearable computers
  - Edge computing



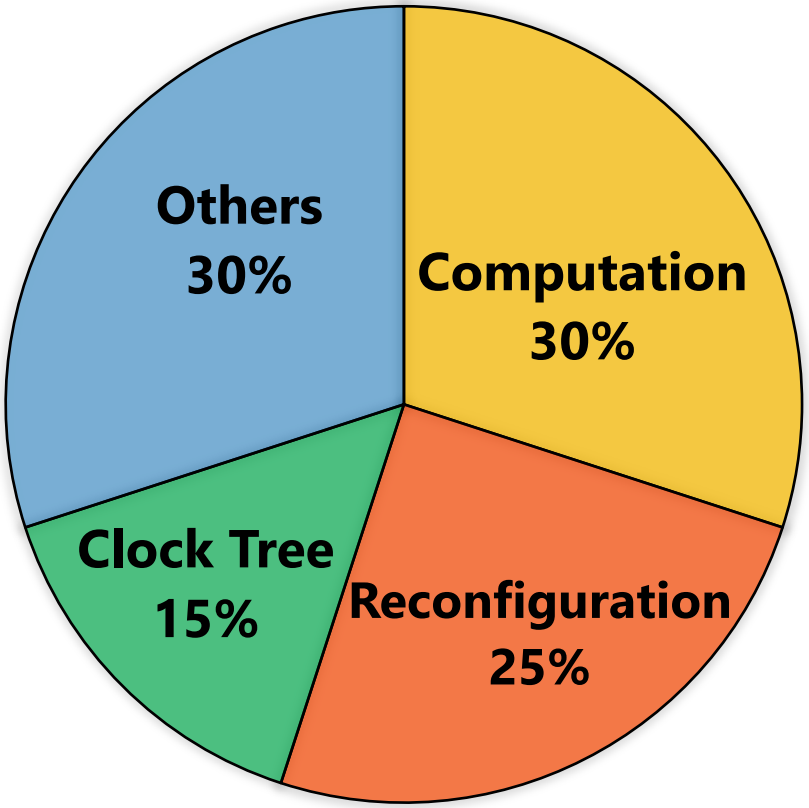
- Challenges for these devices
  - Programmability
    - To satisfy various demands
  - High energy efficiency
    - To extends long battery life

CGRAs: Coarse-Grained Reconfigurable Architectures



- CGRAs
  - Support word-level reconfiguration (↔ bit-level of FPGAs)
  - Have many PEs (Processing Element) in 2D grid
  - Change functionality for each ALU & interconnection between PEs **dynamically** or **statically**

Power-hungy Dynamic Reconfiguration

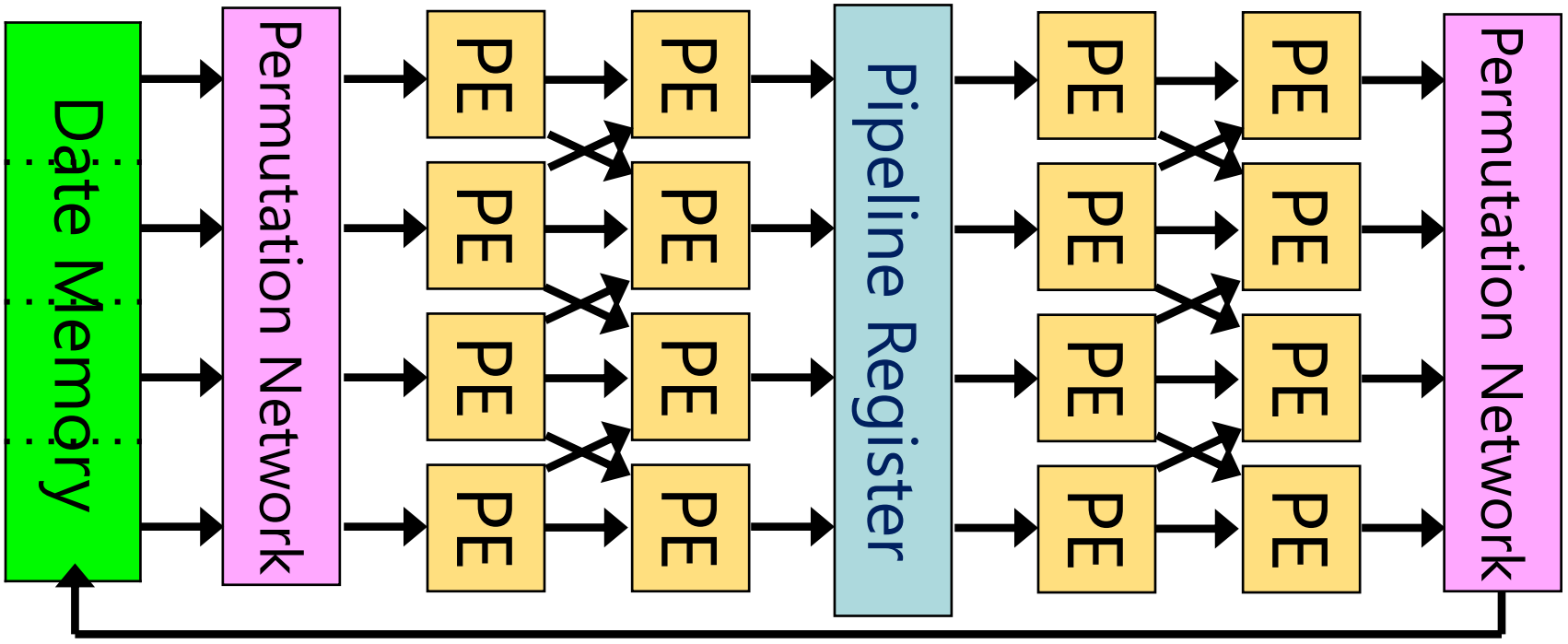


Details of power consumption for a dynamic reconfiguration CGRA[1]

- Dynamic Reconfiguration
  - Changes configuration **cycle-by-cycle**
  - Provides more flexibility
  - Causes large dynamic power consumption

[1] Ozaki, Nobuaki, et al. "Cool mega-arrays: Ultralow-power reconfigurable accelerator chips." IEEE Micro 31.6 (2011): 6-18.

SF-CGRAs: Straight-Forward CGRAs

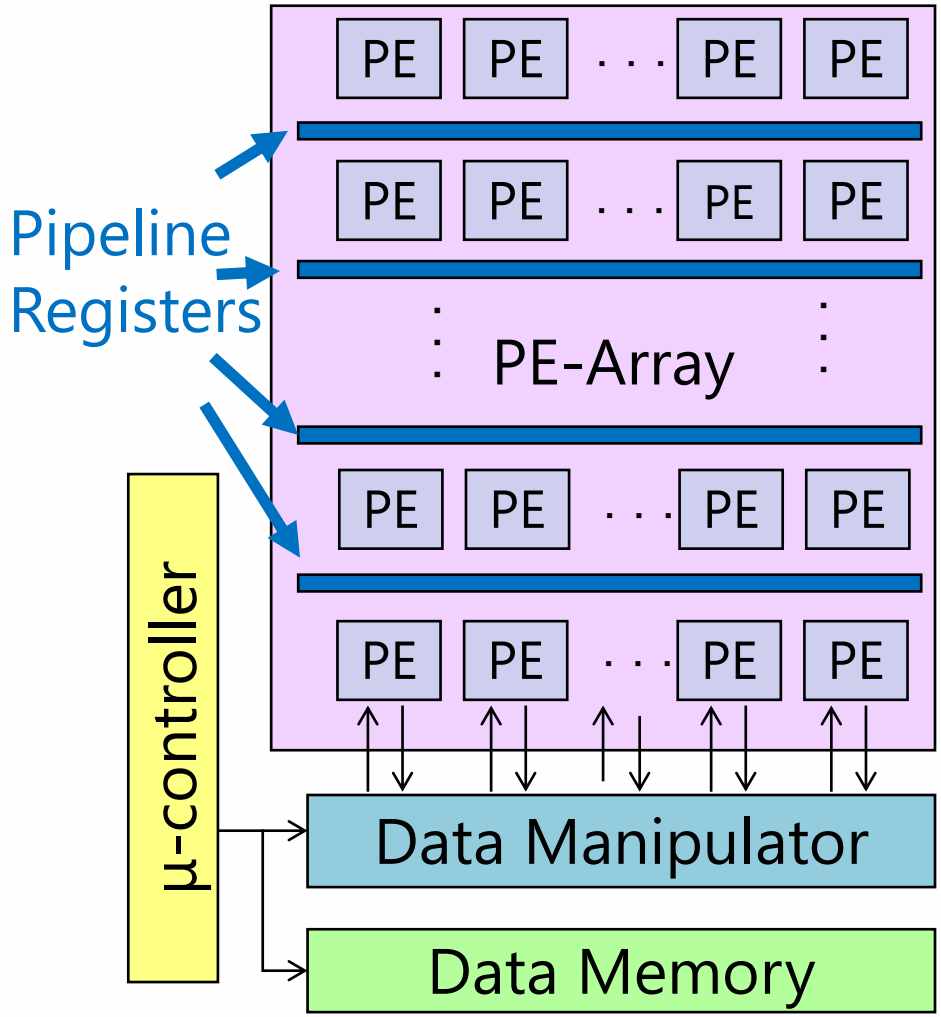


- Key features of straight-forward CGRAs
  - Limited data flow direction
  - Pipelined PE array
  - Less frequent reconfiguration
  - High energy efficiency
    - Piperench[2]
    - XPP[3]
    - EGRA[4]
    - RSPA[5]

[2] H. Schmit, *et al*, CICC 2002 [3] M.Petrov, *et al,* FPL 2004 [4] G. Ansaloni, *et al*, TVLSI 2011

[5] Yoon, Jonghee W., *et al*. *ASP-DAC*, 2008.

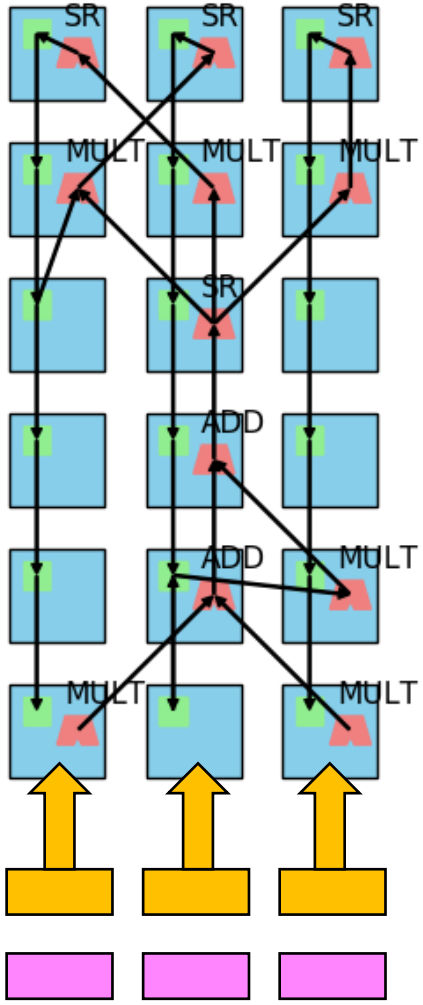
VPCMA: Variable Pipelined Cool Mega Array [2]



- PE array consists of
  - 8 x 12 PEs
  - 7 pipeline registers
- PE has
  - No Register file
  - No clock tree
- Pipeline register works in
  1. latch mode or
  2. bypass mode
- µ-Controller
  - Controls data transfer data mem. ↔ PE array

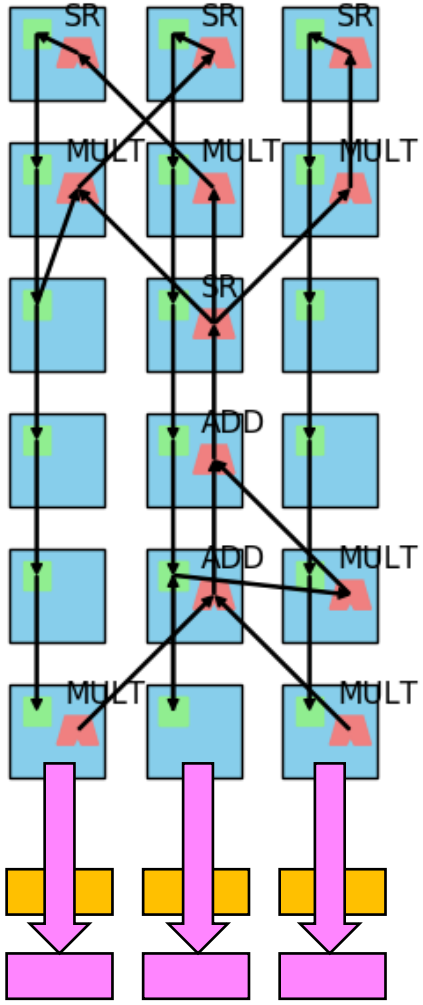
[2] N.Ando, *et al.* "Variable pipeline structure for Coarse Grained Reconfigurable Array CMA." *Field-Programmable Technology*, 2016.

Computation on the PE array



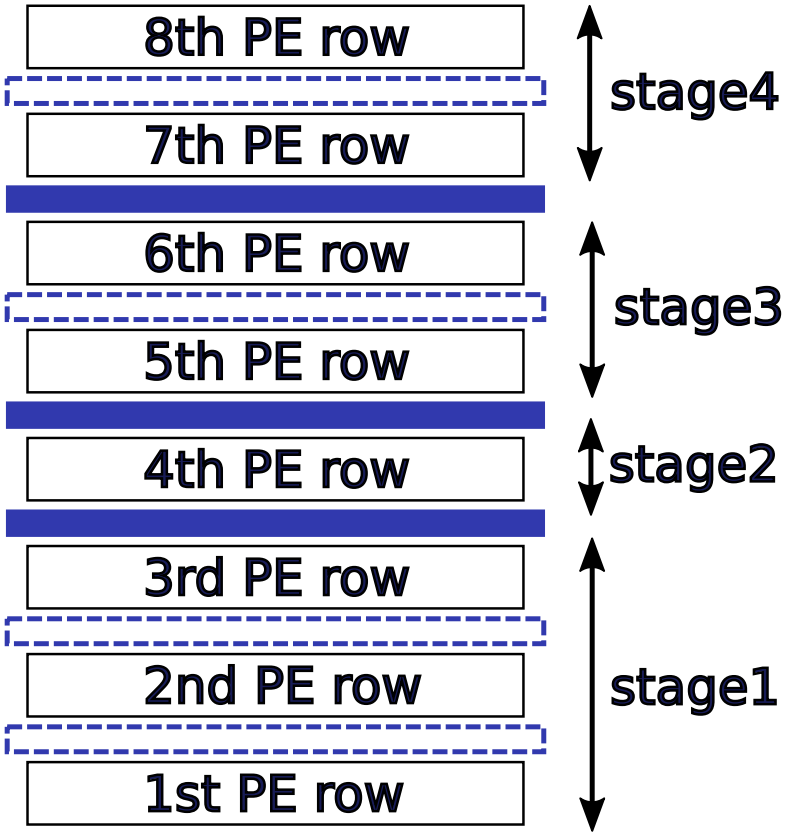
- Fetch registers are connected to input of the PE array
- Gather registers are connected to output of the PE array
- The micro-controller
  - Writes data to the fetch registers
  - Read result from the gather registers

Computation on the PE array



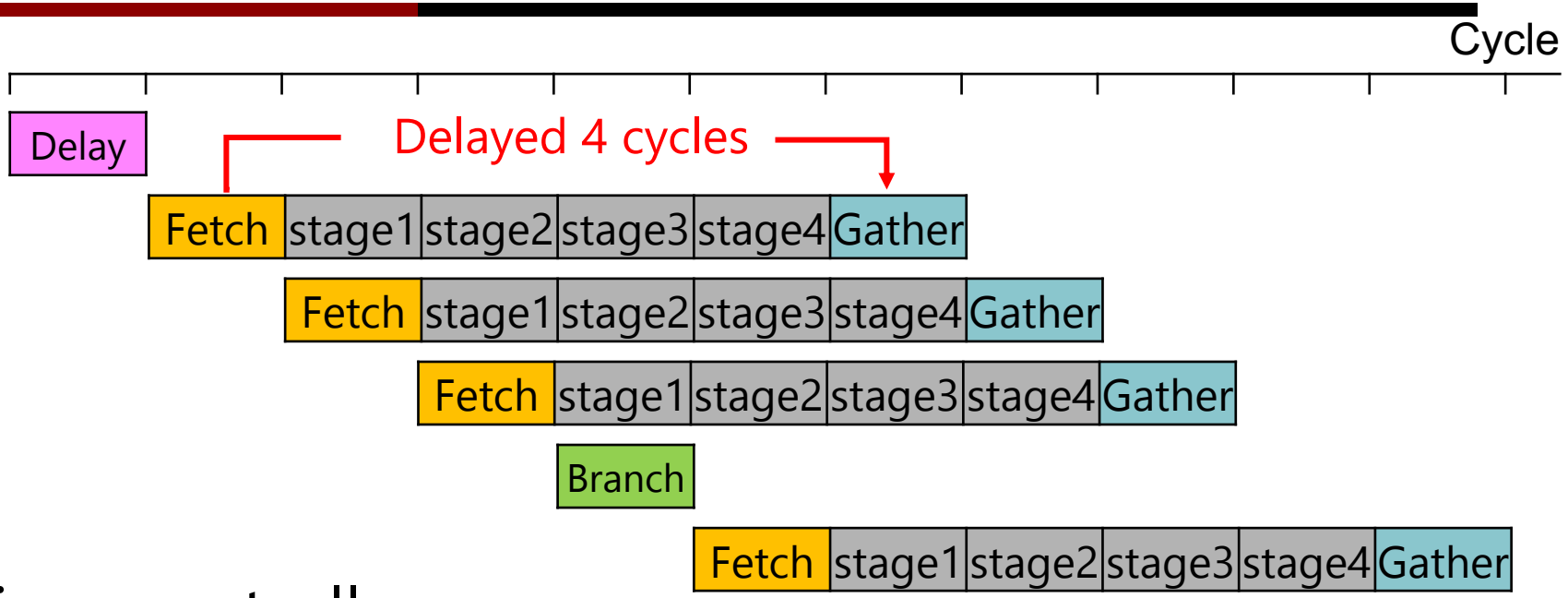
- Fetch registers are connected to input of the PE array
- Gather registers are connected to output of the PE array
- The micro-controller
  - Writes data to the fetch registers
  - Read result from the gather registers

Variable Pipeline Structure



- No registers in each pipeline stage
  → Pure combinational circuit
- Clock tree only for activated pipeline registers
- Variable pipeline structure depending on application

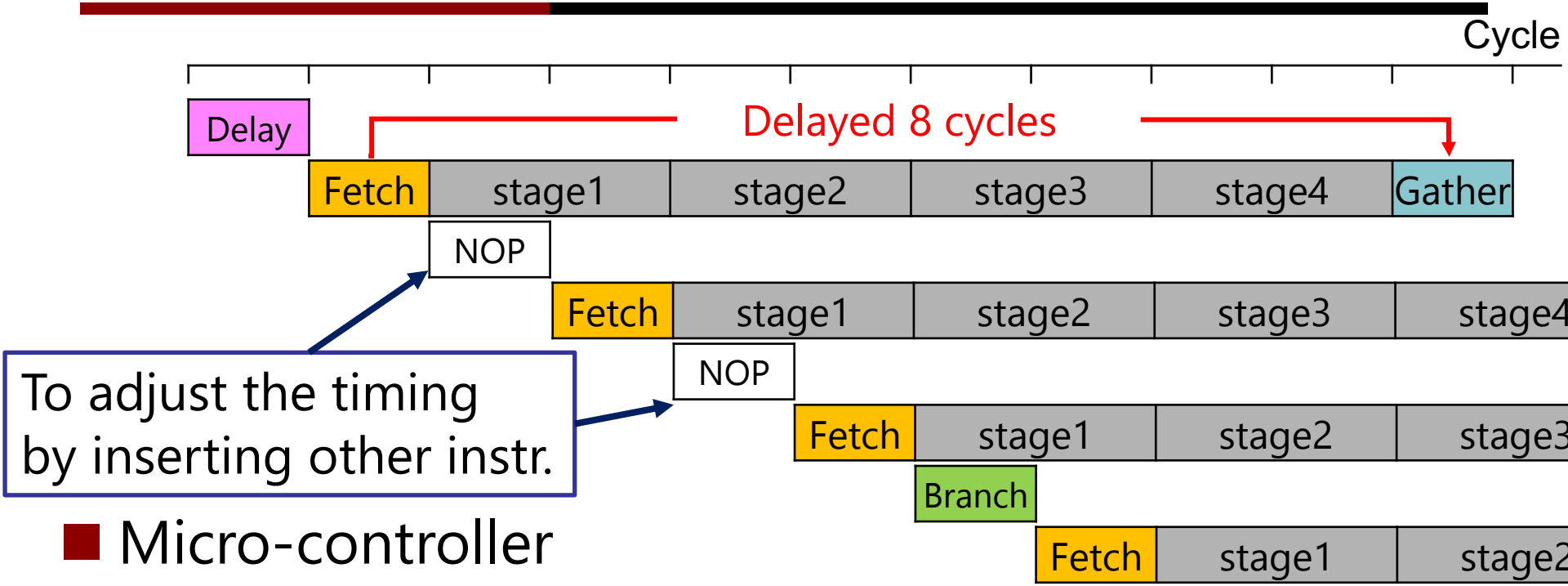
Multi-cycle Execution on PE Array



- Micro-controller
  - A custom tiny RISC processor controls the processing
  - "Fetch" op kicks off the execution
  - "Gather" op writes back the results
  - "Delay" op specifies delay time of "Gather" execution
  - "Branch" op makes a loop

Fused into an instruction

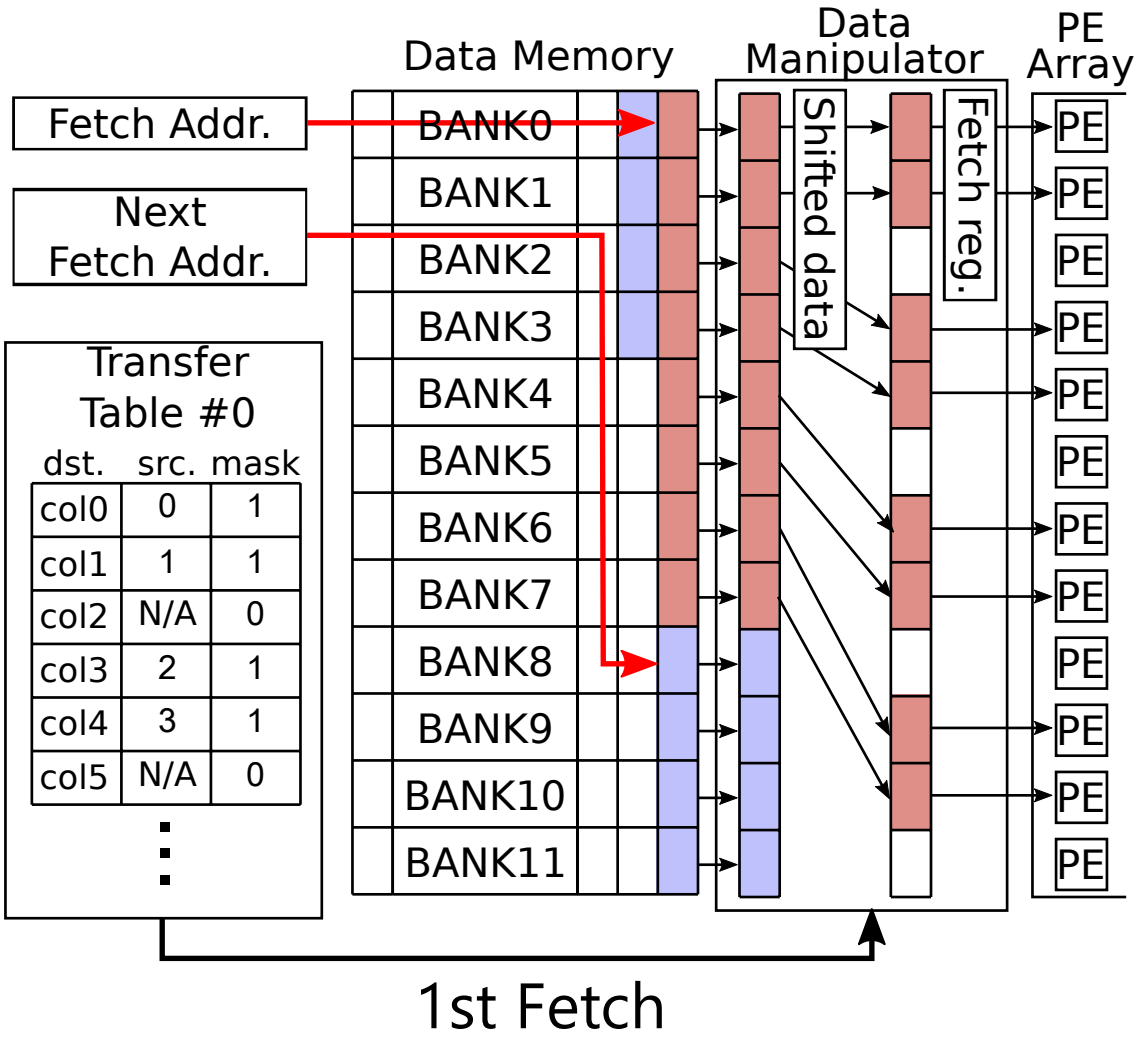
Multi-cycle Execution on PE Array



- Micro-controller
  - A custom tiny RISC processor controls the processing
  - "Fetch" op kicks off the execution
  - "Gather" op writes back the results

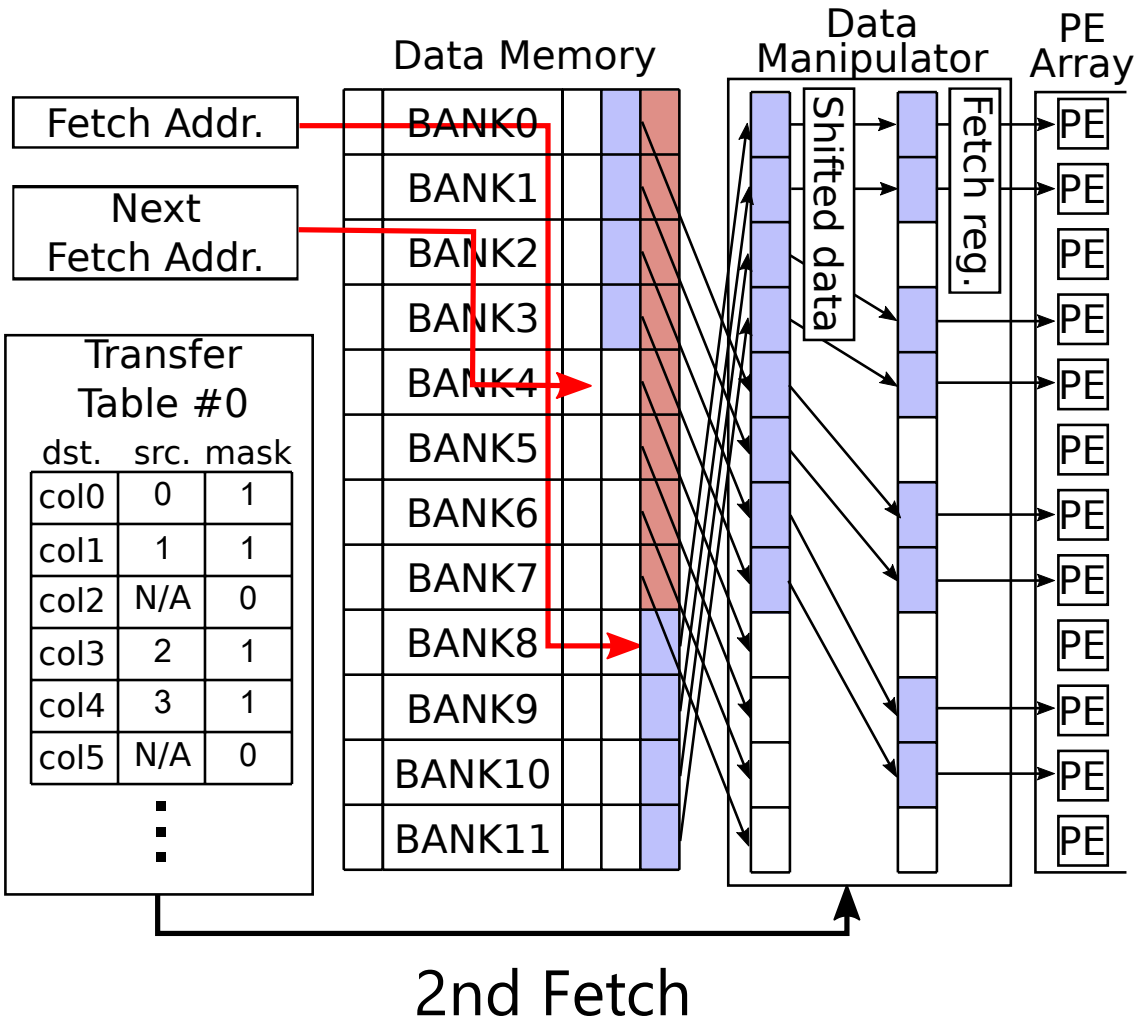
  Fused into an instruction
  - "Delay" op specifies delay time of "Gather" execution
  - "Branch" op makes a loop

Data Manipulator of VPCMA



- Data manipulator
  - Placed between Dmem & PE array
  - Transfers any input data to any outputs
  - Loads at most consecutive 12 data from 12 mem banks
  - Increments addr. automatically for next fetch

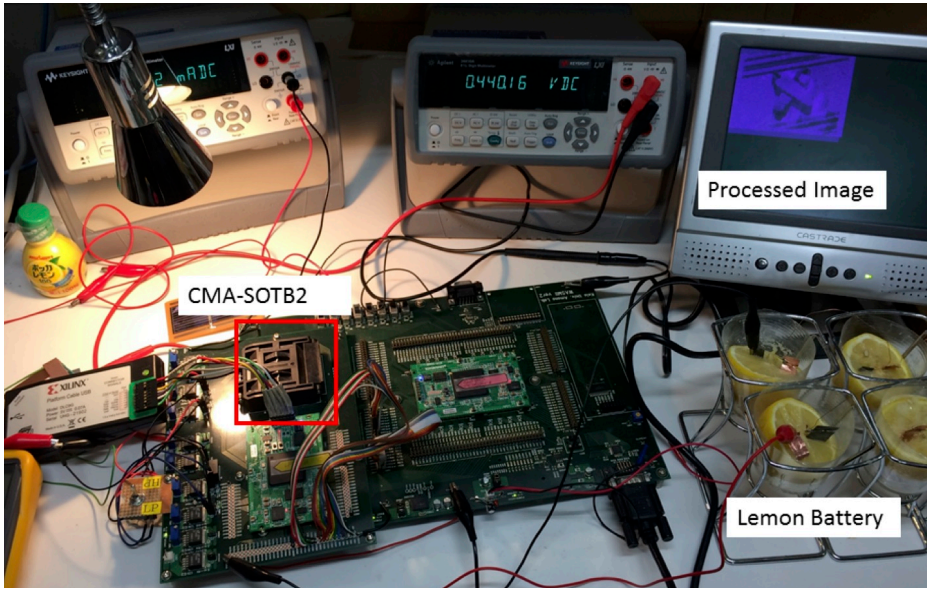
Data Manipulator of VPCMA



- Data manipulator
  - Placed between Dmem & PE array
  - Transfers any input data to any outputs
  - Loads at most consecutive 12 data from 12 mem banks
  - Increments addr. automatically for next fetch

Ultra Low Power Consumption of CMA

- No-Pipelined version of CMA[6]
  - Works with Lemon battery
  - Achieves 743 MOPS/mW (297MOPS/0.4mW)



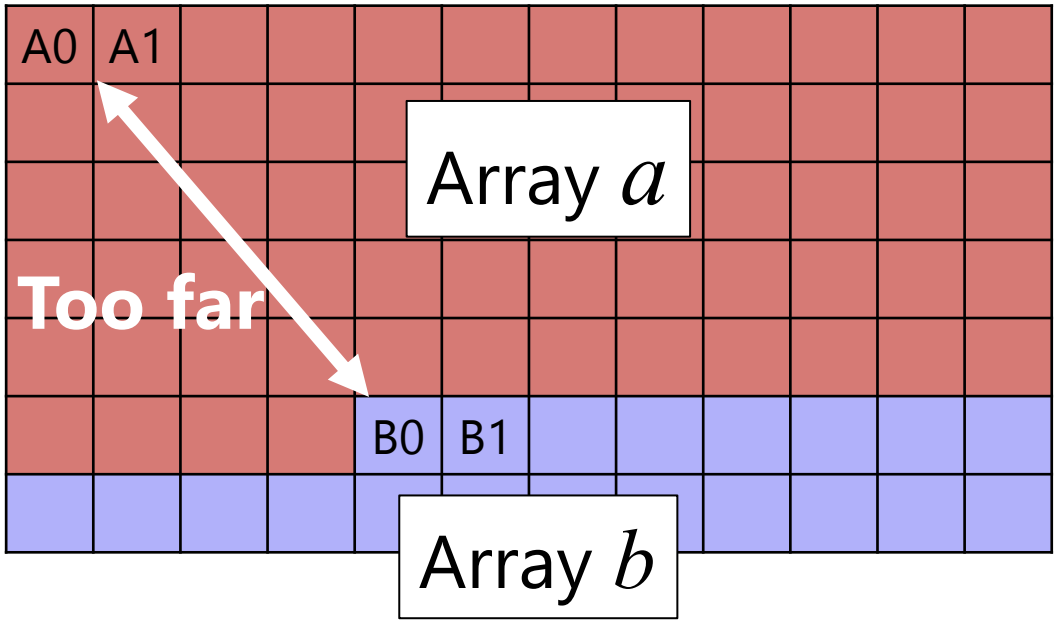
- VPCMA
  - Keeps the same energy efficiency
  - Achieves 4x higher peek performance
- Problem
  - Less flexibility because of saving too much energy

[6] M.Koichiro, *et al.* "A 297mops/0.4 mw ultra low power coarse-grained reconfigurable accelerator CMA-SOTB-2." *2015 International Conference on ReConFigurable Computing and FPGAs (ReConFig)*

Limitation of data handling in VPCMA

for $i \leftarrow 0$ to $N = 64$ **do**
  $c[i] \leftarrow a[i] \times b[i \bmod 16]$
end for

Loop example



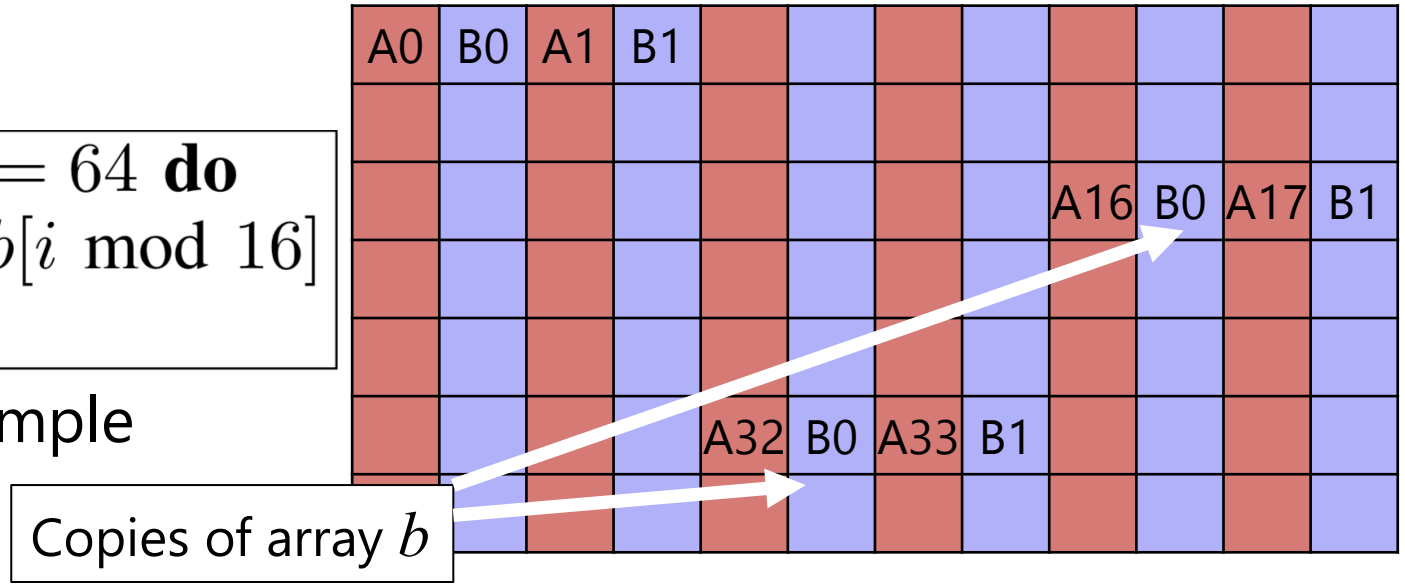
Memory allocation in bank memory

- Data manipulator cannot access multiple data more than 12 step distance simultaneously
  → needs data rearrangement
  → often incurs extra copy of data

Limitation of data handling in VPCMA

for $i \leftarrow 0$ to $N = 64$ **do**
  $c[i] \leftarrow a[i] \times b[i \bmod 16]$
end for

Loop example



Memory allocation in bank memory

- Data manipulator cannot access multiple data more than 12 distance simultaneously
  → needs data rearrangement
  → often incurs extra copy of data

Other limitations of VPCMA

- Also, VPCMA
  1. Suffers from a lack of constant registers for the PE array
     - A PE row (12 PEs) share two const regs.
       or borrows from other rows via interconnection

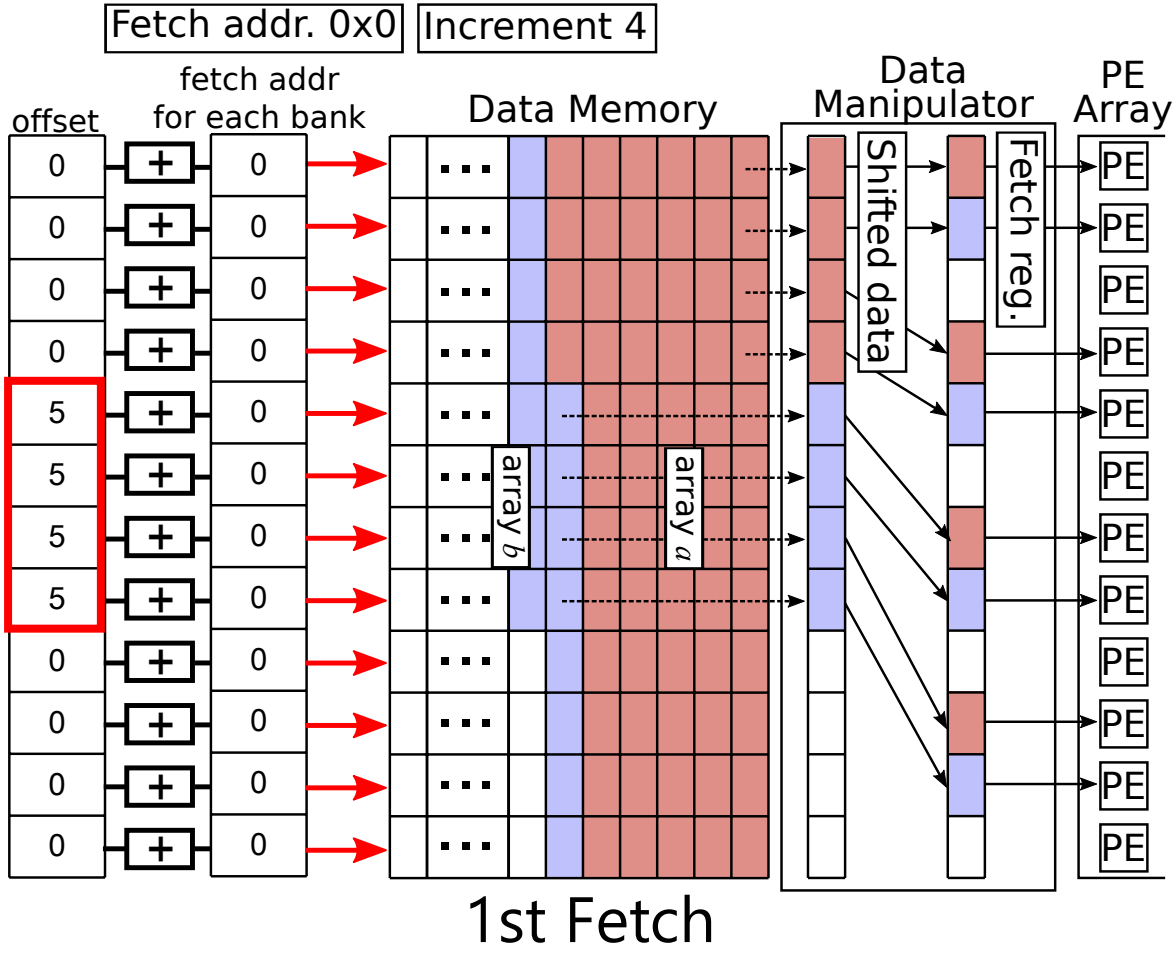
  → Degrades mappability of complex kernels

  2. Depends on a host processor for overall control
     - Micro-controller basically controls data transfer & loop counter
     - All of other controls (e.g. reconfiguration) are carried out by the host processor
       even if trivial change is needed

Proposed architecture

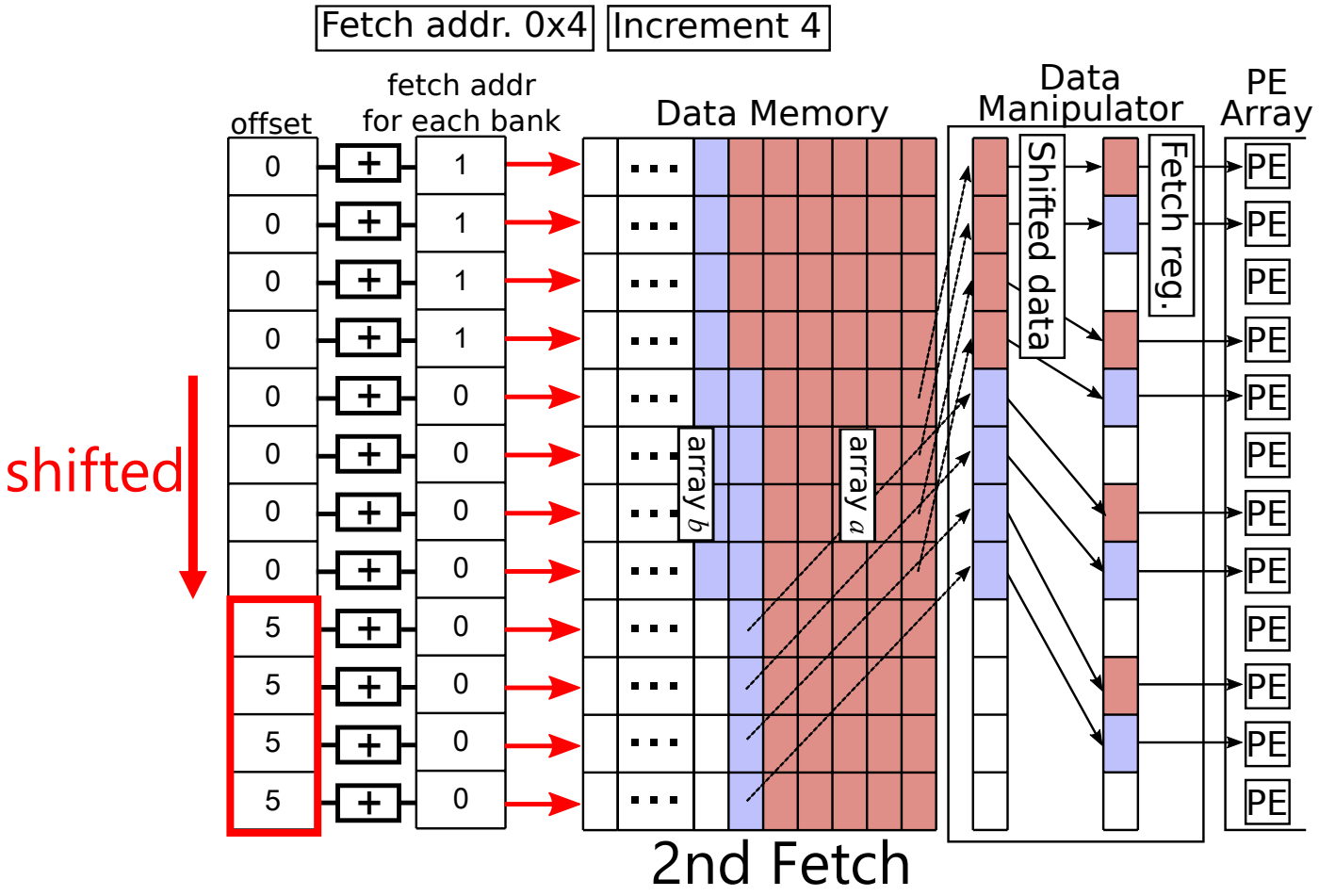
- A new architecture VPCMA2
  - Relaxing aforementioned limitations
  1. Improved bank access by new data manipulator
  2. Refined connectivity of constant registers
     - PE array has 16 constant registers (same as VPCMA)
     - All PE can use any 16 registers
  3. Introduced an extended data bus for micro-controller

New Data Manipulator



- Offset values for each bank is introduced
- Relaxed the limitation of consecutive data access

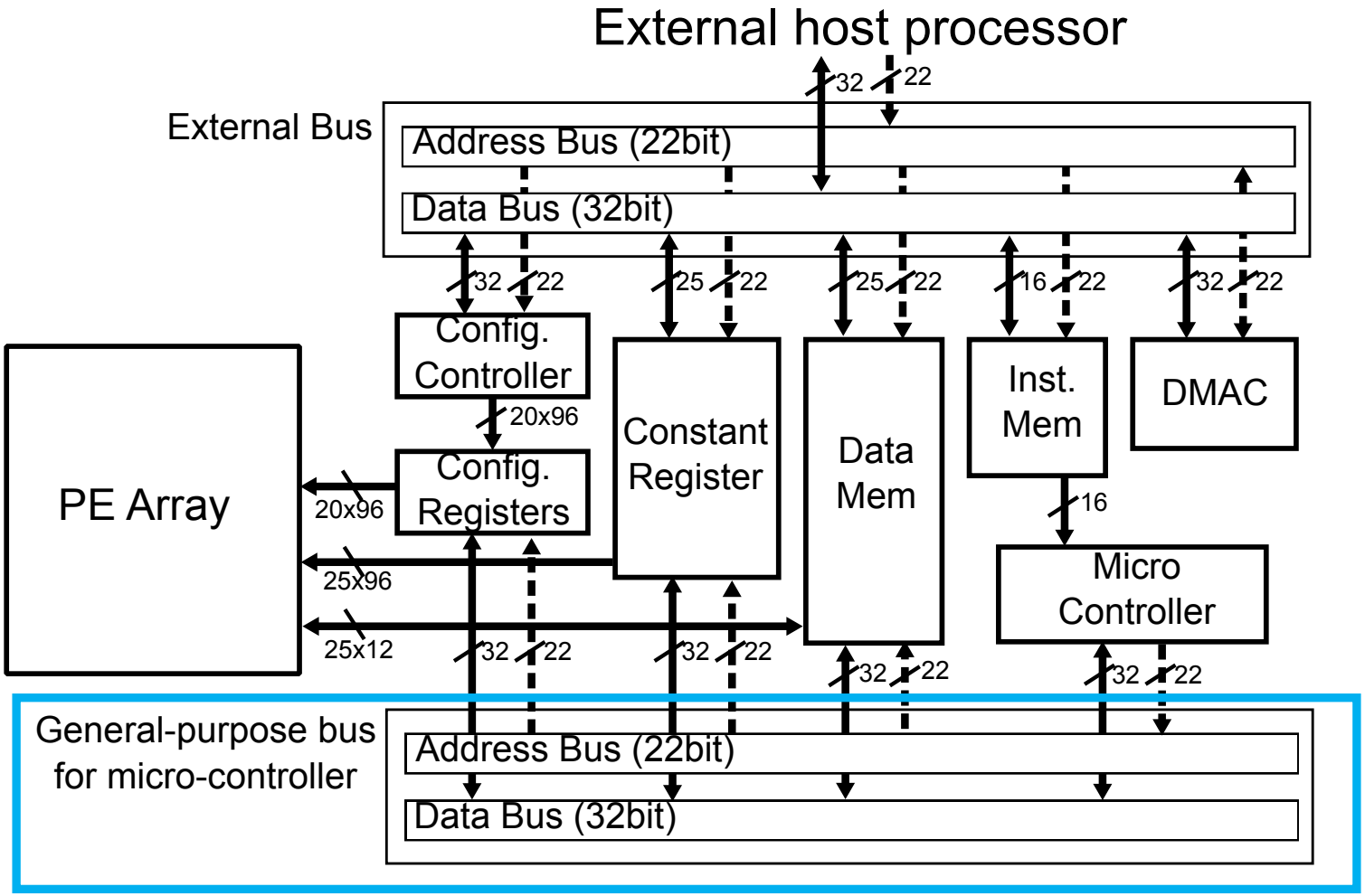
New Data Manipulator



- Offset values for each bank is introduced
- Relaxed the limitation of consecutive data access

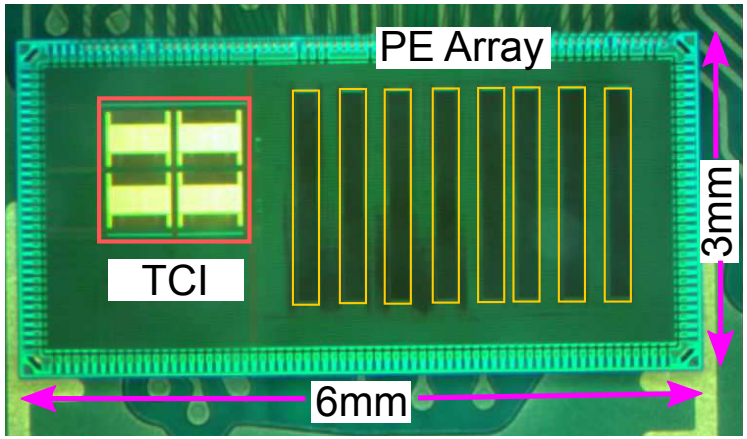
Extended Data Bus

- Micro-controller can handle any data in other modules



Evaluation Setup

- An implementation of VPCMA2
  - Using Renesas SOTB 65-nm technology
    - LSTP (Low STanby Power) version
  - Synthesized by Synopsys Design Compiler 2017
- A real chip of VPCMA[7]
  - Fabricated same technology
    - LP (Low Power) version (75% slower than LSTP)



Chip photo of VPCMA[7]

[7] T. Kojima, *et al*. "Real chip evaluation of a low power CGRA with optimized application mapping," 9th International Symposium on Highly-Efficient Accelerators and Reconfigurable Technologies. ACM, 2018, p. 13.

Hardware overhead

	VPCMA [7]	VPCMA2 1-cycle f/g	VPCMA2 2-cycle f/g
Max Frequency (MHz) (75% scaled)	87.71	95.23 (71.42)	125.0 (93.75)
Cell Area (mm^2) without PE array	10.04	14.55	14.22

- Improved data manipulator could increase critical path delay (i.e. degradation of operating freq.)
  - 2 version of designs are evaluated
  1. Fetch&Gather are performed within 1 cycle (naïve)
  2. Fetch&Gather take 2 cycles (to divide the long critical path)
- 2-cycle f/g
  - Does not have any effects on the frequency
  - Causes 40% increase of cell area

Comparison of Power Consumption

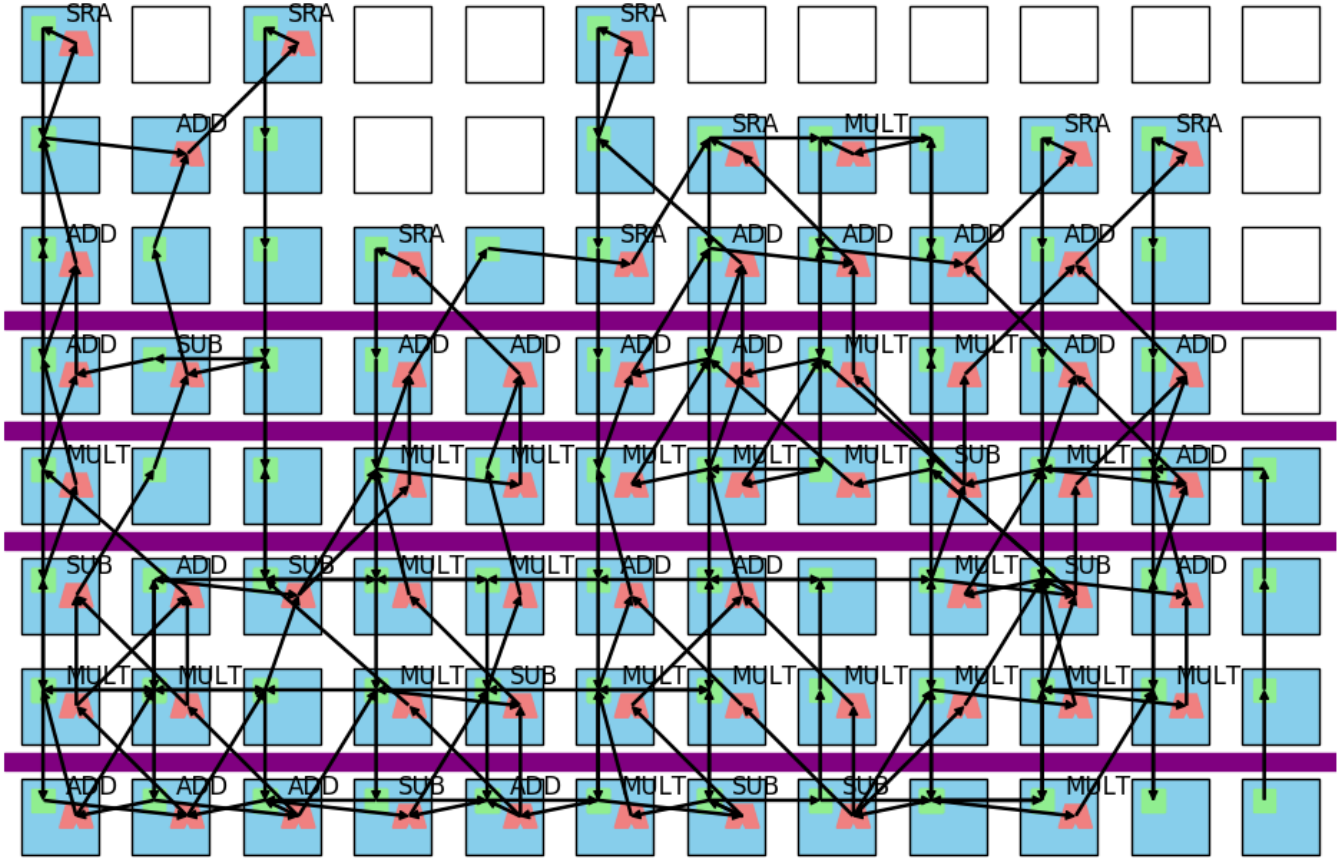
Power Consumption while running gray scale processing at 30MHz

	VPCMA[7]	VPCMA2 (sim)
Process version	LP	LSTP
Standard Voltage	0.55 V	0.75 V
Static Power	0.126 mW	0.0252 mW
Dynamic Power	3.337 mW	4.029 mW
Total Power	3.463 mW	4.053 mW

17% increase

- Compared to VPCMA(real chip), VPCMA2 (simulation)
  - Reduces static power consumption because of process difference (not architecture difference)
  - Increases dynamic power consumption because of the improved functionality and partially due to the higher standard voltage

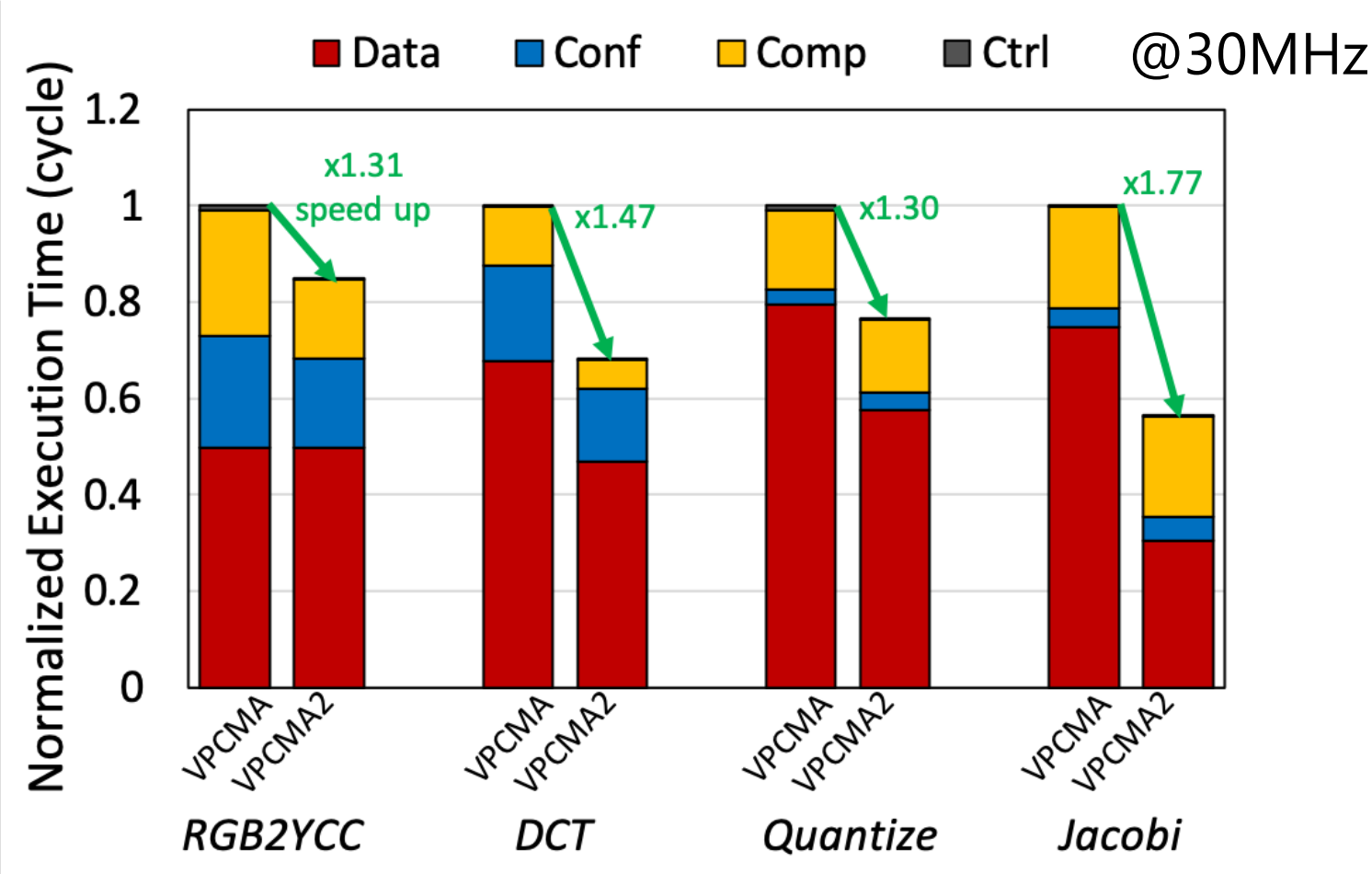
Enhanced Application Mappability



Mapping result of DCT by Genetic algorithm-based mapper[6]

- VPCMA2 can accommodate large & complex kernel but VPCMA cannot

Performance Improvement



- Improved both of mappability & bank access contribute to 1.46x higher performance

Conclusion

- This work points out a problem by data handling limitations of VPCMA
- A new SF-CGRA: VPCMA2 is proposed to relax the limitations
- Evaluation results shows
  - 10% area overhead (as a whole of chip)
  - No degradation of operating frequency
  - 17% power overhead
  - 46% performance improvement
- Future work
  - Analysis of effectiveness for other architectures
  - Evaluation of real chip implementation (under fabrication)

End of presentation
Thank you for your attention

Any questions?