flexlogix®

eFPGA + AI

专为AI应用而优化的eFPGA

王成诚博士
Flex Logix联合创始人
高级副总裁

flexlogix Technologies, Inc.

TSMC IP ALLIANCE MEMBER

FPGA现已广泛应用于AI

- Microsoft DataCenters have FPGAs for acceleration including Deep Neural Networks

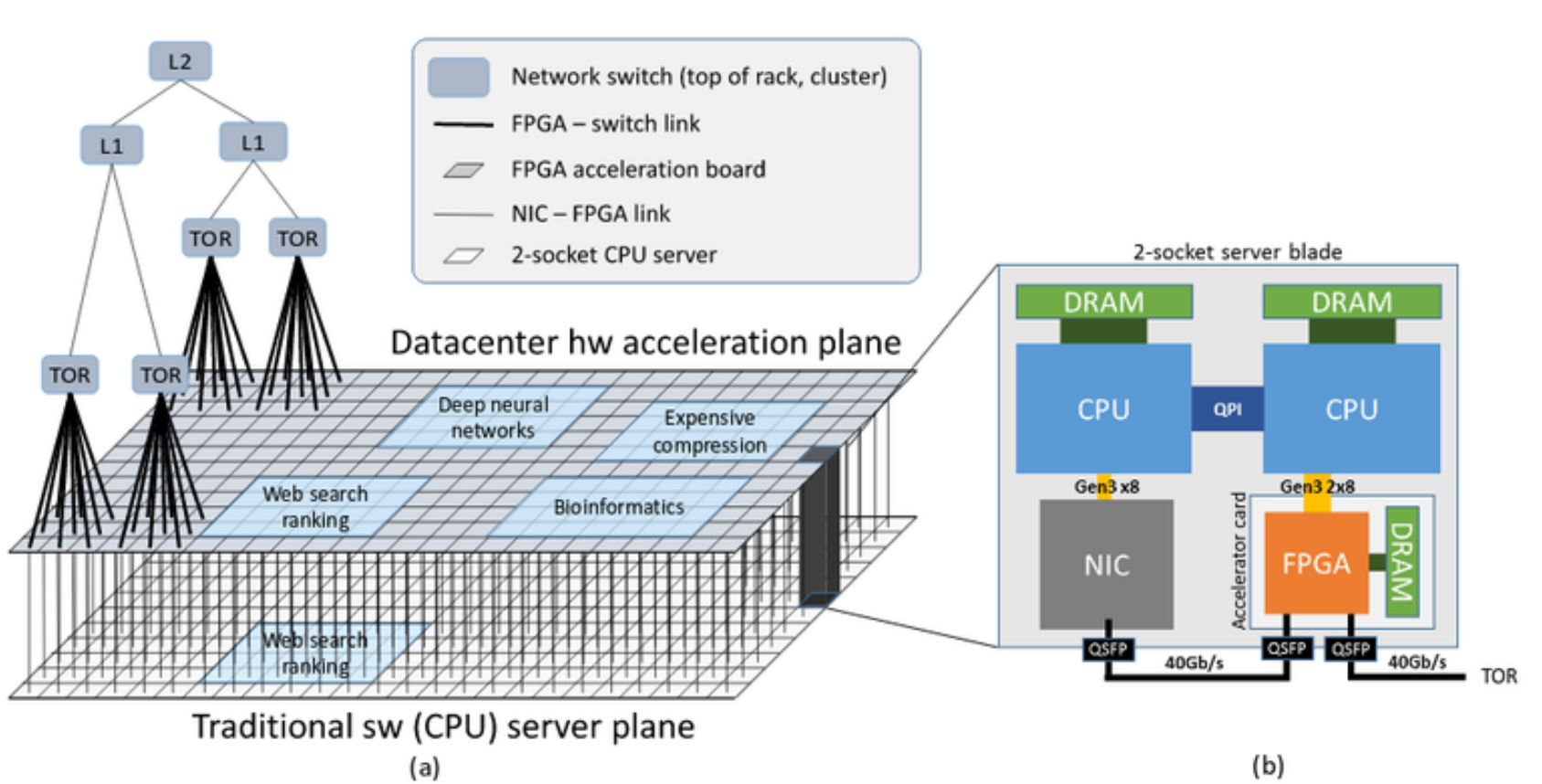
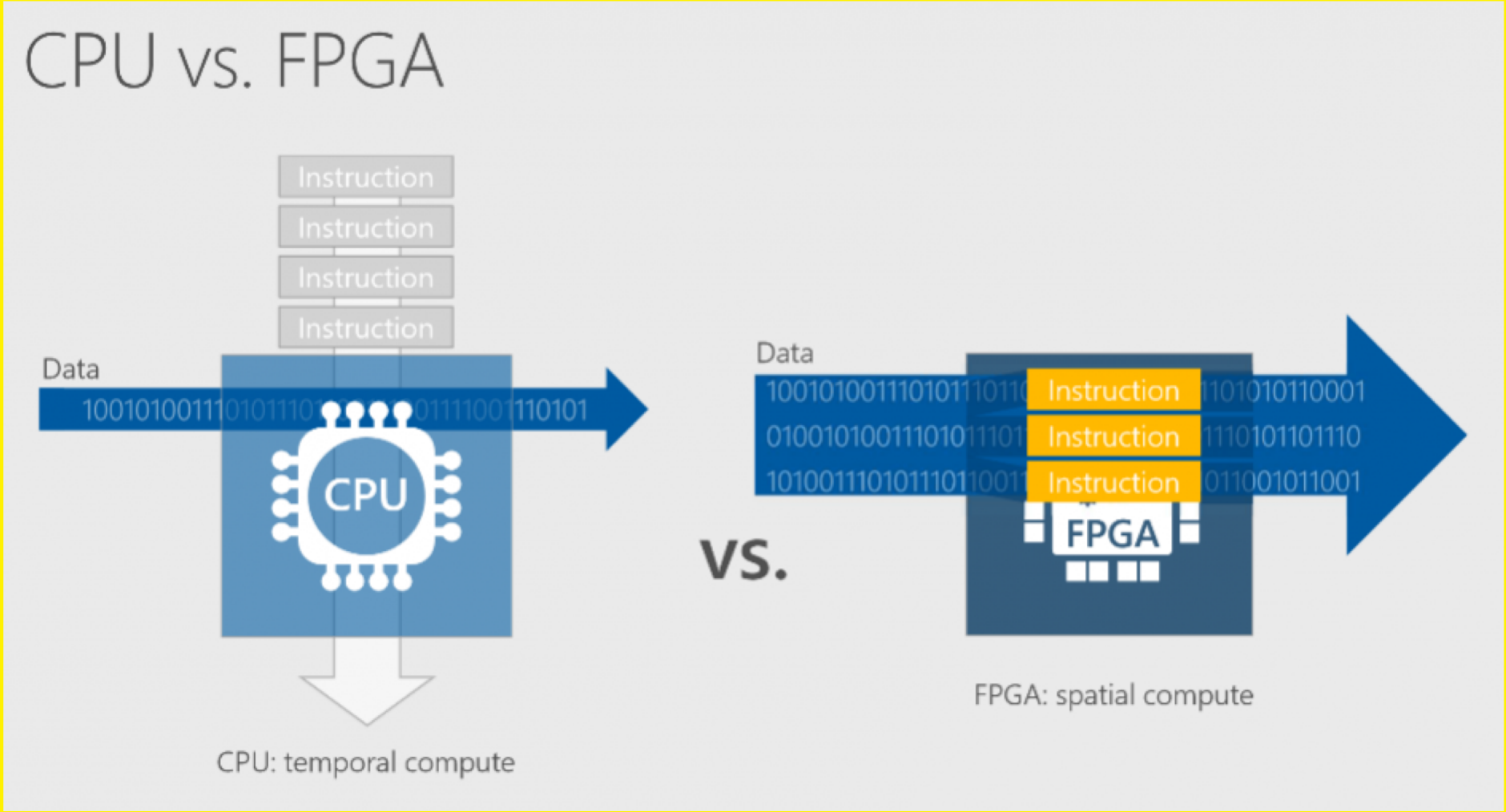


Fig. 1. (a) Decoupled Programmable Hardware Plane, (b) Server + FPGA schematic.

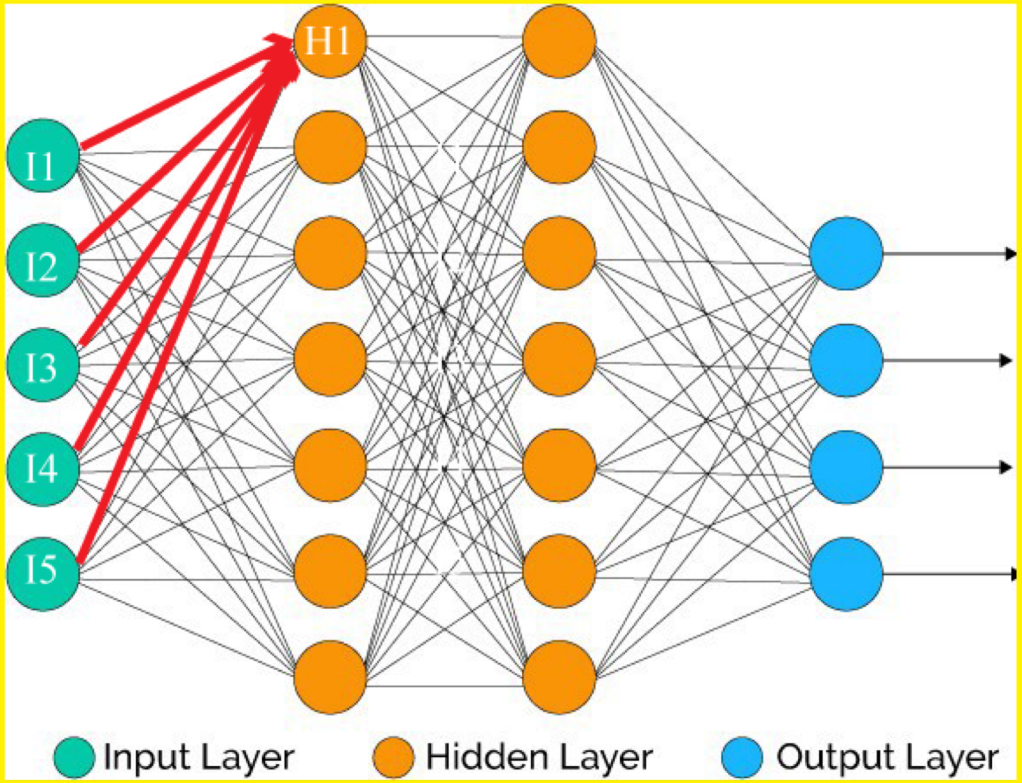
FPGA 是最好的加速器解决方案



flexlogix Technologies, Inc.

TSMC IP ALLIANCE MEMBER

Neural Networks 神经网络的计算主要是矩乘



$$\mathbf{Ax} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \end{bmatrix}.$$

Each layer is a large matrix multiply

flexlogix Technologies, Inc.

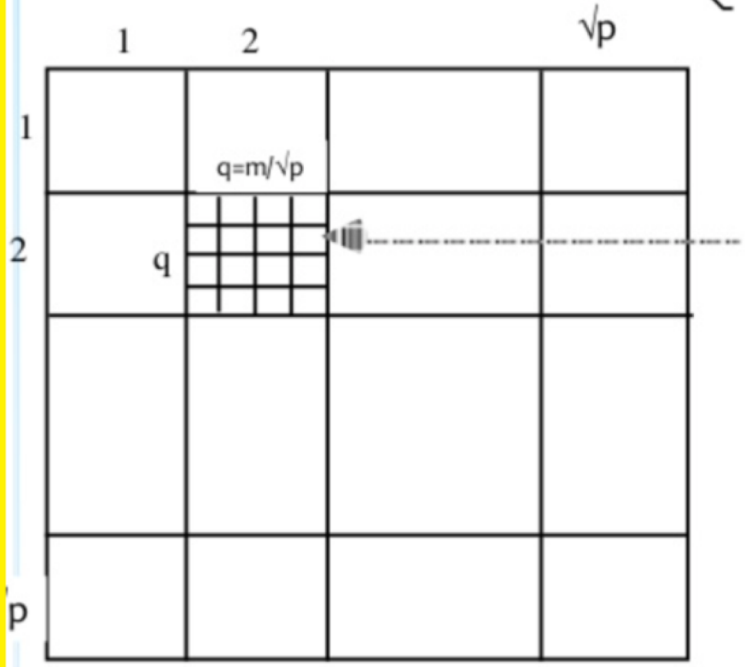
TSMC IP ALLIANCE MEMBER

大型的矩乘利用流水的乘加器来实现

More Efficient Matrix Multiplication (for NUMA)

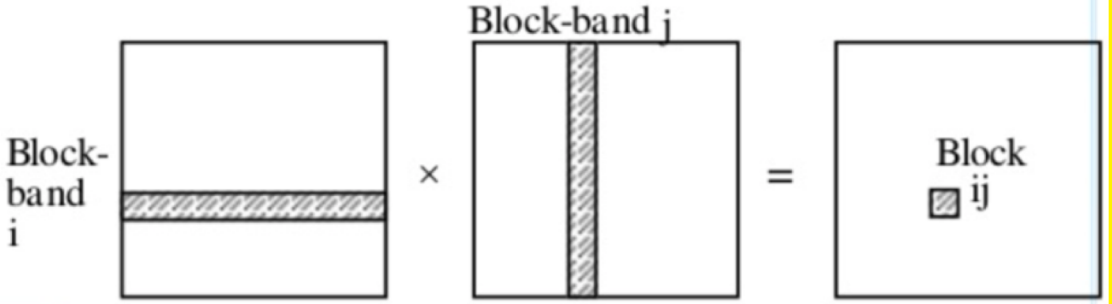
Partition the matrices into *p* square blocks

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} \times \begin{pmatrix} E & F \\ G & H \end{pmatrix} = \begin{pmatrix} AE+BG & AF+BH \\ CE+DG & CF+DH \end{pmatrix}$$



One processor computes these elements of C that it holds in local memory

Block matrix multiplication follows the same algorithm as simple matrix multiplication.



flexlogix Technologies, Inc.

TSMC IP ALLIANCE MEMBER

EFLX eFPGA有更多的DSP/MAC资源

- For EFLX DSP tile, the LUT6:DSP ratio is around 50:1
- For Xilinx devices, the LUT6:DSP ratio is around 100:1
- For Altera devices, the LUT6:DSP ratio is around 200:1

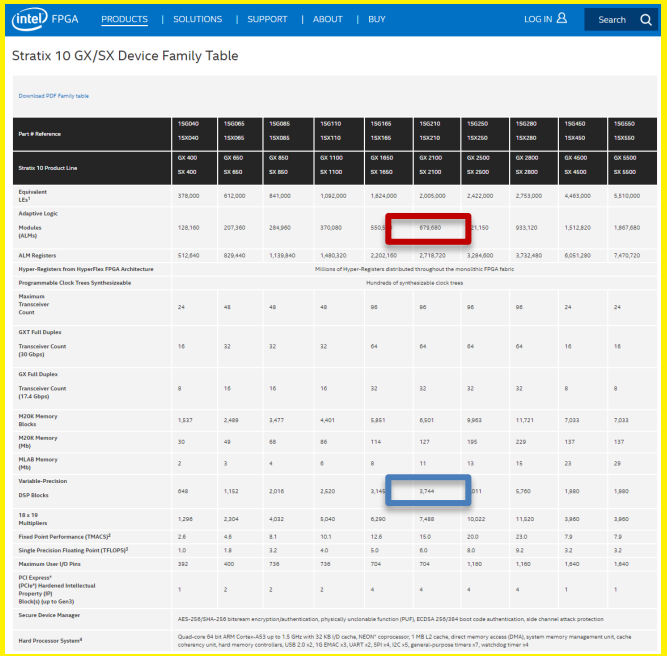
	Device Name	KU3P	KU5P	KU9P	KU11P	KU13P	KU15P
Logic	System Logic Cells (K)	356	475	600	653	747	1,143
	CLB Flip-Flops (K)	325	434	548	597	683	1,045
	CLB LUTs (K)	163	217	274	299	341	523
Memory	Max. Distributed RAM (Mb)	4.7	6.1	8.8	9.1	11.3	9.8
	Total Block RAM (Mb)	12.7	16.9	32.1	21.1	26.2	34.6
	UltraRAM (Mb)	13.5	18.0	0	22.5	31.5	36.0
Clocking	Clock Mgmt Tiles (CMTs)	4	4	4	8	4	11
Integrated IP	DSP Slices	1,368	1,824	2,520	2,928	3,528	1,968
	PCIe® Gen3 x16 / Gen4 x8	1	1	0	4	0	5
	150G Interlaken	0	0	0	1	0	4
	100G Ethernet w/RS-FEC	0	1	0	2	0	4
I/O	Max. Single-Ended HD I/Os	96	96	96	96	96	96
	Max. Single-Ended HP I/Os	208	208	208	416	208	572
	GTH 16.3Gb/s Transceivers	0	0	28	32	28	44
	GTY 32.75Gb/s Transceivers	16	16	0	20	0	32
Speed Grades	Extended[1]	-1 -2 -2L -3	-1 -2 -2L -3	-1 -2 -2L -3	-1 -2 -2L -3	-1 -2 -2L -3	-1 -2 -2L -3
	Industrial	-1 -1L -2	-1 -1L -2	-1 -1L -2	-1 -1L -2	-1 -1L -2	-1 -1L -2

Footprint[2, 3]	Dimensions (mm)	HD I/O, HP I/O, GTH 16.3Gb/s, GTY 32.75Gb/s					
B784[4]	23x23[5]	96, 208, 0, 16	96, 208, 0, 16				
A676[4]	27x27	48, 208, 0, 16	48, 208, 0, 16				
B676	27x27	72, 208, 0, 16	72, 208, 0, 16				
D900[4]	31x31	96, 208, 0, 16	96, 208, 0, 16		96, 312, 16, 0		
E900	31x31			96, 208, 28, 0		96, 208, 28, 0	
A1156[4]	35x35				48, 416, 20, 8		48, 468, 20, 8
E1517	40x40				96, 416, 32, 20		96, 416, 32, 24
A1760	42.5x42.5						96, 416, 44, 32
E1760	42.5x42.5						96, 572, 32, 24

Footprint compatible with 20nm UltraScale Devices with same footprint identifier

Kintex® UltraScale+™ FPGAs

Notes:
1. -2LE (Tj = 0°C to 110°C). For more details, see the Ordering Information section in DS890, UltraScale Architecture and Product Overview.
2. Maximum achievable performance is device and package dependent; consult the associated data sheet for details.
3. For full part number details, see the Ordering Information section in DS890, *UltraScale Architecture and Product* Overview.
4. GTY transceiver line rates are package limited: B784 to 12.5 Gb/s; A676, D900, and A1156 to 16.3 Gb/s. Refer to data sheet for details.
5. The B784 package is only offered in 0.8mm ball pitch. All other packages are 1.0mm ball pitch.



flexlogix Technologies, Inc.

TSMC IP ALLIANCE MEMBER

EFLX4K eFPGA IP Cores

EFLX 4K
Logic
4K LUT4
>1000 I/O Pins

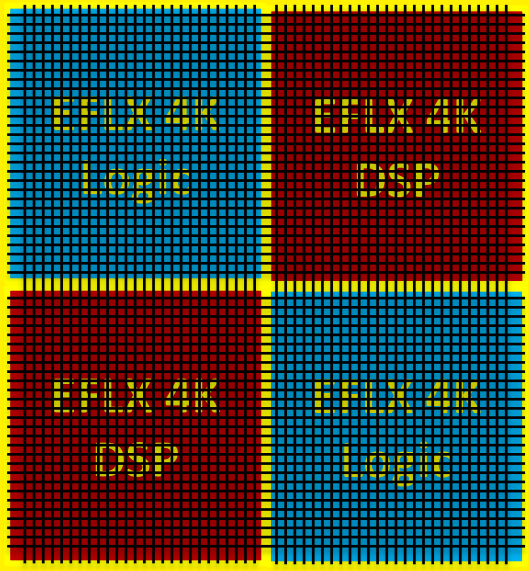
EFLX 4K
DSP
3K LUT4
40 MACs 22x22
>1000 I/O Pins

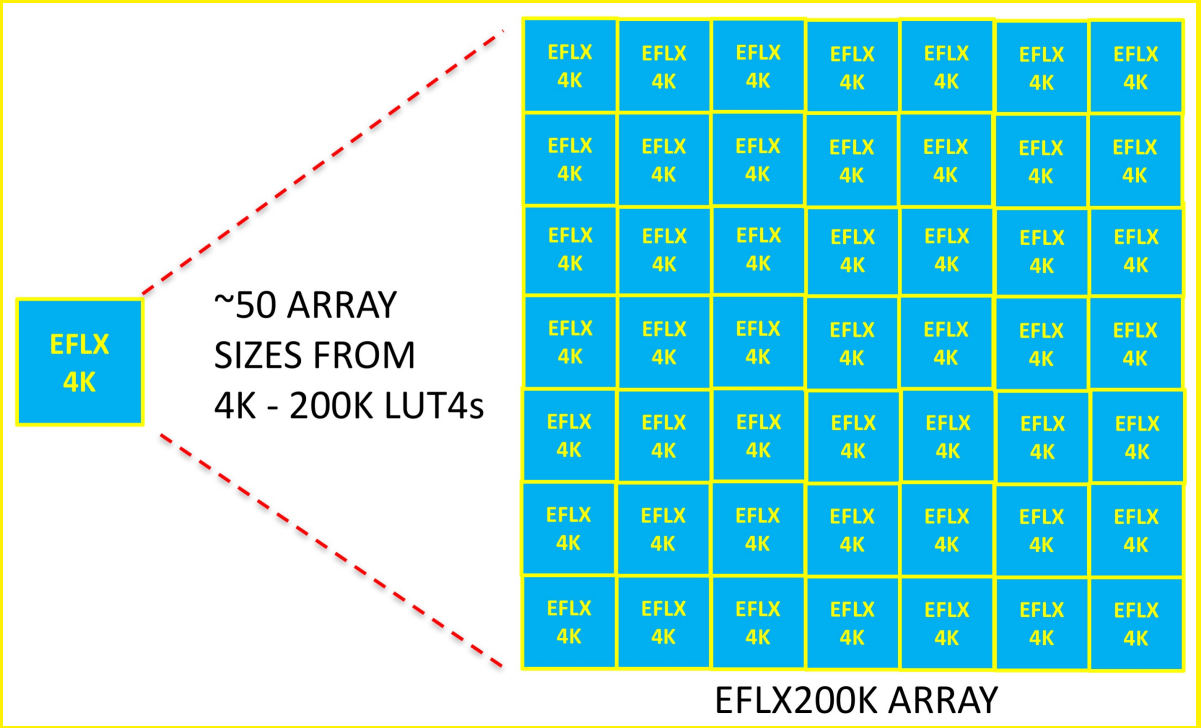
EFLX 4K is a complete eFPGA encircled by a narrow trip of interface pins (a few % of area).

There are two versions

- all-logic and DSP where ~1/4 of the logic is replaced by strips of pipelined MACs
- Both are the same size and exactly the same at the edges

EFLX4K IP核可以任意拼成最大7x7的阵列

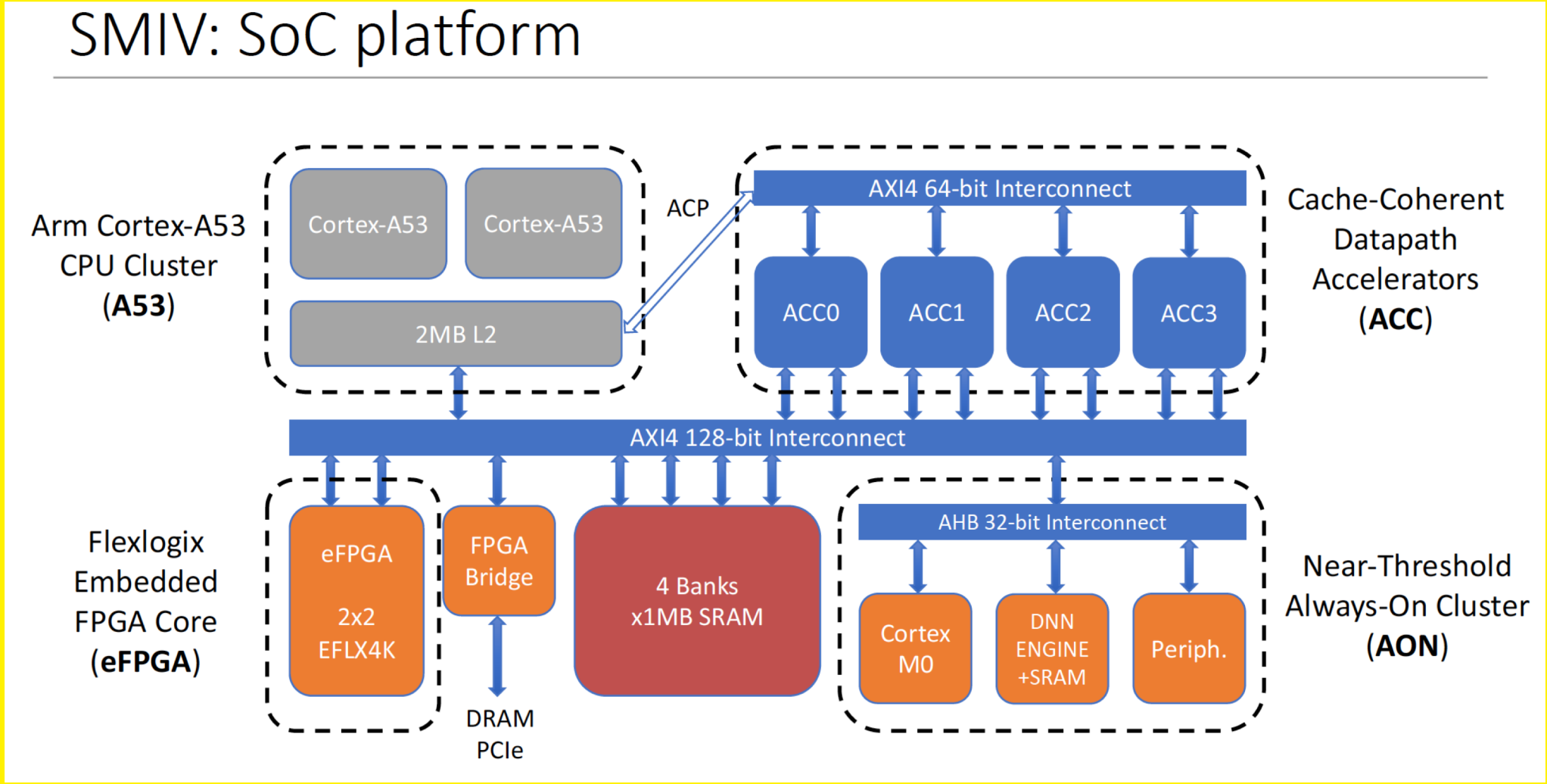




flexlogix
Technologies, Inc.

TSMC IP ALLIANCE MEMBER

哈佛大学的SMIV SoC采用了EFLX4K eFPGA TSMC16FFC



flexlogix Technologies, Inc.

TSMC IP ALLIANCE MEMBER

哈佛大学用EFLX eFPGA来实现 DNN

Embedded FPGA IP

Efficient interconnect enables density and scalability

Density & performance similar to full custom FPGA

Scalable and flexible array size and layout

Logic tile: 2,520 LUT6 + 21 kbits distributed mem

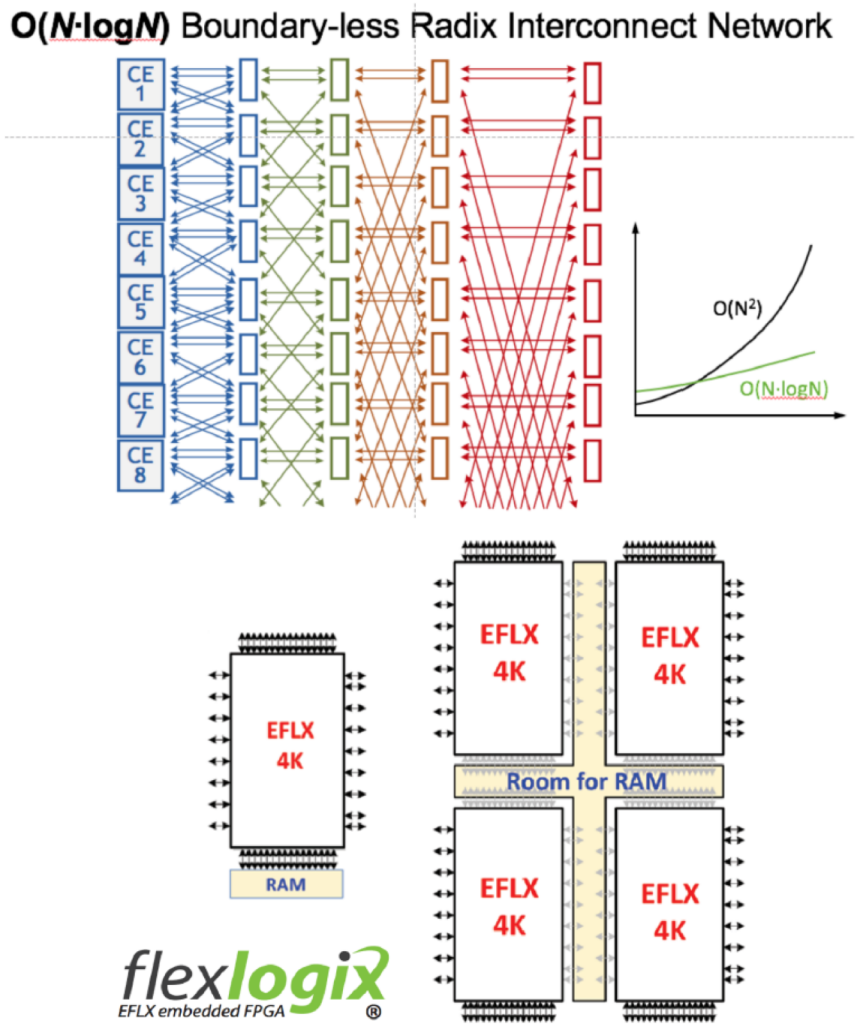
DSP tile: 40 22b hardware MACs + 1,880 LUT6

SMIV contains a 2x2 eFPGA array

2x logic tiles and 2x DSP tiles

Total ~9,000 LUT6 logic, 80 hardware MACs, 44kbits RAM

Attached as a first-class citizen on the SoC interconnect



eFPGA是可编程DNN最好的解决方案

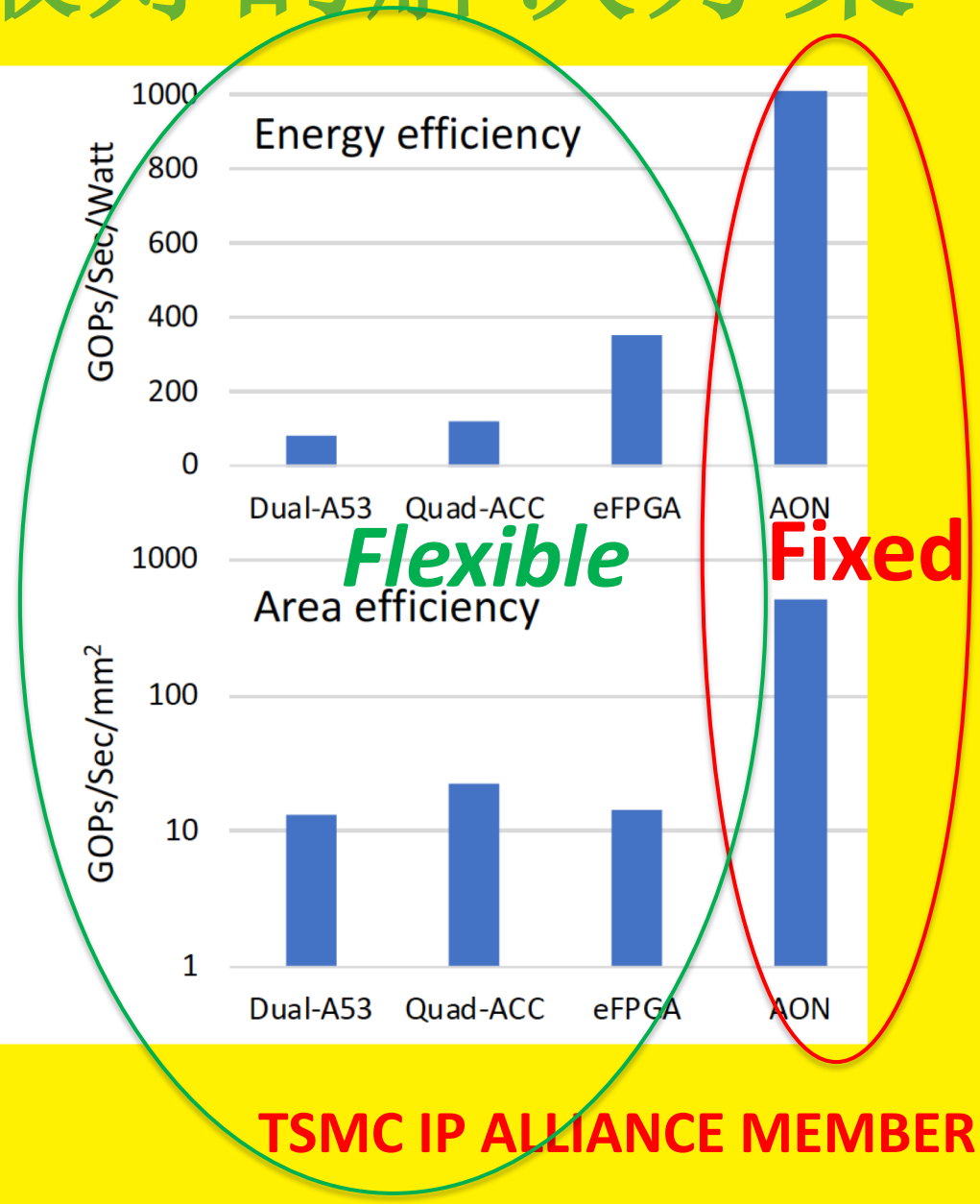
Measured on representative DNN kernels

Energy efficiency range is >10x

- CPU to fixed-function AON
- This also spans the whole flexibility spectrum
- AON will not benefit from new algorithm advances
- eFPGA is 4.5x energy efficiency of CPUs

Area efficiency heavily impacted by SRAM

- Important to share on-chip SRAM resources efficiently
- ACP allows accelerators to use large L2 cache
- eFPGA has the area overhead of reconfiguration



flexlogix Technologies, Inc.

TSMC IP ALLIANCE MEMBER

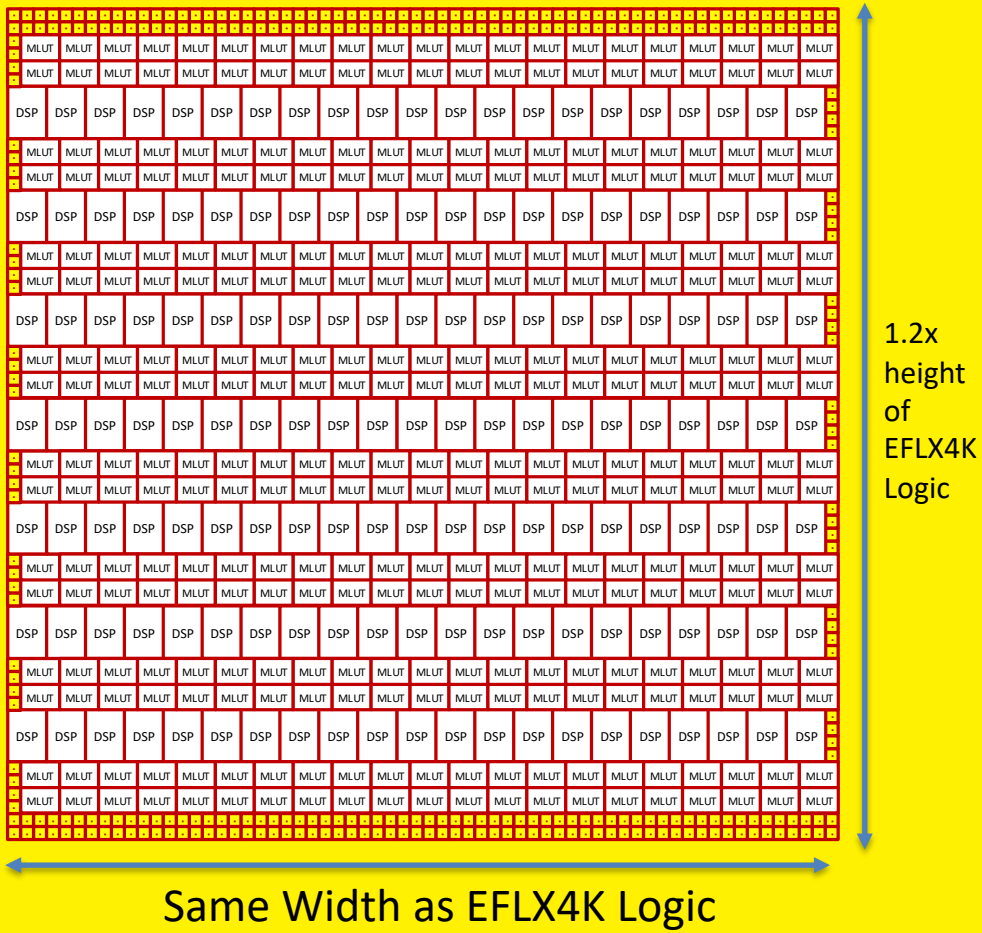
现有FPGA的乘加器需要优化

- EFLX DSP is 22bx22b MAC
- Xilinx DSP is 25bx18b MAC
- Altera DSP is 18bx18b MAC
- BUT, most NN applications use 8b x 8b MAC, and uses millions of them for inference (millions of weights)
  - Option to perform 16x8b at half-rate, or 16bx16b at quarter-rate
- EFLX can fit 6 x 8-bit MAC into footprint of 1 traditional DSP

EFLX-AI核：比现有FPGA多10倍以上的MAC

- Optimized for AI
  - 8/16 bit MACs
  - More MACS, fewer LUTs
  - >10x MACs/second vs EFLX4K
  - More memory (MLUTS=RBBM)
- 441 MAC8s/tile, >11x EFLX4K DSP
  - ~1GHz worst case operation in TSMC16FFC
  - **~440 GMACs/second for 1 core**
  - **22 TeraMACs/second for 7x7 array**
- 16x16, 16x8 and 8x16 modes too
- Reconfigurable any time
- ~1.2 mm^2 in TSMC16FFC
- Available on any TSMC process in 6-8 months

Reconfigurable building block for 8 or 16bit Matrix Multipliers for AI of any size



flexlogix Technologies, Inc.

TSMC IP ALLIANCE MEMBER

而更高级的应用需要支持Tensorflow/Caffe

- Neural Network models 直接map到eFPGA，无需通过HLS或RTL

```
layer {
  name: "ip1"
  type: "InnerProduct"
  bottom: "pool2"
  top: "ip1"
  param {
    lr_mult: 1
  }
  param {
    lr_mult: 2
  }
  inner_product_param {
    num_output: 500
    weight_filler {
      type: "xavier"
    }
    bias_filler {
      type: "constant"
    }
  }
}
layer {
  name: "relu1"
  type: "ReLU"
  bottom: "ip1"
  top: "ip1"
}
```

Not used for inference

Not used for inference

```
layer {
  name: "ip1"
  type: "InnerProduct"
  bottom: "pool2"
  top: "ip1"
  inner_product_param {
    num_output: 500
  }
}
layer {
  name: "relu1"
  type: "ReLU"
  bottom: "ip1"
  top: "ip1"
}
```

敬请期待： NMAX™ Neural Network Inference Engine

- 将在十月31日的Linley Processor Conference正式推出
- NMAX is an IP core, like EFLX
  - Maps directly from Caffe/Tensorflow to NMAX without translating to HLS or RTL
  - Reconfigurable, scalable high-speed NN accelerator with low DRAM bandwidth
  - EFLX eFPGA control logic reconfigured at each stage
  - Novel high-bandwidth on-chip memory architecture using XFLX/ArrayLinx interconnect
  - Result is high MAC utilization and low DRAM bandwidth
  - Targeted at edge inference NN applications
  - More data available under NDA for strategic customers

flexlogix®

eFPGA + AI

专为AI应用
而优化的eFPGA

flexlogix Technologies, Inc.

TSMC IP ALLIANCE MEMBER