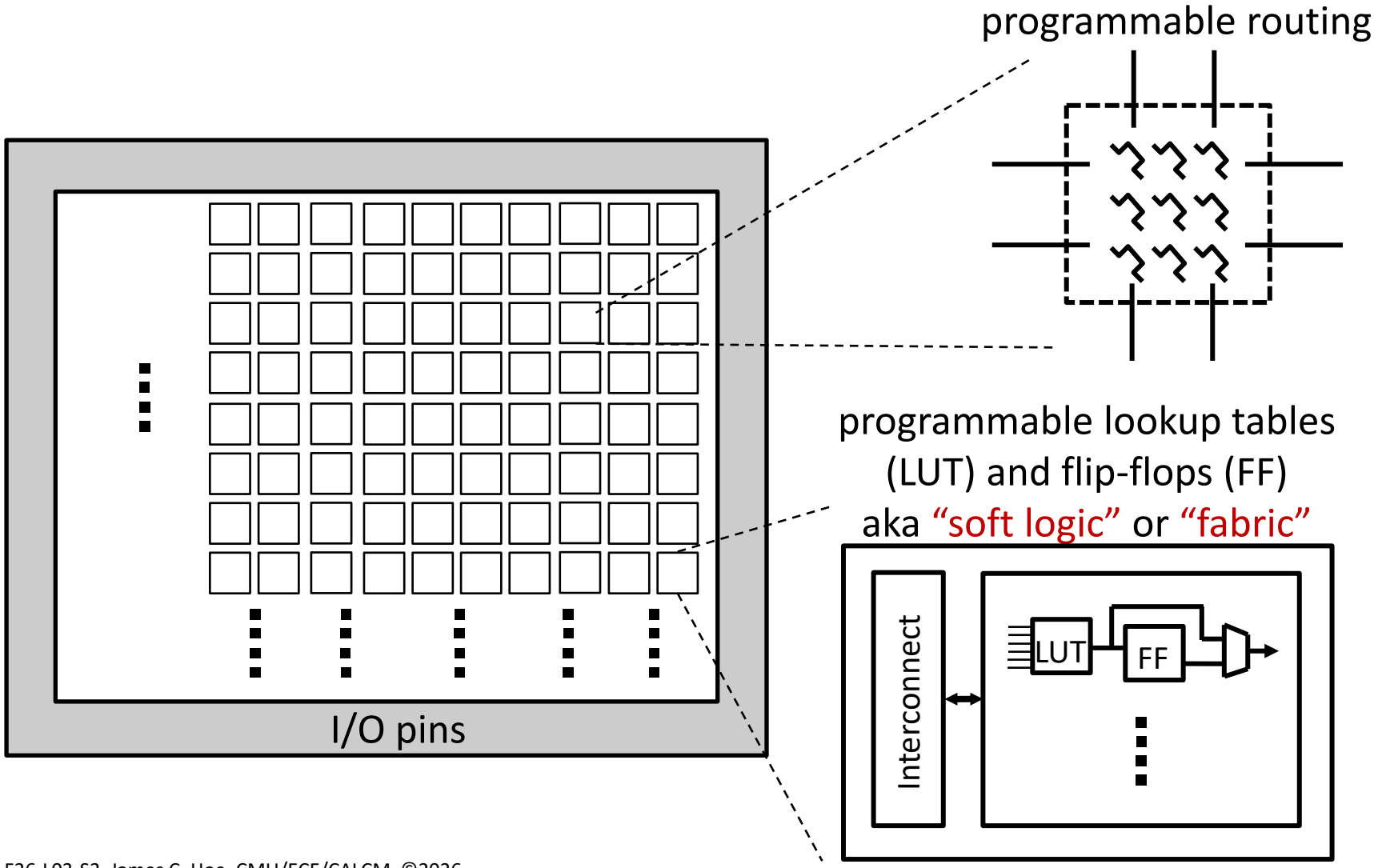
18-643 Lecture 3: FPGA on Moore's Law

James C. Hoe
Department of ECE
Carnegie Mellon University

Housekeeping

- Your goal today: get caught up on 3 decades of progress (upto 2010'ish)
- Notices
  - Complete survey on Canvas, Wed 9/2
  - Lab 0, due noon, Mon 9/7
  - Handout #4: Term Project Intro
  - Handout #5: Review Paper Selection
- Readings (see lecture schedule online)
  - skim [Boutros, et al., 2021]
  - for next time: skim [Ahmed, et al., 2016] and [Chromczak, et al., 2020]

Where we stopped last time: FPGA as Universal Fabric



Fast-forward through Moore's Law

Part Number	Logic Capacity (gates)	Config-urable Logic Blocks	User I/Os	Config-uration Program (bits)
XC2064	1200	64	58	12038
XC2018	1800	100	74	17878

1985

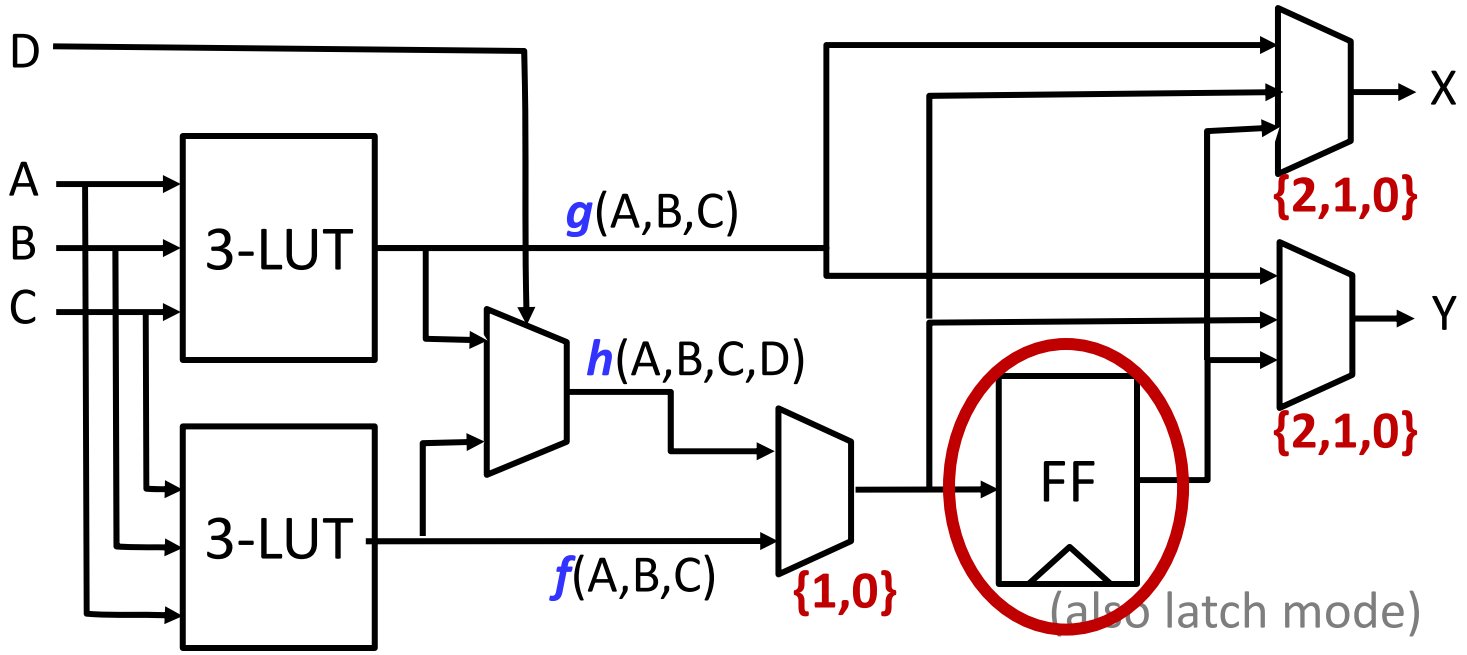
2026

		VP1002	VP1052	VP1102	VP1202	VP1402	VP1502	VP2502	VP1552	VP1702	VP1802	VP2802	VP1902
AI Engine	AI Engines Tiles	–	–	–	–	–	–	472	–	–	–	472	–
	AI Engine Data Memory (Mb)	–	–	–	–	–	–	118	–	–	–	118	–
Programmable Logic	System Logic Cells (K)	833	1,186	1,575	1,969	2,233	3,763	3,738	3,837	5,558	7,352	7,326	18,507
	LUTs	380,800	542,080	719,872	900,224	1,020,928	1,720,448	1,708,672	1,753,984	2,540,672	3,360,896	3,349,120	8,460,288
	DSP Engines	1,140	1,572	1,904	3,984	2,672	7,440	7,392	7,392	10,896	14,352	14,304	6,864
	NoC Master / NoC Slave Ports	22	22	30	28	42	52	52	52	76	100	100	192
	Super Logic Regions (SLRs)[1]	–	–	–	–	–	2	2	2	3	4	4	4
Memory	Distributed RAM (Mb)	12	17	22	27	31	53	52	54	78	103	102	258
	Block RAM (Mb)	19	26	49	47	70	89	89	89	132	174	174	239
	UltraRAM (Mb)	97	138	127	190	181	366	366	366	541	717	717	619
	Multiport RAM (Mb)	80	80	–	–	–	–	–	–	–	–	–	–
	Total PL Memory (Mb)	208	261	198	264	282	508	507	509	751	994	994	1116
	DDR4 Memory Controllers	2	2	3	4	3	4	4	4	4	4	4	14
	DDR Bus Width	128	128	192	256	192	256	256	256	256	256	256	896
Processing System	Application Processing Unit	Dual-core Arm® Cortex®-A72, 48 KB/32 KB L1 Cache w/ parity & ECC; 1 MB L2 Cache w/ ECC											
	Real-Time Processing Unit	Dual-core Arm Cortex-R5F, 32 KB/32 KB L1 Cache, and 256 KB TCM w/ECC											
	Memory	256 KB On-Chip Memory w/ECC											
	Connectivity	Ethernet (x2); UART (x2); CAN-FD (x2); USB 2.0 (x1); SPI (x2); I2C (x2)											
Serial Transceivers	GTY Transceivers	8	8	–	–	–	–	–	–	–	–	–	–
	GTYP Transceivers	–	–	8	28[2]	8	28[2]	28[2]	68[2]	28[2]	28[2]	28[2]	128
	GTM Transceivers (58G (112G))	24 (12)	36 (18)	64 (32)	20 (10)	96 (64)[3]	60 (30)	60 (30)	20 (10)	100 (50)	140 (70)	140 (70)	32 (16)
Integrated Protocol IP	PCIe® w/DMA (CPM4)	x Gen4x4	2 x Gen4x4	–	–	–	–	–	–	–	–	–	–
	PCIe w/DMA (CPM5)	–	–	–	2 x Gen5x8	–	2 x Gen5x8	2 x Gen5x8	2 x Gen5x8	2 x Gen5x8	2 x Gen5x8	2 x Gen5x8	–
	PCIe (PL PCIE4)	x Gen4x8	1 x Gen4x8	–	–	–	–	–	–	–	–	–	–
	PCIe (PL PCIE5)	–	–	2 x Gen5x4	2 x Gen5x4	2 x Gen5x4	2 x Gen5x4	2 x Gen5x4	8 x Gen5x4	2 x Gen5x4	2 x Gen5x4	2 x Gen5x4	16 x Gen5x4
	100G Multirate Ethernet MAC	3	5	6	2	6	4	4	4	6	8	8	12
	600G Ethernet MAC	2	3	7	1	11	3	3	1	5	7	7	4
	600G Interlaken	1	2	–	–	–	1	1	–	2	3	3	–
	400G High-Speed Crypto Engine	1	1	3	1	4	2	2	2	3	4	4	–
Ordering Information	Extended[4]	-1MSE, -1LSE, -2MSE, -2MLE, -2LSE, -2LLE		-1MSE, -1LSE, -2MSE, -2MLE, -2LSE, -2LLE, -3HSE									
	Industrial[4]	-1MSI, -1MLI, -1LSI, -1LLI, -2MSI, -2MLI, -2LLI, -2HSI		-1MSI, -1MLI, -1LSI, -1LLI, -2MSI, -2MLI									-

What happened is more than Moore

30 Years of Becoming Hardwired

LUT-based Configurable Logic Block (simplified sketch)



- 2 fxns (*f* & *g*) of 3 inputs OR 1 fxn (*h*) of 4 inputs
- hardwired FFs (too expensive/slow to fake)
- Just 10s of these in the earliest FPGAs

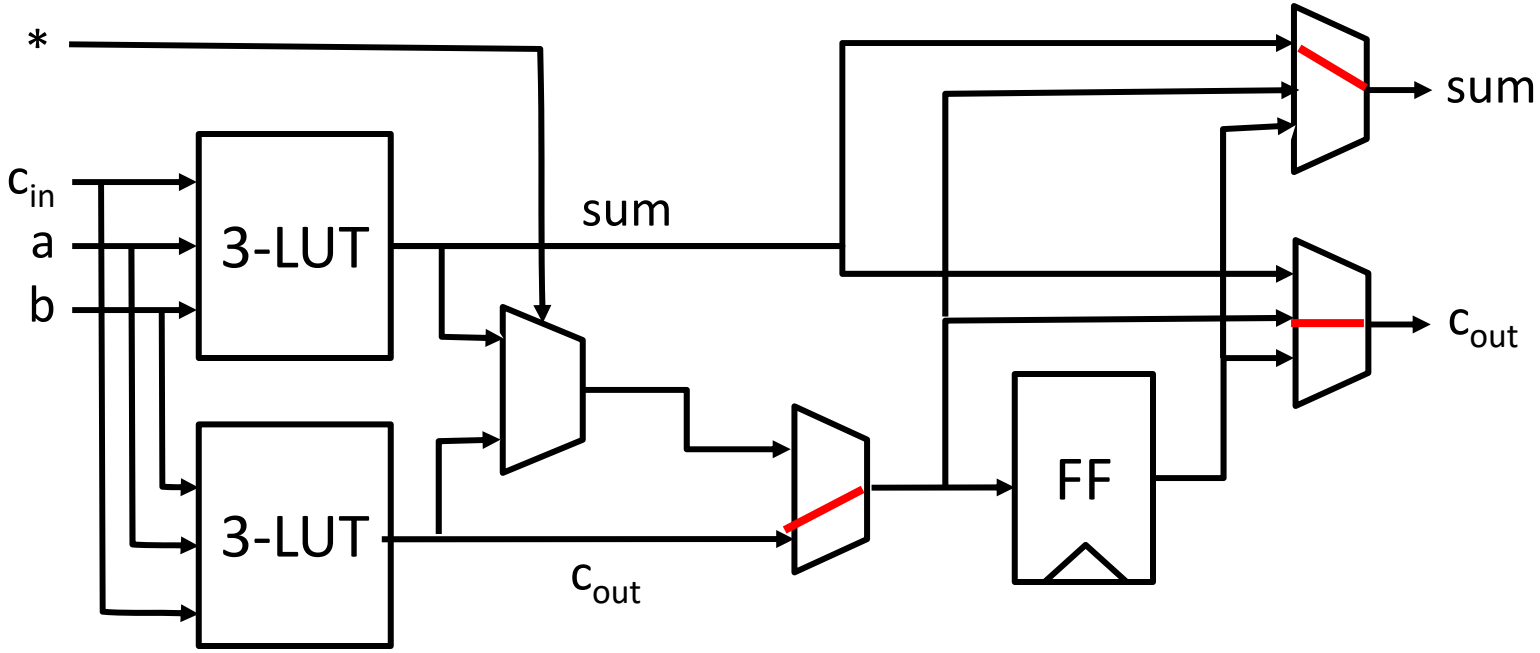
Recall

Why Hardwired Logic

- LUTs already can do everything (digital)
- Revisit: why hardwired flip-flop in CLB?
  - would take 4 LUTs to make 1 M-S flip-flop
  - LUT-built FFs have atrocious setup/hold time
  - almost all designs affected in cost and speed
- Makes sense to hardwire a functionality
  - needed by everyone (or by the big customers)
  - expected benefits outweigh displaced LUT area, i.e.,
    - much more expensive/slow in LUTs
    - easy/cheap to ignore when not in use

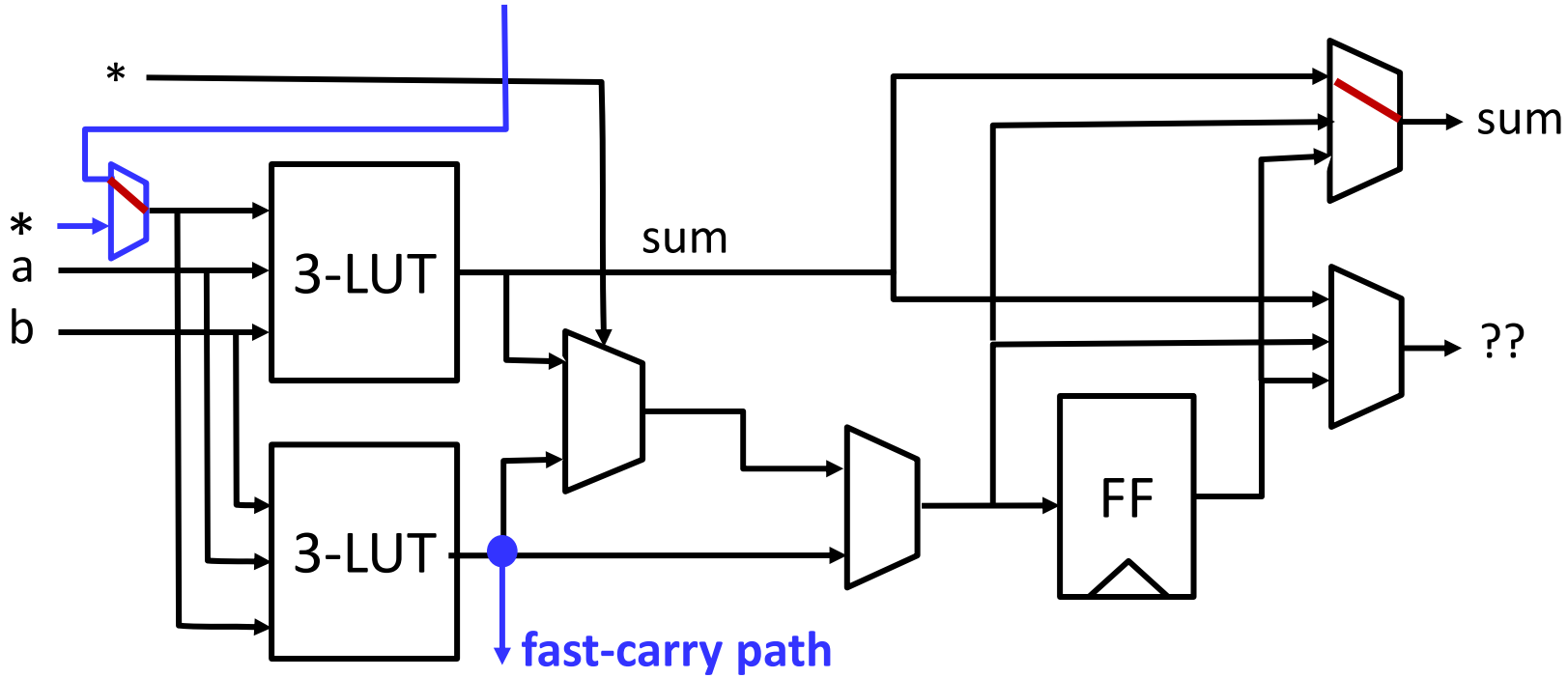
Hardwiring is a great thing if it is usable and is used

E.g., Special Support for Addition



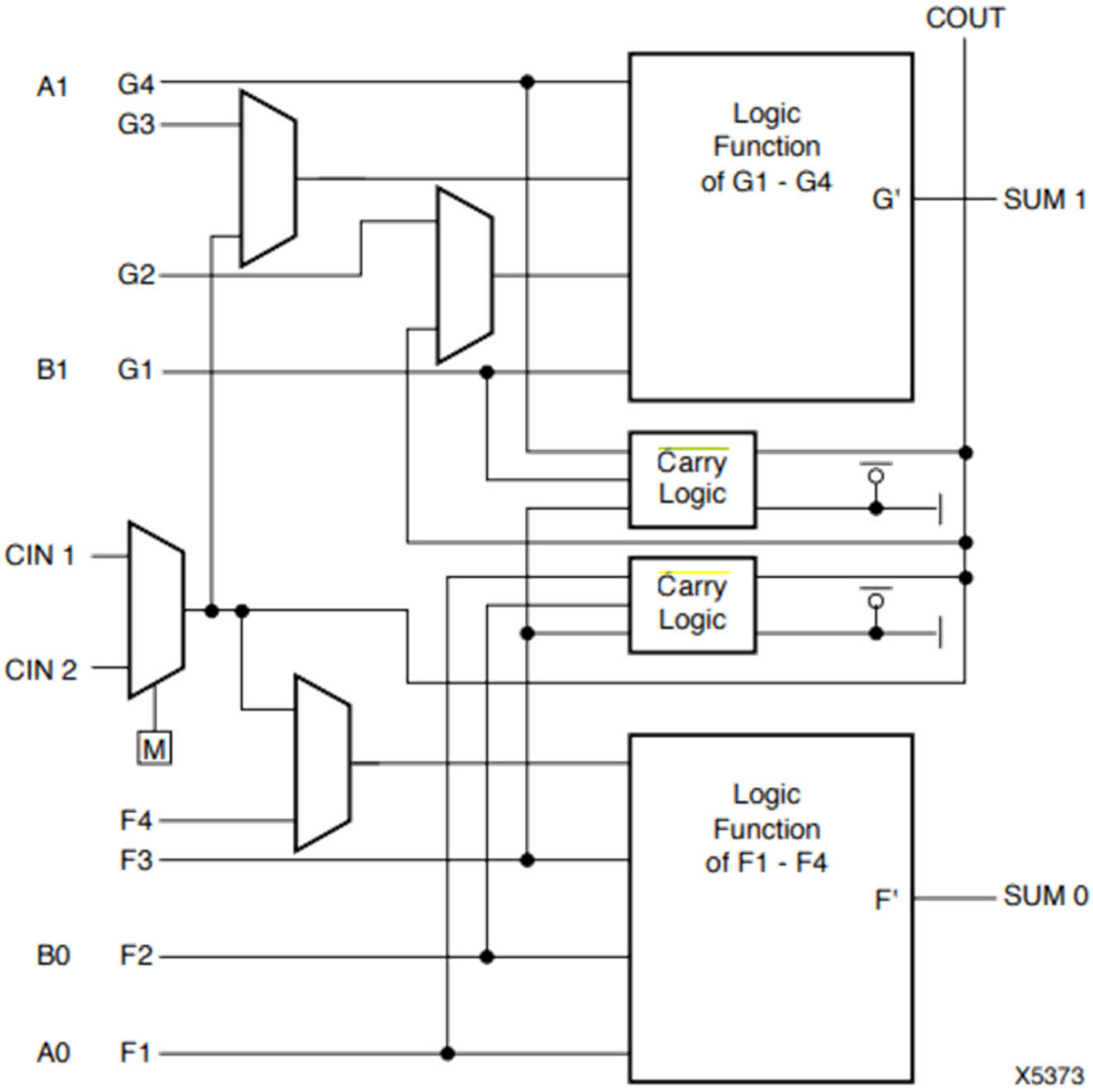
- A full-adder fits perfectly in 1 CLB with 2x3LUTs
- But carry propagation slow---flow through multiple configurable connections, not simple wires
- Addition is pretty important to most designs

Specialized Logic for Fast Carry



- Cost = 1 (real) wire and 1 mux
- Big win in adder performance

Xilinx XC4000 (1990s)



[XC4000, XC4000A, XC4000H Logic Cell Array Families]

Why not full-fledged adders?

- How many bits wide?
- How many (in ratio to # CLB)?
- Where to place the adders?
  - evenly sprinkled or clustered (for delay)

How about multipliers?

Hard Multipliers (2000s)

- Motivating forces
  - DSP became an important domain
  - very expensive and slow to multiply in LUTs
  - dies large enough to spare some area
- Virtex-II hardwired multiplier "macro" blocks
  - 18-bit inputs, full 36-bit product
  - explicit instantiation or inferable from RTL
  - relatively cheap (since native implementation)
  - Still no hard adders

Adders came later as a part of MAC in DSP slices

Meanwhile, multiply faster/cheaper than add!!

An Early Multiplier Block
Xilinx Virtex-II, circa 2000

[Virtex-II Pro and Virtex-II Pro X Platform FPGAs]

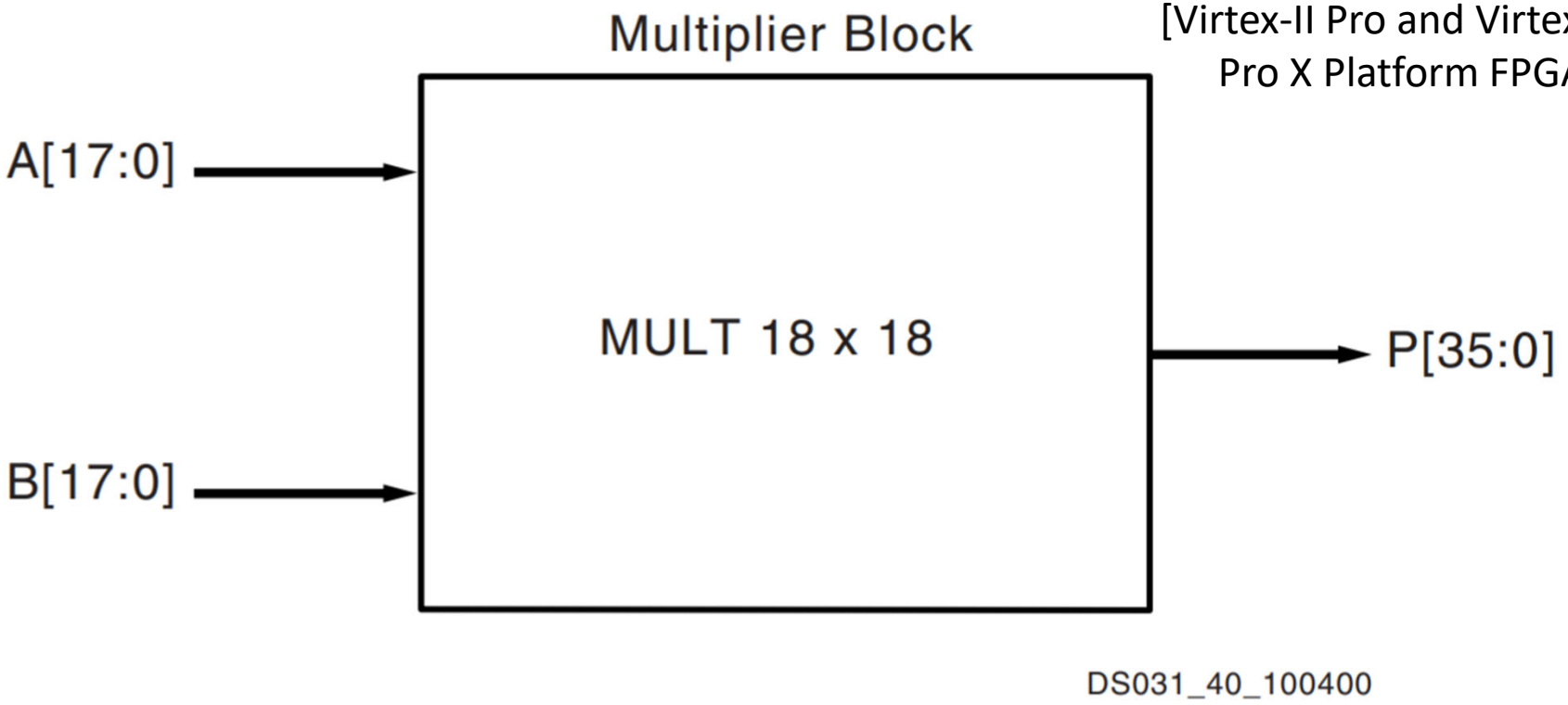
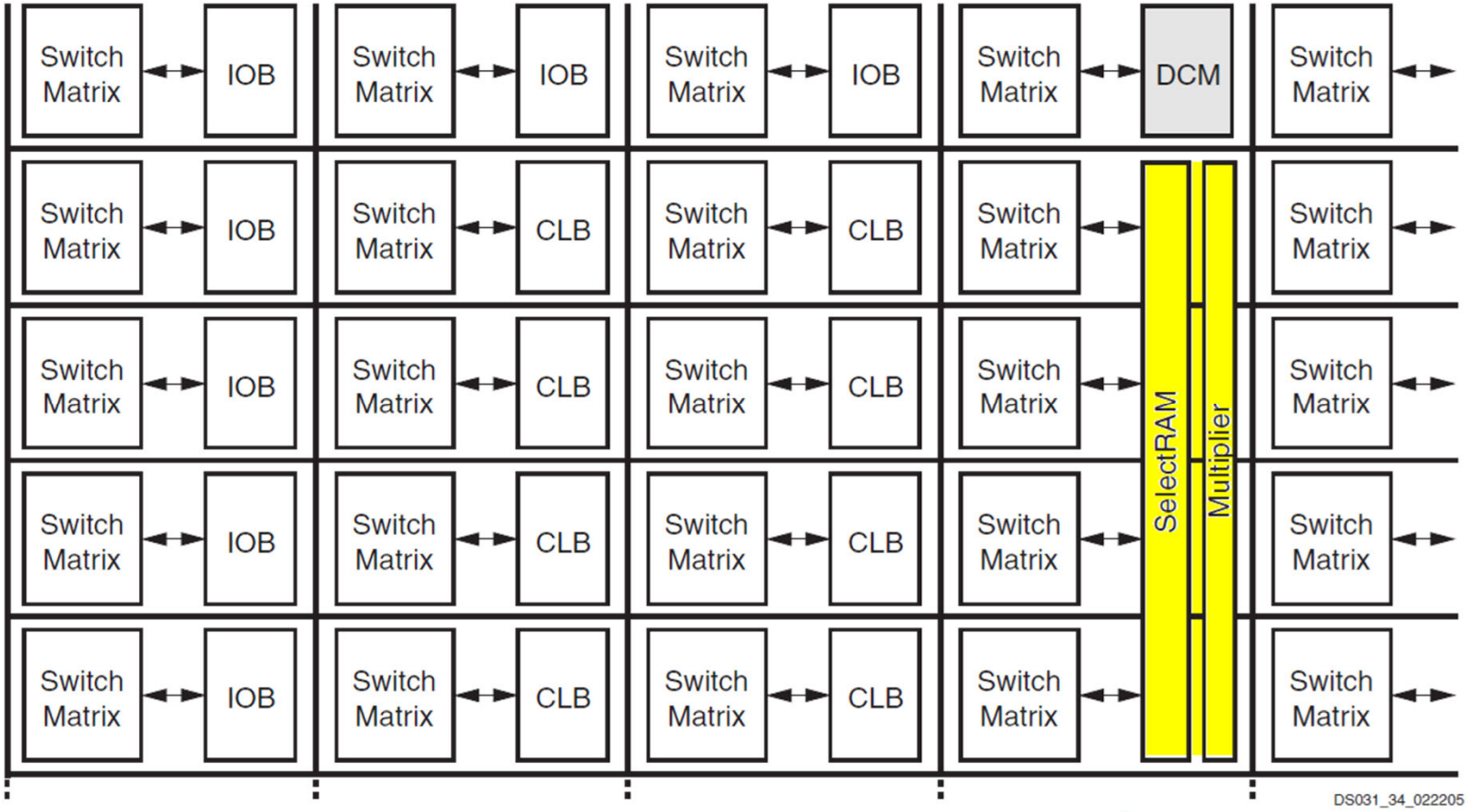


Figure 54: **Multiplier Block**

Where are these hard DSP slices?
How to get to them?

MACROs: a disturbance in the force . . .

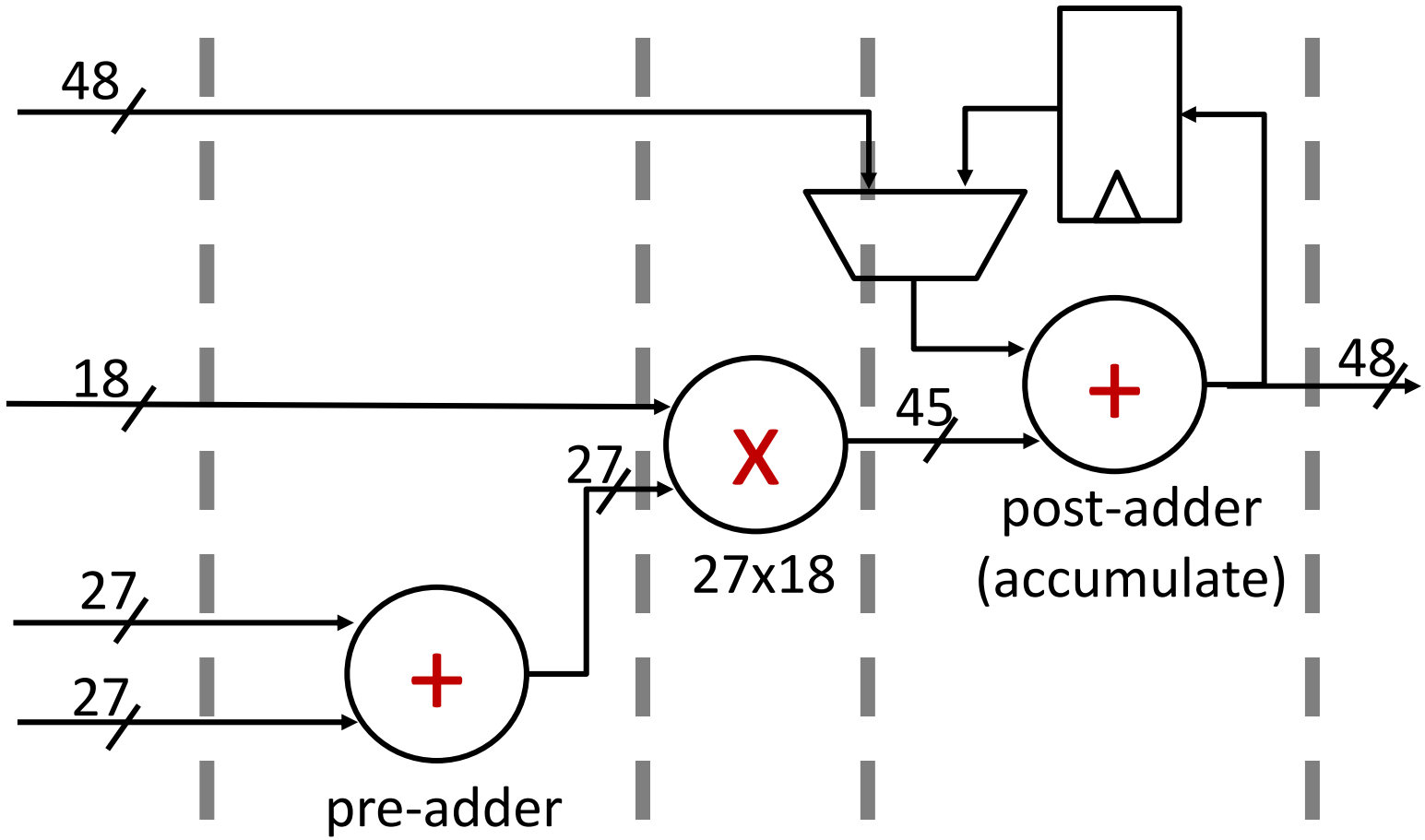


Too much vs not enough?

Benefit of using macro outweigh cost of getting to one?

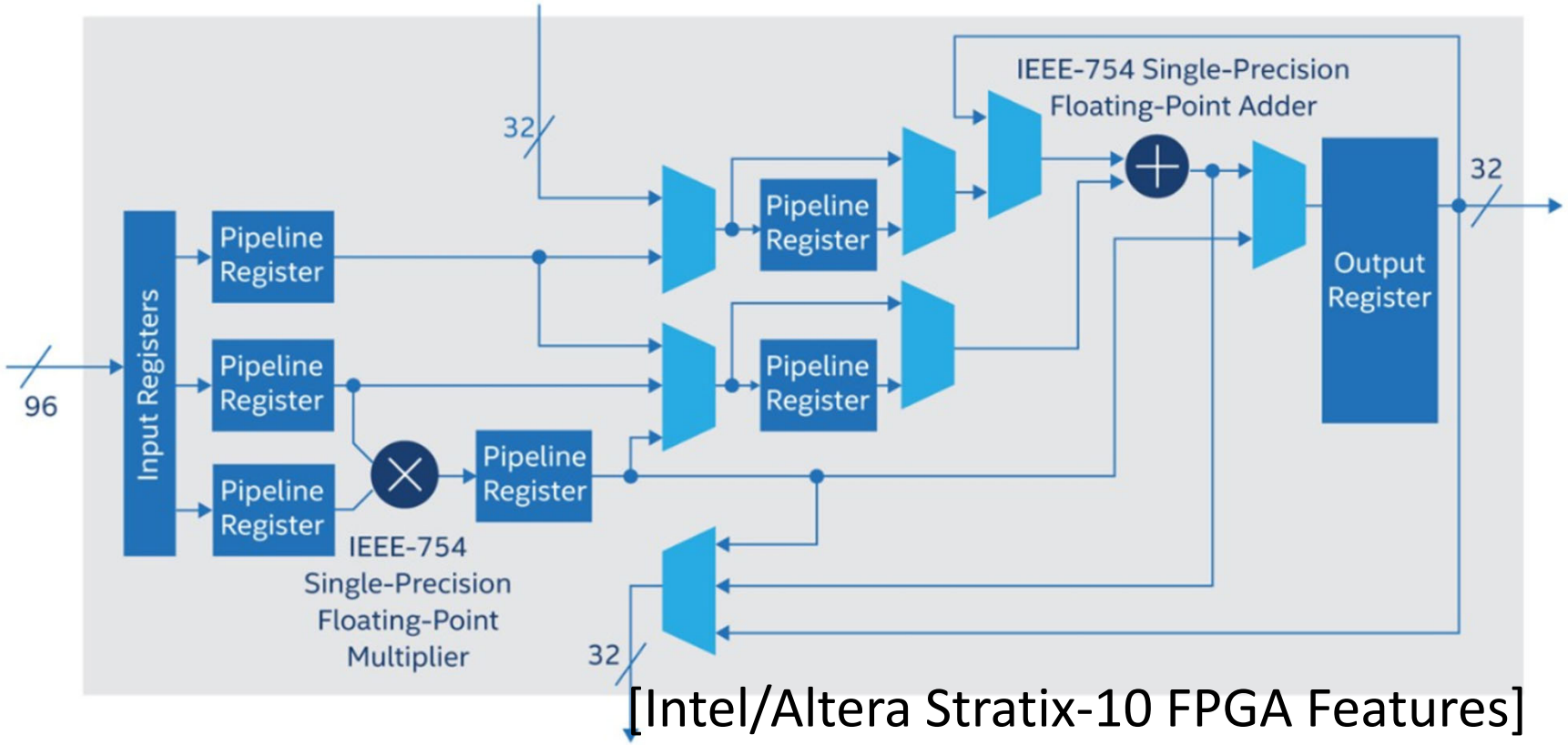
[Figure 48: Virtex-II Platform FPGAs: Complete Data Sheet]

Ultrascale DSP48E2



optional pipeline stages
inferable from RTL and retiming

Stratix/Arria-10 IEEE-754 DSPs



[Intel/Altera Stratix-10 FPGA Features]

Intel® Stratix® 10 Device DSP Block: Single-Precision Floating Point

Can floating-point units be inferred automatically?

Memory

- Flip-flops relatively scarce (only 1-bit per CLB)
- Need more storage when applications moved beyond FSM controllers and glue logic
- Option A: LUTs repurposable as 16x1-bit SRAMs
- Option B: 4Kb (now 32Kb) 2-ported SRAM blocks
  - very compact, very fast because native in silicon
  - explicit instantiation or inferable from RTL

    (tool can even decide which SRAM option to use)
  - configurable and combinable to a wide range of sizes and aspect ratios

Where are they? How to connect up to them?

FPGA fabric not true blank slate

- FPGA Macros (especially RAM and DSP)
  - coarse functions and structures
  - some powerful but arbitrarily specific features
  - penalty is too huge to not get it right
- Inferable from RTL but . . . .
  - a hard macro only does what it does
  - tools cannot recognize all "functional equivalent" descriptions
  - good idea to check inference report

Straight out-of-the-box ASIC RTL likely suboptimal, sometimes not-mappable

Example: Flip-Flops Inference

- Use asynch set or reset

  $\Rightarrow$ not all FFs have async reset; prevents DSP retiming

- Use both set and reset

  $\Rightarrow$ no FF has this; emulated externally with LUTs

- Use set and reset operationally

  $\Rightarrow$ set/reset cannot use special global lines

- Active-low set/reset and enables

  $\Rightarrow$ need LUTs to turn active-high

How could you know this?

- BRAM cannot be used if combinational read
- Shift registers can be made out of LUTs

  BUT! no set/reset and can't read middle bits
- Registers will retime into multiplier and DSP (if no asynch reset)
- Use "initial" for power-on reset
- Timing analysis doesn't do "latches"
- Many, many more like this. . .

Always want to RTMF!

Becoming Heterogenous "SoC"

"The Virtex brand has been the number one choice of designers worldwide, based on its industry-leading capacity, performance, capabilities, and cost-effectiveness. In each generation of families, our mission is to deliver significantly enhanced functionality while driving down prices. We continue this strategy in the Virtex-II Pro family by including state-of-the-art multi-gigabit serial transceivers and high-performance PowerPC RISC CPUs, in the standard feature set, at no charge. Thus, we enable the creation of next-generation systems designs at previously unattainable price points."

Erich Goetting — Vice President and General Manager of the Xilinx Advanced Products Division

[Xcell Journal, Issue 45, Q1 2003]

Processor Cores

- Not everything needs to be in hardware; not everything improves when made into hardware
- Augment fabric with simple embedded CPUs
  - provide universality of functionality
  - easy handling of irregular, sequential operations
  - easy handling anything that doesn't need to be fast
- Interests developed in early 2000s when FPGA applications grew to systems with DRAM, video, and Ethernets, etc.

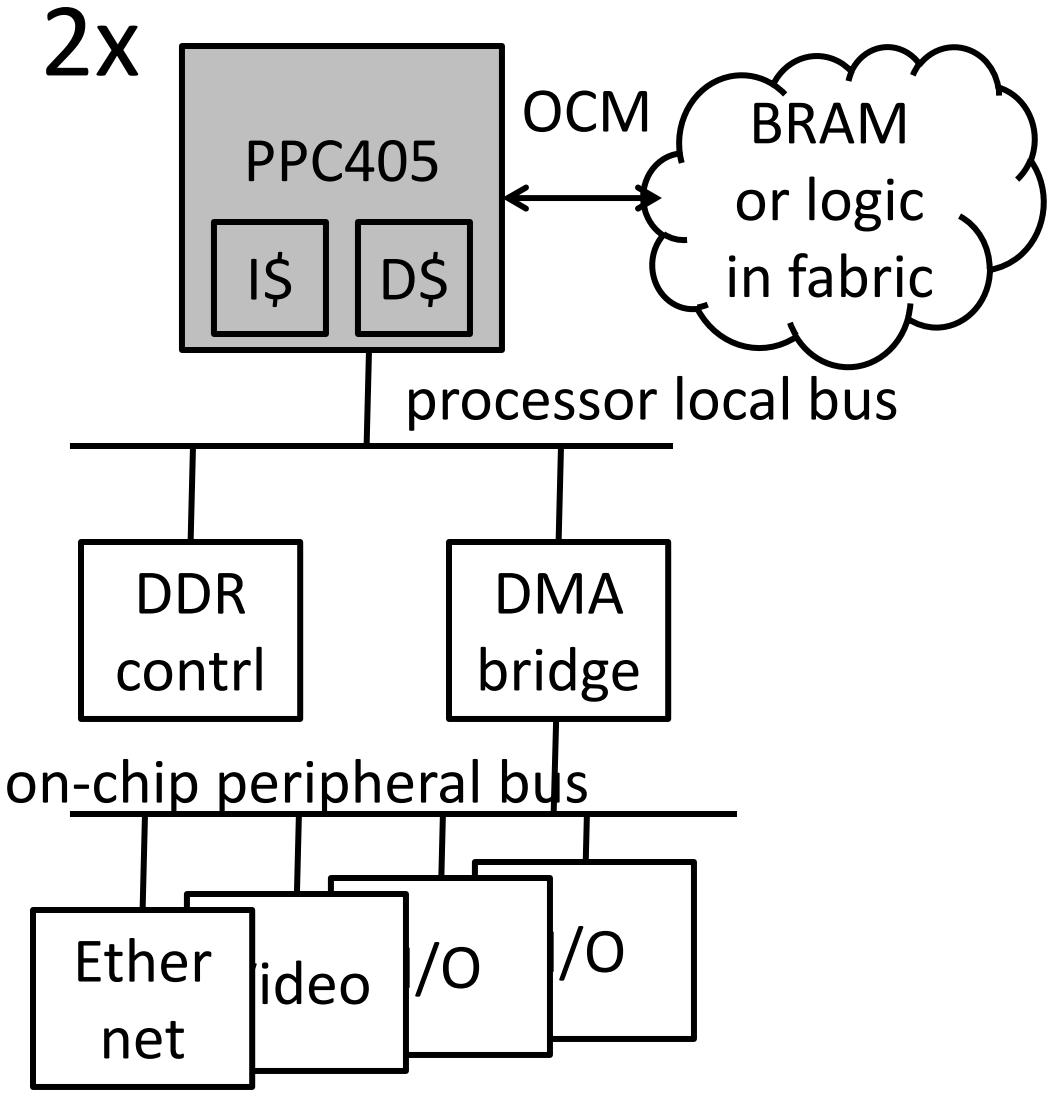
Hard or soft core?

Hardcore vs Softcore

- First came PowerPC hardcores on Virtex-II
  - you got 2 whether you needed them or not
  - new tool promoted IP-based system building
  - entirely soft-logic built surroundings: busses and IPs (DRAM controller, Ethernet, video, ...)
- MicroBlaze softcores took over in later rounds
  - Xilinx proprietary ISA (runs OS, gcc and all that)
  - configurable for cost-performance tradeoff
  - available in RTL to some folks
  - by this time, softcore footprint and performance was acceptable

Several 3rd-party softcores existed in that era, e.g., LEON SPARC

Embedding PowerPC in Fabric



- everything else is soft
- two hierarchies of soft-logic busses
- special on-chip memory (OCM) port allows ld/st directly into fabric
- CoreGen Library of IPs to hang off the busses

[Xilinx Vertex II, early 2000]

Hardcores Return in Virtex7 (~2010)

- This time in a complete, full-speed, fully-capable, two-core Cortex-A9 system
- Latest Ultrascale uses 64-bit ARMv8 Cortex-A53 + ARM R5 + Mali GPU
- Why ARMs?

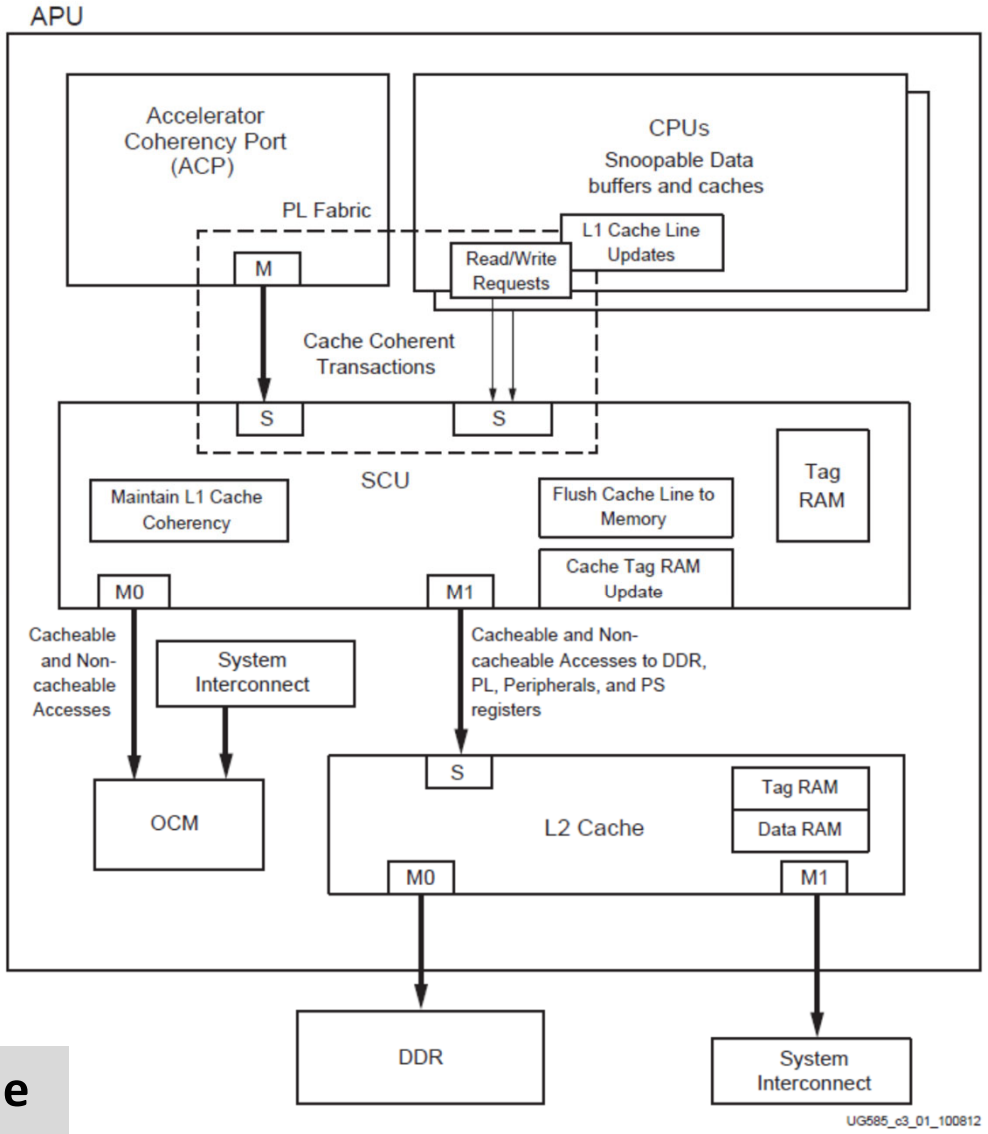


Figure 3-1: **APU Block Diagram**

[Figure 3-1, Zynq-7000 All Programmable SoC Technical Reference Manual]

Hardcore vs Softcore

- Table 4.2: The Zynq Book

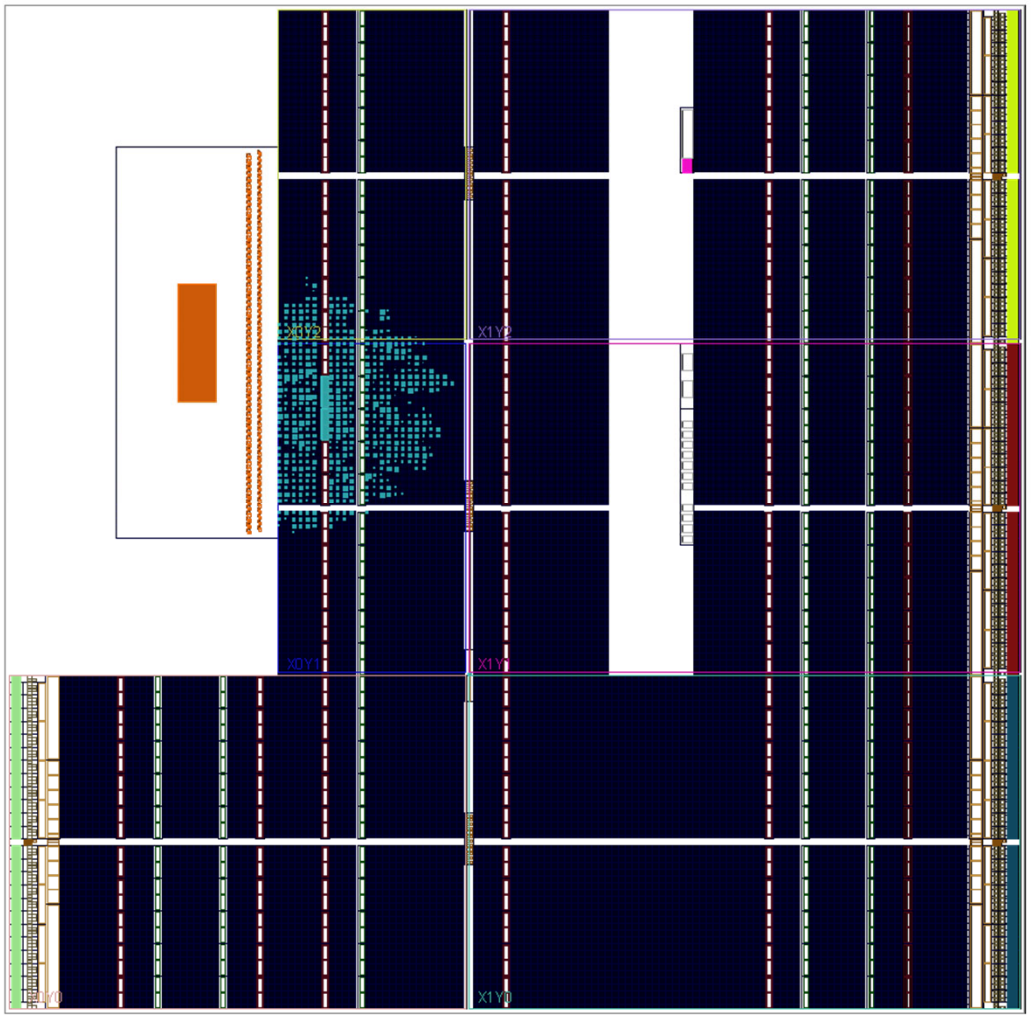
Processor	Configuration	DMIPs
MicroBlaze 900LUT/700FF/ 2BRAM to 3800LUT/3200FF/ 6DSP/21BRAM	area optimized (3-stage)	196
	perf. optimized (5-stage) with branch optimizations	228
	perf. optimized (5-stage) without branch optimizations	259 ??from book
ARM Cortex-A9	1GHz; both cores combined	5000

- Table 4.3: The Zynq Book

Processor	Configuration	CoreMark
MicroBlaze	125MHz; 5-stage (Virtex-5)	238
ARM Cortex-A9	1GHz; both cores combined	5927
ARM Cortex-A9	800MHz; both cores combined	4737

PPC405 about 1/5 of ARM in Figure 4.3 of The Zynq Book

Die Area "Return on Investment"

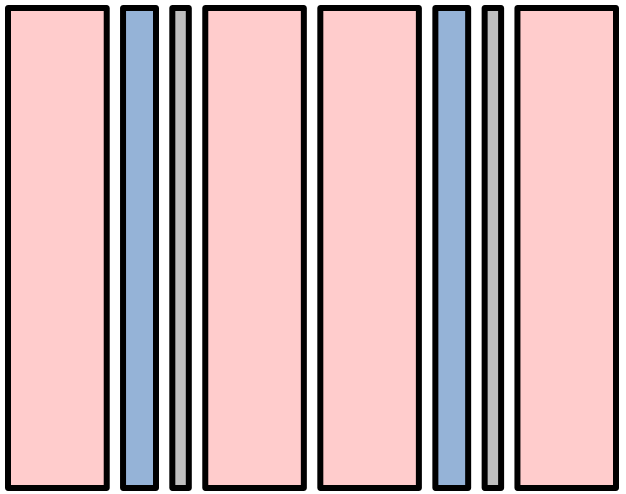


[Vivado screenshot XC7z020]

Soft-logic dominates die area, but compute/storage concentrated in DSP and BRAM—consider what if 100% soft or 100% hard

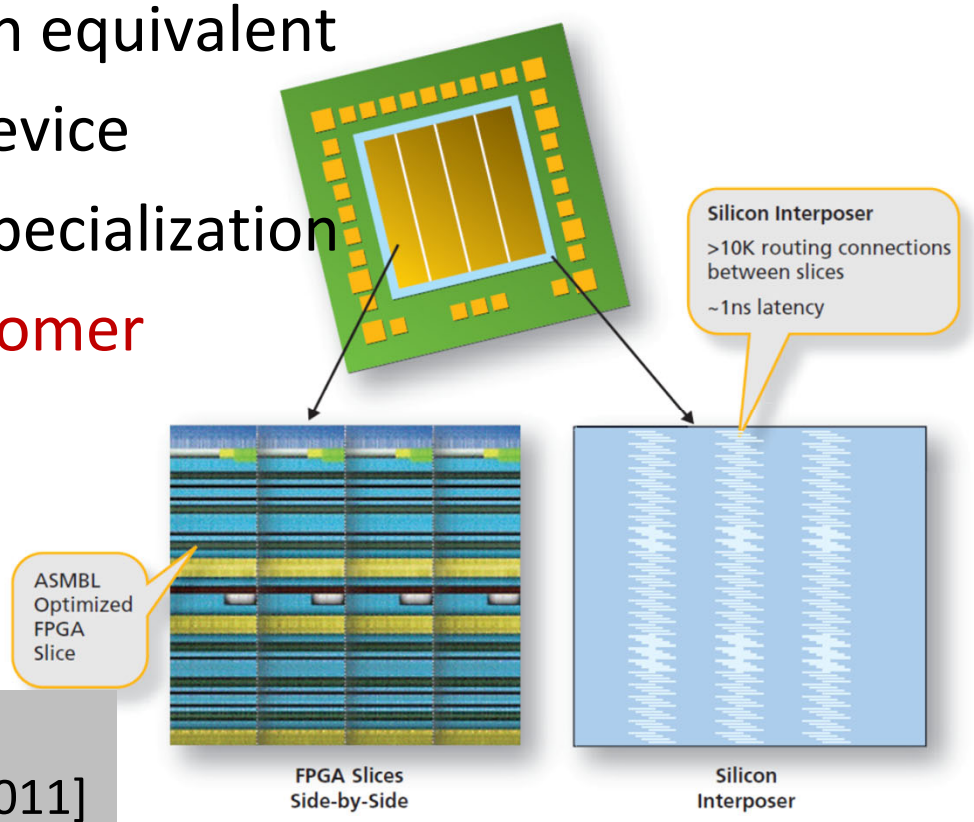
Xilinx ASMBL Architecture
(Application Specific Modular Block Arch.)

- Xilinx fabric assembled from composable tall-and-thin strip types, CLB, BRAM, DSP, I/O, etc.
- Derivative products at the cost of just new masks
  - vary capacity by composing more or fewer strips
  - domain-specialization by varying ratios of strips e.g., {DSP+IP} vs logic for DSP vs ASIC replacement market
  - variations handled by parameterization in design tool algorithms



Stacked Silicon Interconnect (SSI)

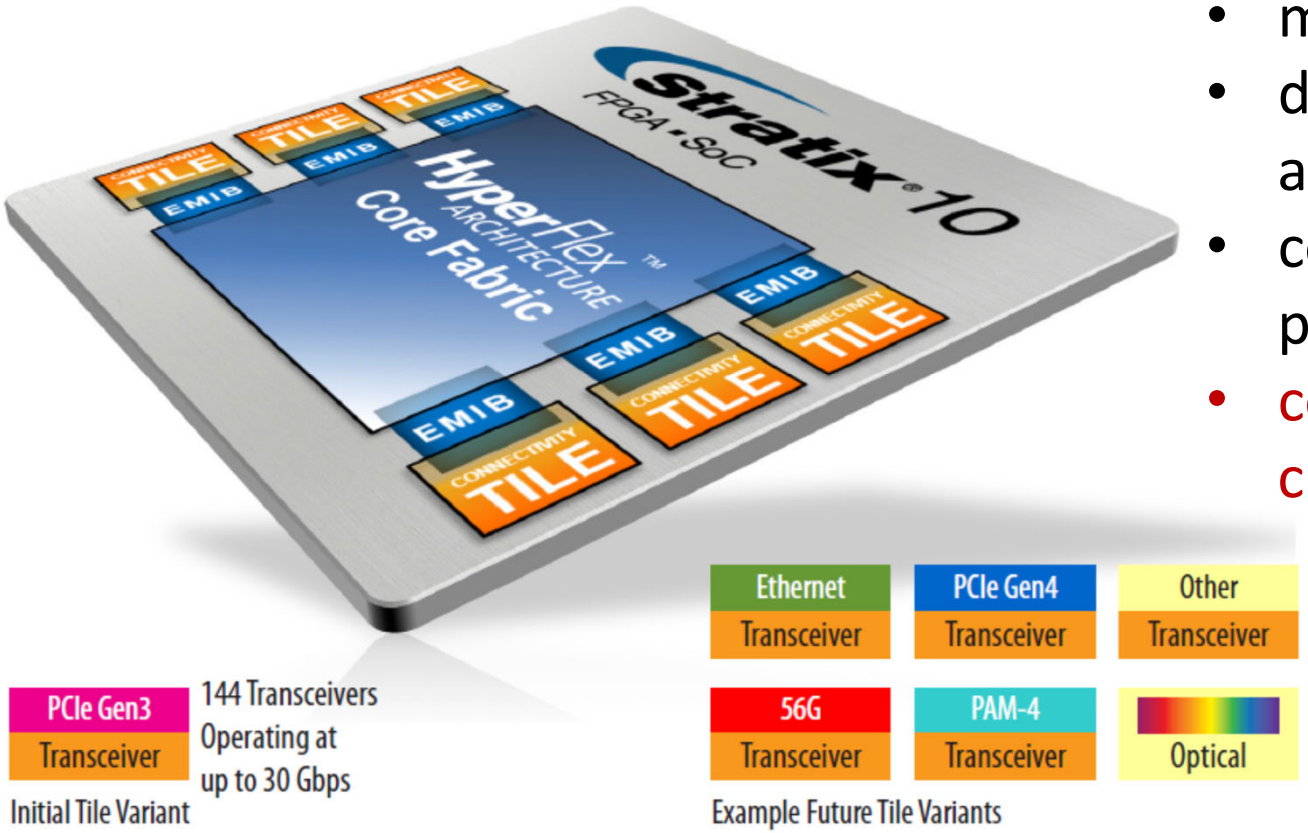
- 2.5D stacking: multiple dies on passive interposer
  - lower latency, higher bandwidth, lower power than crossing package
  - much better yield than equivalent capacity monolithic device
  - mix dies for domain-specialization
  - possible to insert customer proprietary dies?



[Figure 1, Stacked & Loaded: Xilinx SSI, 28-Gbps I/O Yield Amazing FPGAs, Xcell, Q1 2011]

Intel's take on 2.5D with EMIB

Figure 8. Enhanced Flexibility and Scalability with Separate Transceiver Tiles



- monolithic fabric
- displace noisy, hot analog IPs
- connect same-package HBMs
- connect 3rd-party chiplets?

[Figure 8, Enabling Next-Generation Platforms Using Altera's 3D System-in-Package Technology]

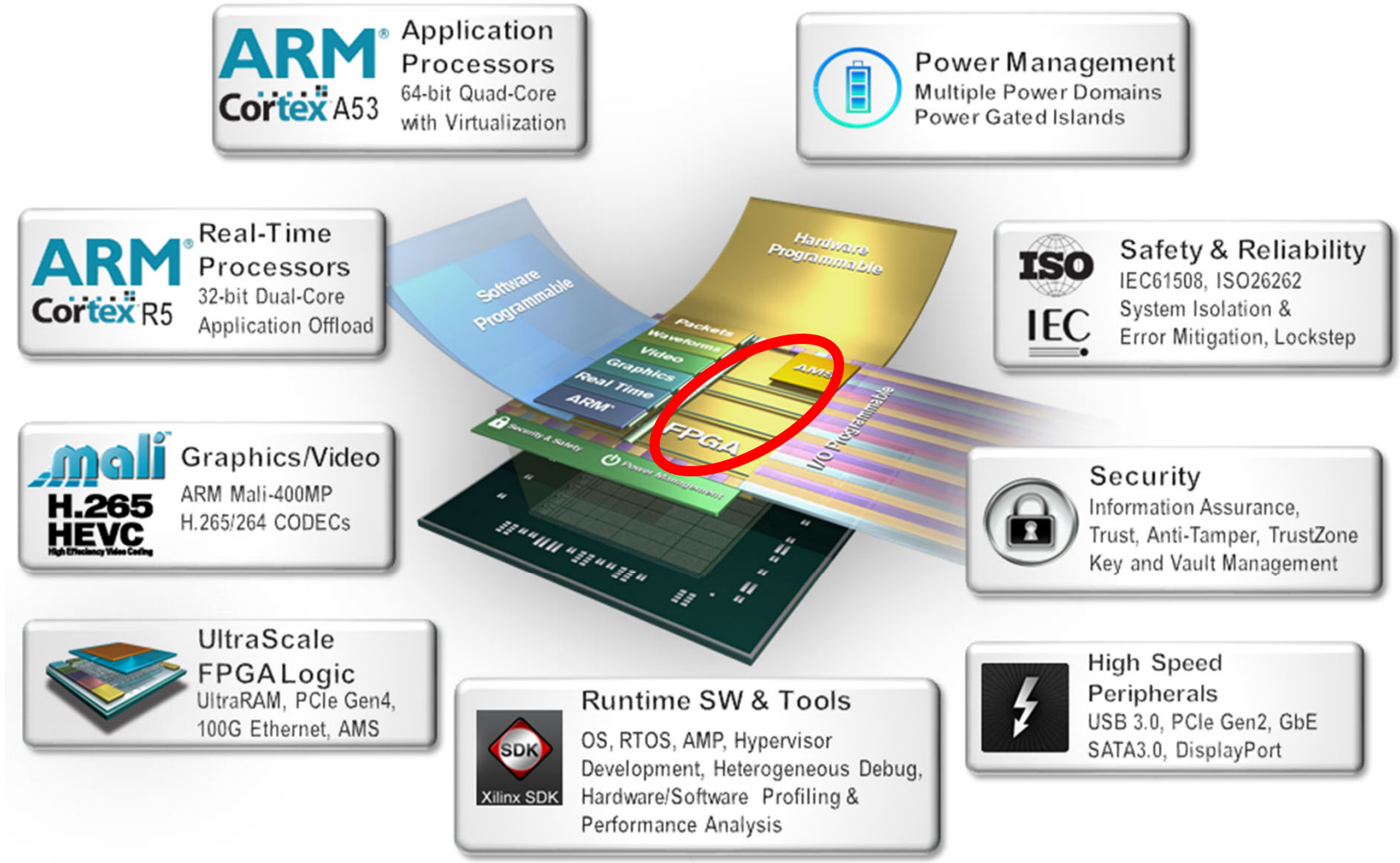
Reviewing Hard IPs Added Over Time

- 1990s
  - fast carry
  - LUT RAM
  - block RAM
- 2000s
  - programmable clock generator
  - PowerPC core
  - gigabit transceiver
  - multiplier and DSP slices
  - Ethernet and PCI-E
- 2010s
  - system monitor
  - ADC
  - power management
  - ARM cores and GPU
  - DRAM controller
  - floating point arithmetic
  - **"UltraRAM" hierarchy (up to 500Mbits)**
  - **HBM controllers**
- 2020s . . . . *next lecture*

Chicken or Egg First?

- 1990s: glue logic, embedded cntrl, interface logic
  - reduce chip-count, increase reliability
  - rapid roll-out of “new” products
- 2000s: DSP and HPC
  - strong need for performance
  - abundant parallelism and regularity
  - low-volume, high-valued
- 2010s: communications and networking
  - throughput performance
  - fast-changing designs and standards
  - price insensitive
  - $value in field updates and upgrades

SoC with reconfigurable fabric (2010s)



[http://www.xilinx.com/products/silicon-devices/soc/zynq-ultrascale-mpsoc.html]

2026 AMD Versal Premium Offerings

		VP1002	VP1052	VP1102	VP1202	VP1402	VP1502	VP2502	VP1552	VP1702	VP1802	VP2802	VP1902	ultra96 ZU3EG
AI Engine	AI Engines Tiles	–	–	–	–	–	–	472	–	–	–	472	–	--
	AI Engine Data Memory (Mb)	–	–	–	–	–	–	118	–	–	–	118	–	--
Programmable Logic	System Logic Cells (K)	833	1,186	1,575	1,969	2,233	3,763	3,738	3,837	5,558	7,352	7,326	18,507	154K
	LUTs	380,800	542,080	719,872	900,224	1,020,928	1,720,448	1,708,672	1,753,984	2,540,672	3,360,896	3,349,120	8,460,288	71K
	DSP Engines	1,140	1,572	1,904	3,984	2,672	7,440	7,392	7,392	10,896	14,352	14,304	6,864	360
	NoC Master / NoC Slave Ports	22	22	30	28	42	52	52	52	76	100	100	192	--
	Super Logic Regions (SLRs)[1]	–	–	–	–	–	2	2	2	3	4	4	4	--
	Distributed RAM (Mb)	12	17	22	27	31	53	52	54	78	103	102	258	1.8
Memory	Block RAM (Mb)	19	26	49	47	70	89	89	89	132	174	174	239	7.6
	UltraRAM (Mb)	97	138	127	190	181	366	366	366	541	717	717	619	--
	Multiport RAM (Mb)	80	80	–	–	–	–	–	–	–	–	–	–	--
	Total PL Memory (Mb)	208	261	198	264	282	508	507	509	751	994	994	1116	9.4
	DDR4 Memory Controllers	2	2	3	4	3	4	4	4	4	4	4	14	1
	DDR Bus Width	128	128	192	256	192	256	256	256	256	256	256	896	64
Processing System	Application Processing Unit	Dual-core Arm® Cortex®-A72, 48 KB/32 KB L1 Cache w/ parity & ECC; 1 MB L2 Cache w/ ECC												similar
	Real-Time Processing Unit	Dual-core Arm Cortex-R5F, 32 KB/32 KB L1 Cache, and 256 KB TCM w/ECC												similar
	Memory	256 KB On-Chip Memory w/ECC												similar
	Connectivity	Ethernet (x2); UART (x2); CAN-FD (x2); USB 2.0 (x1); SPI (x2); I2C (x2)												similar
Serial Transceivers	GTY Transceivers	8	8	–	–	–	–	–	–	–	–	–	–	--
	GTYP Transceivers	–	–	8	28[2]	8	28[2]	28[2]	68[2]	28[2]	28[2]	28[2]	128	--
	GTM Transceivers (58G (112G))	24 (12)	36 (18)	64 (32)	20 (10)	96 (64)[3]	60 (30)	60 (30)	20 (10)	100 (50)	140 (70)	140 (70)	32 (16)	--
Integrated Protocol IP	PCIe® w/DMA (CPM4)	2 x Gen4x4	2 x Gen4x4	–	–	–	–	–	–	–	–	–	–	Gen2 x4
	PCIe w/DMA (CPM5)	–	–	–	2 x Gen5x8	–	2 x Gen5x8	2 x Gen5x8	2 x Gen5x8	2 x Gen5x8	2 x Gen5x8	2 x Gen5x8	–	
	PCIe (PL PCIE4)	1 x Gen4x8	1 x Gen4x8	–	–	–	–	–	–	–	–	–	–	
	PCIe (PL PCIE5)	–	–	2 x Gen5x4	2 x Gen5x4	2 x Gen5x4	2 x Gen5x4	2 x Gen5x4	8 x Gen5x4	2 x Gen5x4	2 x Gen5x4	2 x Gen5x4	16 x Gen5x4	
	100G Multirate Ethernet MAC	3	5	6	2	6	4	4	4	6	8	8	12	
	600G Ethernet MAC	2	3	7	1	11	3	3	1	5	7	7	4	
	600G Interlaken	1	2	–	–	–	1	1	–	2	3	3	–	
	400G High-Speed Crypto Engine	1	1	3	1	4	2	2	2	3	4	4	–	
Ordering Information	Extended[4]	-1MSE, -1LSE, -2MSE, -2MLE, -2LSE, -2LLE		-1MSE, -1LSE, -2MSE, -2MLE, -2LSE, -2LLE, -3HSE										
	Industrial[4]	-1MSI, -1MLI, -1LSI, -1LLI, -2MSI, -2MLI, -2LLI, -2HSI		-1MSI, -1MLI, -1LSI, -1LLI, -2MSI, -2MLI									-	

[UltraScale FPGA Product Tables and Product Selection Guide (XMP102)]

2026 Altera Agilex-7 Offerings

	Product Line	AGI 019	AGI 023	AGI 022	AGI 027	AGI 035	AGI 040	AGI 041
Resources	Specification Options[1]	A, B, C, D	A, B, C, D	A, B	A, B	A, C	A, C	A, B, C, D
	Logic elements (LEs)	1,918,975	2,308,080	2,208,075	2,692,760	3,540,000	4,047,400	4,000,672
	Adaptive logic modules (ALMs)	650,500	782,400	748,500	912,800	1,200,000	1,372,000	1,356,160
	ALM registers	2,602,000	3,129,600	2,994,000	3,651,200	4,800,000	5,488,000	5,424,640
	High-performance crypto blocks	2	2	0	0	4	4	4
	eSRAM memory blocks	1	1	0	0	3	3	2
	eSRAM memory size (Mb)	18	18	0	0	54	54	36
	M20K memory blocks	8,500	10,464	10,900	13,272	14,931	19,908	17,136
	M20K memory size (Mb)	166	204	212	259	292	389	335
	MLAB memory count	32,525	39,120	37,425	45,640	60,000	68,600	67,808
	MLAB memory size (Mb)	20	24	23	28	37	42	42
	Fabric PLL	5	5	12	12	6	6	8
	I/O PLL	10	10	16	16	12	12	16
	Variable-precision digital signal processing (DSP) blocks	1,354	1,640	6,250	8,528	9,594	12,792	0
	18 x 19 multipliers	2,708	3,280	12,500	17,056	19,188	25,584	0
	Single-precison or half-precision tera floating point operations per second (TFLOPS)	2.4 / 4.9	2.4 / 4.9	9.4 / 18.8	12.8 / 25.6	14.3 / 28.7	19.1 / 38.3	0

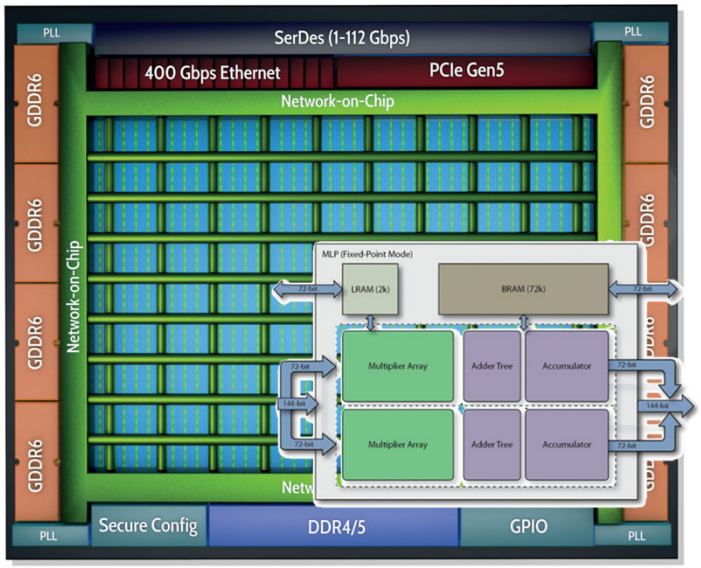
Tile Resources		- Bifurcateable 200 Gb hard IP block (10/25/50/100/200 Gbps FEC/PCS) 300G Interlaken IEEE 1588 v2 support PMA direct
	E-Tile	Transceiver channel count : Up to 24 channels at 28.9 Gbps (NRZ) / 12 channels at 58 Gbps (PAM4) - RS & KP FEC[1] Networking support : - 400GbE (4 x 100GbE hard IP blocks (10/25 GbE FEC/PCS/MAC)) IEEE 1588 v2 support PMA direct
	P-Tile	PCIe hard IP block (4.0 x16) or bifurcateable 2x PCIe 4.0 x8 (EP) or 4x 4.0 x4 (RP) SR-IOV 8PF / 2kVF VirtIO support Scalable IOV

[Altera FPGA Product Catalog]

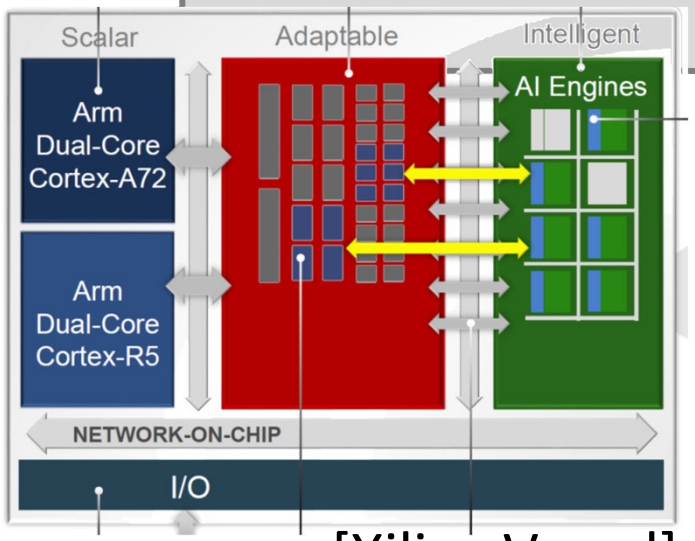
Today's Diverging Architectures

Are they FPGAs?

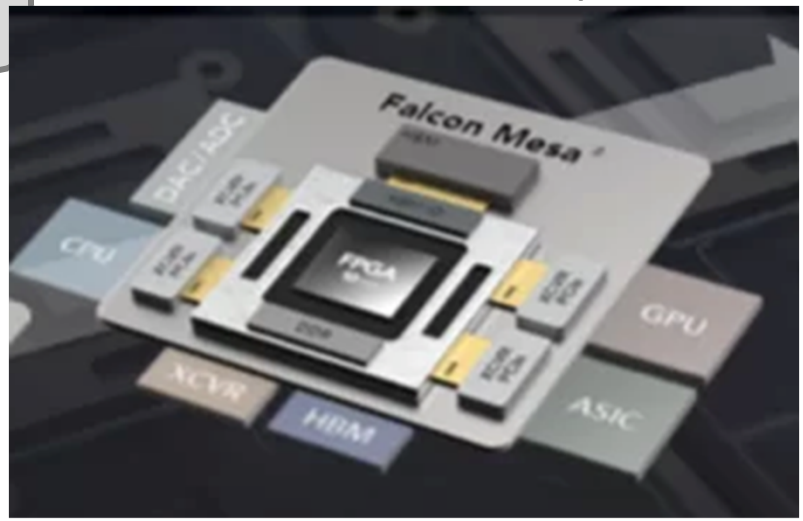
- spatial data/compute
  - highly concurrent
  - finely controllable
  - reprogrammable



[Achronix Speedster]

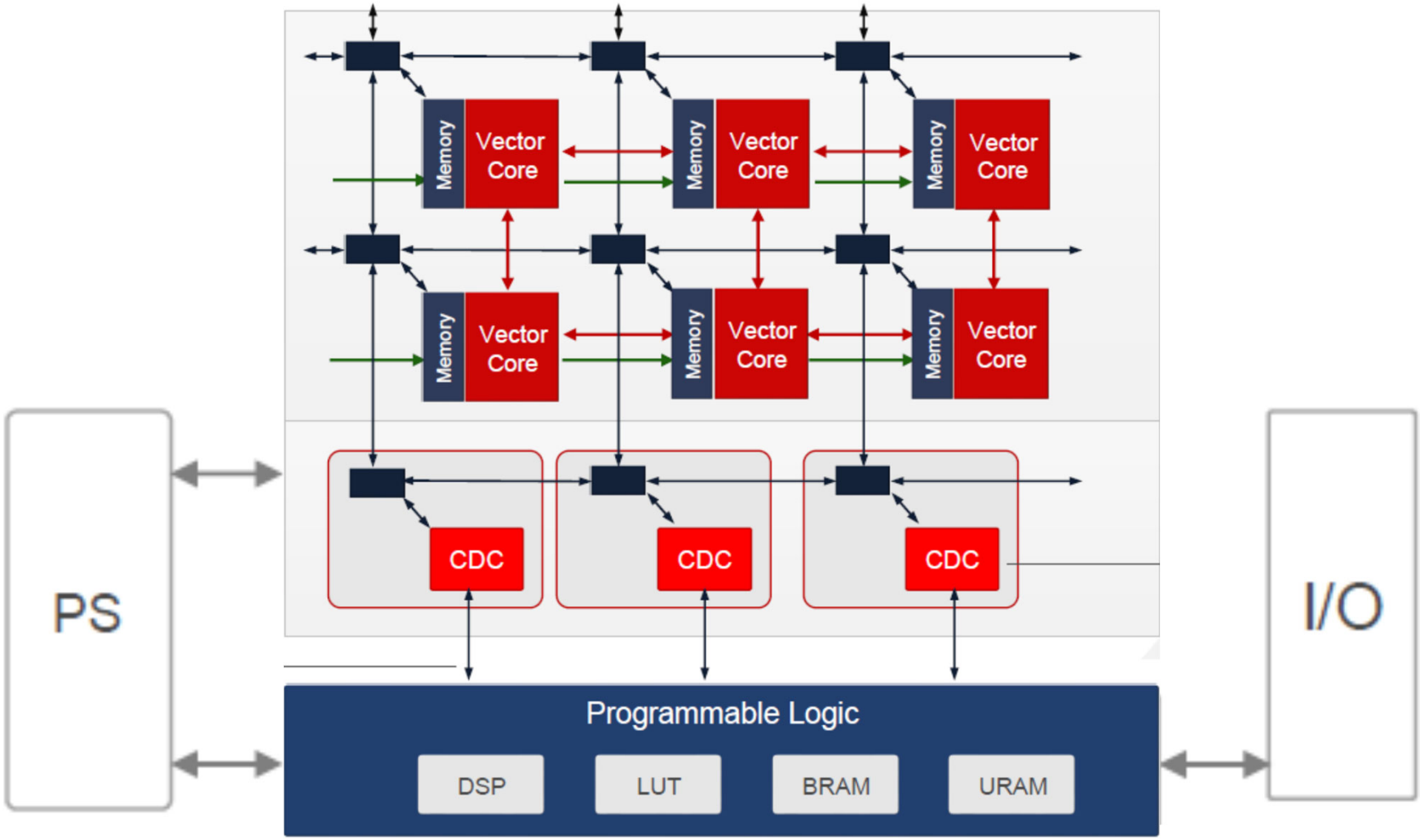


[Xilinx Versal]



[Altera Agilex]

Xilinx Versal AI Engines



HotChips 2018, “HW/SW Programmable Engine”

If not RTL then what?

Parting Thoughts

- FPGAs steadily moved away from universal fabric
  - efficiency of hardwired logic (driven by application demands) complements flexibility of reconfig. logic
  - architected deliberately to play up this advantage
- Retain a high degree of regularity to ease design and manufacturing
  - fastest way to use up transistors from Moore's Law
  - power and performance advantage by just being first on new process
- Architectural evolution both push-and-pull with applications