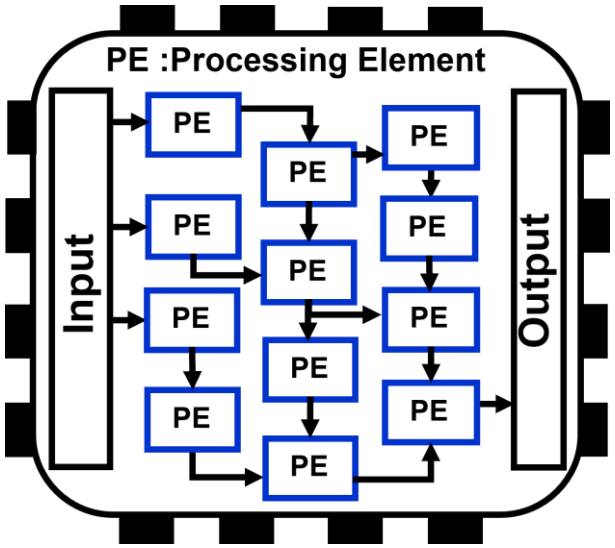
A 1.2nJ/Classification Fully Synthesized All-Digital Asynchronous Wired-Logic Processor Using Quantized Non-linear Function Blocks in 0.18μm CMOS

Department of Electrical Engineering and Information System, The University of Tokyo

Rei Sumikawa, Kota Shiba, Atsutake Kosuge, Mototsugu Hamada, Tadahiro Kuroda

Reduction of Energy Consumption of DNN Chip

- Challenge of DNN chip is large energy consumption
- Wired-logic architecture saves energy by eliminating memory accesses
  - Implement plenty of processing elements
- Chip size is an issue for wired-logic architecture

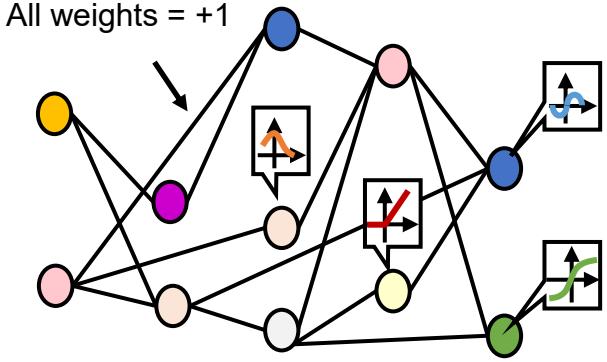


To implement in 1 chip
- Reduction of the number of PE
- Improvement of area efficiency of PE

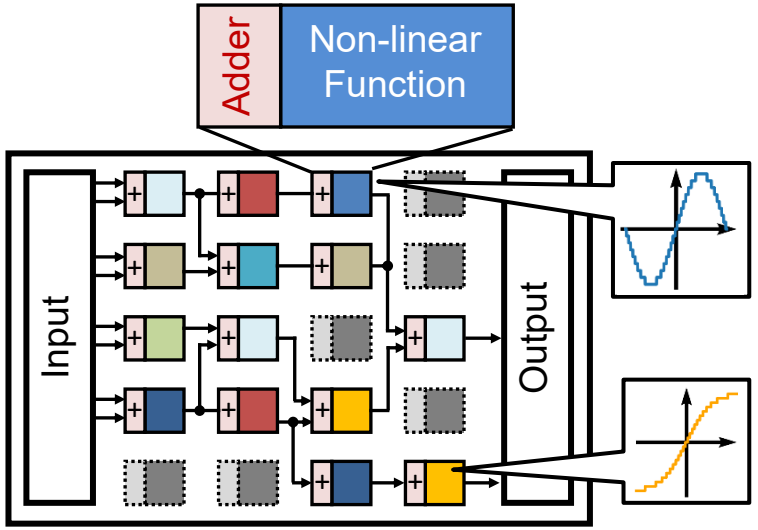
Both are required!

Non-Linear Neural Network (NNN)

- **NNN reduces the number of neurons and synapses**
  - **Activation function is optimized for each neuron**
- **Most hardware is used in non-linear function blocks**
  - **Minimization of this part is important**



Overview of NNN



Wired-logic NNN chip

Logically Compressed Non-Linear Function Blocks

- **Non-linear functions are coarsely quantized**
  - ◆Degradation of accuracy by quantization is small
- **8b functions are described by HDL and synthesized**
  - ◆Realized blocks with minimum amount of digital circuits

```
module TANH_4b(input[3:0] X, output[3:0]Y);

  function[3:0] DATA;
  input [3:0] FIN;
  begin
      case(FIN)
            4'b1000: DATA = 4'b1110;
            4'b1001: DATA = 4'b1110;
            4'b1010: DATA = 4'b1110;
            4'b1011: DATA = 4'b1110;
            4'b1100: DATA = 4'b1110;
            4'b1101: DATA = 4'b1110;
            4'b1110: DATA = 4'b1110;
            4'b1111: DATA = 4'b1111;
            4'b0000: DATA = 4'b0000;
            4'b0001: DATA = 4'b0001;
            4'b0010: DATA = 4'b0010;
            4'b0011: DATA = 4'b0010;
            4'b0100: DATA = 4'b0010;
            4'b0101: DATA = 4'b0010;
            4'b0110: DATA = 4'b0010;
            4'b0111: DATA = 4'b0010;
      endcase
  end
  endfunction

  assign Y = DATA(X);

endmodule
```

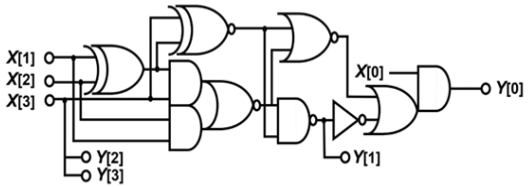
(a)

```
module TANH_4b ( X, Y );
  input [3:0] X;
  output [3:0] Y;
  wire   N6, N7, N8, N9, N10;
  assign Y[2] = X[3];
  assign Y[3] = X[3];

  OA21D0BWP7T U10 ( .A1(N6), .A2(N7), .B(X[0]), .Z(Y[0]) );
  CKND0BWP7T U11 ( .I(Y[1]), .ZN(N7) );
  CKND2D0BWP7T U12 ( .A1(N8), .A2(N9), .ZN(Y[1]) );
  NR2D0BWP7T U13 ( .A1(N8), .A2(N9), .ZN(N6) );
  XNR2D1BWP7T U14 ( .A1(N10), .A2(X[3]), .ZN(N9) );
  AOI22D0BWP7T U15 ( .A1(X[2]), .A2(X[1]), .B1(N10), .B2(X[3]), .ZN(N8) );
  CKXOR2D0BWP7T U16 ( .A1(X[2]), .A2(X[1]), .Z(N10) );

endmodule
```

(b)



(c)

HDL of Tanh function (a) before synthesis, (b) after synthesis and its (c) schematic

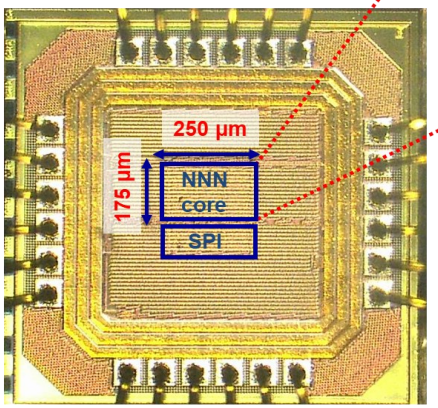
Circuit area of Tanh function

	LUT(SRAM)	LC-NLF(8b)
Area (μm²)	>10,000	717.8
FoM	1	1/13.9

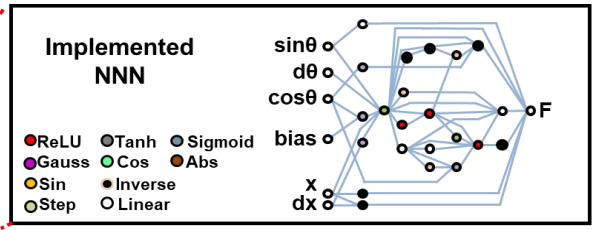
FoM : Figure of Merit

Experimental Results

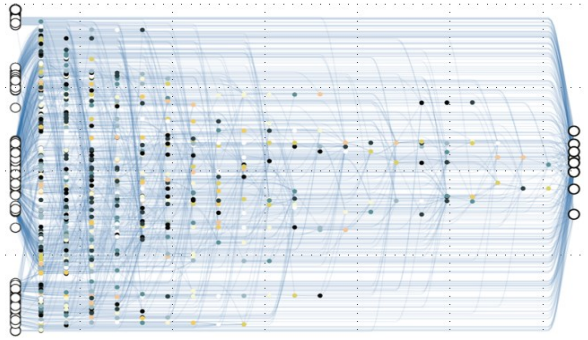
- **Prototyped the proposed wired-logic processor**
- **Estimated energy consumption of MNIST task by scaleup simulation**
  - **Predicted 1.2 nJ/classification, 2.6x improvement compared with ReRAM based wired-logic CiM chip**



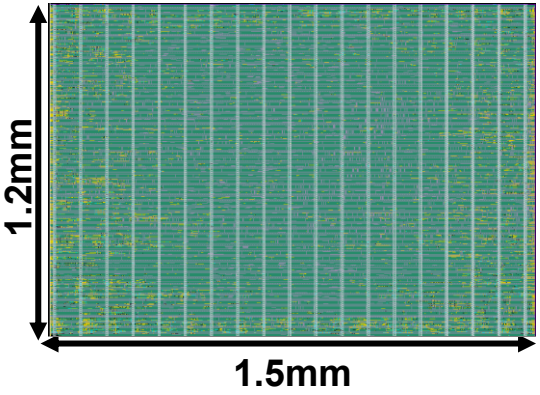
Prototyped Chip



Process	TSMC 0.18 μm
Area	175 μm × 250 μm
Supply voltage	0.9-1.8 V
# of neurons	23
# of synapses	51
Power efficiency per neuron	0.8 pJ/NOPS @0.9 V
Power efficiency per synapse	0.3 pJ/SOPS @0.9 V



NNN for MNIST



Please come and listen to my poster presentation!