RAW2020, May 18-19, 2020

Kyutech

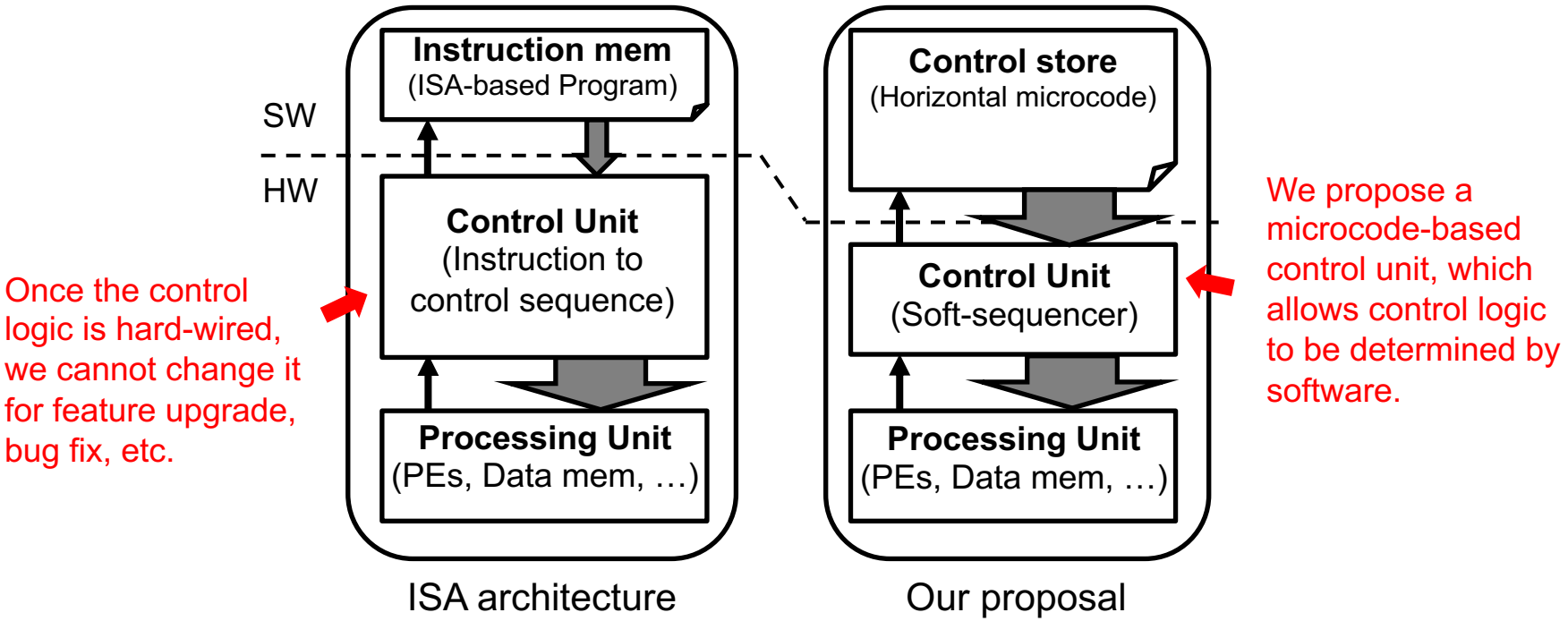
A Microcode-based Control Unit for Deep Learning Processors

Qian Zhao[1] (cho@ai.kyutech.ac.jp) ,
Yasuhiro Nakahara[2], Motoki Amagasaki[2],
Masahiro Iida[2], and Takaichi Yoshida[1]

[1] Kyushu Institute of Technology, [2] Kumamoto University

Japan

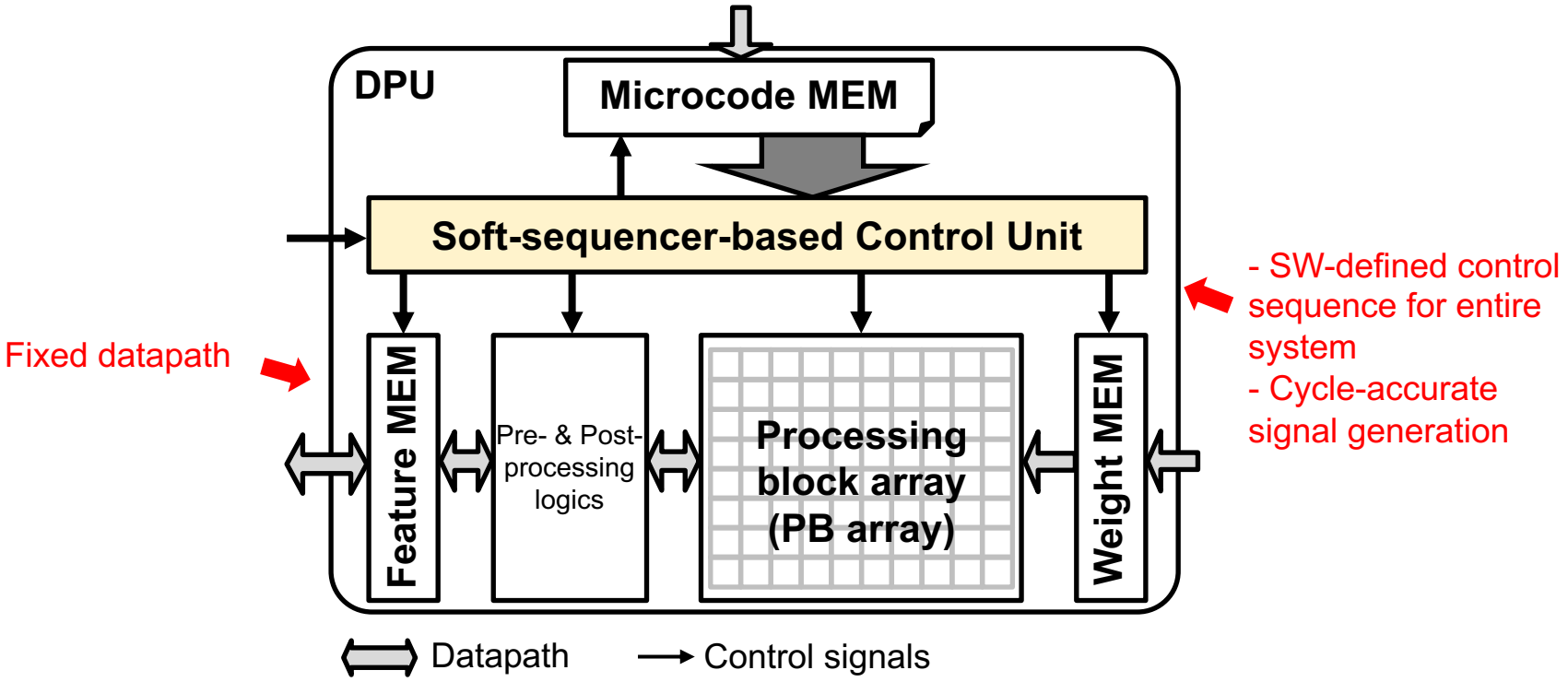
A fixed ISA-based controller may limit the flexibility of DPU



Provide full control over the DPU in a writable firmware, and thus delivers high flexibility and shorter dev. life cycle.

ISA: Instruction set architecture, DPU: Deep learning processing unit

DPU architecture using our approach

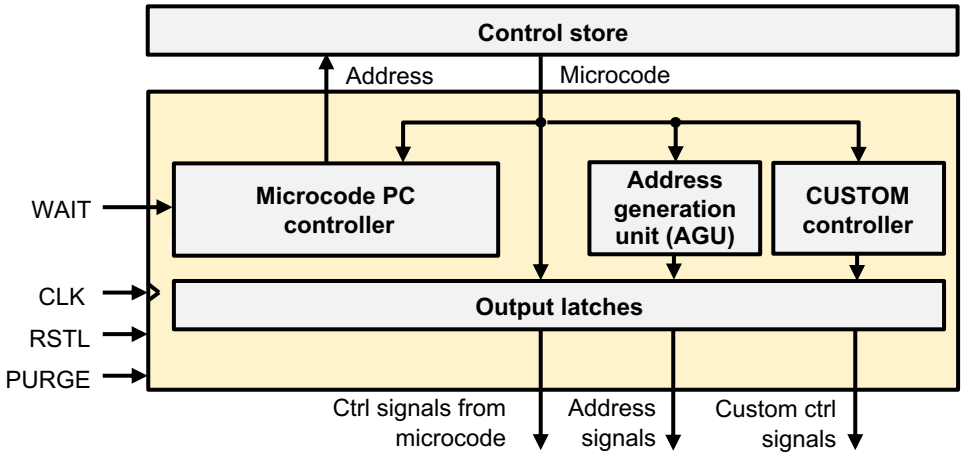


Microcode-based control unit is suitable for DPU because:

- **DPUs have fewer instructions than general-purpose processors.**
- **The control flow and data flow of DPUs are much less complex and more predictable than those of general-purpose processors.**

Soft-sequencer control unit design

Soft-sequencer structure



Microcode format

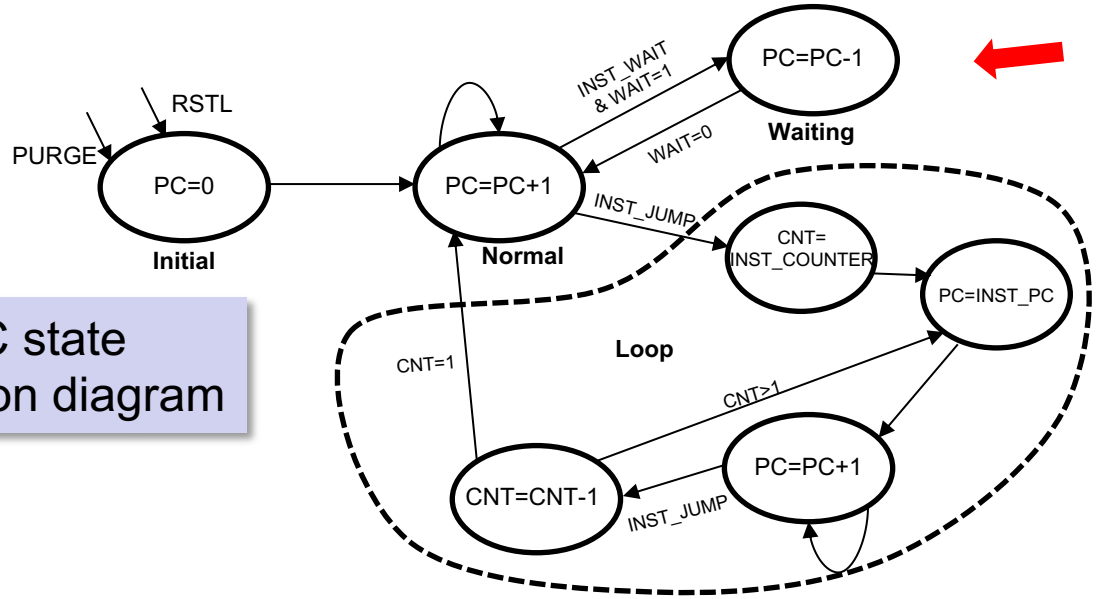
Bit	Field name	Function
0-3	INST_COUNTERn_WE	Write enable of CNTn[1)]
4-7	INST_JUMP_COUNTERn	Loop jump if CNTn > 1
8-10	INST_ADDRn_WE	Write enable of ADDRn[2)]
11-12	INST_ADDRn_BAK	Backup enable of ADDRn
13-15	(Reserved)	-
16-47	INST_PC	Immediate for PC[3)]
48-111	INST_COUNTERn	Immediate for COUNTERn
112-159	INST_ADDRn	Immediate for ADDRn
160-217	INST_CTRLn	A control signal per bit
217-255	(Unused)	-

1)Four 16-bit CNT registers for loop control.
2)Three 16-bit ADDR registers for AGU.
3)The PC regsiter is 32-bit.

Soft-sequencer fetches a microcode and generates control signal sequences according to the binary status saved in it.

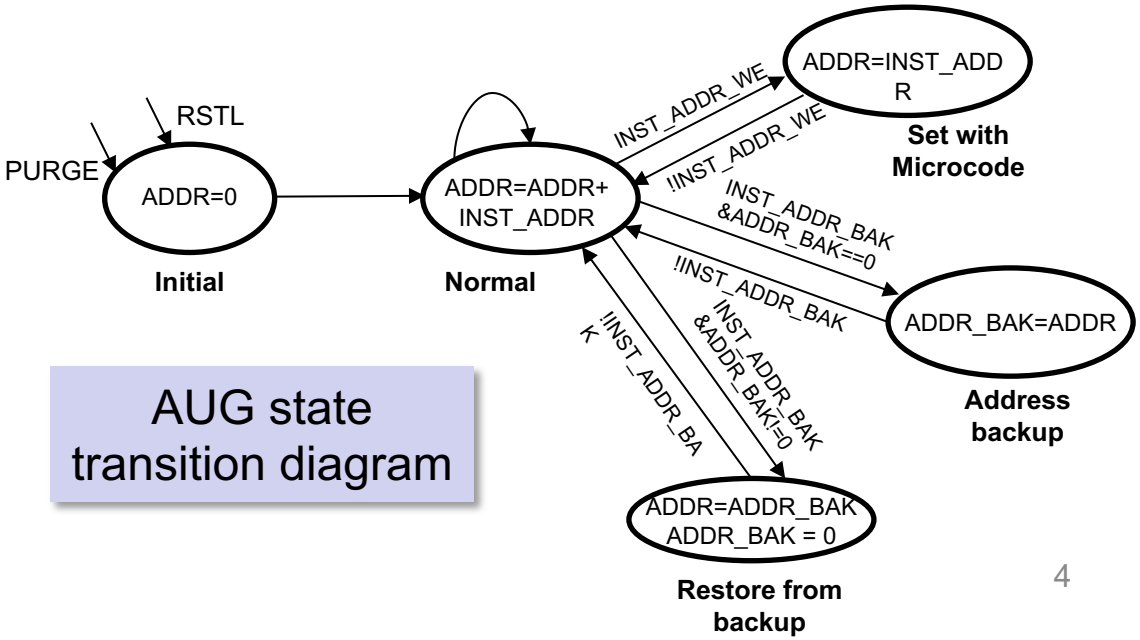
- **Simple flow control like loops are implemented in the PC controller, allowing the size of the microcode to be significantly shortened.**
- **The AGU holds and computes addresses in user memory.**
- **A designer can add a custom controller to generate signals by combining other control signals.**

State transition of PC and AGU design



PC state transition diagram

- In our DPU design, three addresses for feature memory write, feature memory read, and weight memory read are generated by the AGU.
- Flexible address sequence generation, like:
  - 1, 2, 3, 4
  - 1, 3, 5, 7
  - 1, 2, 3, 4, 2, 3, 4, 5
    (using backup & restore)



AUG state transition diagram

Implementation case study

FPGA implementation

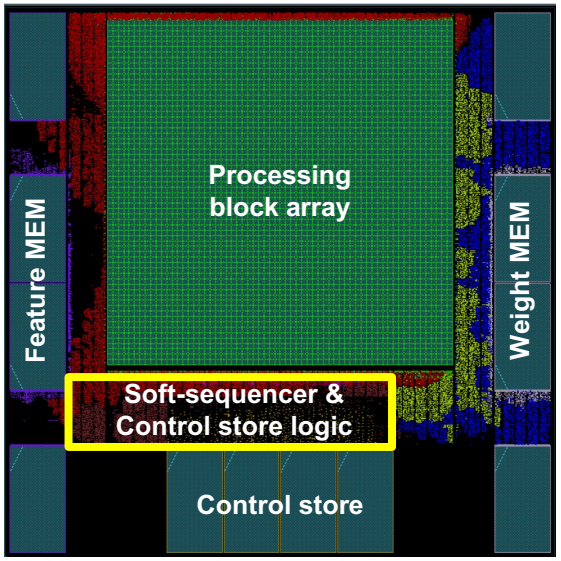
- Device: ADM-PCIE-KU3
- DPU: 32*32 PB array
- Freq.: 100MHz
- Power: 1.255W in total, control unit(<1%), control store(<3%)

Resource utilization

	LUTs	FFs	BRAMs
Soft sequencer	17,085 (8.4%)	440 (0.4%)	0
Control store	361 (0.2%)	3 (<0.1%)	114 (43.2%)
DPU total	202,562	106,754	264

ASIC implementation

- Process: TSMC 40-nm
- DPU: 64*64 PB array
- Freq.: 350MHz
- Area: control unit (5%), control store (5.2%)
- Power: 0.516W in total, control unit(<0.1%), control store(5.2%)



DPU layout

Microprogramming

- Demo DNN: 4 CNN layers, 2 FC layers, CIFAR-10 task
- Lines-of-code: 759 lines of microprogram

Conclusion

- **Problem:** A fixed ISA-based controller may limit the flexibility of DPU.
- **Approach:** We propose a microcode-based control unit, which allows control logic to be determined by SW.
- **Pros:** Full control over the DPU in a writable firmware, and thus delivers high flexibility and shorter dev. life cycle.
- **Result:** We evaluate the proposed control unit using two DPU design cases implemented by FPGA and ASIC. The results show that high flexibility of the proposed control unit can be achieved without obvious performance overhead.