Brief Introduction of FPGA And Neural Network Deployment

Zhang Bin

SmartMore 思谋

FPGA---Field Programmable Gate Array

- Field:
  - In the field
- Programmable:
  - Re-Configurable - Change Logic Functions
- Gate Array:
  - Reference to ASIC (Application specific integrated circuits) internal architecture

Why FPGA?

Parallelism!

Parallelism!

Parallelism!

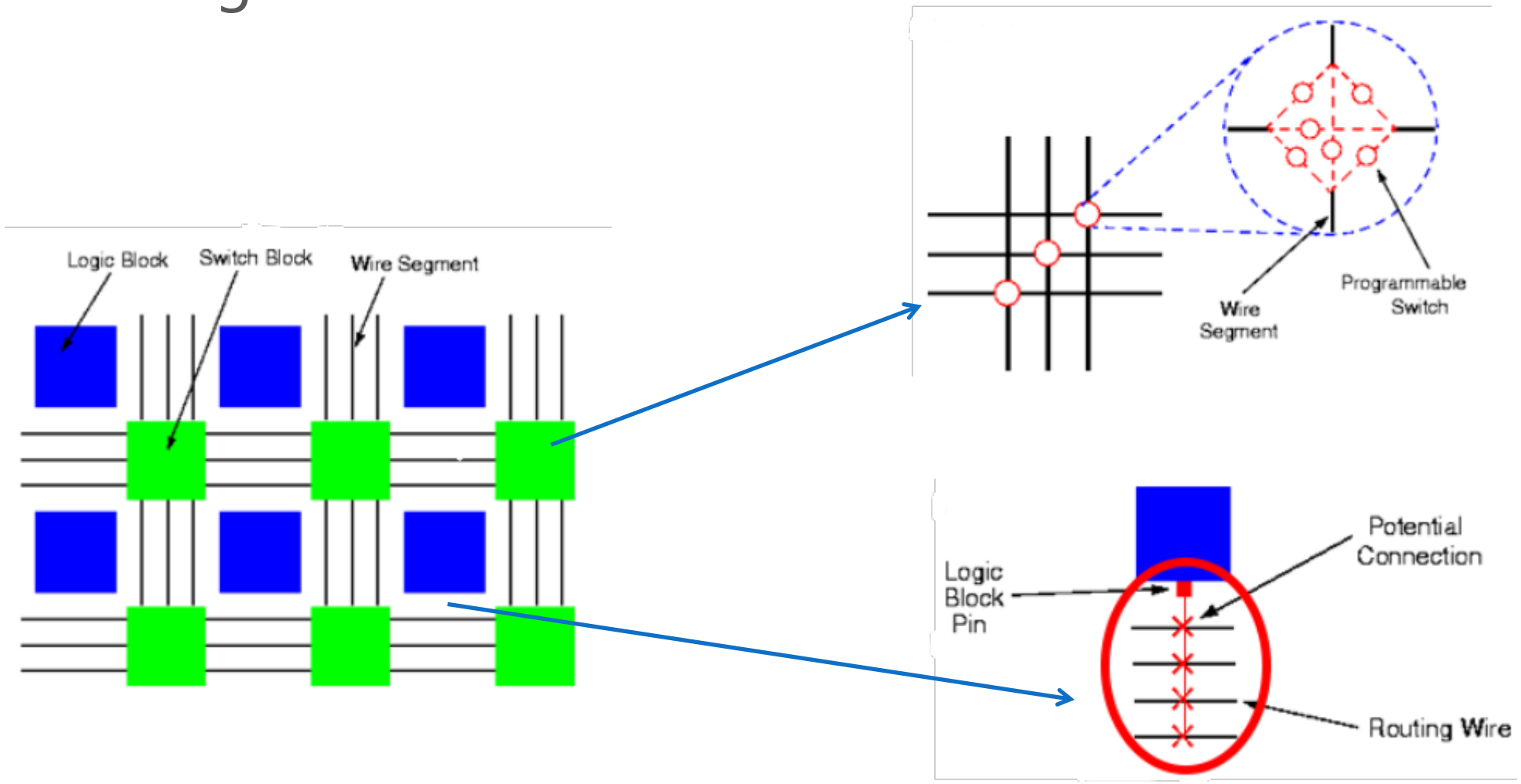
Parallelism!

Programmable? How?

Logic elements

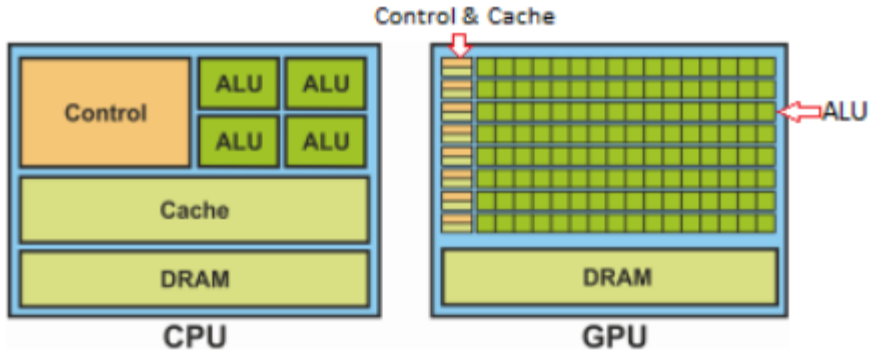
FGPA Implementation

Programmable? How?



FPGA vs. CPU vs. GPU

--Architecture



- CPU
  - Complex internal structure
- GPU
  - A large-scale data which is highly unified and mutually independent and normally operate in a pure computing environment with limited external interrupts.

FPGA vs. GPU (Image processing)

SmartMore 思谋

- GPU
  - Suitable for a real time image processing system requires high resolution and complex calculations.
  - SIMD
- FPGA
  - If the system only requires convolution based on low level algorithms, FPGA becomes more suitable.

Compare to FPGA, GPU has higher power consumption, but better calculation ability.

FPGA

- GPU
  - Suitable for a real time image processing system requires high resolution and complex calculations.
  - SIMD
- FPGA
  - If the system only requires convolution based on low level algorithms, FPGA becomes more suitable.

Compare to FPGA, GPU has higher power consumption, but better calculation ability.

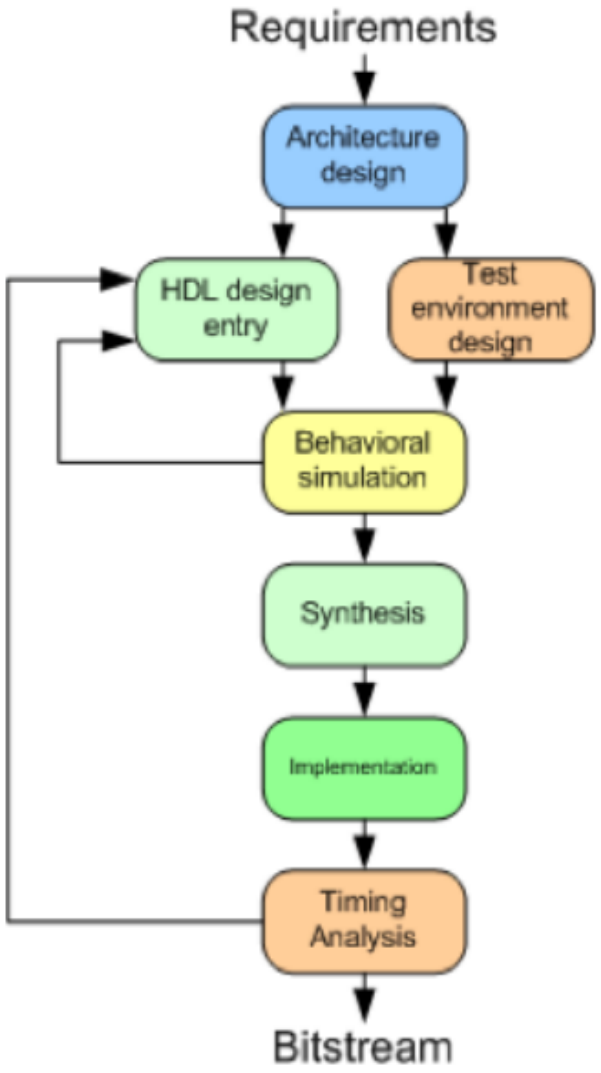
Embeded system & FPGA SoC

SmartMore 思谋

Processing System (PS)

Features	Zynq-7000S	Zynq-7000	
Devices	Z-7007S, Z-7012S, Z-7014S	Z-7010, Z-7015, Z-7020	Z-7030, Z-7035, Z-7045, Z-7100
Processor Core	Single-core ARM® Cortex™-A9 MPCore™	Dual-core ARM Cortex-A9 MPCore	
Maximum Frequency	Up to 766MHz	Up to 866 MHz	Up to 1GHz
External Memory Support	DDR3, DDR3L, DDR2, LPDDR2		
Key Peripherals	USB 2.0, Gigabit Ethernet, SD/SDIO		
Dedicated Peripheral Pins	Up to 128	Up to 128	128

FPGA Develop Workflow



FPGA Pros & Cons

Disadvantages

- Limited resources. Need to choose different FPGA models based on design needs.
- Compare to ASIC:

  a. High power consumption

  b. Only suitable for low quality production, because of the high expense
- Compare to GPU:

  a. Relatively low calculation ability.

  b. Low generality.
- HDL (Hardware Description Language). Long development period.

FPGA Pros & Cons

SmartMore 思谋

Advantages

- Flexibility

  a. Programmable after manufacturing.

  b. Customization

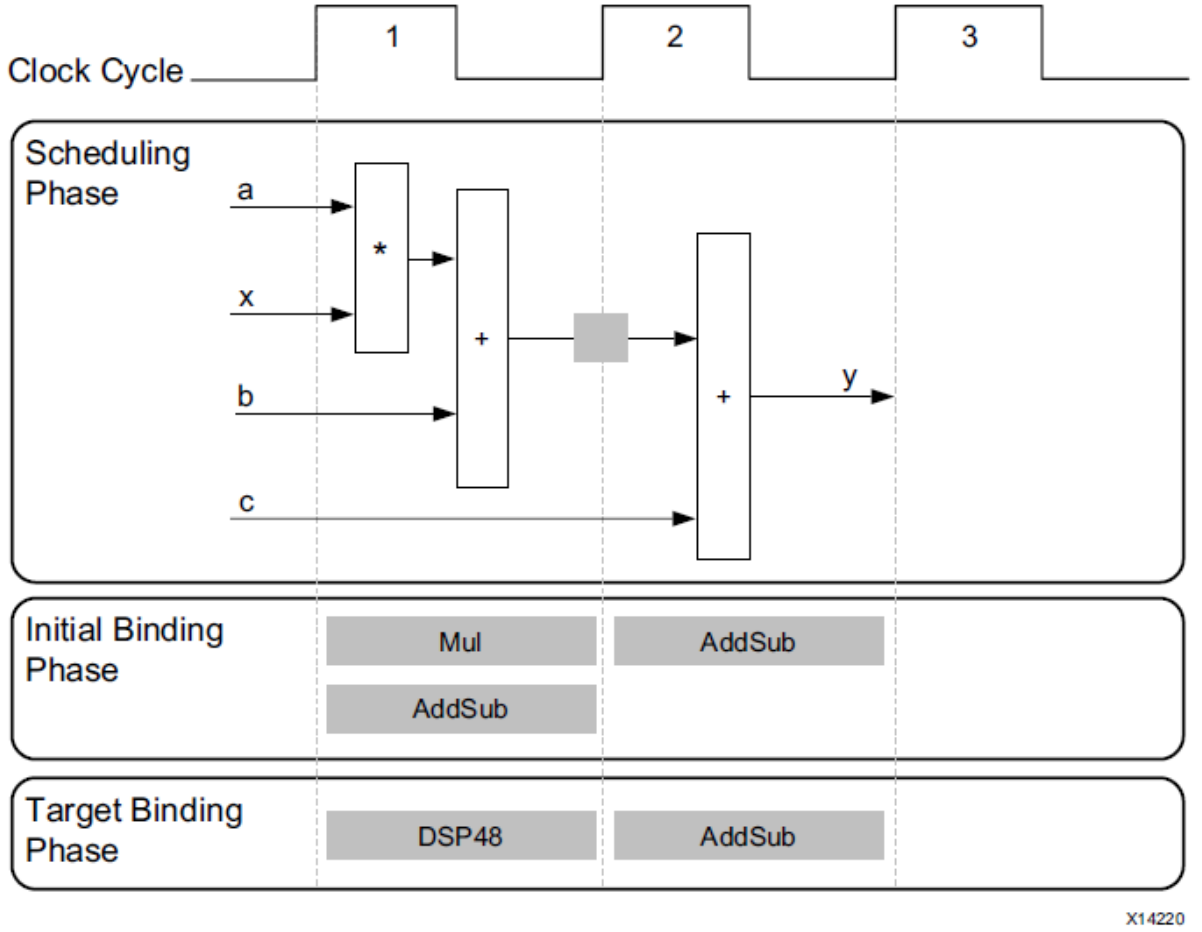
- Massively parallel

- Low latency

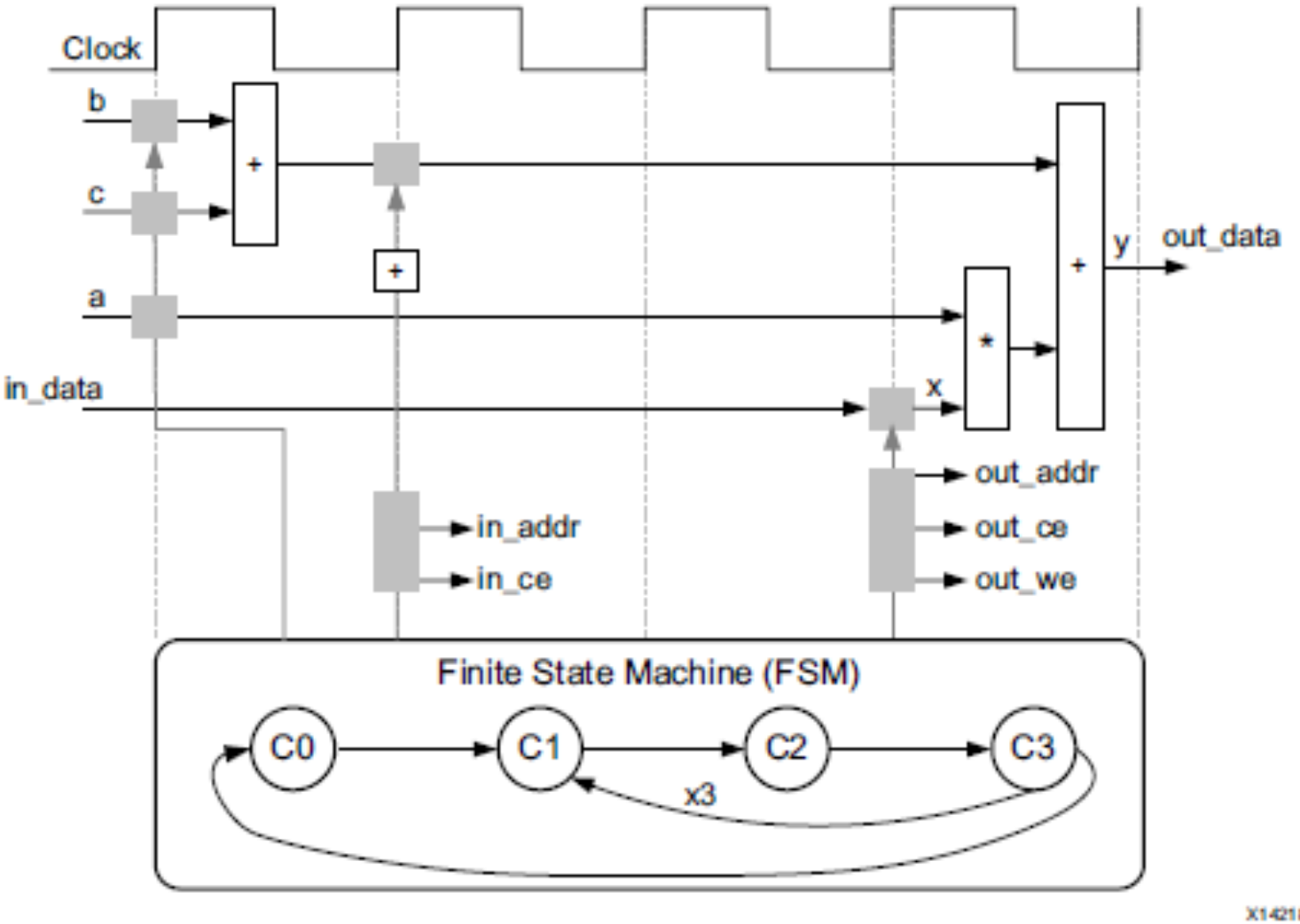
Mission-critical applications require very low-latency. (autonomous vehicles and manufacturing operations)

Example 1

```
int foo (char x, char a, char b, char c) {

    char y;

    y = x*a+b+c;

    return y

    }
```



```
void foo(int in[3], char a, char b, char c, int out[3])

{

        int x,y;

        for(int i = 0; i < 3; i++) {

                x = in[i];

                y = a*x + b + c;

                out[i] = y;

                }

        }
```



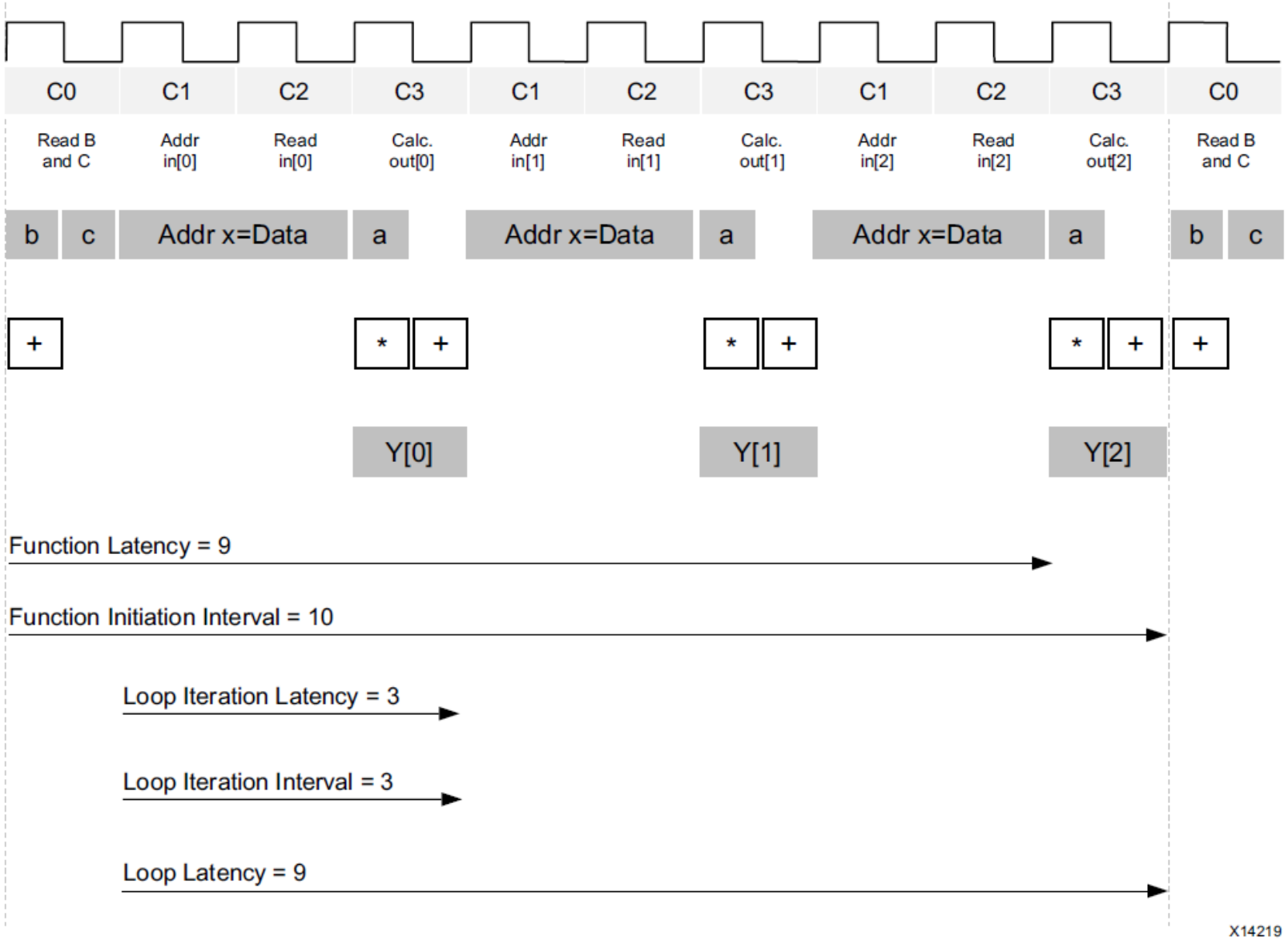


Figure 1-3: **Latency and Initiation Interval Example**

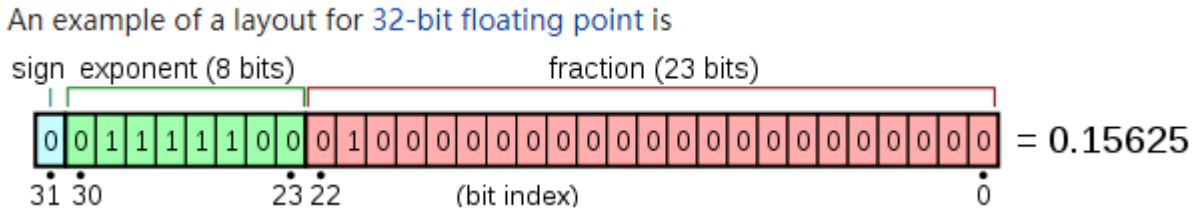
Example 2

Calculation of an addition of two 32-bits floating-point numbers

C = A + B

Functions need to be designed.

- Same exponents -- compare unit, shift unit
- 2's complement -- adder
- add {1, fractions}-- adder
- carry? -- adder
- overflow or underflow
- Infinity, Zero or Nan -- IEEE 754



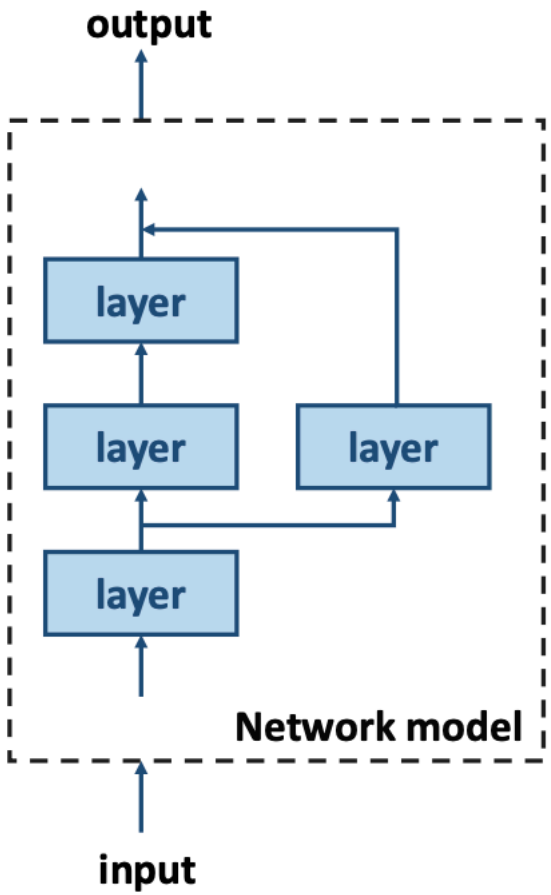
Example 2

Design flow

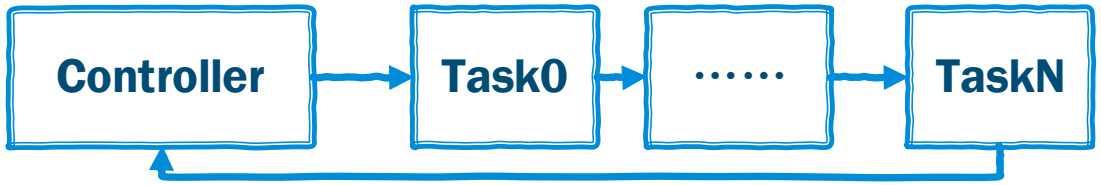
- Arrangement of operation functions
  - Special numbers bypass
  - Predict overflow/underflow
- Optimize
- Timing analysis

- Same exponents -- compare unit, shift unit
- 2's complement -- adder
- Add {1, fractions}-- adder
- Carry? -- adder
- Overflow/Underflow
- Infinity, Zero or Nan -- IEEE 754

FPGA Deployment of Neural Network

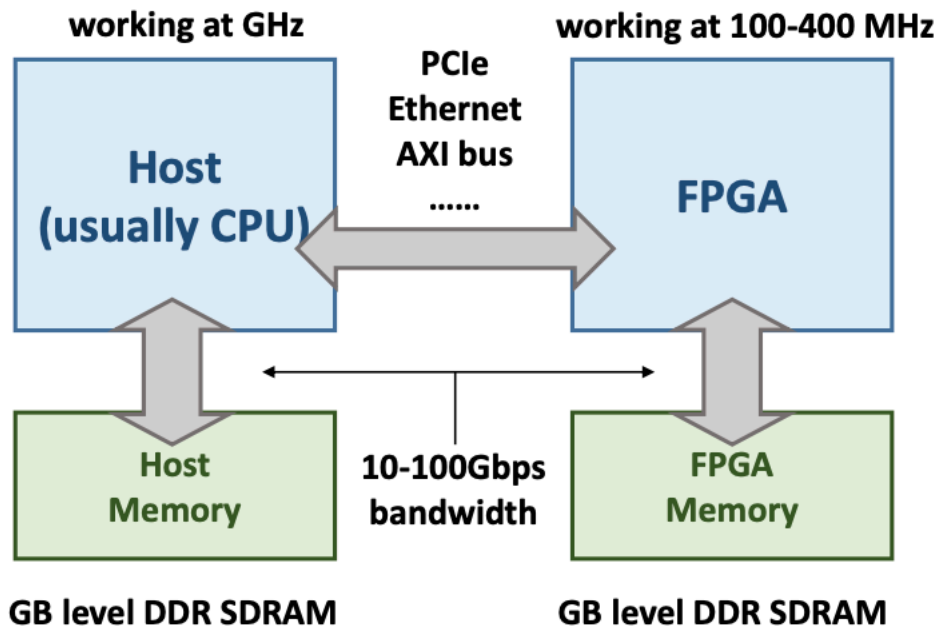


Task pipeline --- Efficient Improvement

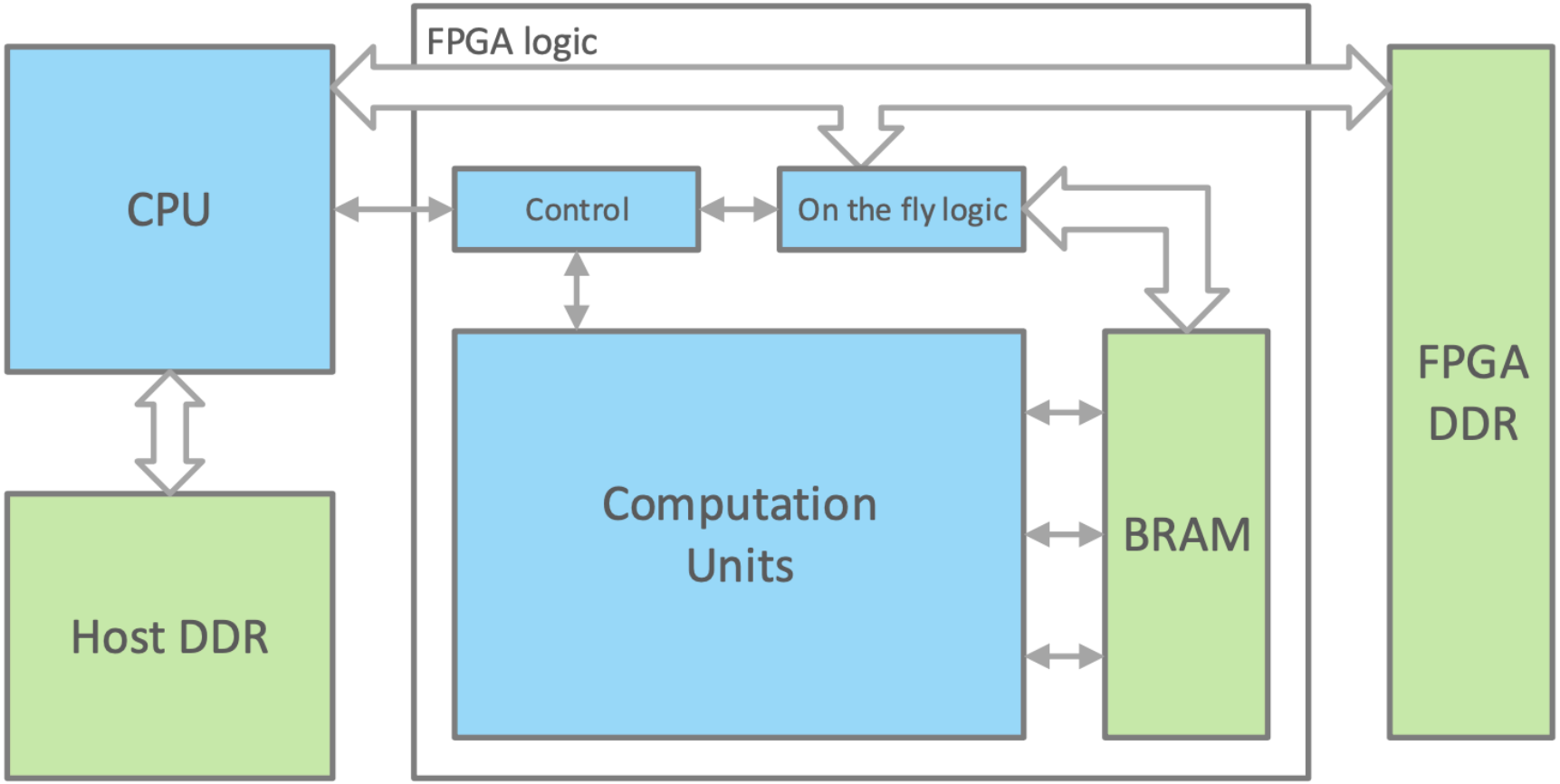


Controller design?

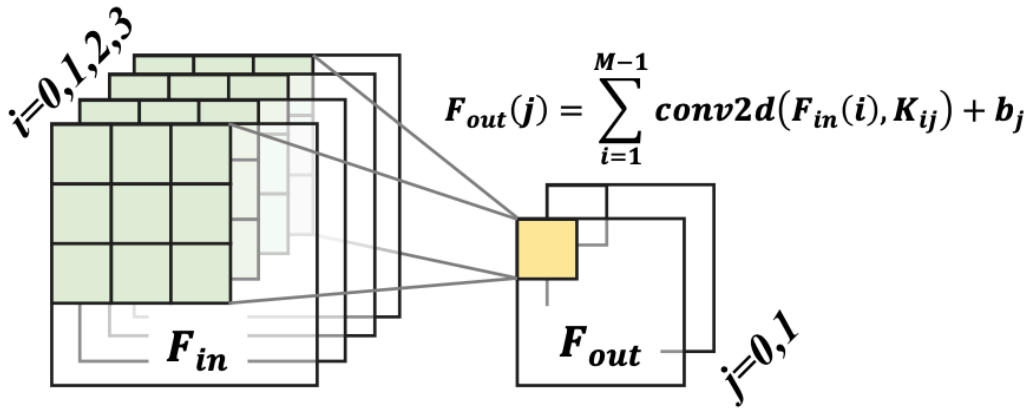
FPGA + ARM

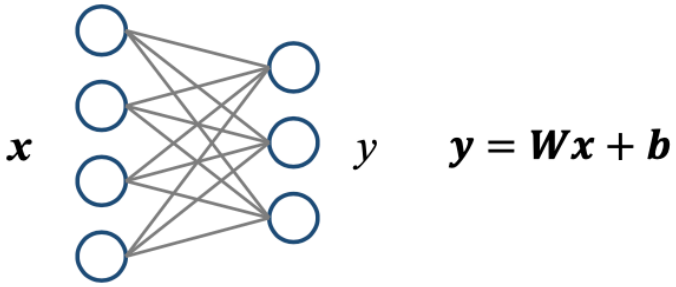

working at GHz
working at 100-400 MHz
PCIe
Ethernet
AXI bus
.....
Host
(usually CPU)
FPGA
Host
Memory
10-100Gbps
bandwidth
FPGA
Memory
GB level DDR SDRAM
GB level DDR SDRAM

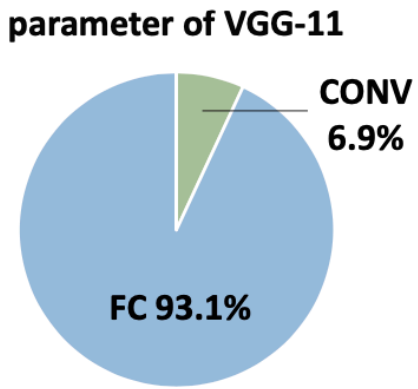
A typical FPGA-based neural network accelerator system

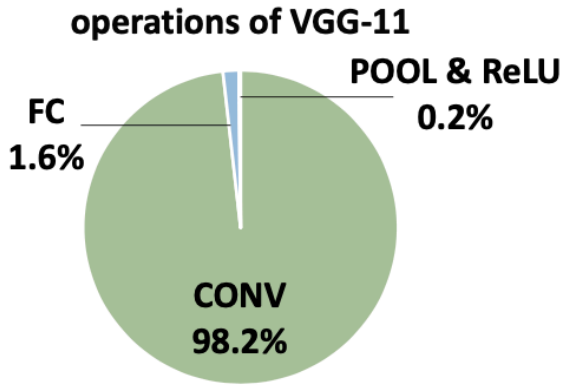


Computation and parameter of a typical NN model









Convolution and FC is the mainly computation.

Matrix Multiplication

Resource consumption of mul-add operation

SmartMore 思谋

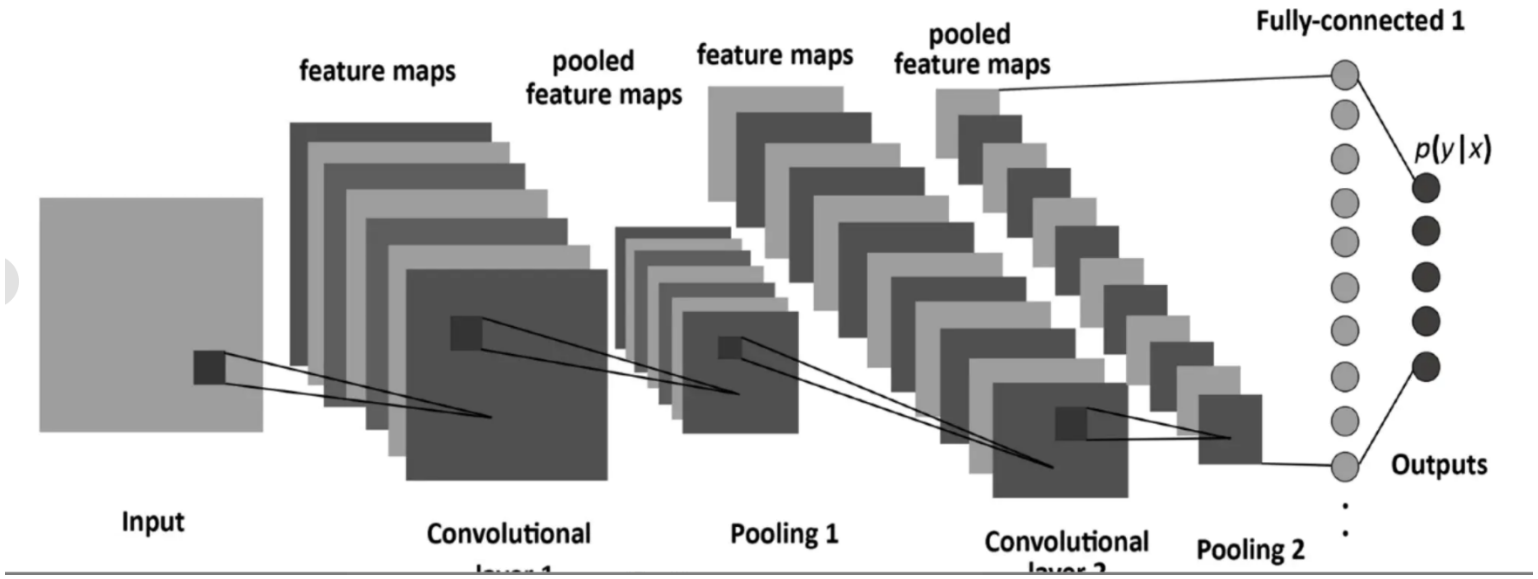
	Xilinx Logic				Xilinx DSP		
	multiplier		adder		multiply & add		
	LUT	FF	LUT	FF	LUT	FF	DSP
fp32	708	858	430	749	800	1284	2
fp16	221	303	211	337	451	686	1
fixed32	1112	1143	32	32	111	64	4
fixed16	289	301	16	16	0	0	1
fixed8	75	80	8	8	0	0	1
fixed4	17	20	4	4	0	0	1

Floating-point number computation cost is much higher than fixed-point number

Neural Networks Implementation

SmartMore 思谋

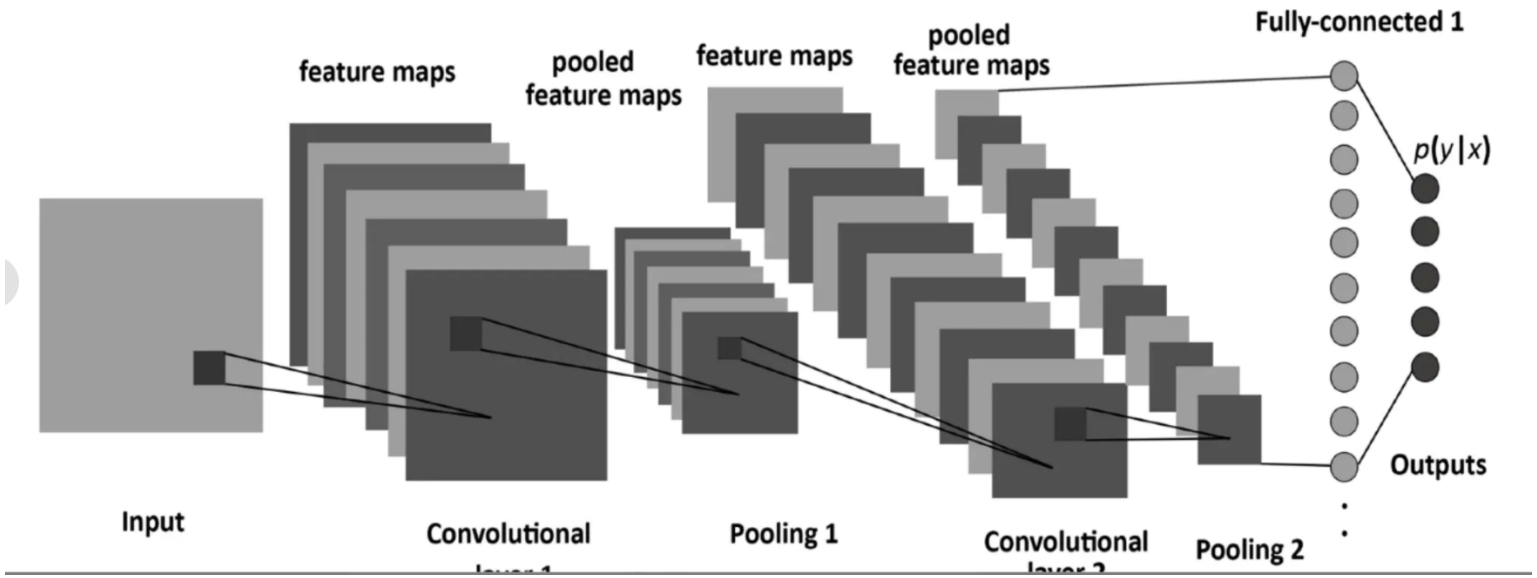
Build operations in the network



Neural Networks Implementation

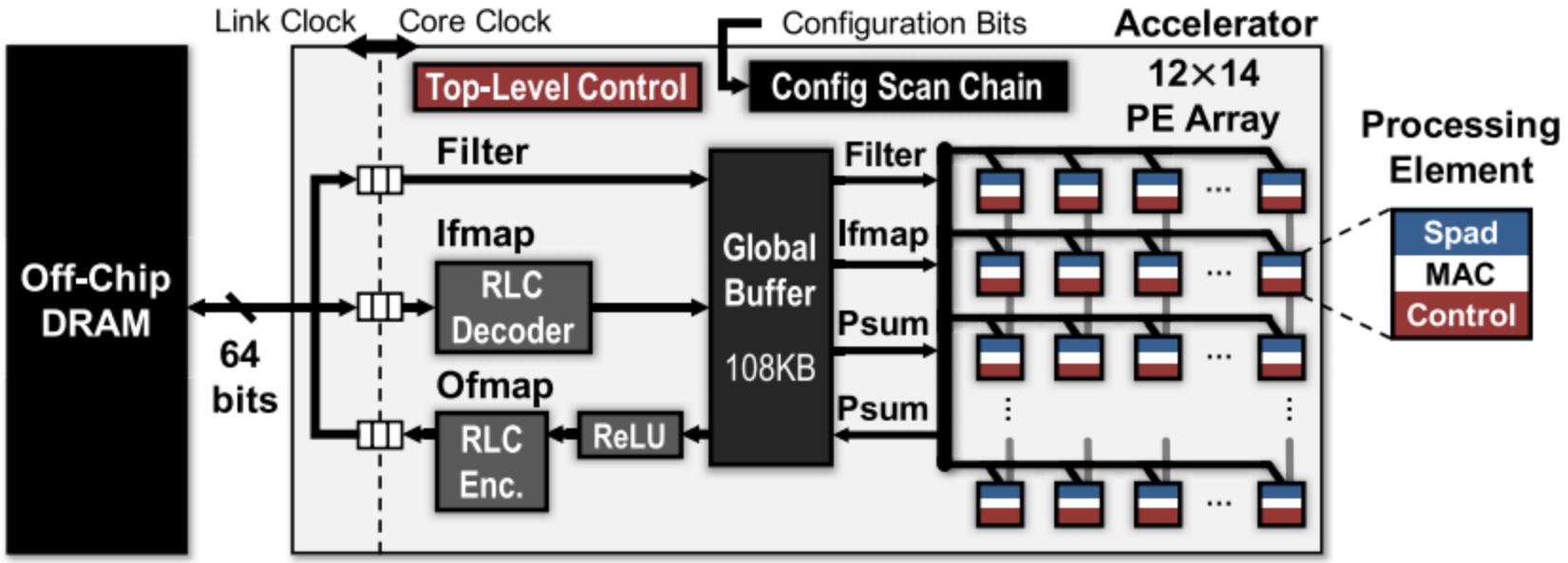
SmartMore 思谋

Parallel operations (Operation channels)



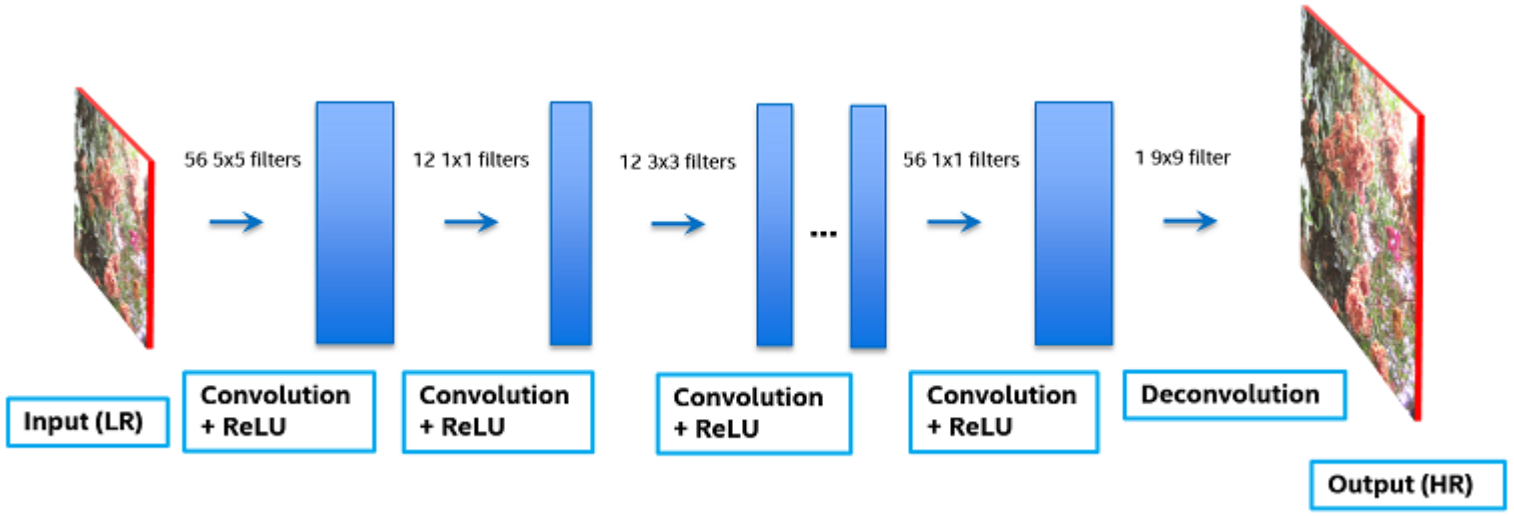
Neural Networks Implementation

Data transactions, storage. (Calculation data, parameters)



Challenges

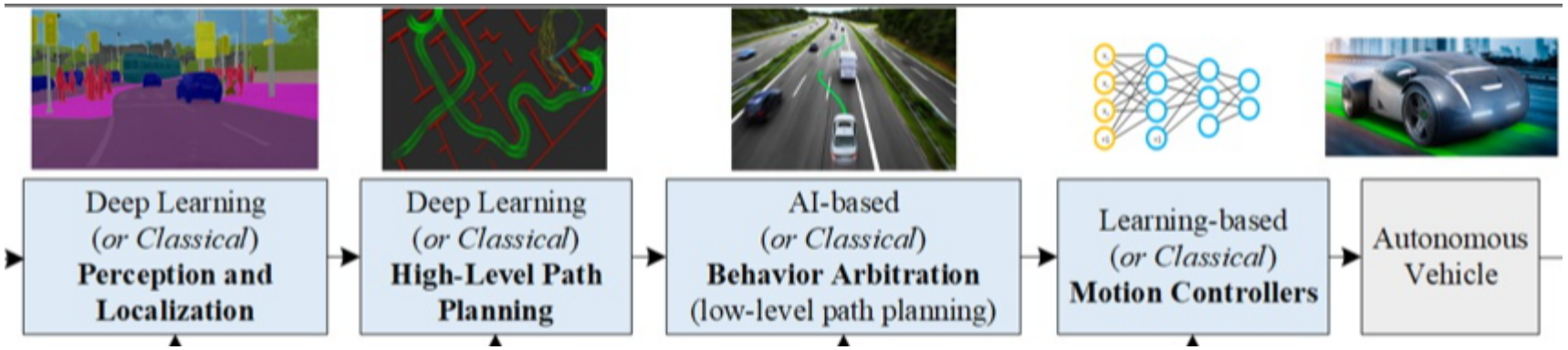
- Variety networks.
  - SR: High data bandwidth. High calculation resources needed.



Challenges

Low-latency models. E.g. autonomous vehicles and manufacturing operations.

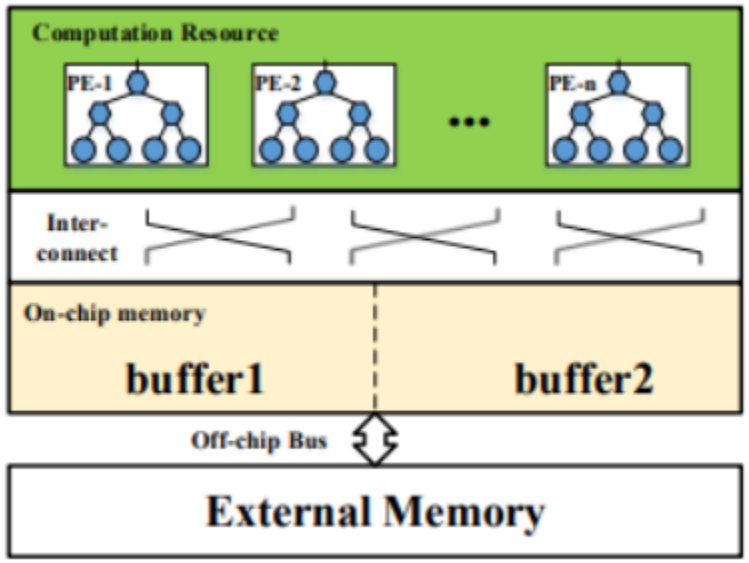
- Hight frequency.
- Low latency design.



Challenges

Performance and Resources balance

- Computation-limited
- Memory bandwidth-limited



Resolutions

SmartMore 思谋

Building FPGA operation library.

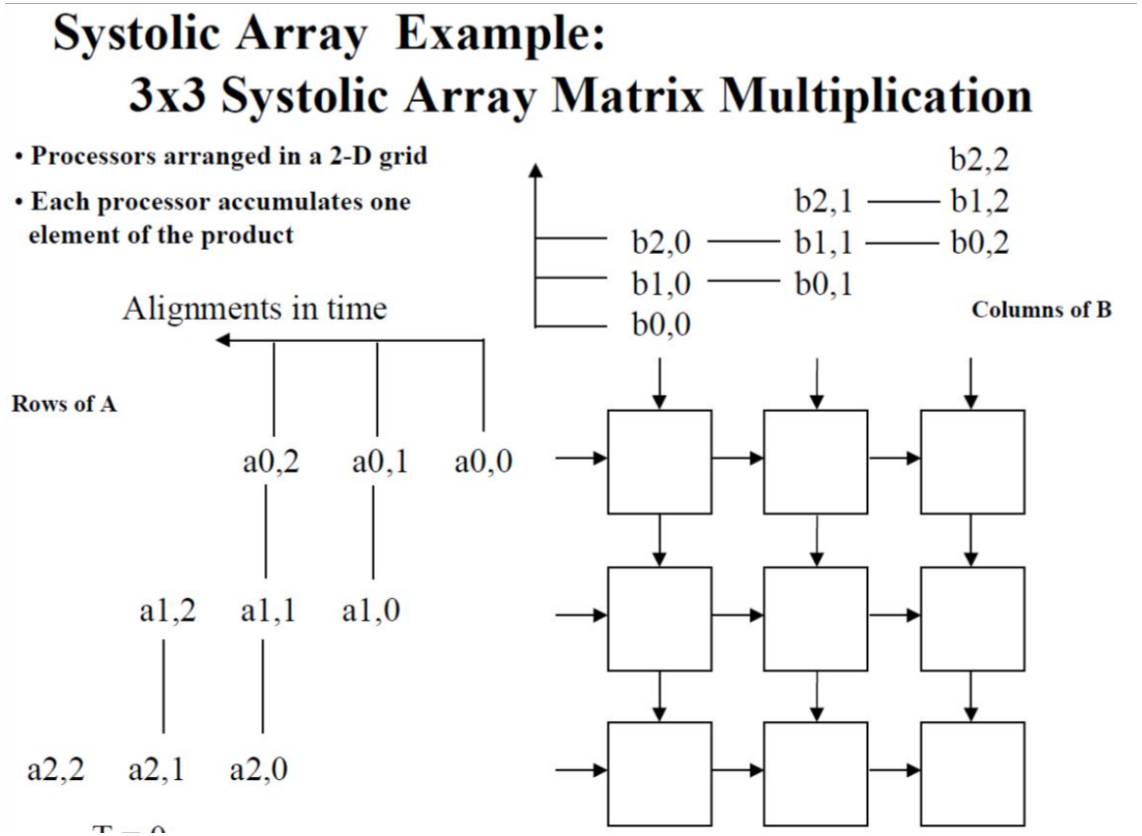
- Conv
- FC
- ReLu
- Upsample
- Resize
- Pooling
- ...

Model of FPGA Implementation

Resolutions

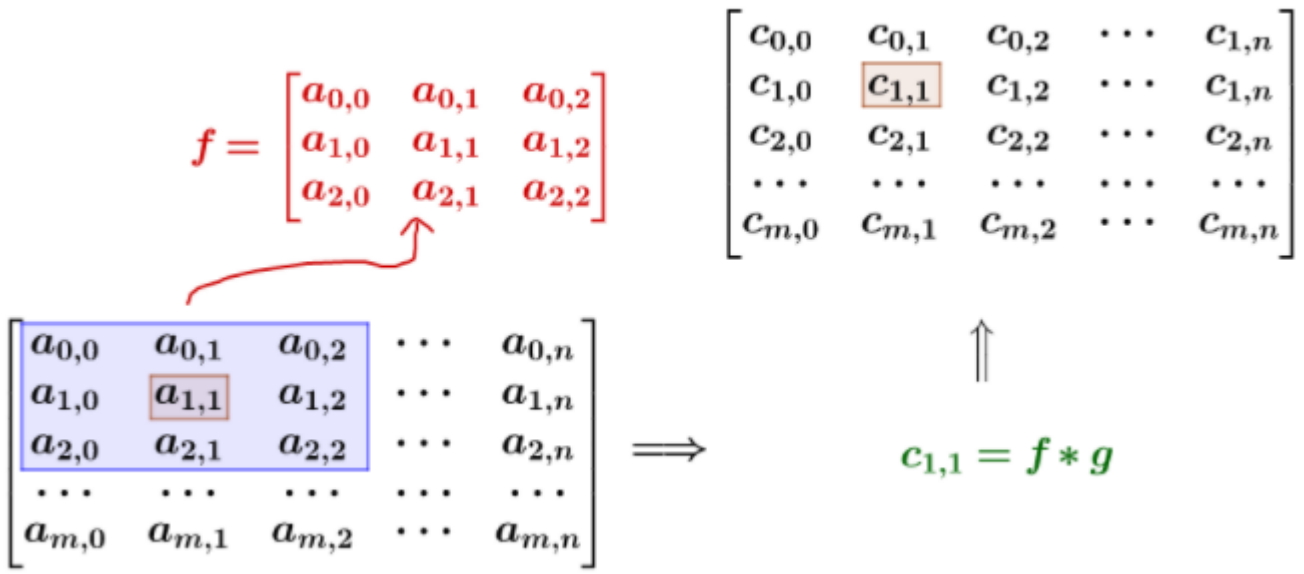
Optimize operation implementation.

- Better structure. (e.g. Mat-mul: systolic array)



Resolutions

Optimize storage strategy



Resolutions

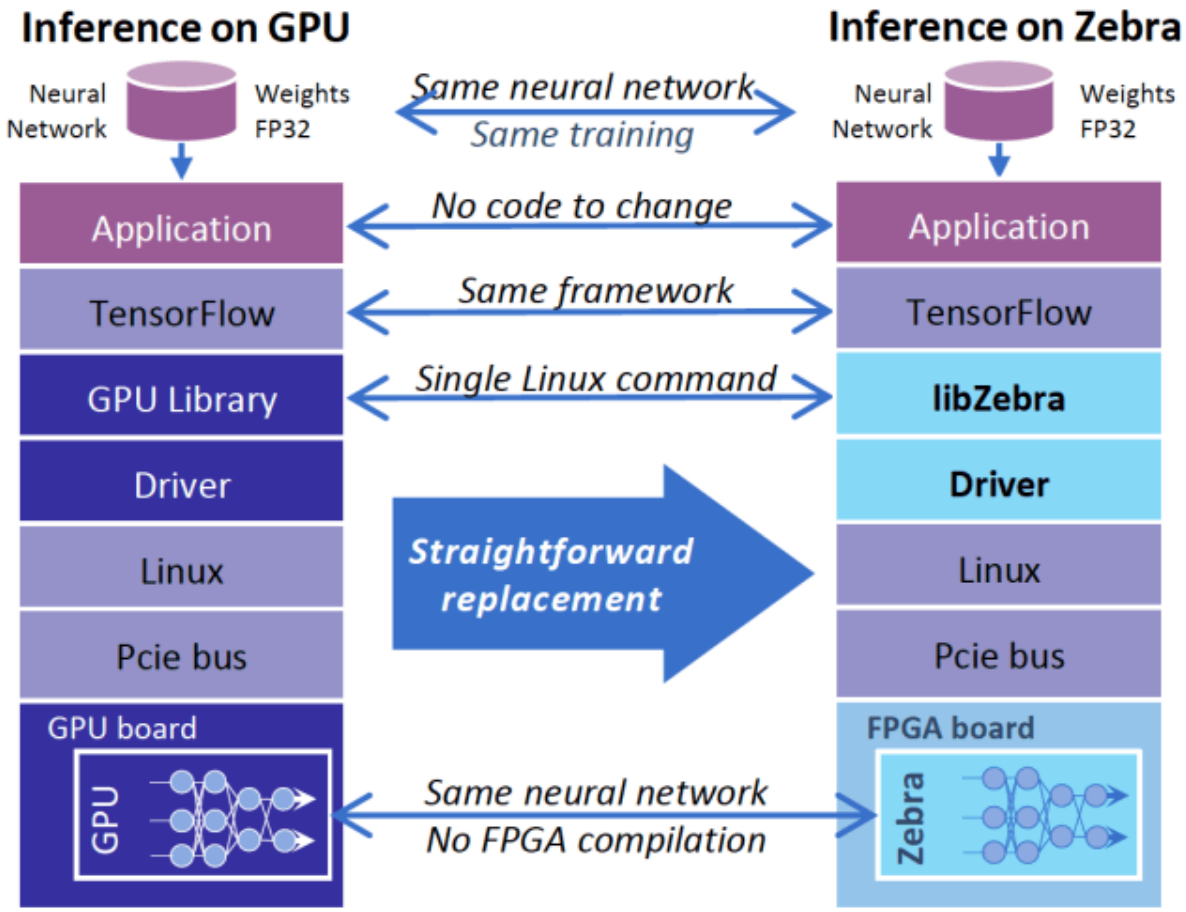
Optimize based on network needs.

- Speed
- Resources usage
- Power consumption
- ...

Resolutions

Design automation (under exploration)

- Model to Verilog (Zebra)

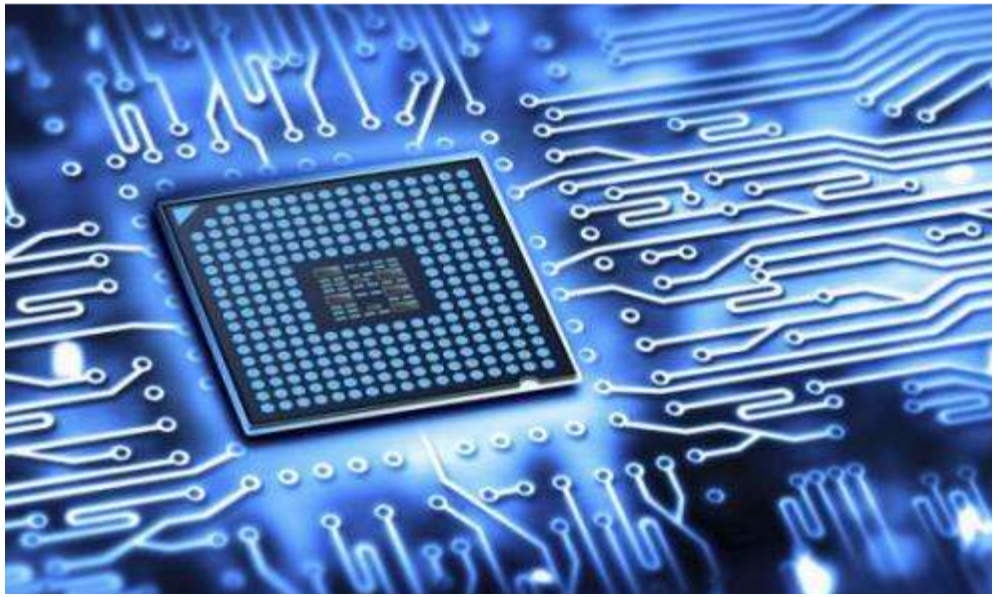


Role in product development

SmartMore 思谋

Model -> RTL design -> FPGA -> ASIC

The Goal Is ASIC



Q&A

Feel free to contact me if you have any further questions.

vince.zhang@smartmore.com

Thank you!