



ECE 6775
High-Level Digital Design Automation
Fall 2024

Field-Programmable Gate Arrays (FPGAs)



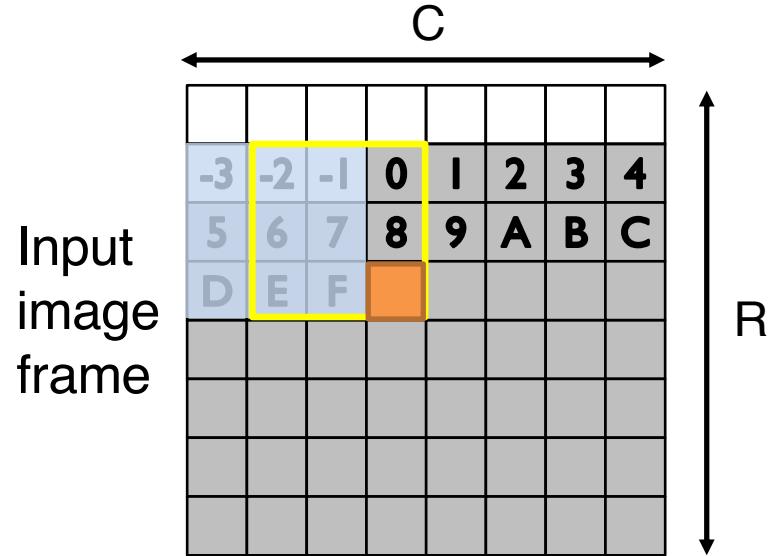
Cornell University



Announcements

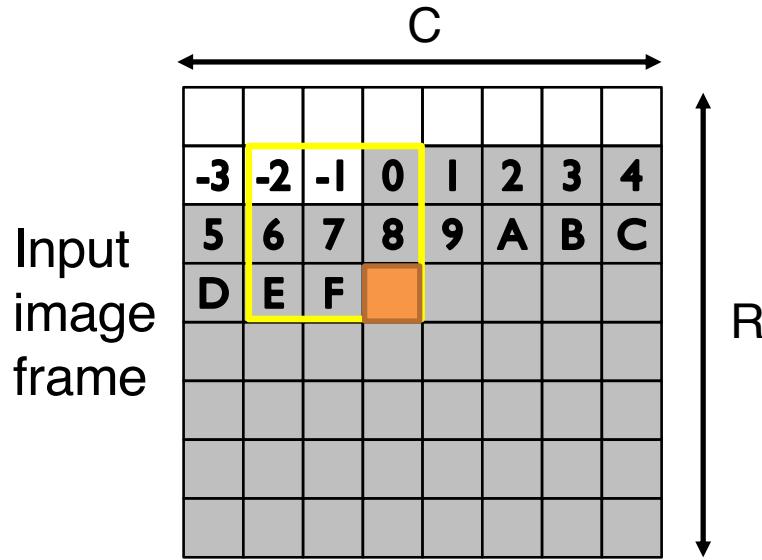
- ▶ TA-led hands-on tutorial on HLS next Tuesday
 - Bring your laptop

Exercise: OI Analysis of 2D Convolution



Estimate the OI for the 2D convolution kernel both without and with data use

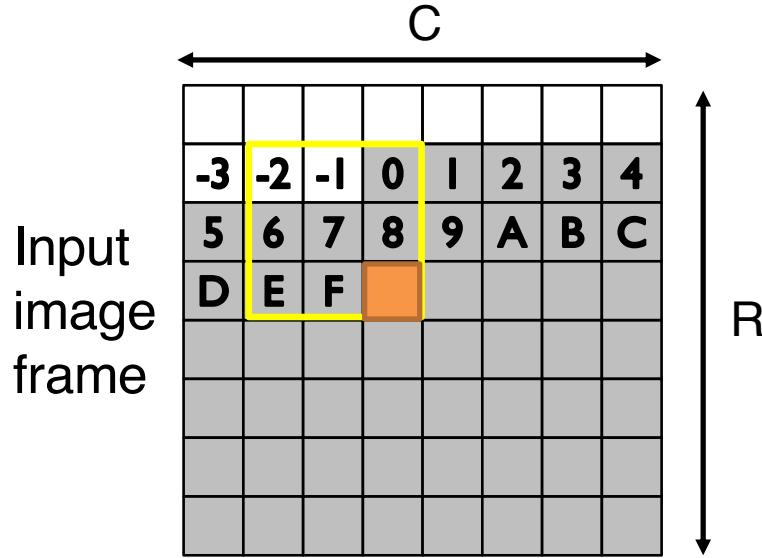
OI Analysis of 2D Convolution w/o LineBuffer



```
for (r = 1; r < R; r++)  
    for (c = 1; c < C; c++)  
        for (i = 0; i < 3; i++)  
            for (j = 0; j < 3; j++)  
                out[r][c] += img[r+i-1][c+j-1] * f[i][j];
```

- ▶ OI without data reuse
 - Number of operations = $C \cdot R \cdot 9 \cdot 2$
(1 multiply + 1 add per pixel)
 - External mem accesses (bytes) = $C \cdot R \cdot 9$
(assuming 1 byte per pixel in grayscale)
 - Resulting OI = 2

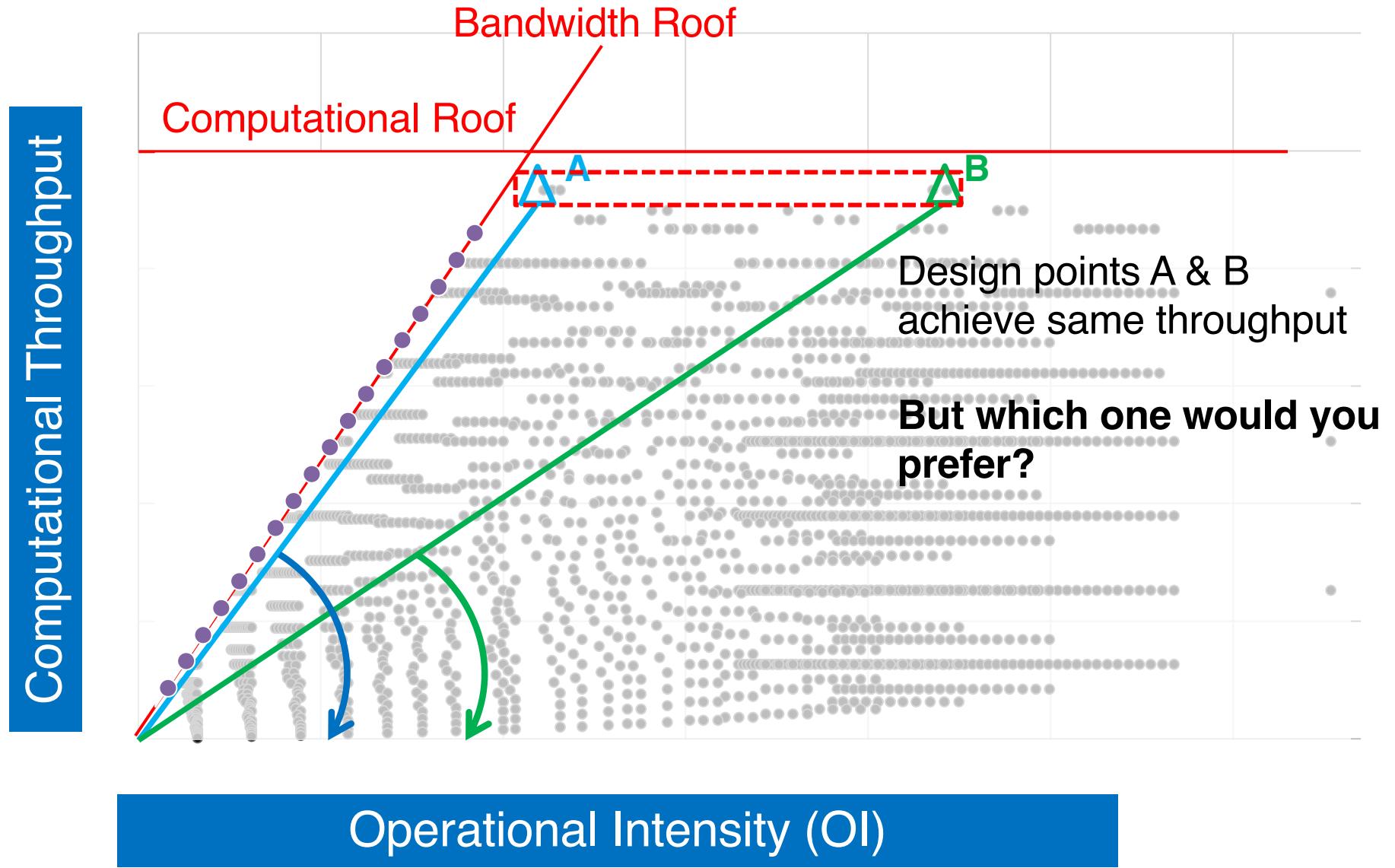
OI Analysis of 2D Convolution w/ LineBuffer



```
for (r = 1; r < R; r++)  
    for (c = 1; c < C; c++)  
        for (i = 0; i < 3; i++)  
            for (j = 0; j < 3; j++)  
                out[r][c] += img[r+i-1][c+j-1] * f[i][j];
```

- ▶ OI with data reuse using line buffer
 - Number of operations = $C * R * 9 * 2$
 - External mem accesses (bytes) = $C * R$
 - OI = 18

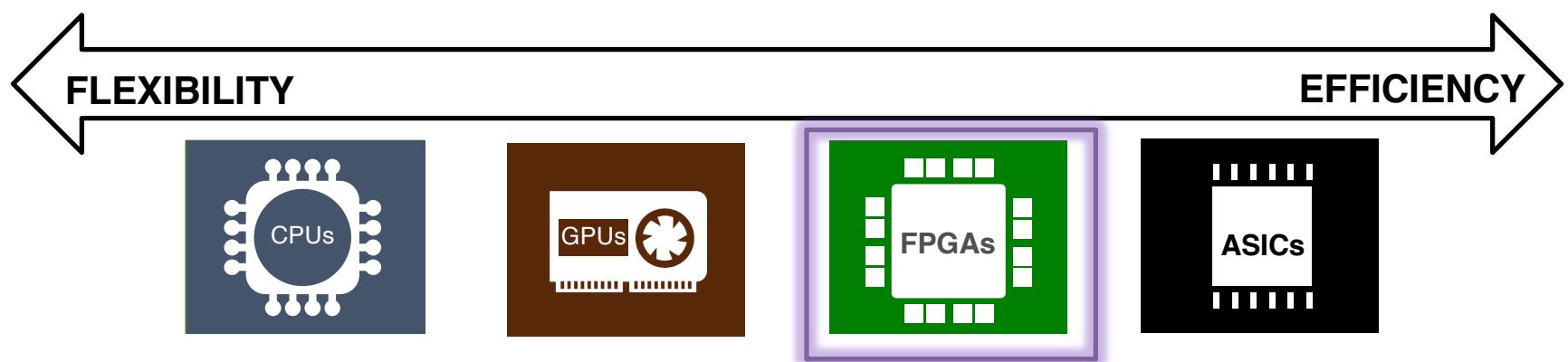
Recap: Design Space Exploration with Roofline



Agenda

- ▶ FPGA introduction
 - Basic building blocks
 - Classical homogeneous FPGA architectures
 - Modern heterogeneous FPGA architectures

Tradeoff between Compute Efficiency and Flexibility



What Are FPGAs

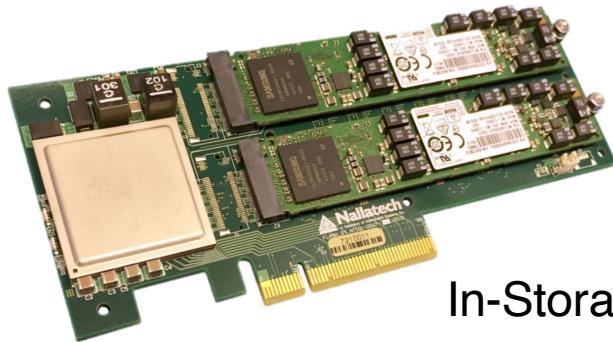
- ▶ Field-programmable gate array
 - Can be configured to act like any circuit after manufacturing
 - Can do many things – we focus on computation acceleration



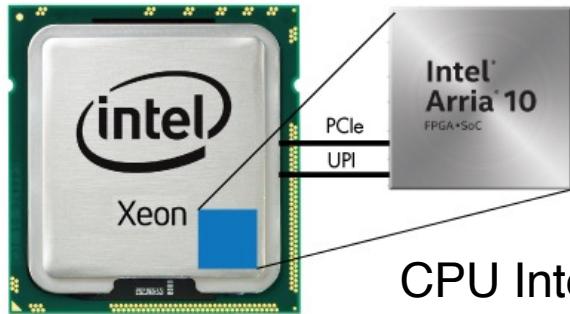
FPGAs Come In Many Forms



PCIe-Attached



In-Storage



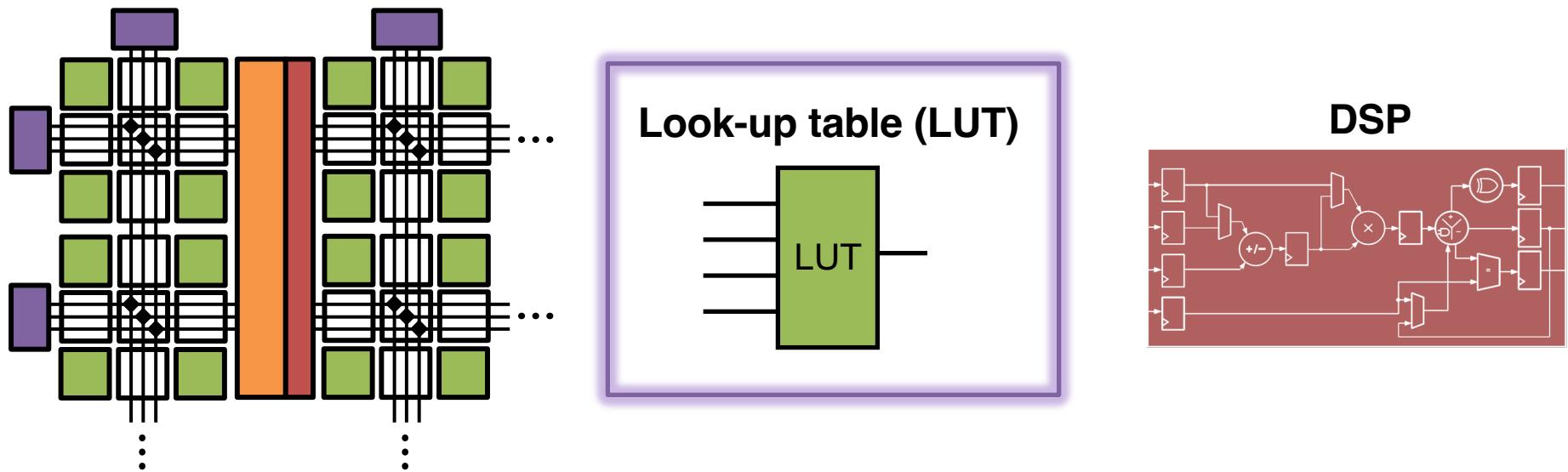
CPU Integrated



In-Network

Building Blocks of Modern FPGA Architectures

- ▶ A programmable array of logic blocks (LUT, FF), interconnects, I/Os, and dedicated blocks (BRAM, DSP)



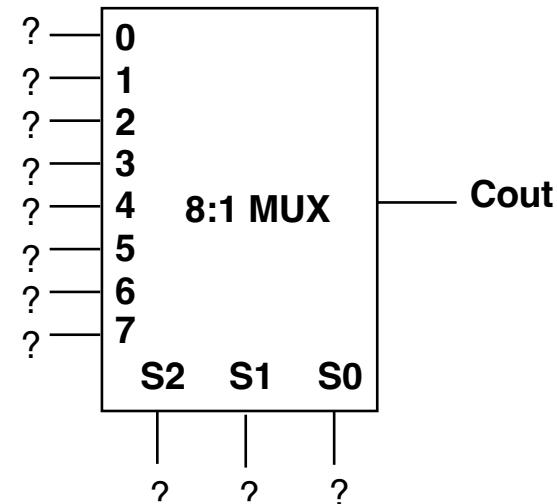
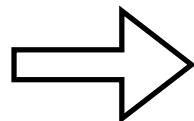
Counting Boolean Functions

- ▶ How many distinct 2-input 1-output Boolean functions exist?
- ▶ What about K inputs?

Multiplexer as a Universal Gate

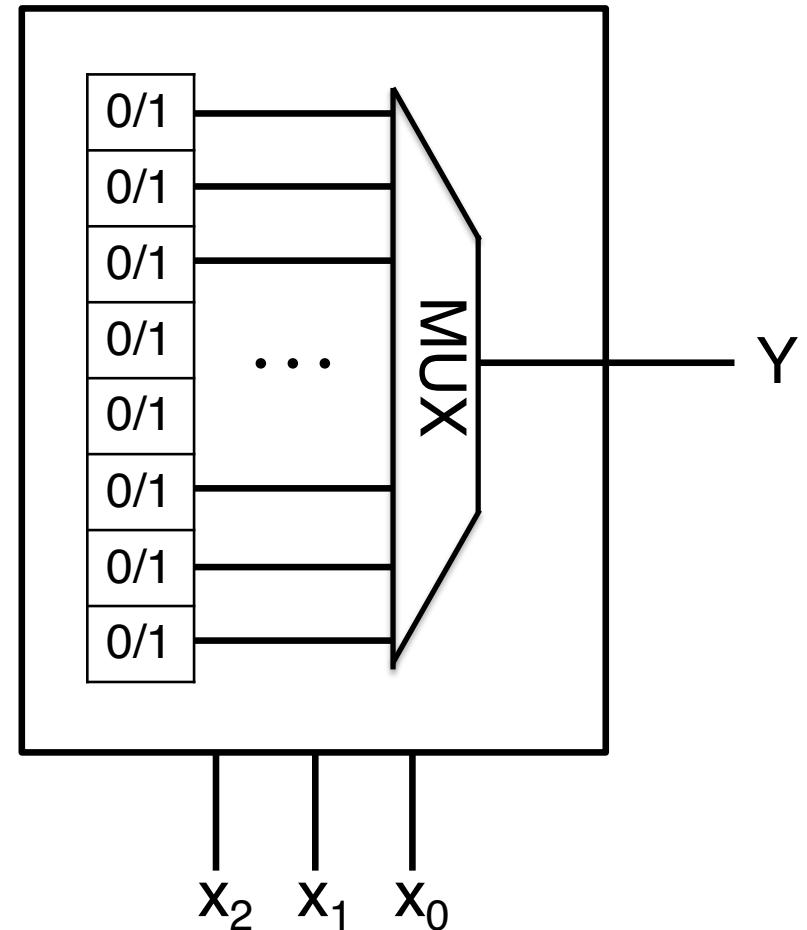
- Any function of k variables can be implemented with a $2^k:1$ multiplexer

A	B	Cin	S	Cout
0	0	0	0	0
0	0	1	1	0
0	1	0	1	0
0	1	1	0	1
1	0	0	1	0
1	0	1	0	1
1	1	0	0	1
1	1	1	1	1



Look-Up Table (LUT)

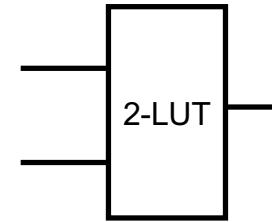
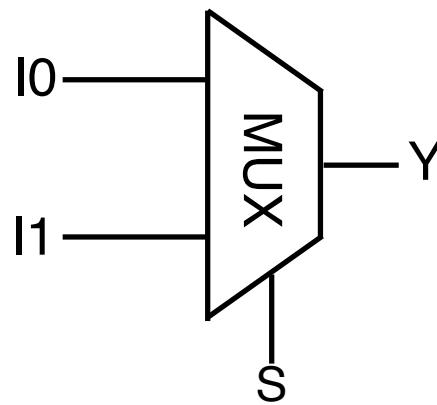
- A k-input LUT (k-LUT) can be configured to implement any k-input 1-output combinational logic
 - 2^k SRAM bits
 - Delay is independent of logic function



A 3-input LUT

Exercise: Implementing Logic with LUTs

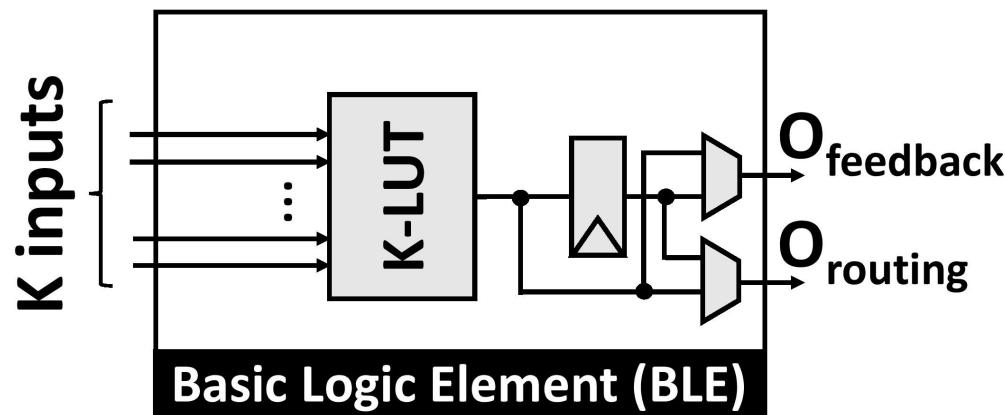
- ▶ Implement a 2:1 MUX using a network of 2-input LUTs. Use the minimum number of LUTs



Building block:
2-input LUT

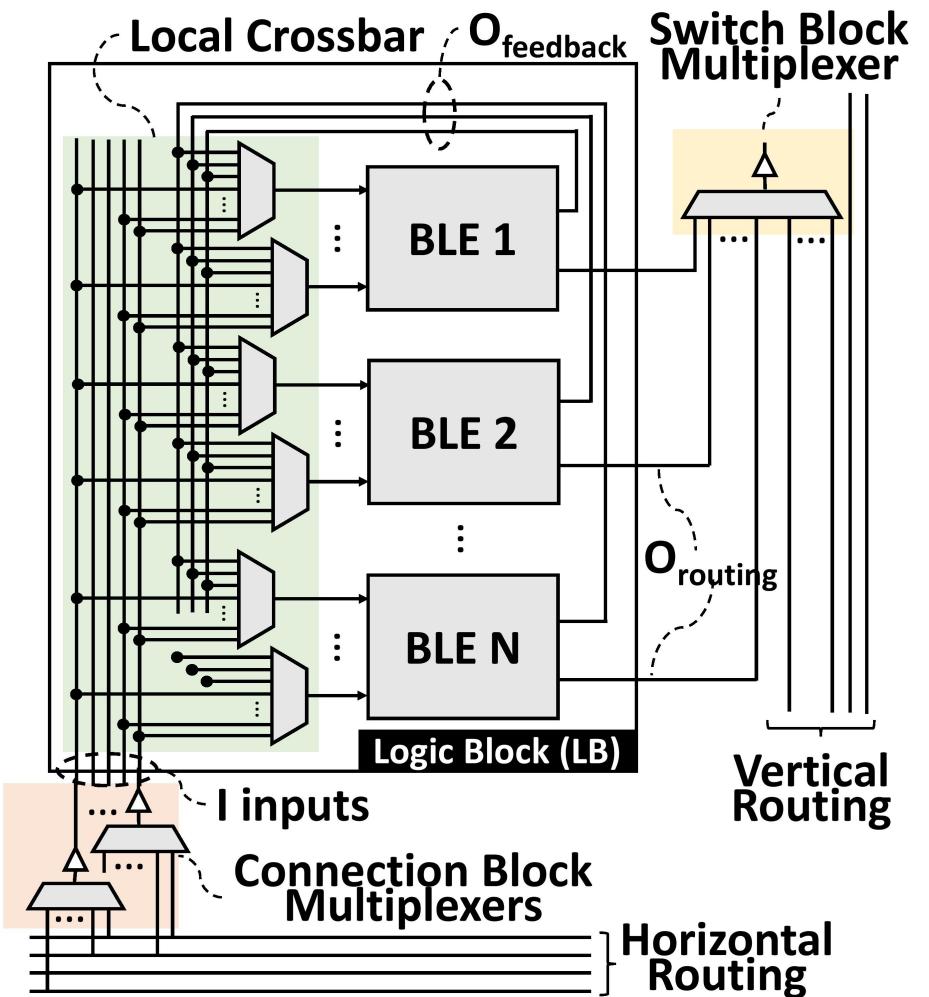
A Logic Element

- ▶ A k-input LUT is usually followed by a flip-flop (FF) that can be bypassed
- ▶ The LUT and FF combined form a logic element

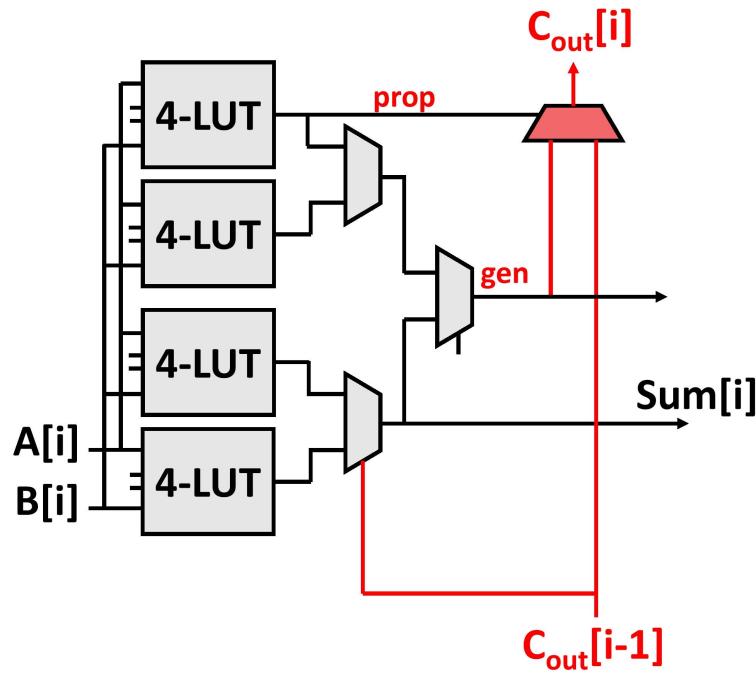


A Logic Block

- ▶ A logic block clusters multiple logic elements

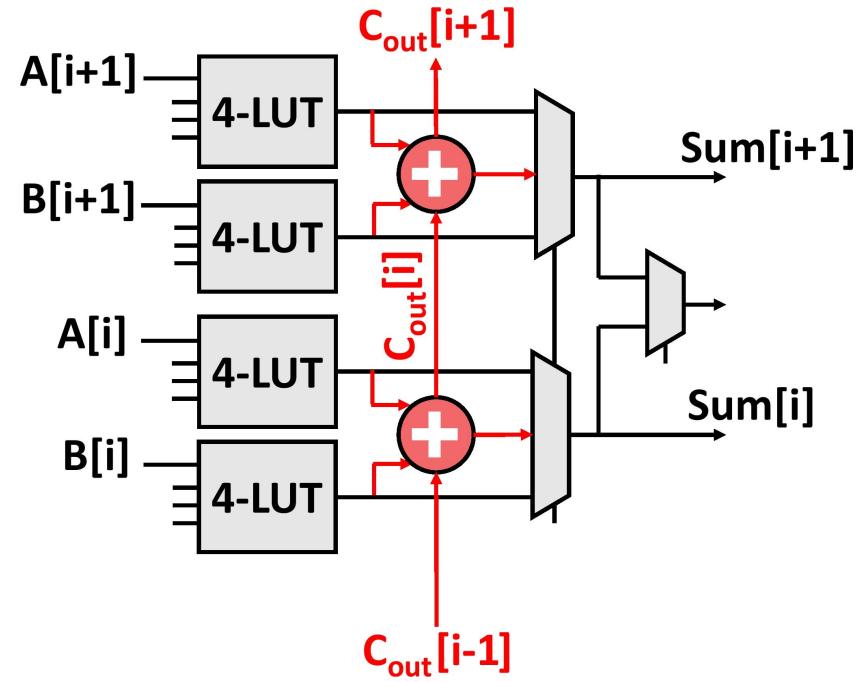


Arithmetic Circuitry in Logic Block



Xilinx (now AMD)

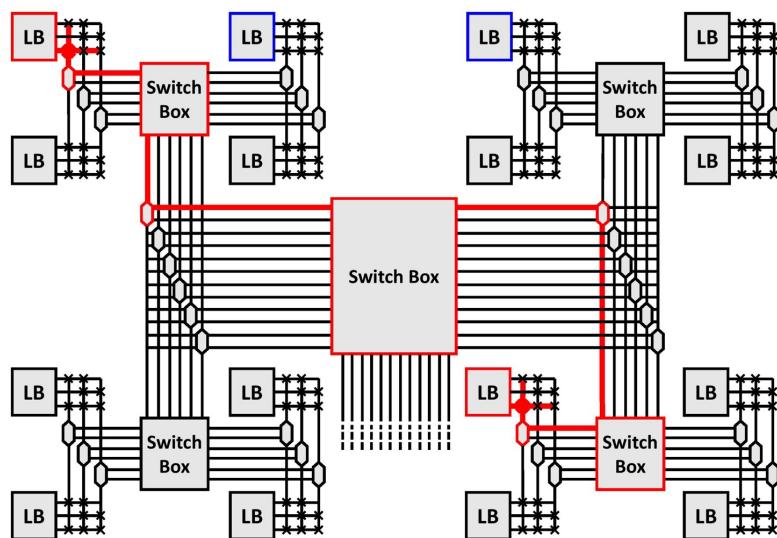
LUTs implement carry propagate and generation logic



Intel/Altera

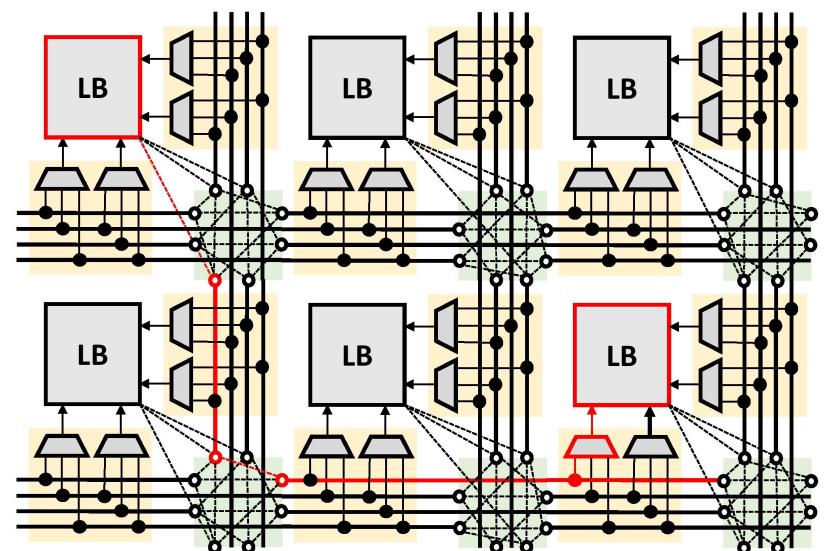
LUTs pass inputs to hardened adders

Routing Architecture



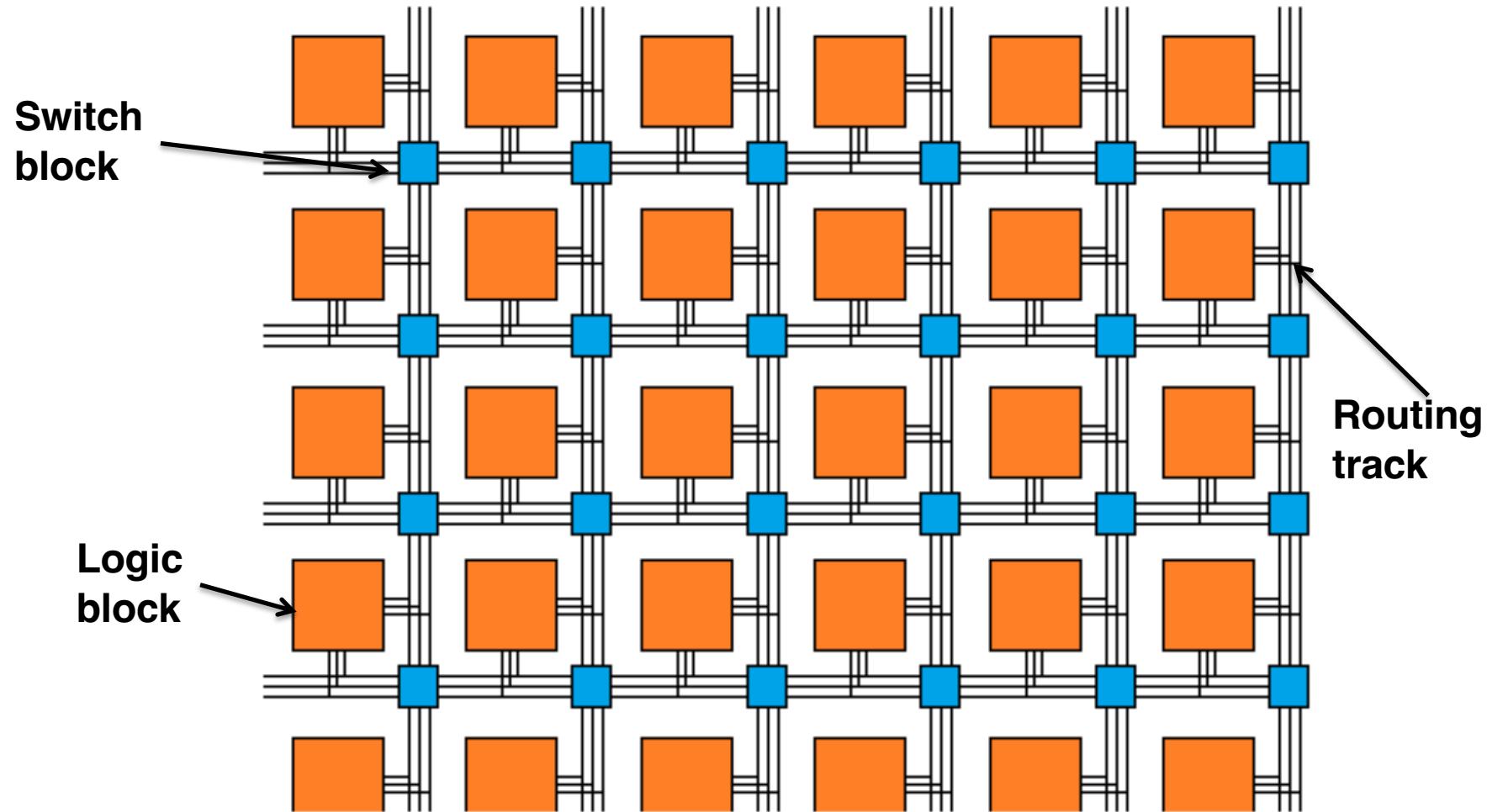
Hierarchical routing architecture

VS.



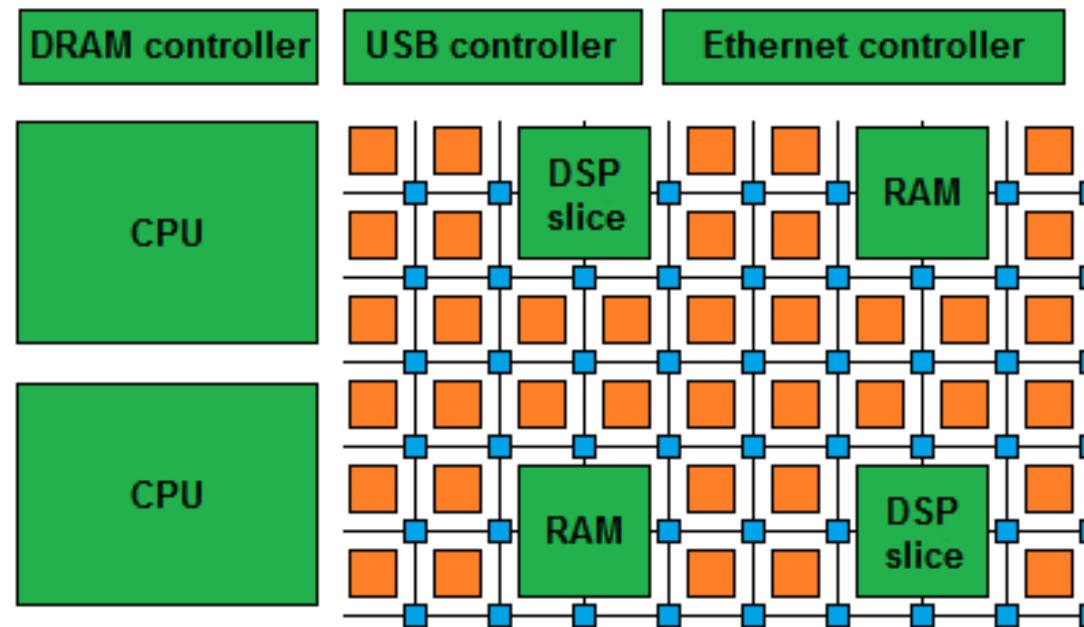
Island-style routing architecture

Traditional Homogeneous FPGA Architecture



Modern Heterogeneous Field-Programmable System-on-Chip (SoC)

- ▶ Island-style configurable mesh routing
- ▶ Lots of dedicated components
 - Memories/multipliers, I/Os, processors
 - Specialization leads to higher performance and lower power

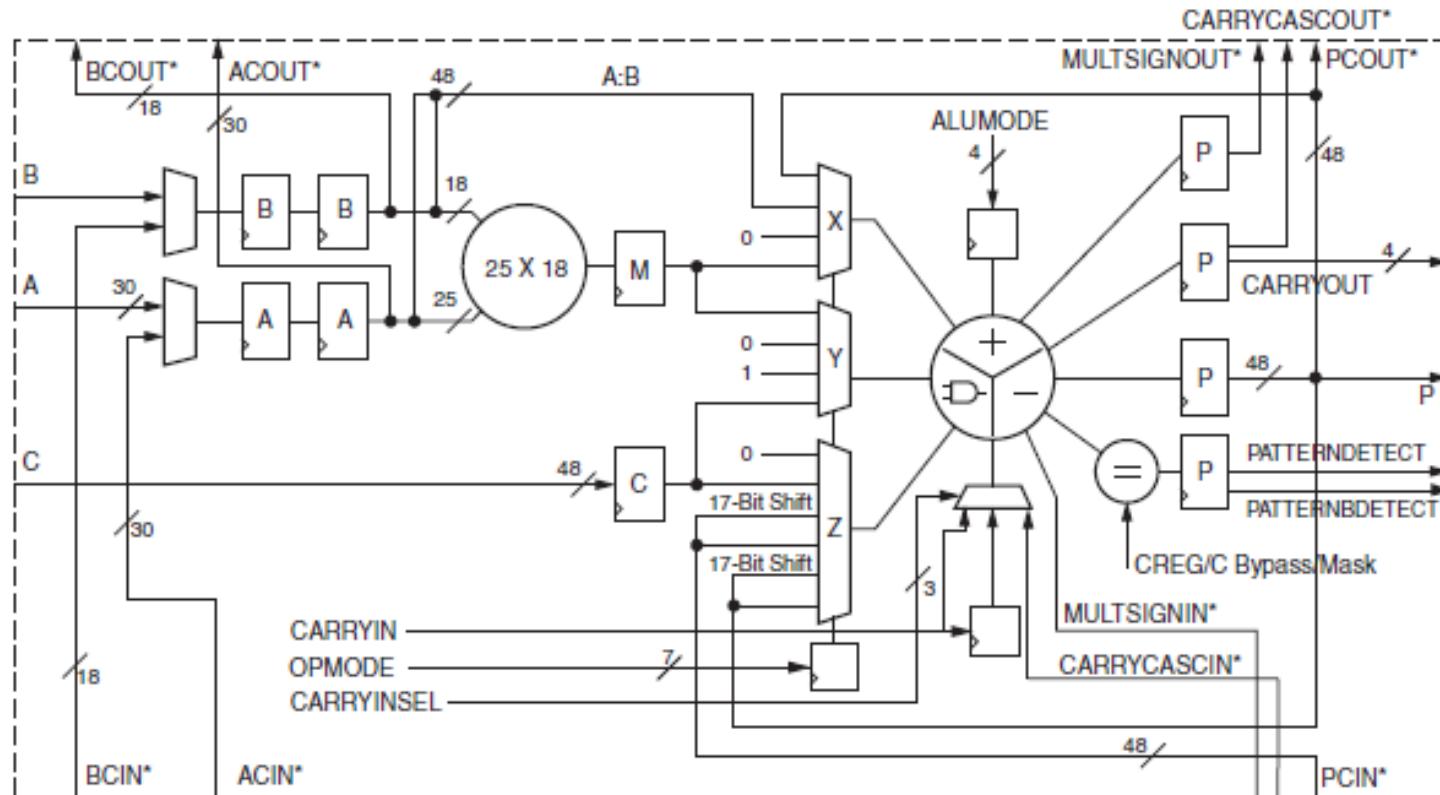


[Figure credit: embeddedrelated.com]

Dedicated DSP Blocks

- ▶ Built-in components for fast arithmetic operation optimized for DSP applications
 - Essentially a multiply-accumulate core with many other features
 - Fixed logic and connections, functionality may be configured using control signals at run time
 - Much faster than LUT-based implementation (ASIC vs. LUT)

Example: Xilinx DSP48E Slice



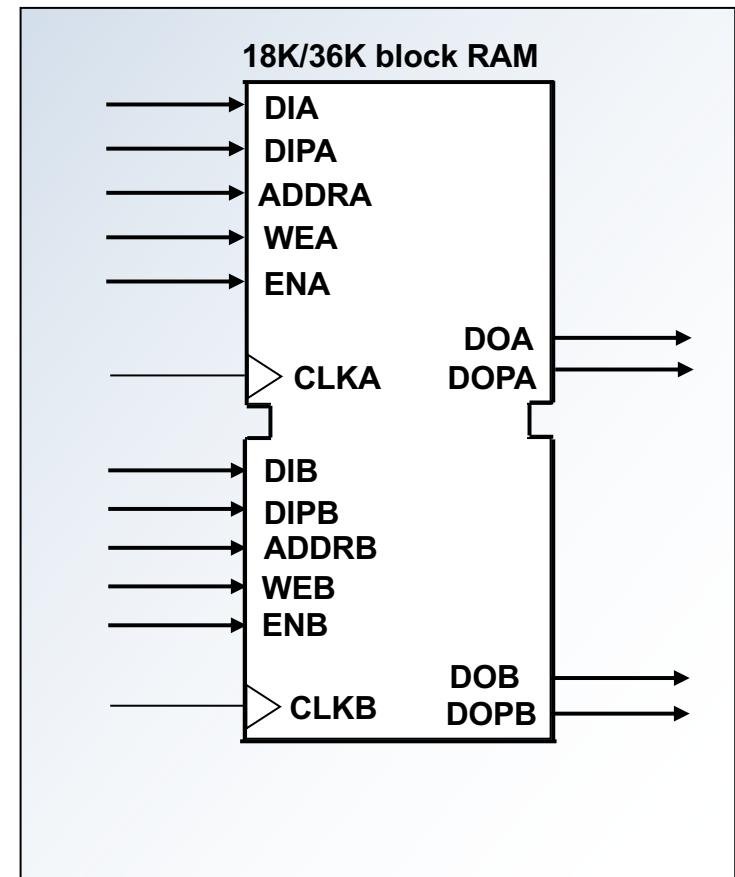
*These signals are dedicated routing paths internal to the DSP48E column. They are not available as pins on the device package.

- 25x18 signed multiplier
- 48-bit add/subtract/accumulate
- 48-bit logic operations
- SIMD operations (12/24 bit)
- Pipeline registers for high speed

Dedicated Block RAMs (BRAMs)

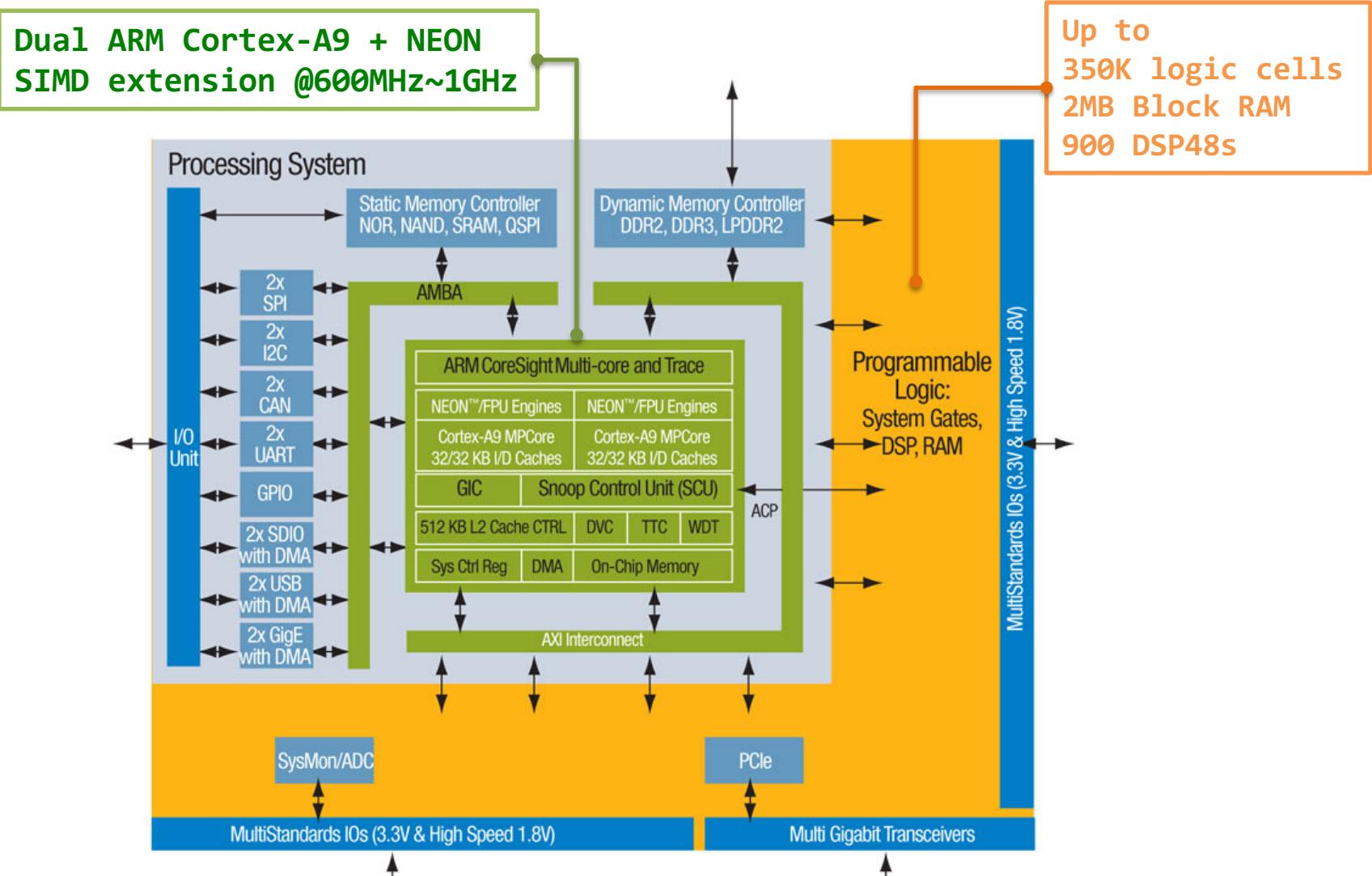
- ▶ Example: Xilinx 18K/36K block RAMs

- 32k x 1 to 512 x 72 in one 36K block
- Simple dual-port and true dual-port configurations
- Built-in FIFO logic
- 64-bit error correction coding per 36K block



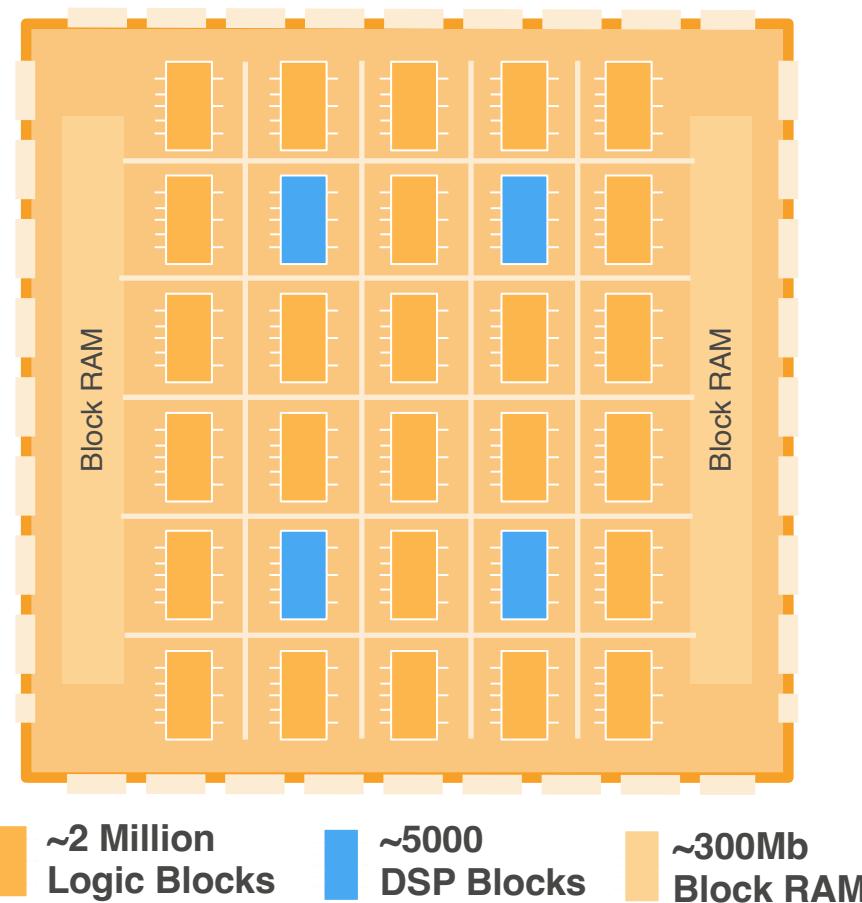
[source: AMD Xilinx]

An Embedded FPGA SoC



Xilinx Zynq All Programmable System-on-Chip
[Source: AMD Xilinx]

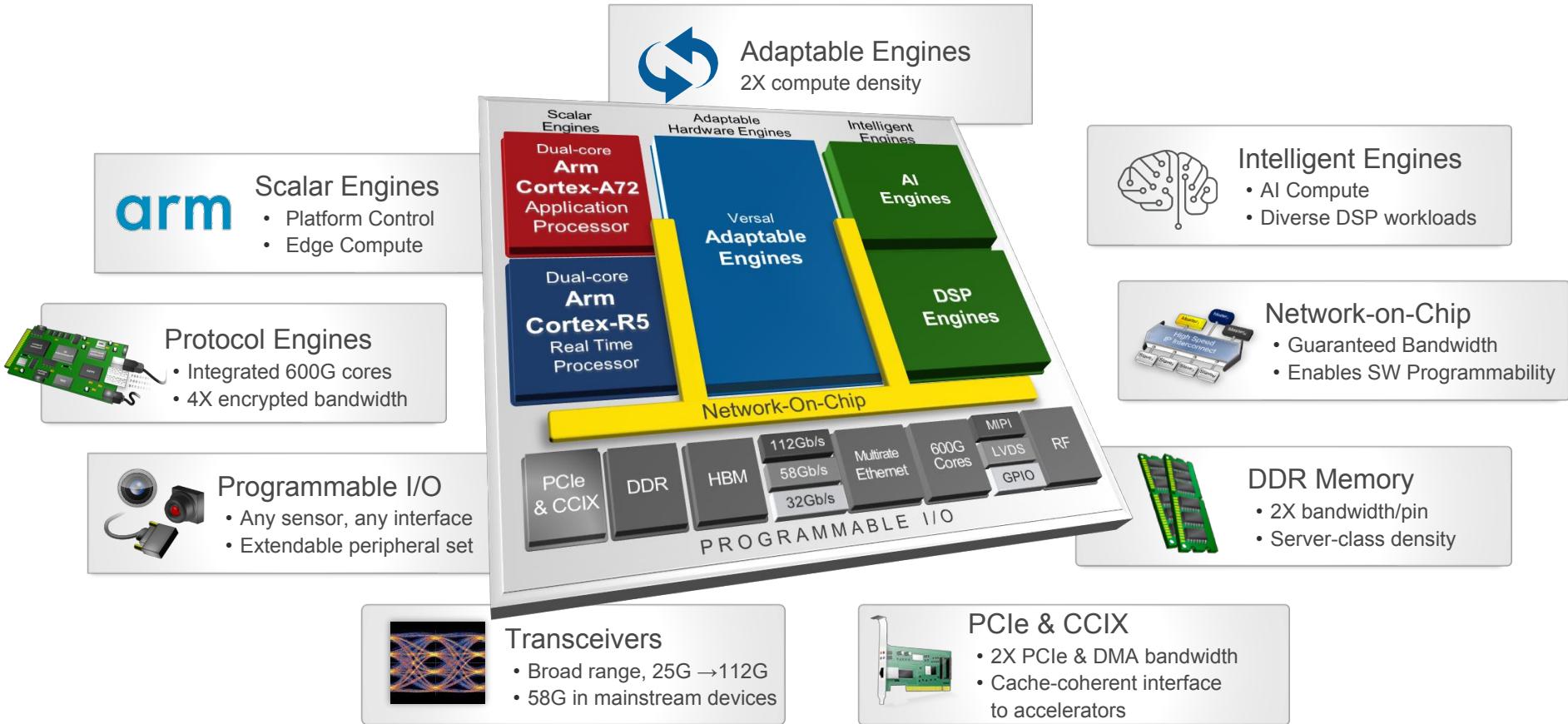
A Cloud FPGA Instance



AWS F1 instance: AMD Xilinx UltraScale+ VU9P
[Figure source: David Pellerin, AWS]

An Even More Heterogeneous (FPGA) Accelerator

AMD Xilinx Versal Architecture

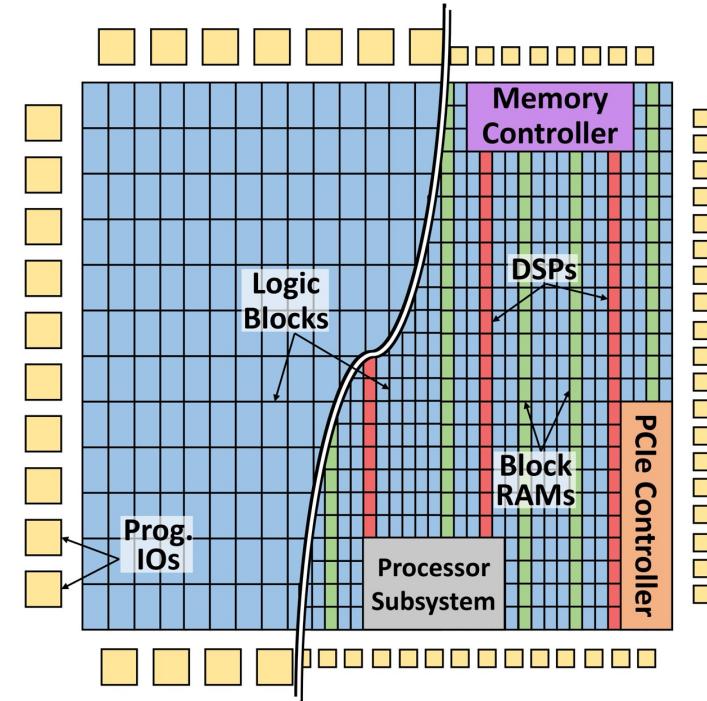


[source: AMD Xilinx]

Key Advantages of FPGA-Based Computing

- ▶ **Massive amount of fine-grained parallelism**
 - Highly parallel and/or deeply pipelined architecture
- ▶ **Silicon (re)configurable to fit the application**
 - Compute at desired numerical accuracy
 - Customized memory hierarchy

⇒ low (and predictable) latency
⇒ higher energy efficiency



Next Lecture

- ▶ Analysis of Algorithms

Acknowledgements

- ▶ These slides contain/adapt materials developed by
 - Prof. Jason Cong (UCLA)
 - Andrew Boutros and Prof. Vaughn Betz (Univ. of Toronto)
 - UCI CS295 by Prof. Sang-Woo Jun