拡張型 SIMD RISC-V アーキテクチャによるアイドル状態の演算コア削減とその高効率化

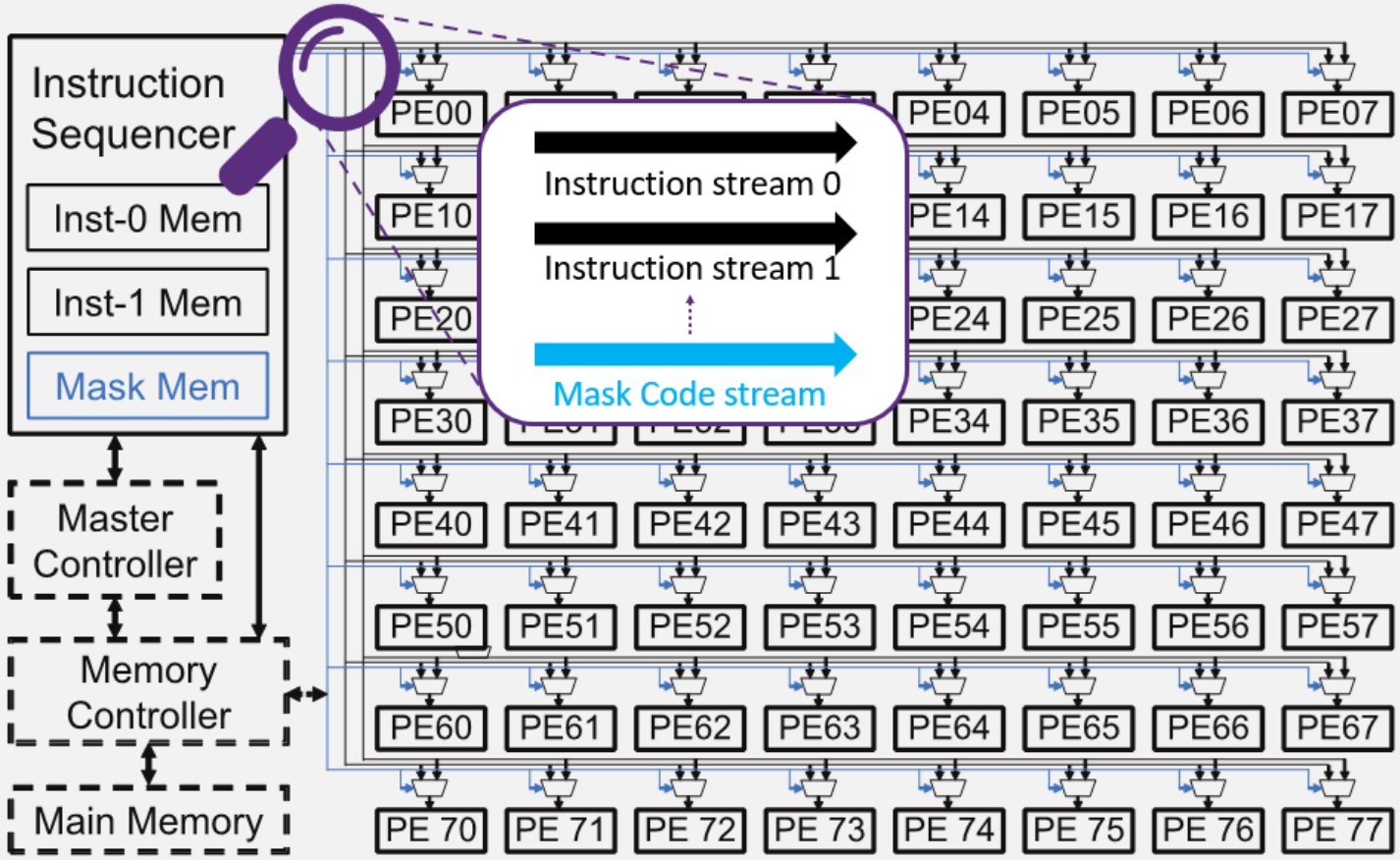
Optimizing Efficiency in Extended SIMD RISC-V Based Architectures through Minimization of Idle Computational Cores

Masaru NISHIMURA, Yuxi TAN, Yoshiki Yamaguchi
University of Tsukuba

Motivation & Target

- SIMD processors often suffer from **Idle time** caused by **data transfer** & limited number of instructions at same time. How can we deal with them ?
- Introduce **DIMD (Dual Instruction Multiple-Data)**, **Torus/Mesh Interconnect**, and **HBM (High Bandwidth Memory) and Memory Hierarchy**.
- Adopt **RISC-V** as PE(Processing Element)'s ISA that can be utilize matured ecosystems for software programming.
- Application range: **HPC**, **AI Computing**, and Embedded systems, etc.

Architecture Overview - DIMD



DIMD (Dual Instruction Multiple-Data)

Issue 2 instructions and **mask code** to all PE at same cycle.

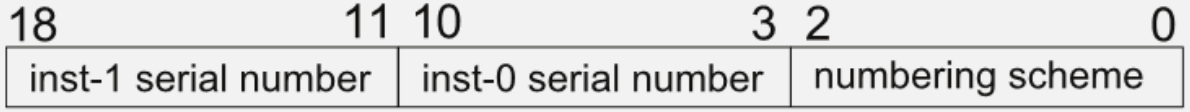
Then each PE determine which instruction should be executed.

Benefits

Reduce idle time caused by limited number of instructions at SIMD. Effective for stencil computation, FFT, etc.

Format of mask code

Mask code consists of **serial numbers** and **numbering scheme**.



000: row-based

0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4
5	5	5	5	5	5	5	5
6	6	6	6	6	6	6	6
7	7	7	7	7	7	7	7

001:column-based

0	1	2	3	4	5	6	7
0	1	2	3	4	5	6	7
0	1	2	3	4	5	6	7
0	1	2	3	4	5	6	7
0	1	2	3	4	5	6	7
0	1	2	3	4	5	6	7
0	1	2	3	4	5	6	7
0	1	2	3	4	5	6	7

010: center

0	0	0	0	0	0	0	0
0	1	1	1	1	1	1	0
0	1	2	2	2	2	1	0
0	1	2	3	3	2	1	0
0	1	2	3	3	2	1	0
0	1	2	2	2	2	1	0
0	1	1	1	1	1	1	0
0	0	0	0	0	0	0	0

011: diagonal

0	1	2	3	4	5	6	7
1	0	1	2	3	4	5	6
2	1	0	1	2	3	4	5
3	2	1	0	1	2	3	4
4	3	2	1	0	1	2	3
5	4	3	2	1	0	1	2
6	5	4	3	2	1	0	1
7	6	5	4	3	2	1	0

100: 45 degree

0	1	2	3	4	5	6	7
1	1	2	3	4	5	6	7
2	2	2	3	4	5	6	7
3	3	3	3	4	5	6	7
4	4	4	4	4	5	6	7
5	5	5	5	5	5	6	7
6	6	6	6	6	6	6	7
7	7	7	7	7	7	7	7

101: 135 degree

7	6	5	4	3	2	1	0
7	6	5	4	3	2	1	1
7	6	5	4	3	2	2	2
7	6	5	4	3	3	3	3
7	6	5	4	4	4	4	4
7	6	5	5	5	5	5	5
7	6	6	6	6	6	6	6
7	7	7	7	7	7	7	7

110: 225 degree

7	7	7	7	7	7	7	7
7	6	6	6	6	6	6	6
7	6	5	5	5	5	5	5
7	6	5	4	4	4	4	4
7	6	5	4	3	3	3	3
7	6	5	4	3	2	2	2
7	6	5	4	3	2	1	1
7	6	5	4	3	2	1	0

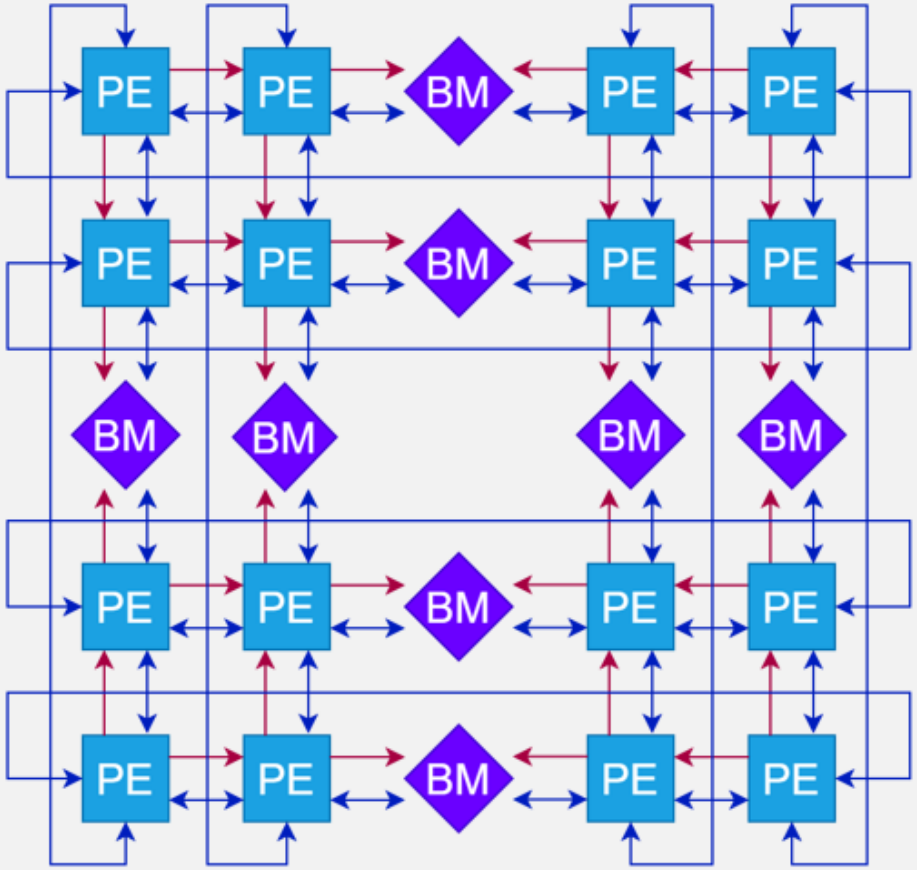
111: 315 degree

7	7	7	7	7	7	7	7
6	6	6	6	6	6	6	7
5	5	5	5	5	5	6	7
4	4	4	4	4	5	6	7
3	3	3	3	4	5	6	7
2	2	2	3	4	5	6	7
1	1	2	3	4	5	6	7
0	1	2	3	4	5	6	7

Example of Operation

01100000 00011111 100 Inst-0 Inst-1 Inst-idle

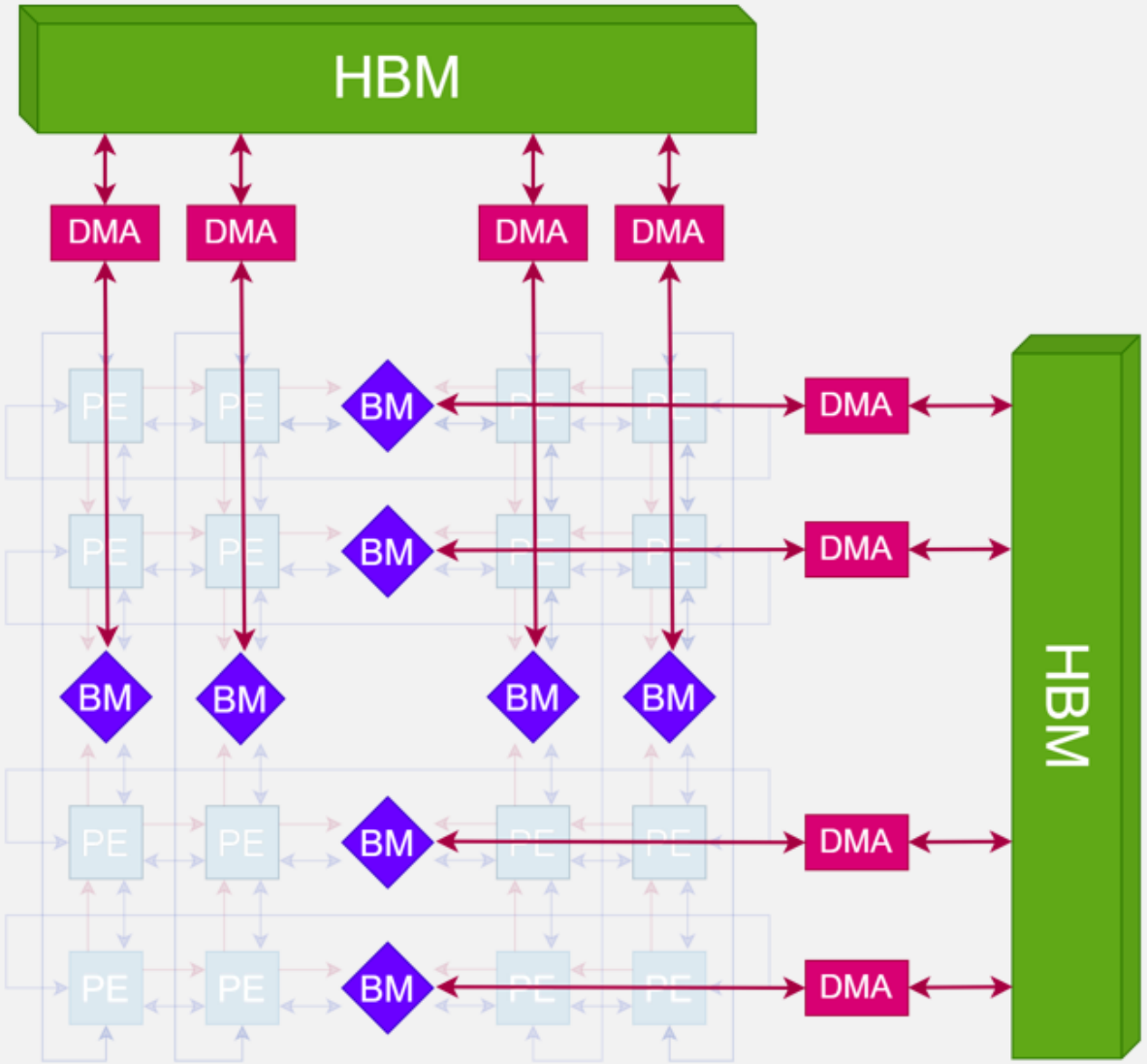
Torus/Mesh Interconnect



Address line : Simplex Mesh

Data line : Duplex Torus

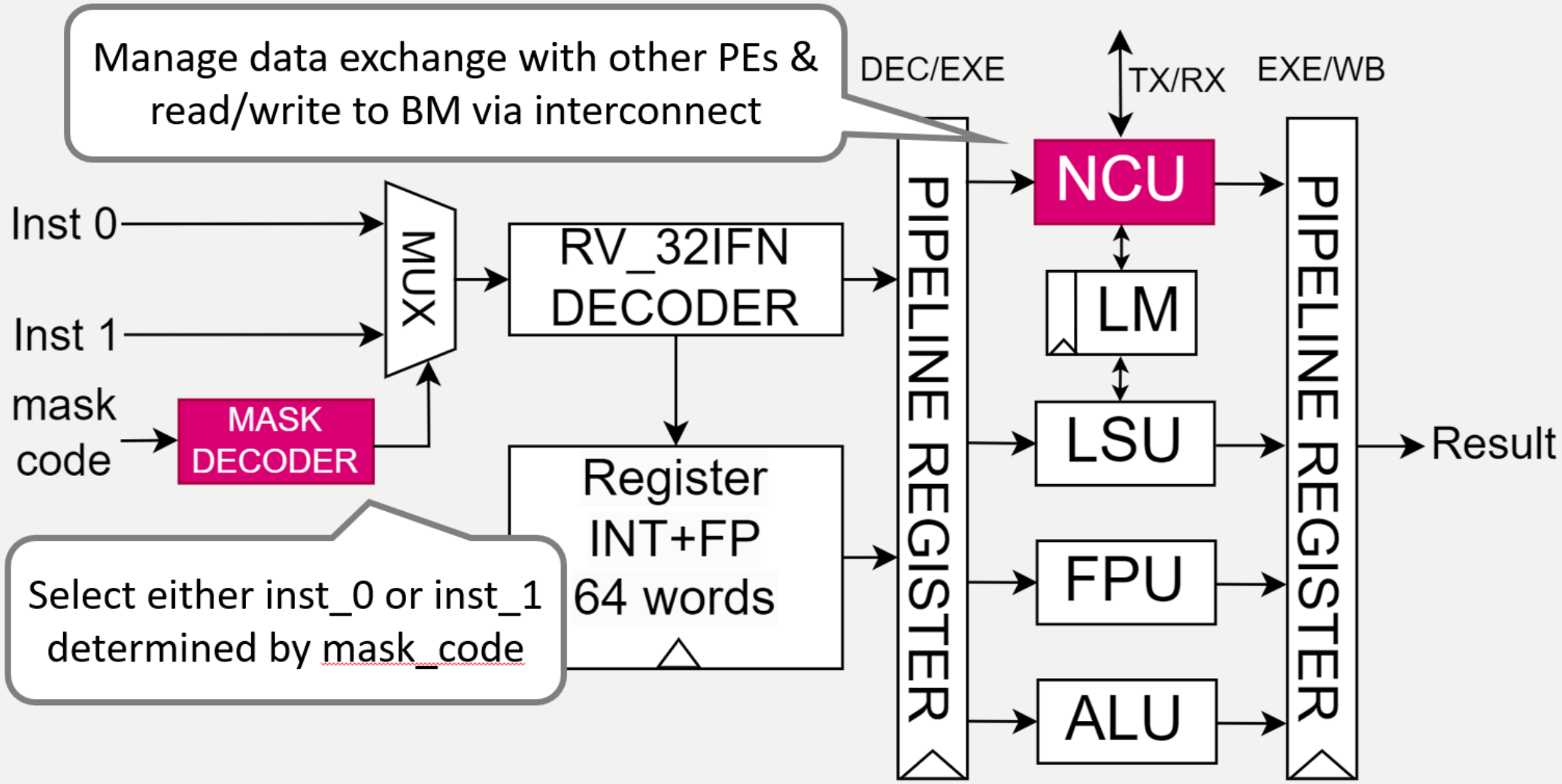
HBM and Memory Hierarchy



HBM : High Bandwidth Memory

BM : Broadcast Memory

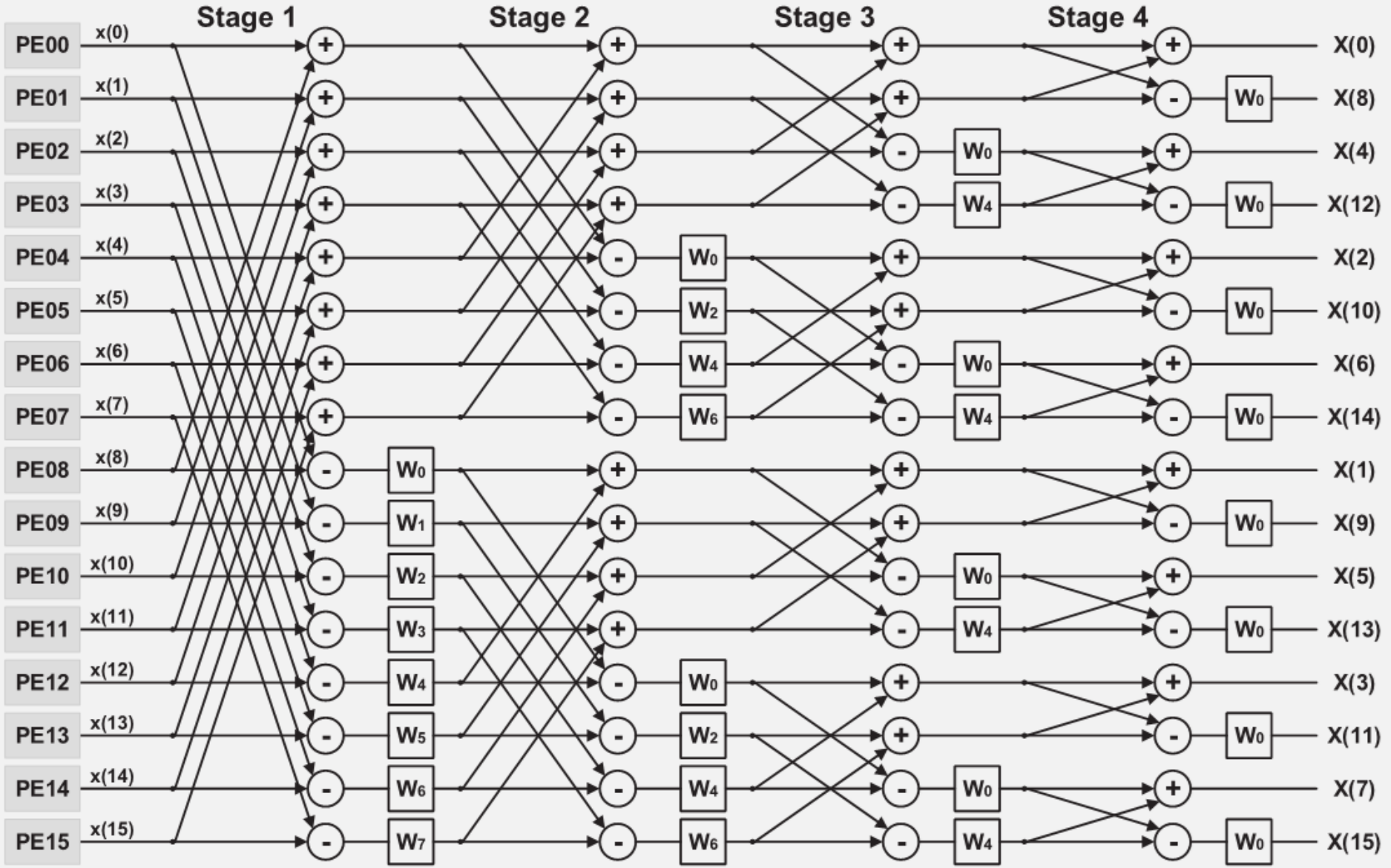
3. Architecture of the Processing Element

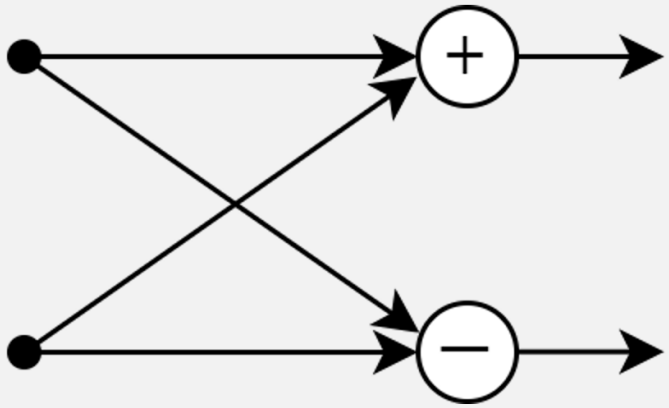


Adopt **RISC-V** as PE's ISA. We've fully designed processor dedicated to PE operations.

NCU : Neighbor Communication Unit
LM : Local Memory
N extension : support inst. corresponding to NCU.

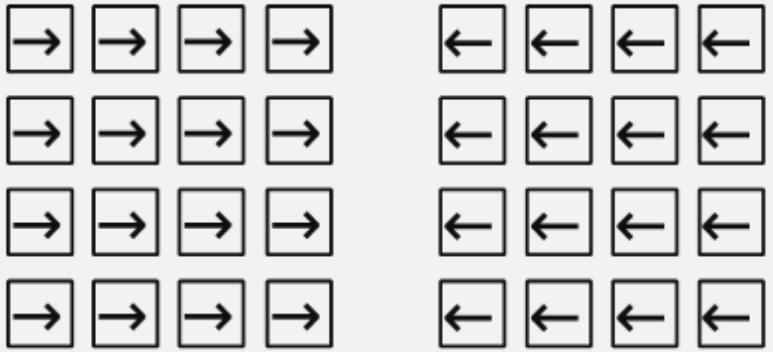
4. Comparison of FFT performance with SIMD



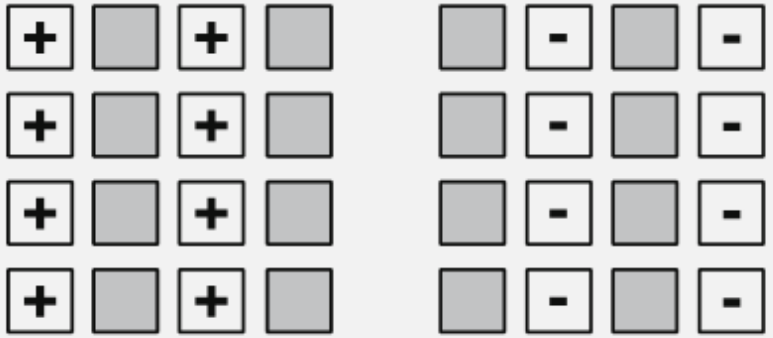


“Butterfly” diagram

“Butterfly” at FFT on SIMD



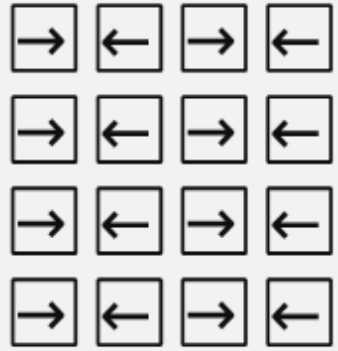
leed least **2 cycles** for data-exchange



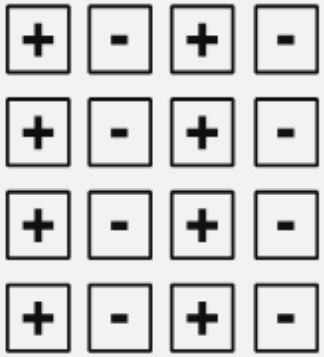
Then also need **2 cycles** for operations

Need 4 cycles totally

“Butterfly” at FFT on DIMD



Need least **1 cycle** for data-exchange

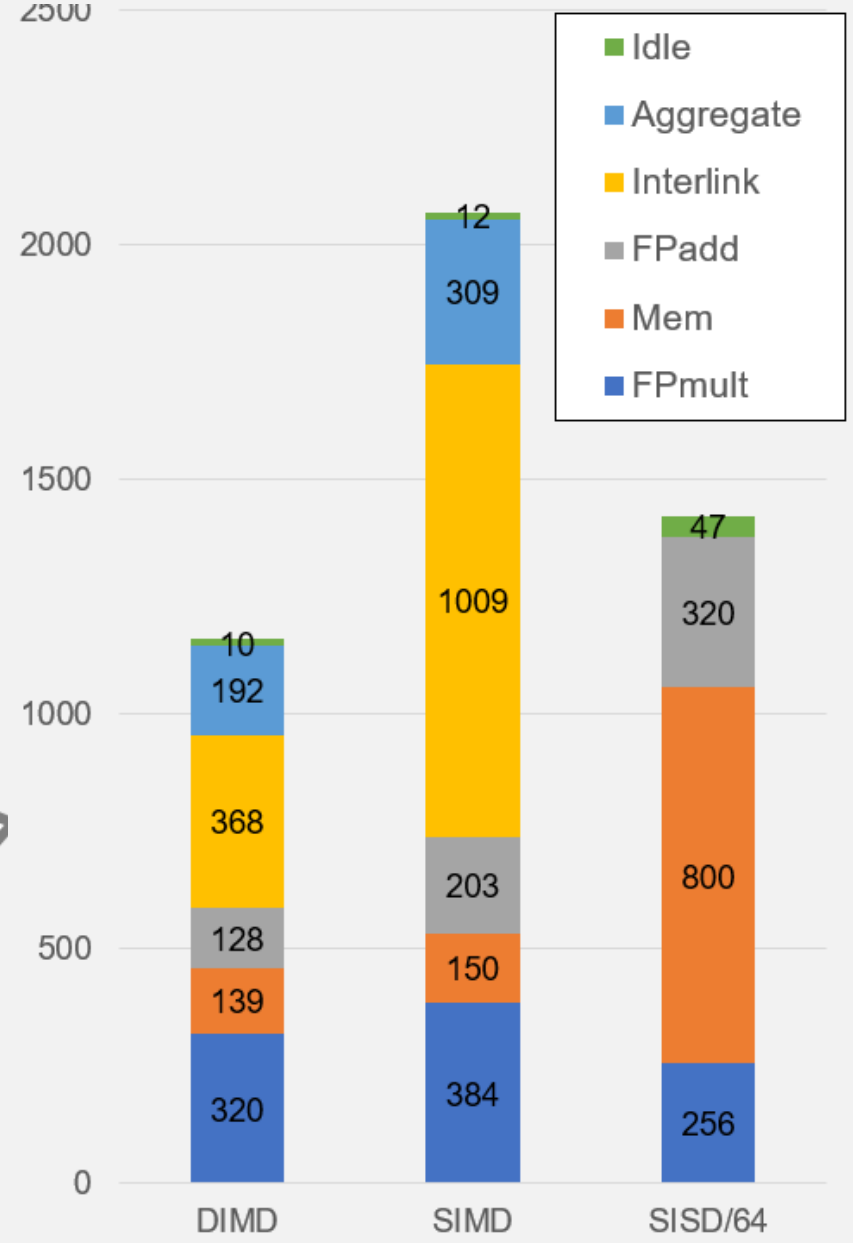


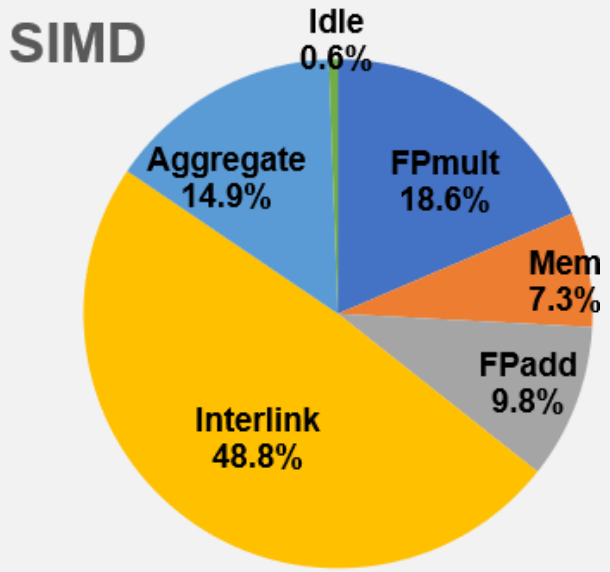
Need **1 cycles** for operations

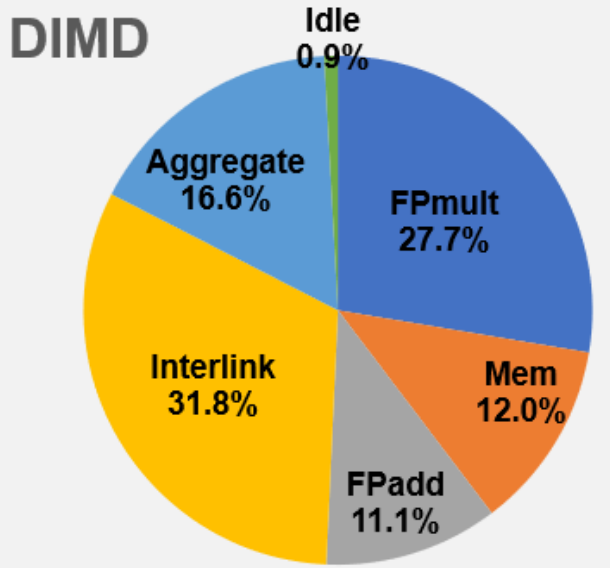
Need only 2 cycles totally

DIMD vs SIMD speedups

178% faster







1024-point 2D FFT on 8x8 array

We are pursuing to implement this architecture on high-end FPGA, Xilinx ALVEO U280.

Estimating we can achieve more than 400 PEs scale many core processors.