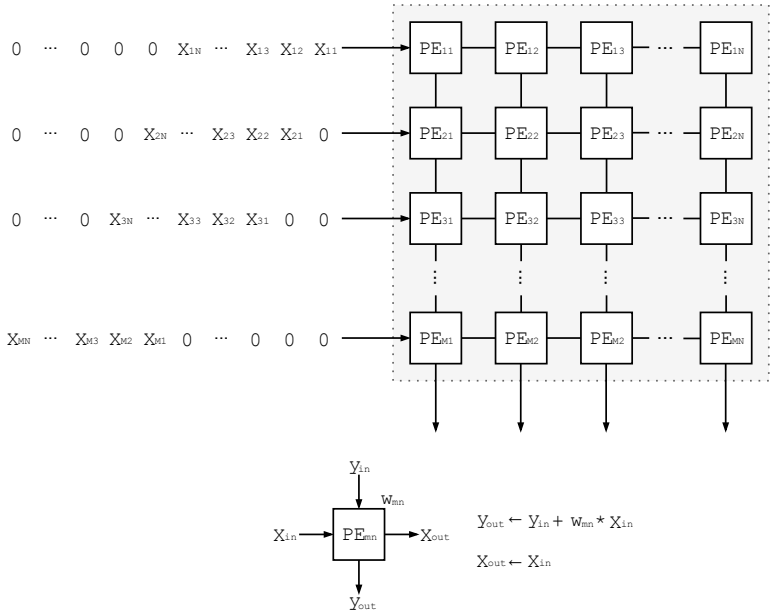
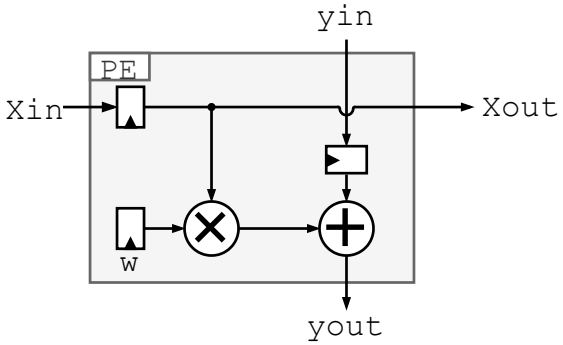
SUSTech Southern University of Science and Technology

Pipeline Systolic Reconfigurable Fixed-Point Accelerator

文章计划

Background





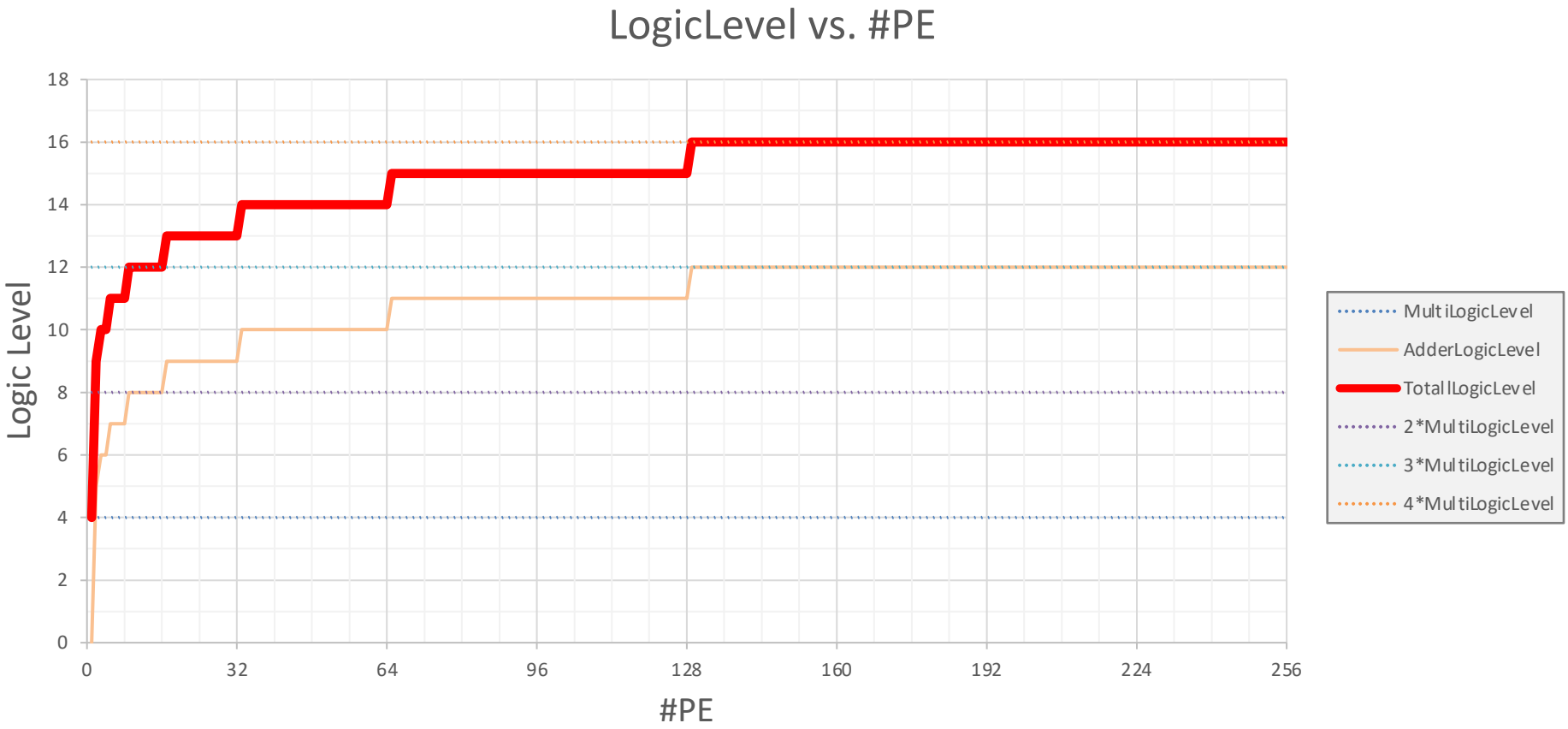

```
Xin   : a bit
Xout  : a bit
w     : b bit
yin   : a+b+⌈log2(n-1)⌉ bit
yout  : a+b+⌈log2(n)⌉ bit
```

```
Multiplier

input1:a bit
input2:b bit
output:a+b bit

Levels of Logic:
  a+b
```

```
Adder

input1:a+b bit
input2:a+b+⌈log2(n-1)⌉ bit
output:a+b+⌈log2(n)⌉ bit

Levels of Logic:
  a+b+⌈log2(n)⌉
```

- 如上图，脉动阵列的Psum长度会随着多次累加的长度而上升（对数上升），其使得加法器的一个input及output长度的增加，从而增加了Logic-Level，增加延时。
- 混合精度加速器的基础W/X带宽为2bit，而log2(256)=8。对于低精度脉动阵列而言，这部分增加的延时对乘加速度有很大影响。

Background



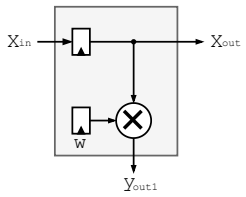
- 以2bit*2bit为基本MAC单元的阵列，当阵列宽度达到N=256时，**加法**的Logic-Level（延时）是**乘法**的3倍。
- 若对MAC采取分段式运行4-states Pipeline，则能提高4倍的运行速度。

Architecture of PE

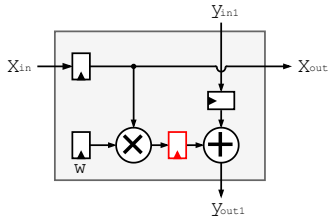
```
n >= 1  (从1开始计数)
t = (a+b+⌈log2(n)⌉) / (a+b+1)
PEmn has ⌈t⌉ states pipeline
Specially, when ⌈t⌉=t
PEmn has Cout
```

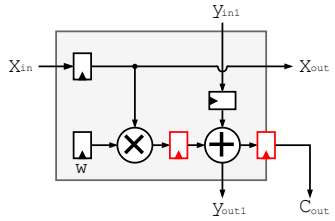
```
Xin  : a bit
Xout : a bit
w    : b bit
yin  : a+b+[log2(n-1)] bit
yout : a+b+[log2(n)] bit

yin1  ... yin[t-1]   : a+b  bit
yout1 ... yout[t-1]  : a+b  bit

yin⌈t⌉  : (a+b+⌈log2(n-1)⌉)%(a+b) bit
yout⌈t⌉ : (a+b+⌈log2(n)⌉)  %(a+b) bit

Cout : 1  bit

yin  = [      yin⌈t⌉  ... yin2  yin1  ]
yout = [ Cout yout⌈t⌉ ... yout2 yout1 ]
```

1 states:

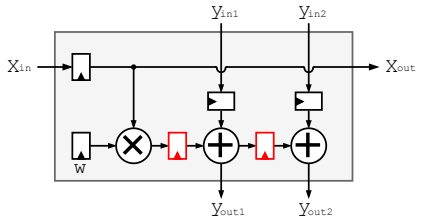


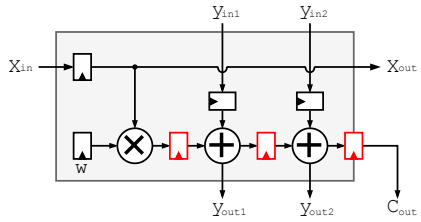
2 states:



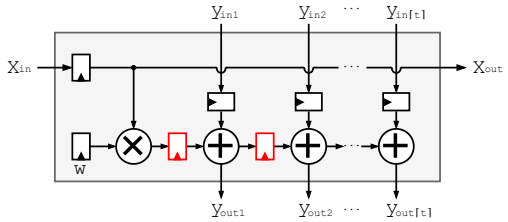


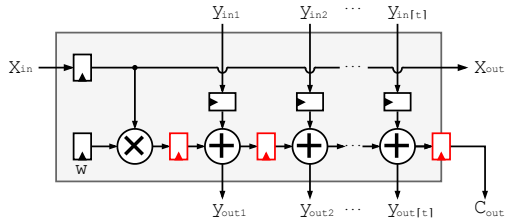
3 states:





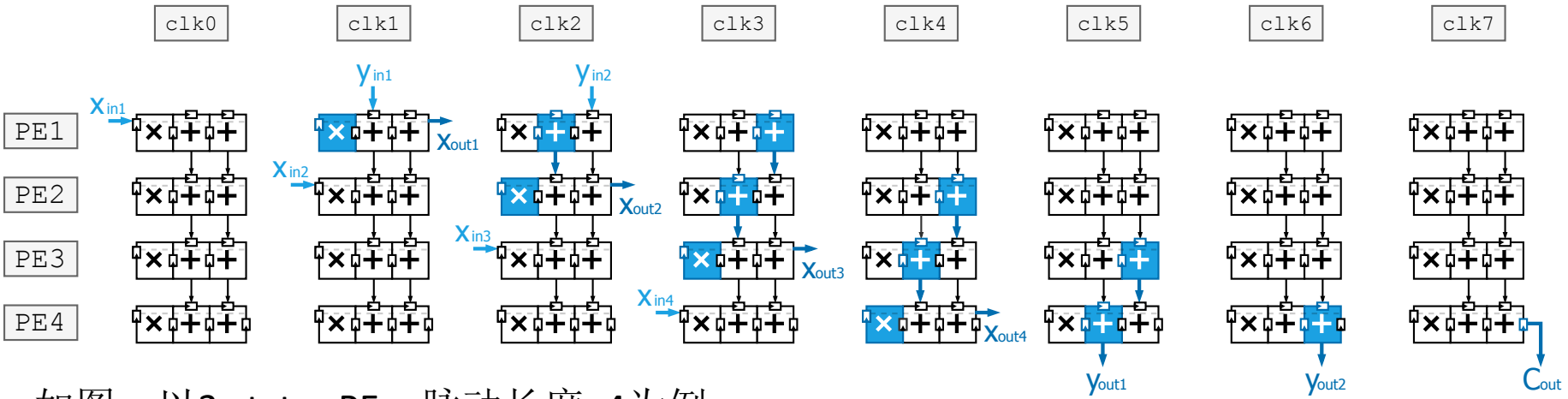
⌈t+1⌉states:





- 左图是多级pipeline的设计
- PE内部的pipeline长度依据y_{in}的长度积累
- 阵列中第n排PE的架构相同，每当log2(n)为整数时，psum位宽+1。
- 位宽没增加a+b bit，新增1state的pipeline
- 乘法器-加法器的寄存器为a+b bit
- 加法器-加法器的寄存器为1bit
- 从2states开始，每增加一级pipeline，其边际成本=1bit寄存器

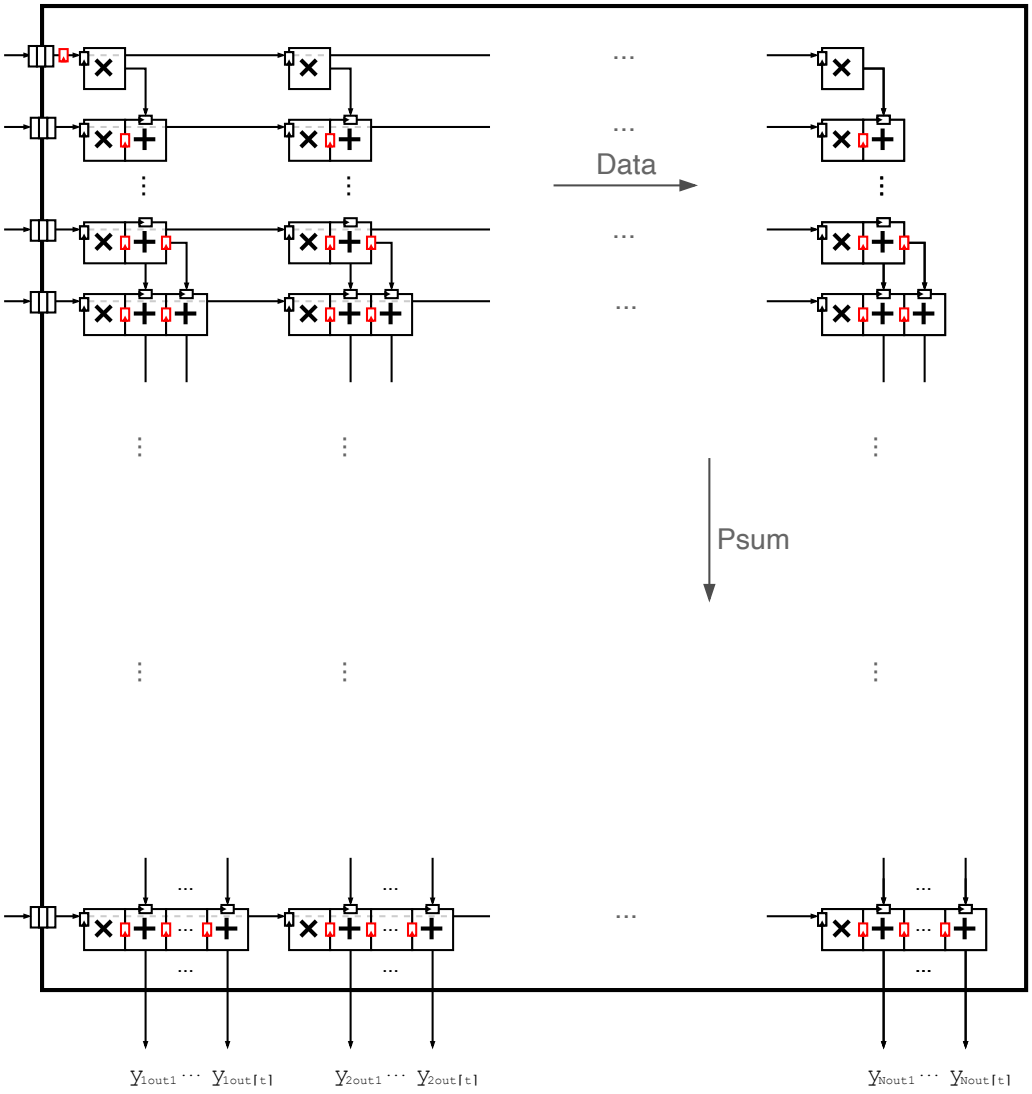
Pipeline + Systolic



如图，以3-states PE，脉动长度=4为例

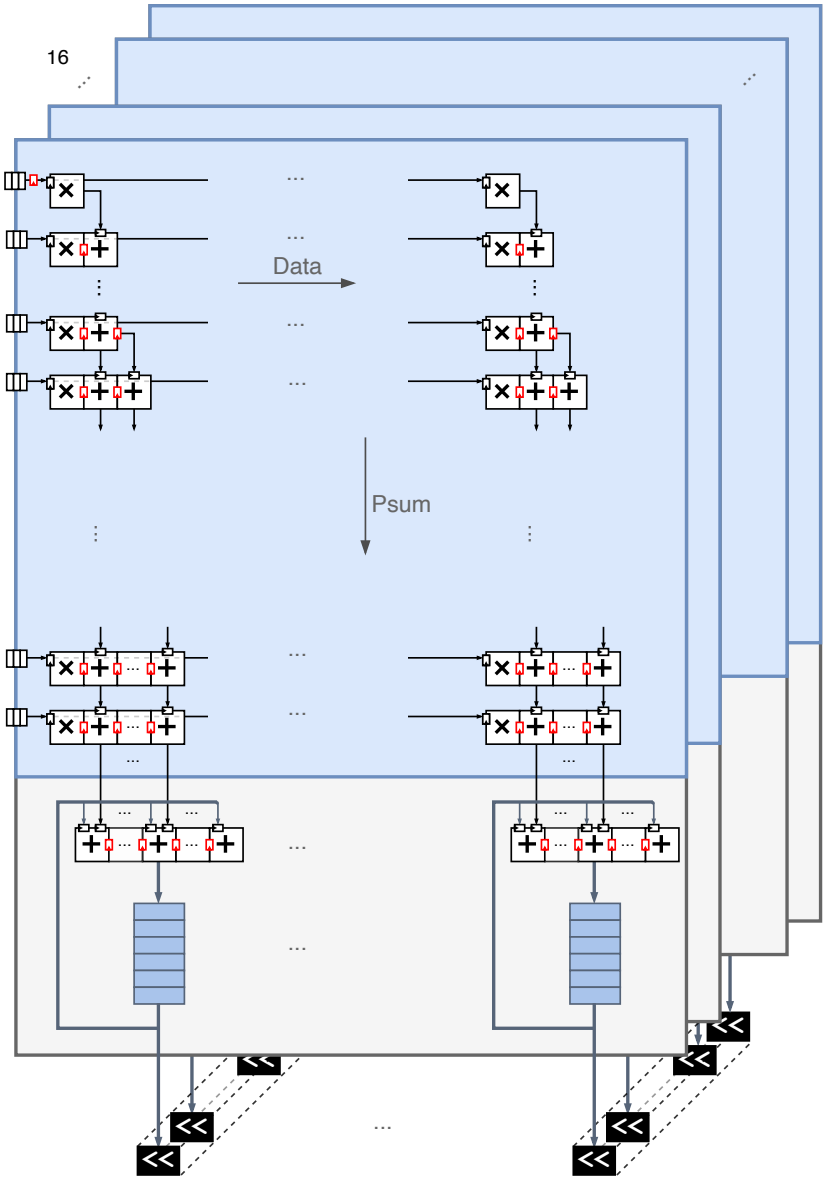
- X_{in1} – X_{in4}分别在clk0 – clk3输入给PE1 – PE4。
- y_{in1} ,y_{in1}分别在clk1,clk2输入给PE1。
- X_{in}一个clk后与PE内部的权重W相乘后生成的乘积X*W
- 乘积与y_{in1} 相加的C_{out1}与y_{in2}相加
- 在某些情况（n为2的指数）下， C_{out1}与y_{in2}相加会产生C_{out2}
- y_1, y_2在PE中流动，相隔1clk，分别
- C_{out2}在clk7时从PE4输出

PE Array



- 增设了pipeline的脉动阵列的输入&控制行为与传统脉动阵列相同。
- 由于增设了pipeline，输出数据为多段数据，相邻数据段相隔1clk延迟输出。

Architecture of Reconfigurable Accelerator



- 加速器采用TPU架构，采用先脉动再移位的思想。在加速器层次进行多精度可重构设计。
- 每层阵列的精度为2bit*2bit，共16层阵列。可支持2-8bit混合精度数据进行矩阵乘法计算。
- 阵列模块后接累加模块&移位模块。累加也采用pipeline加法。

Evaluation and Comparison

PE Comparison	**(1) Bit fusion** **(2) Bit serial**
Accelerator Comparison	(1) TPU (2) Eyeriss
Evaluation	Vgg / Alexnet
Evaluation	(1) Power (2) Area (3) Latency/speed (4) Performance (Gops)/Efficiency (Gops/W) (Mixed precision)

投稿相关信息

时间	周俊卓	满昌海	罗君益
3.29-4.11	PE单元实现&测试		
4.12-4.25	PE阵列实现&测试		
4.26-5.16	多层PE阵列多精度实现&测试		SoC系统搭建
5.17-5.23	加速器总线接口设计	测试模型训练&转化	
5.24-6.6	加速器SoC整体测试（仿真）		文章理论部分
6.7-6.20	FPGA综合+测试		
6.21-7.26	完善文章理论部分 编写实验部分		

	abstract deadline	deadline	length
ASPDAC	2021.7.26	2021.7.26	8 pages
DATE	2021.9.14	2021.9.21	6 pages