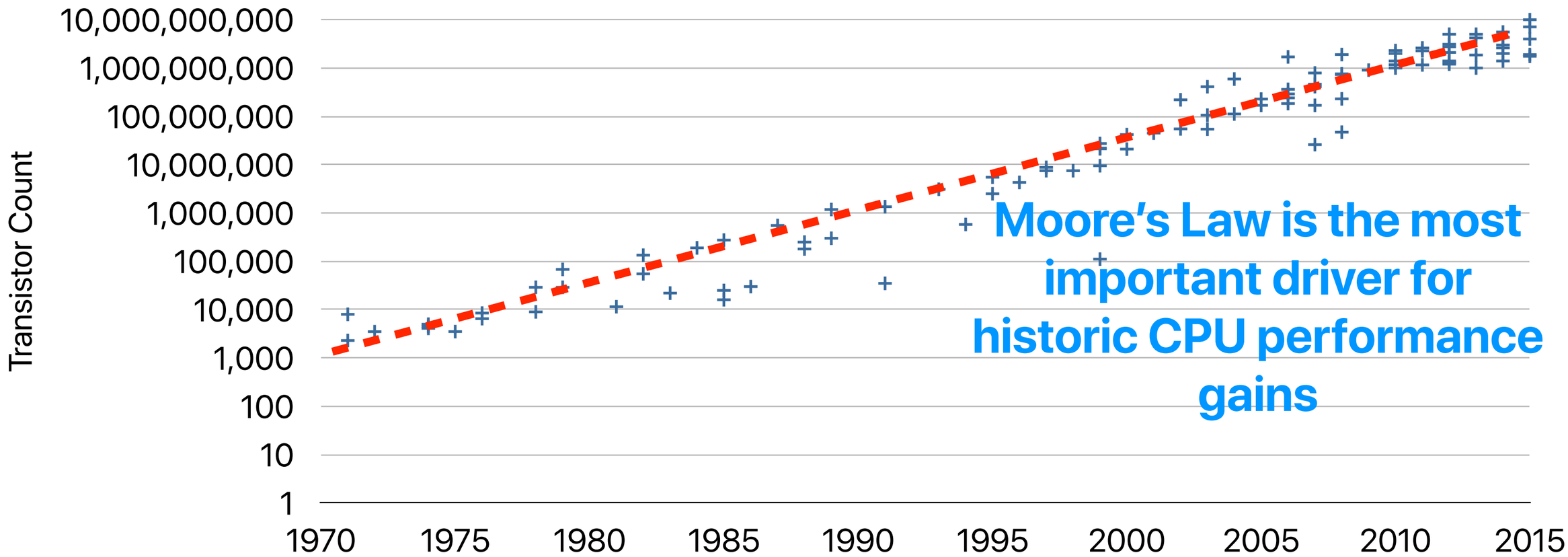
Clock, Power Consumption and the Future Landscape of Computation

Prof. Usagi

Recap: Moore's Law[1]

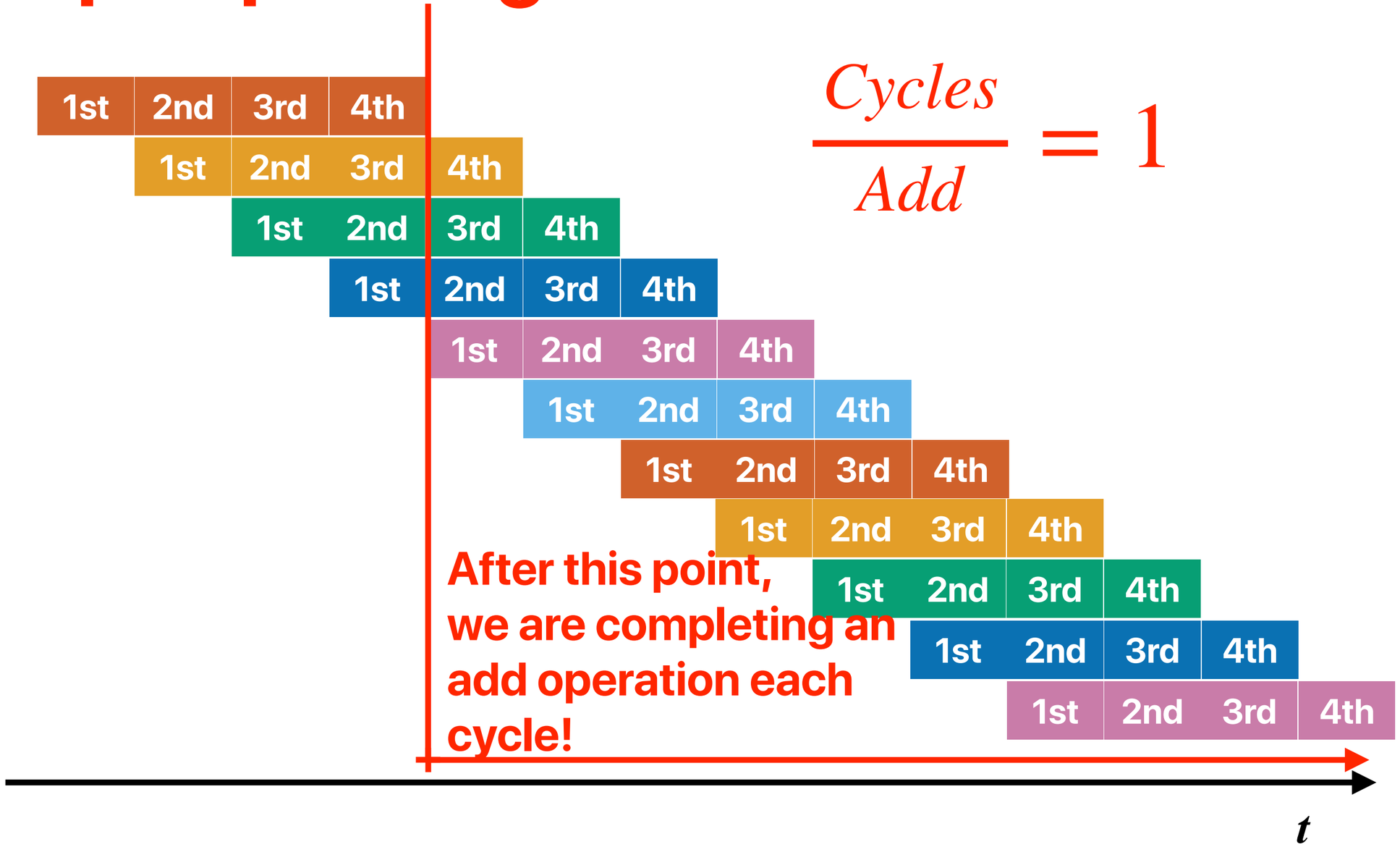
- The number of transistors we can build in a fixed area of silicon doubles every 12 ~ 24 months.



(1) Moore, G. E. (1965), 'Cramming more components onto integrated circuits', Electronics 38 (8) .

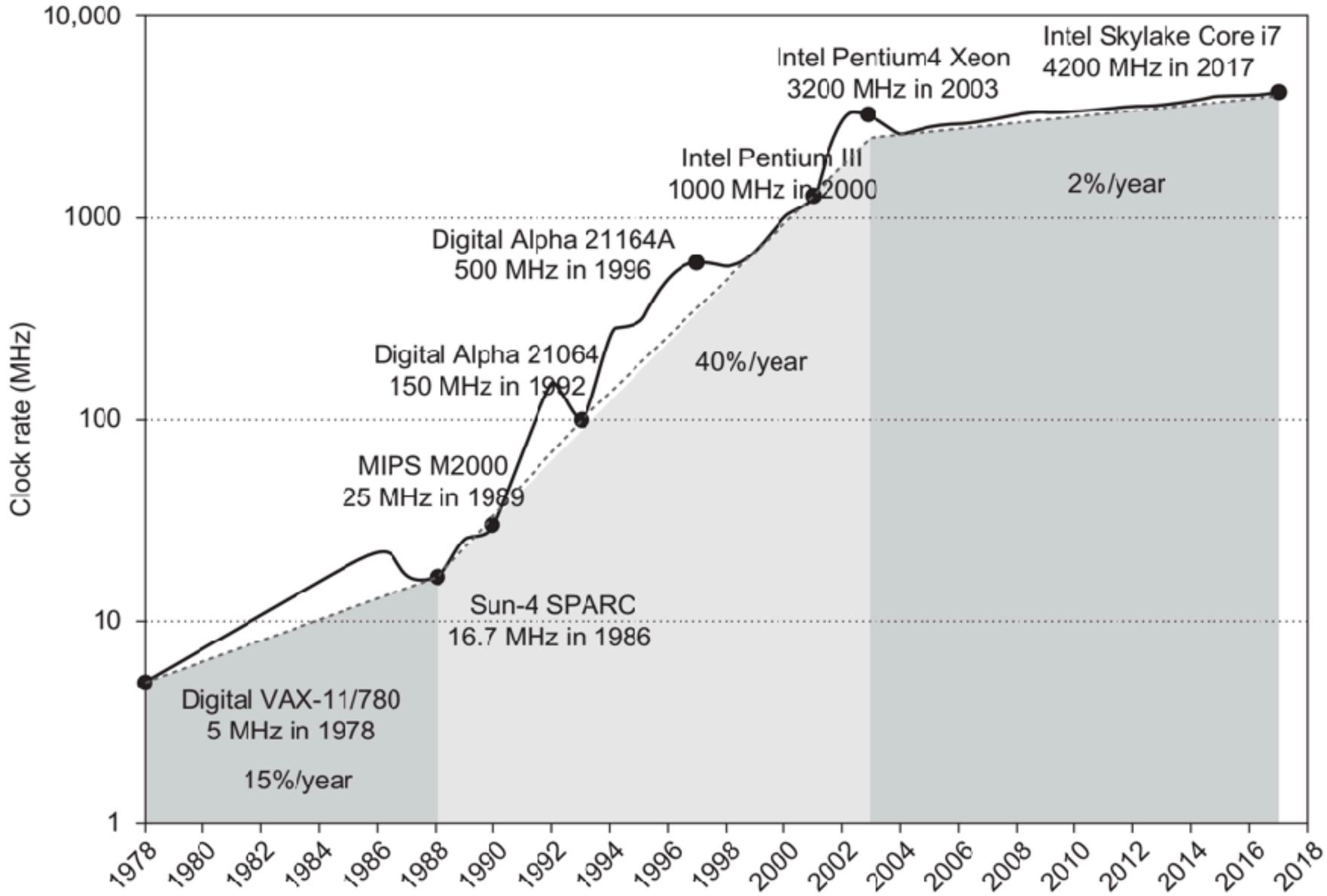
Recap: Pipelining a 4-bit serial adder

```
add a, b
add c, d
add e, f
add g, h
add i, j
add k, l
add m, n
add o, p
add q, r
add s, t
add u, v
```



$$\frac{Cycles}{Add} = 1$$

Recap: The growth of clock rate is slowing down

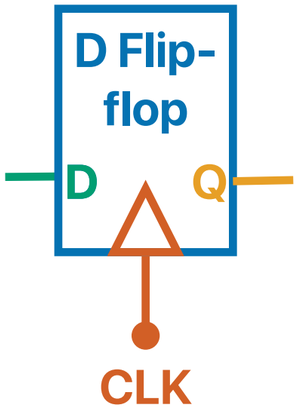


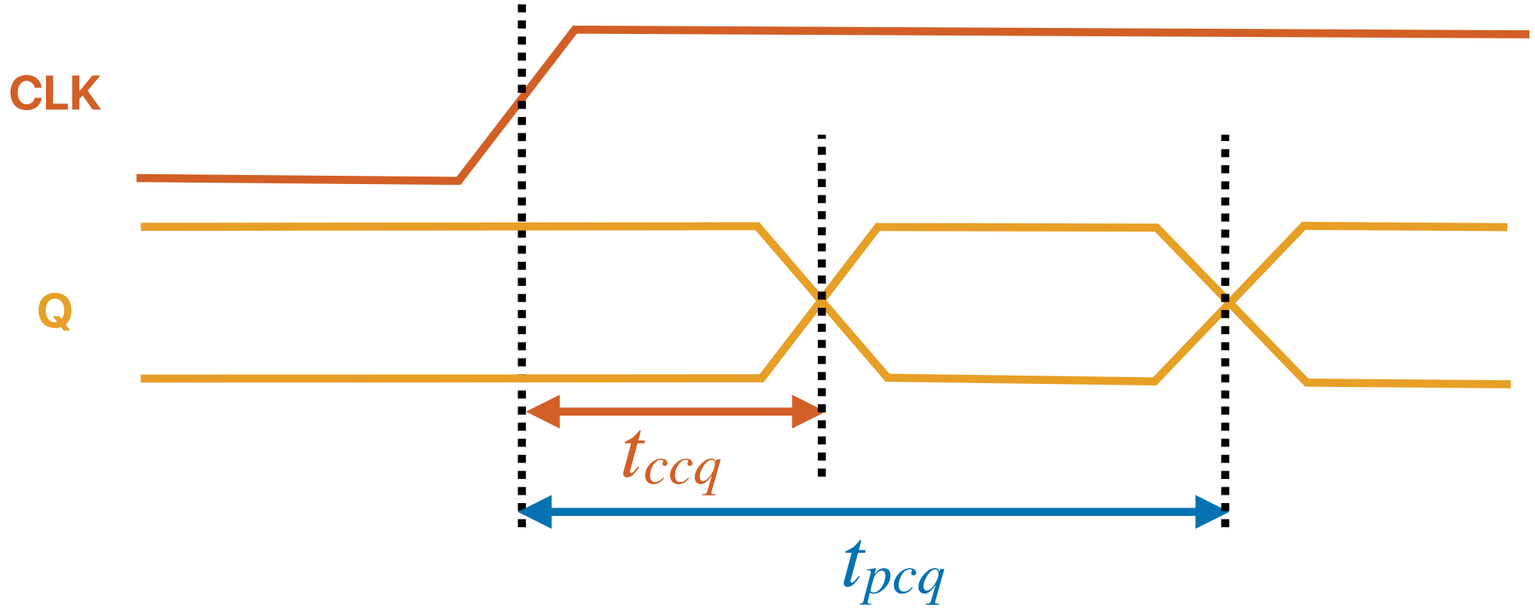
Outline

- What are the basic limits of clock frequency?
- New limit on clock frequency: Power consumption
- Opportunities and the future

Timing constraints

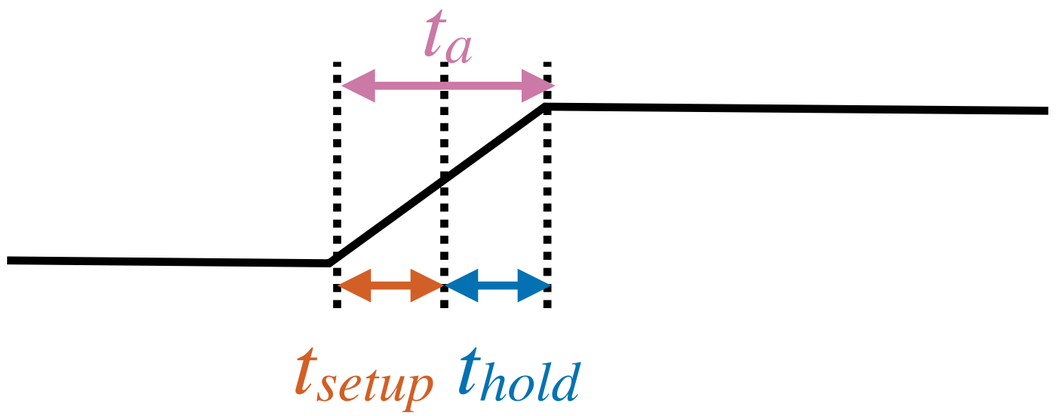
Output Timing Constraints





- Min delay of FF, also called contamination delay or min CLK to Q delay: t_{ccq}
  - Time after clock edge that Q might be unstable (i.e., starts changing)
- Max delay of FF, also called propagation delay or maximum CLK to Q delay: t_{pcq}
  - Time after clock edge that the output Q is guaranteed to be stable (i.e. stops changing)

Setup and hold times for a flip-flop

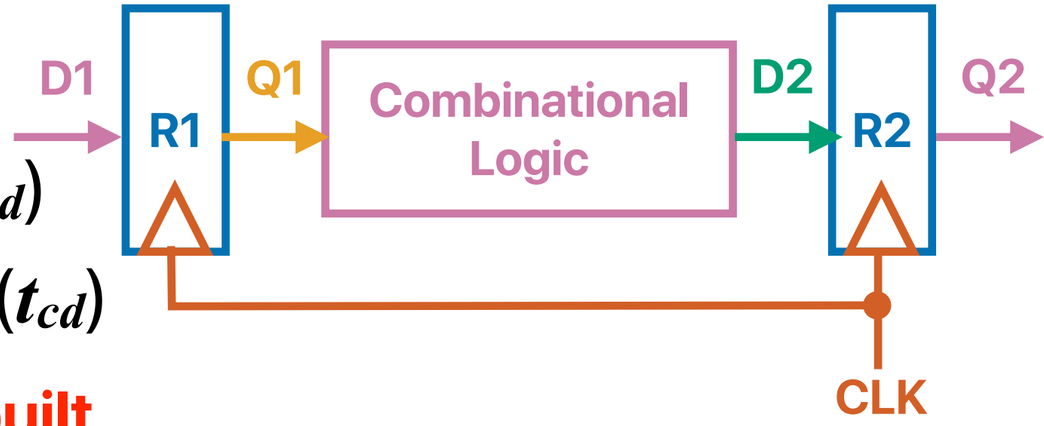


- Setup time: t_{setup}
  - Time before the clock edge that data must be stable (i.e. not change)
  - For the FF to capture the input
- Hold time: t_{hold}
  - Time after the clock edge that data must be stable
  - For the FF to output/store/propagate the data
- Aperture time: t_a
  - Time around clock edge that data must be stable ($t_a = t_{setup} + t_{hold}$)

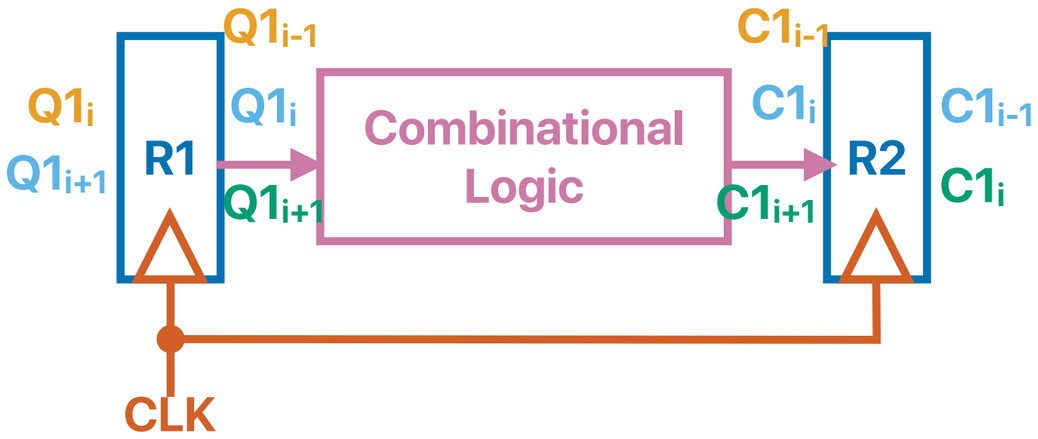
Summary on timing constraints

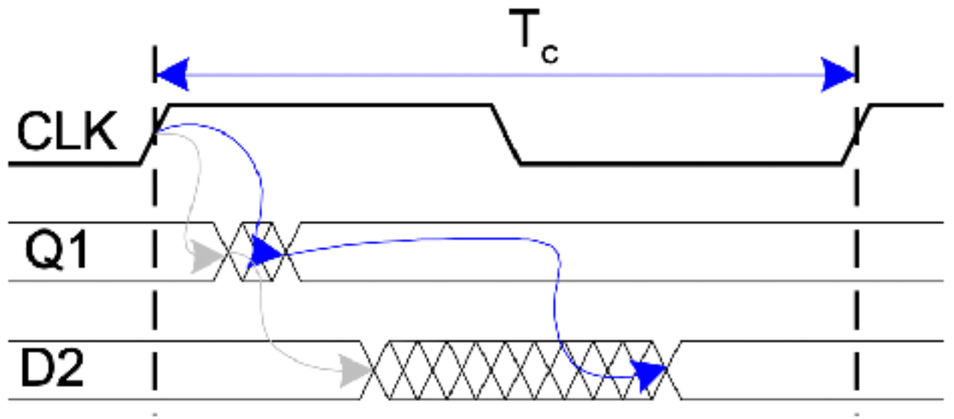
- Combinational:
  - Maximum delay = **P**ropagation **d**elay (t_{pd})
  - Minimum delay = **C**ontamination **d**elay (t_{cd})
- Flip Flops:
  - Input
    - **Setup** time (t_{setup})
    - **Hold** time (t_{hold})
  - Output
    - **P**ropagation **c**lock-to-**Q** time (t_{pcq})
    - **C**ontamination **c**lock-to-**Q** time (t_{ccq})

Once the logic/FFs are built, these timing characteristics are fixed properties



Timing in a circuit



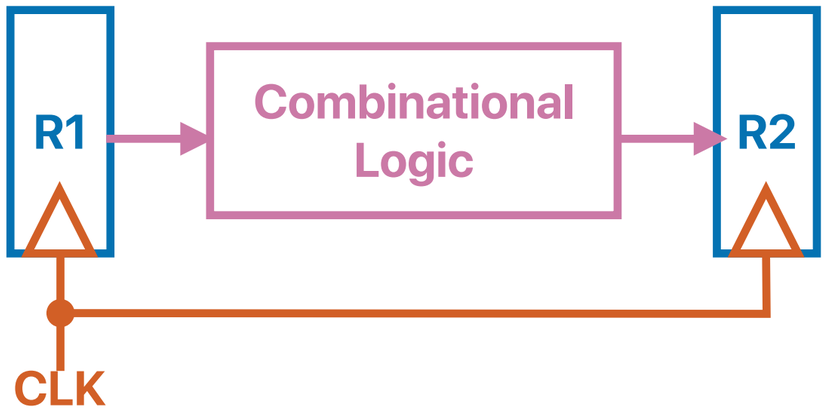


Poll close in 1:30

Causes of Timing Issues in Sequential Circuits

- Input to a FF comes from the output of another FF through a combinational circuit
- The FF and combinational circuit have a min & max delay
- Which of the following violations occurs if max delay of R1 is zero & max delay of the combinational circuit is equal to the clock period?

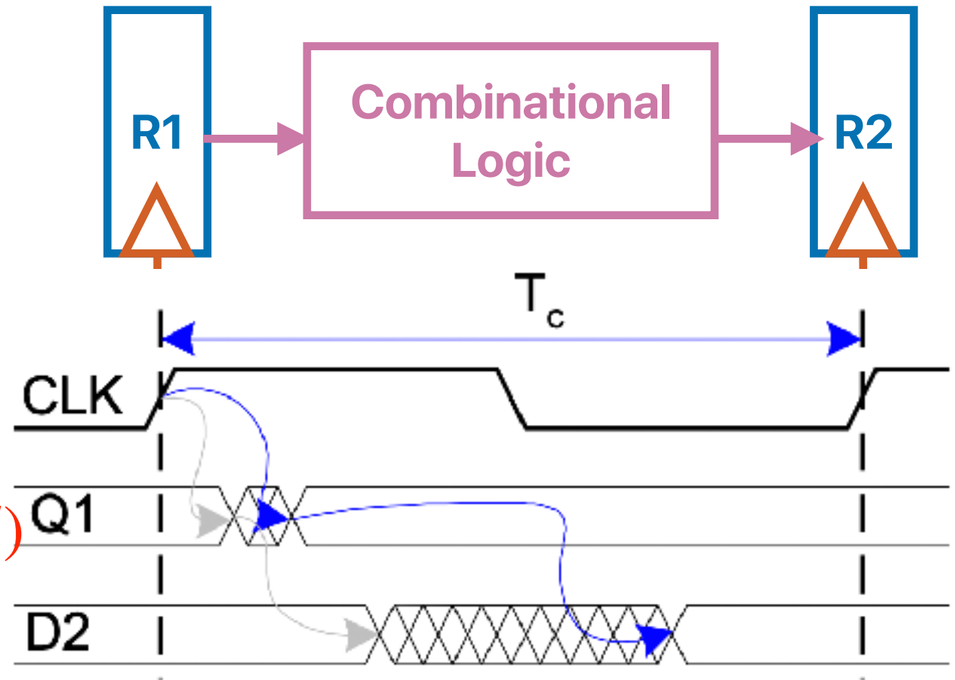
A. Hold time violation for R2
B. Setup violation for R2
C. Hold time violation for R1
D. Setup violation for R1
E. None of the above



Causes of Timing Issues in Sequential Circuits

- Input to a FF comes from the output of another FF through a combinational circuit
- The FF and combinational circuit have a min & max delay
- Which of the following violations occurs if max delay of R1 is zero & max delay of the combinational circuit is equal to the clock period?

A. Hold time violation for R2
B. Setup violation for R2
C. Hold time violation for R1
D. Setup violation for R1
E. None of the above



$$T_c \geq t_{setup} + max\_delay(FF) + max\_delay(combinational)$$

$$T_c \geq t_{setup} + t_{pcq} + t_{pd}$$

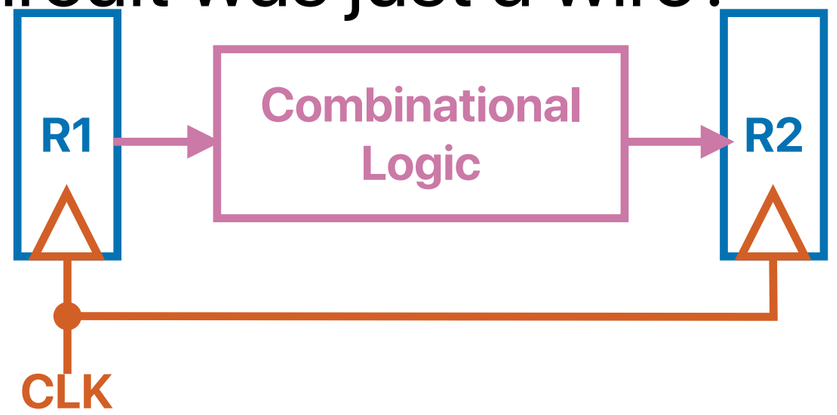
$t_{pcq} = 0$, $t_{pd} = t_c$

Poll close in 1:30

Causes of Timing Issues in Sequential Circuits (2)

- Input to a FF comes from the output of another FF through a combinational circuit
- The FF and combinational circuit have a min & max delay
- Which of the following violations occurs if min delay of R1 is zero & max delay of the combinational circuit was just a wire?

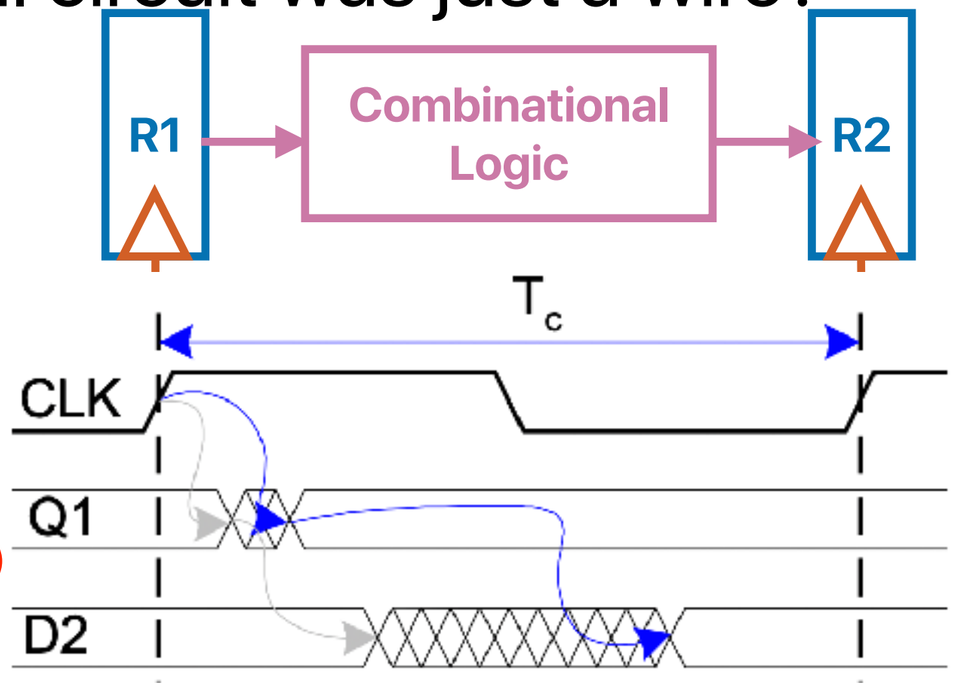
A. Hold time violation for R2
B. Setup violation for R2
C. Hold time violation for R1
D. Setup violation for R1
E. None of the above



Causes of Timing Issues in Sequential Circuits (2)

- Input to a FF comes from the output of another FF through a combinational circuit
- The FF and combinational circuit have a min & max delay
- Which of the following violations occurs if min delay of R1 is zero & max delay of the combinational circuit was just a wire?

A. Hold time violation for R2
B. Setup violation for R2
C. Hold time violation for R1
D. Setup violation for R1
E. None of the above

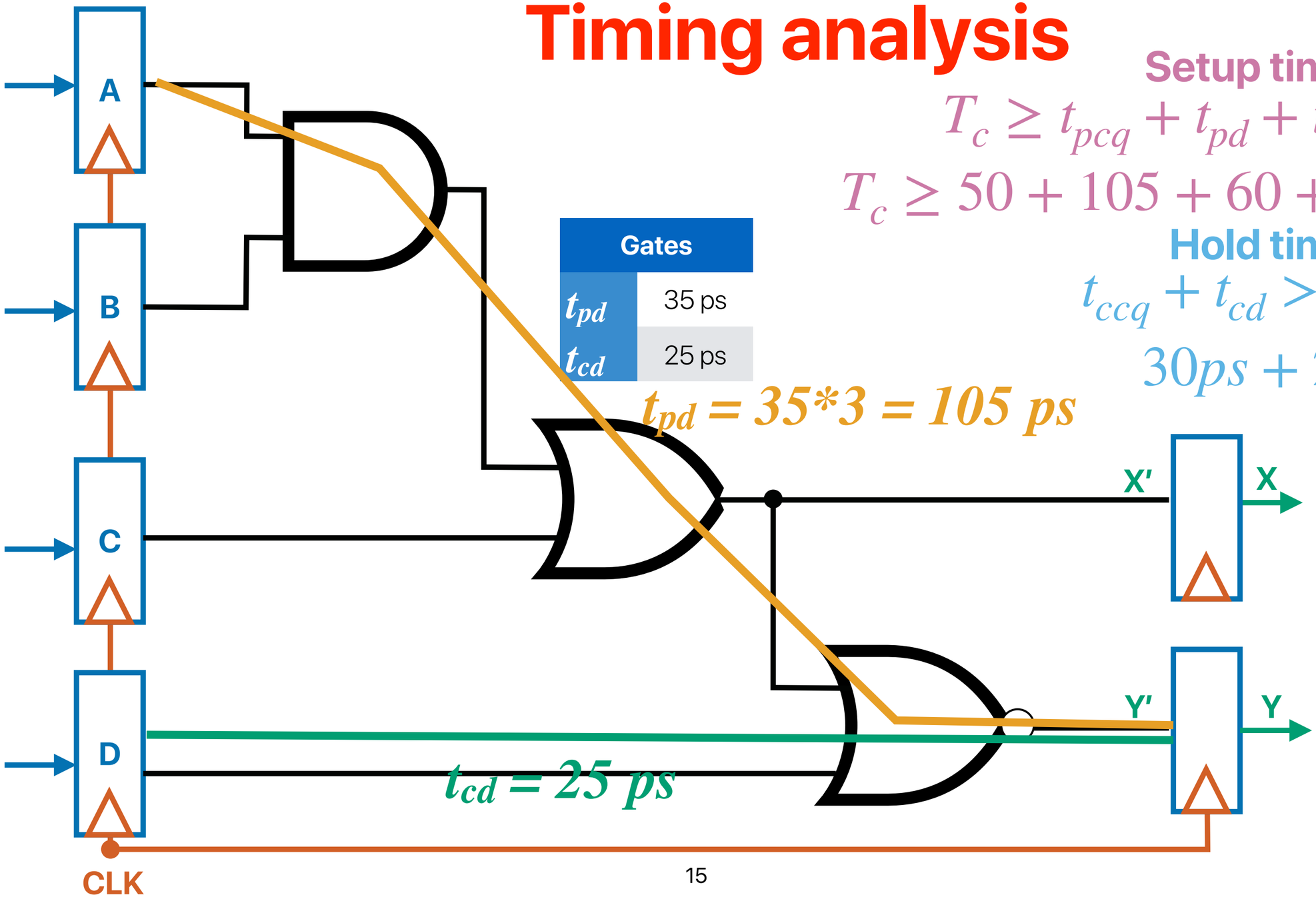


$$T_{hold} \leq min\_delay(FF) + min\_delay(combinational)$$

$$T_{hold} \leq t_{ccq} + t_{cd}$$

$t_{ccq} = 0$, $t_{cd} = 0$

Timing analysis



Gates	
t_{pd}	35 ps
t_{cd}	25 ps

Setup time constraints

$$T_c \geq t_{pcq} + t_{pd} + t_{setup} + t_{skew}$$

$$T_c \geq 50 + 105 + 60 + 0 = 215ps$$

Hold time constraints

$$t_{ccq} + t_{cd} > t_{hold}$$

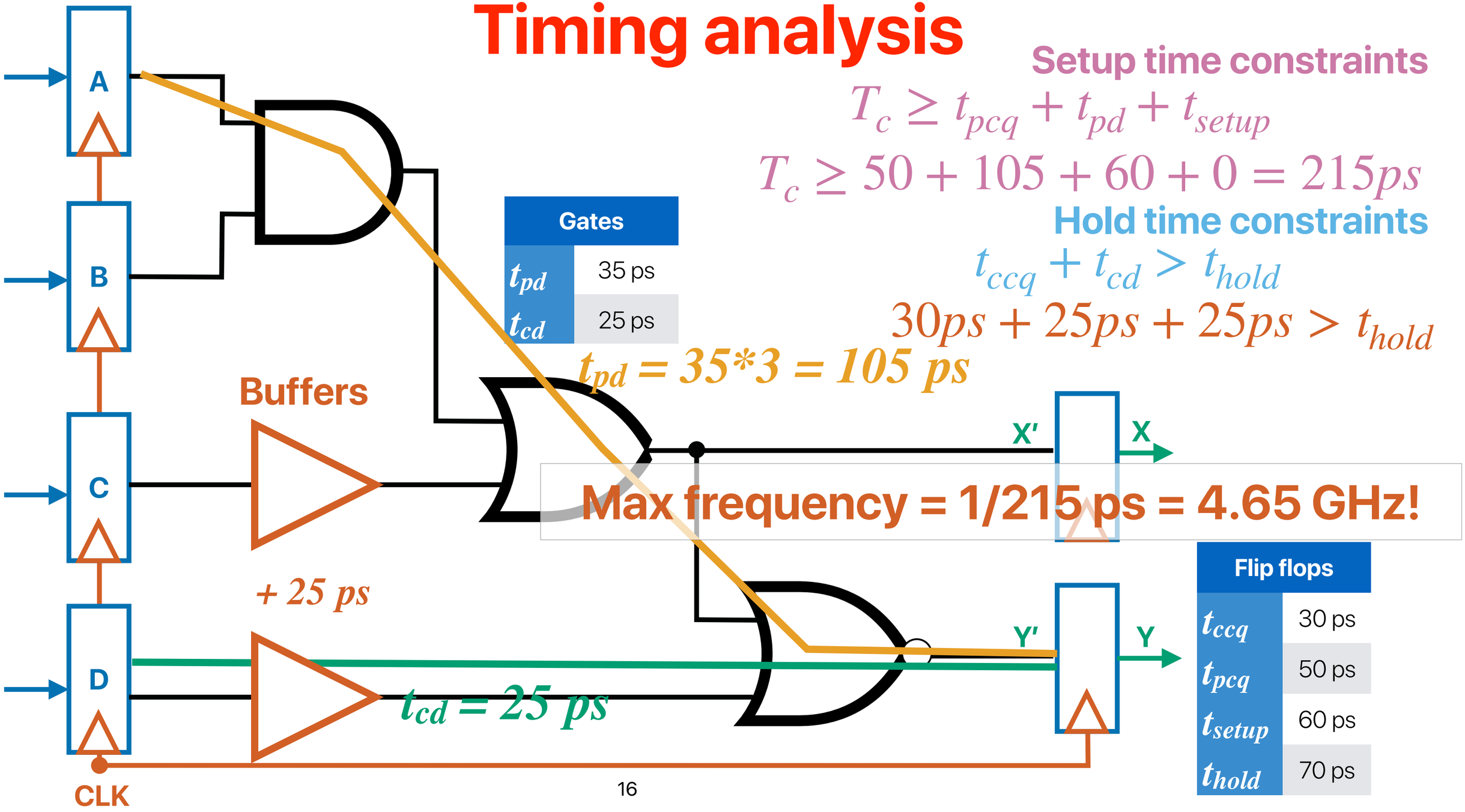
$$30ps + 25ps > t_{hold}$$

t_{hold} **= 70 ps!**

No!!!

Flip flops	
t_{ccq}	30 ps
t_{pcq}	50 ps
t_{setup}	60 ps
t_{hold}	70 ps

Timing analysis



Gates	
t_{pd}	35 ps
t_{cd}	25 ps

Setup time constraints

$$T_c \geq t_{pcq} + t_{pd} + t_{setup}$$

$$T_c \geq 50 + 105 + 60 + 0 = 215ps$$

Hold time constraints

$$t_{ccq} + t_{cd} > t_{hold}$$

$$30ps + 25ps + 25ps > t_{hold}$$

Max frequency = 1/215 ps = 4.65 GHz!

Flip flops	
t_{ccq}	30 ps
t_{pcq}	50 ps
t_{setup}	60 ps
t_{hold}	70 ps

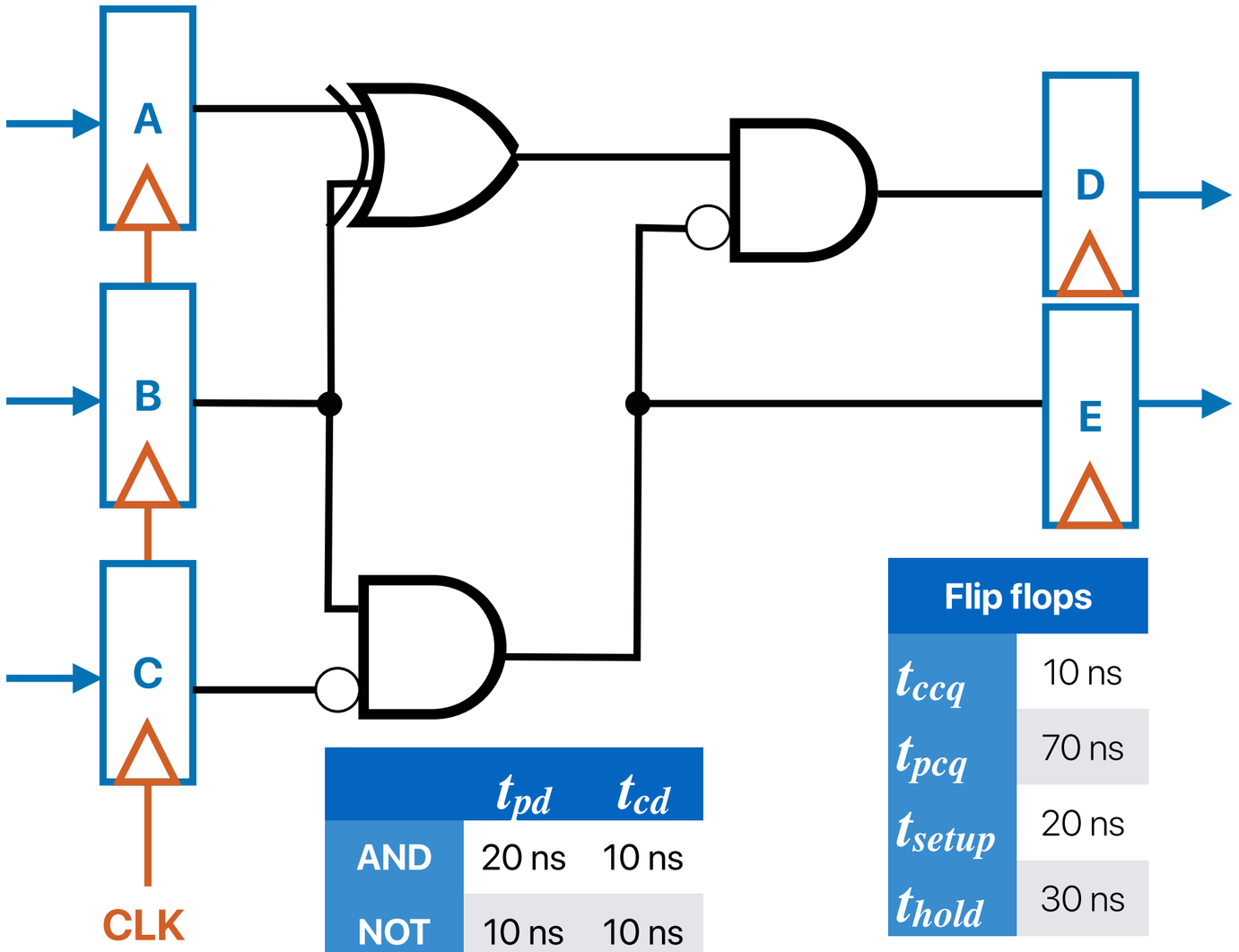
Poll close in 1:30

Example: timing constraints

- What's the maximum frequency?
  - A. 1 / 110ns
  - B. 1 / 220ns
  - C. 1 / 200ns
  - D. 1 / 180ns
  - E. None of the above

$$T_c \geq t_{pcq} + t_{pd} + t_{setup}$$

$$t_{ccq} + t_{cd} > t_{hold}$$



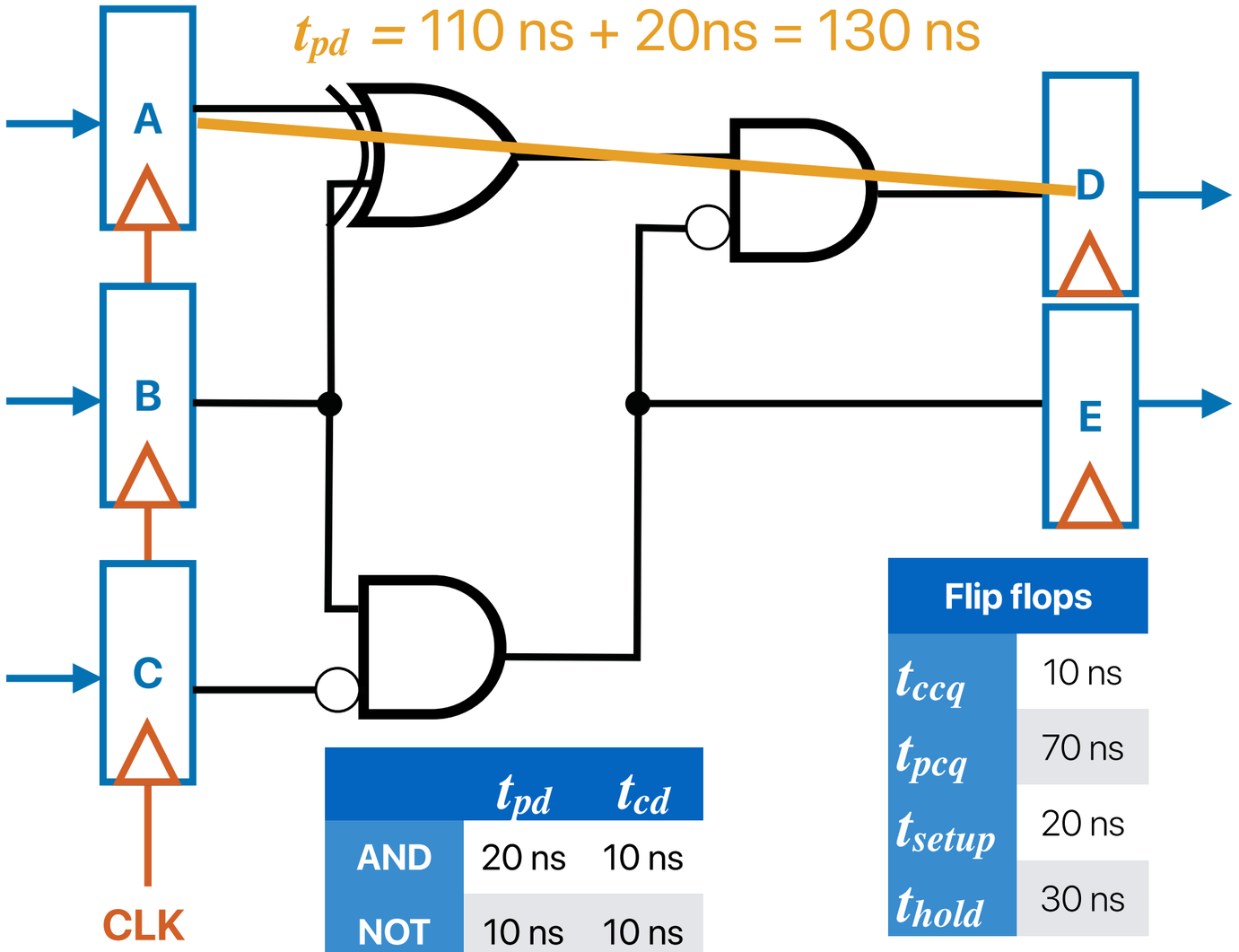
	t_{pd}	t_{cd}
AND	20 ns	10 ns
NOT	10 ns	10 ns
XOR	110 ns	50 ns

Flip flops	
t_{ccq}	10 ns
t_{pcq}	70 ns
t_{setup}	20 ns
t_{hold}	30 ns

Example: timing constraints

- What's the maximum frequency?
  - A. 1 / 110ns
  - B. 1 / 220ns
  - C. 1 / 200ns
  - D. 1 / 180ns
  - E. None of the above

$$T_c \geq t_{pcq} + t_{pd} + t_{setup}$$
$$T_c \geq 70ns + 130ns + 20ns + 0$$

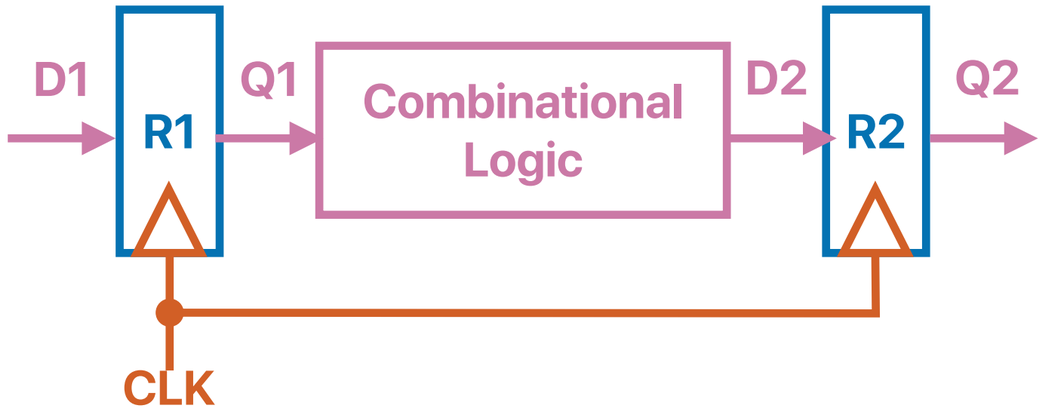


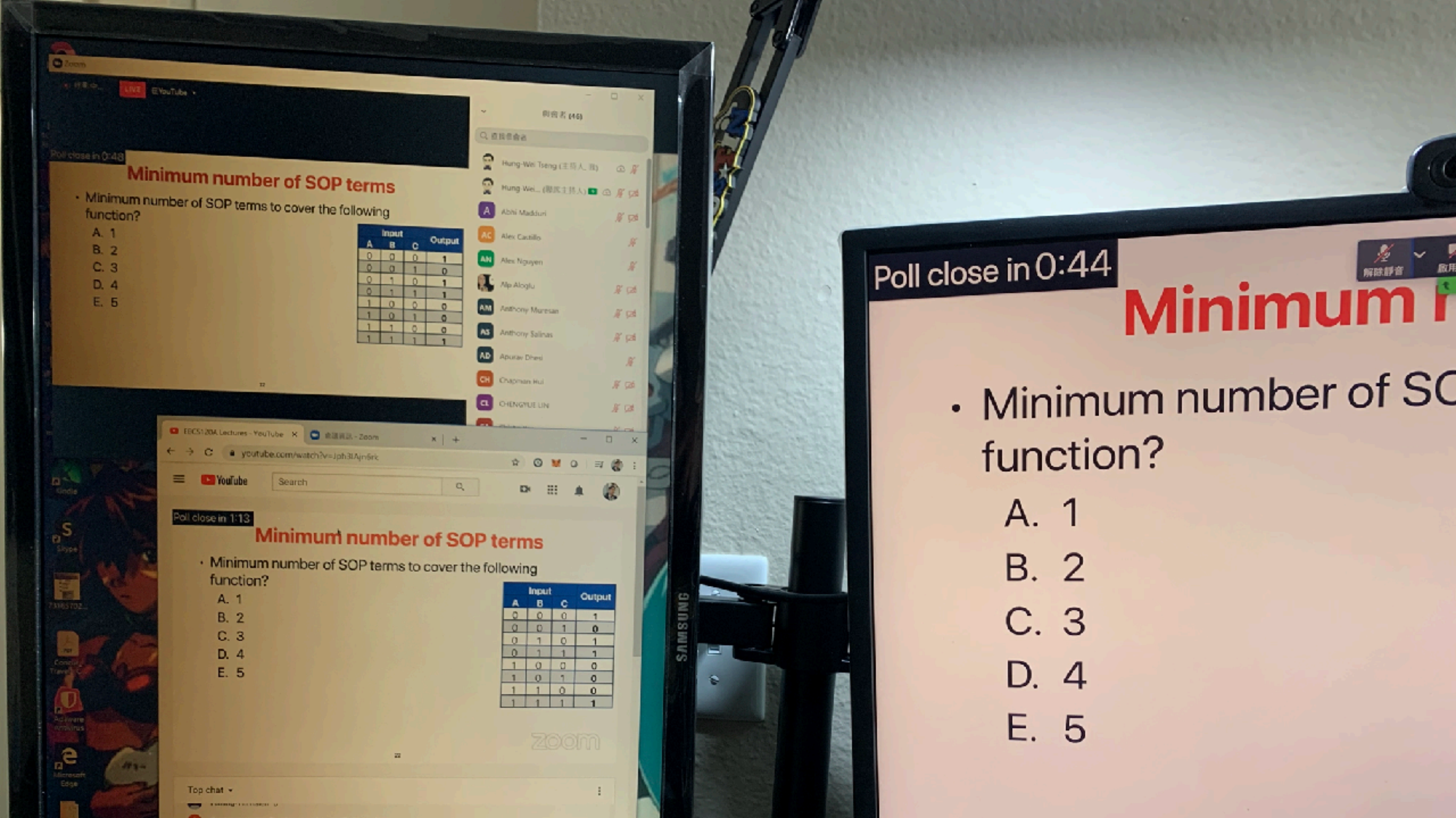
	t_{pd}	t_{cd}
AND	20 ns	10 ns
NOT	10 ns	10 ns
XOR	110 ns	50 ns

Flip flops	
t_{ccq}	10 ns
t_{pcq}	70 ns
t_{setup}	20 ns
t_{hold}	30 ns

FF Timing Parameters

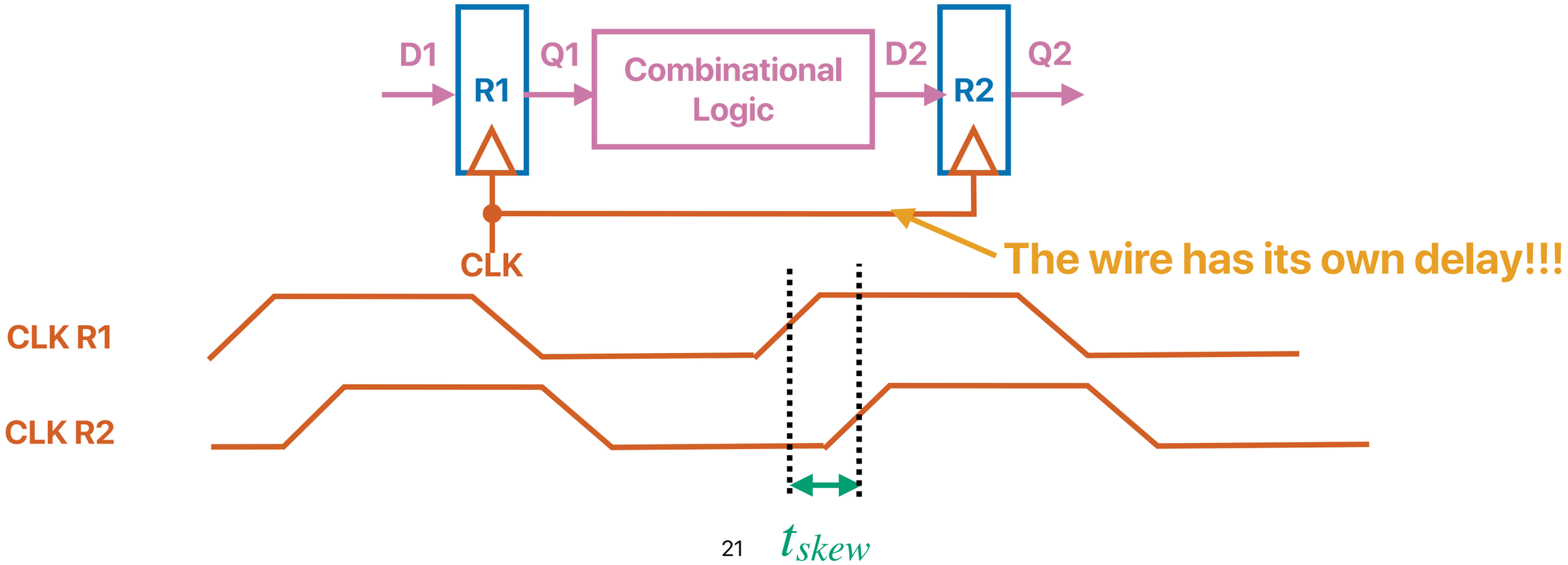
- Once a flip flop has been built, its timing characteristics stay fixed: t_{setup} , t_{hold} , t_{ccq} , t_{pcq}
- What about the clock? Does the clock edge arrive at the same time to all the D-FFs on the chip?




Poll close in 0:44
Minimum number of SOP terms
• Minimum number of SOP terms to cover the following function?
A. 1
B. 2
C. 3
D. 4
E. 5
Poll close in 0:48
Minimum number of SOP terms
• Minimum number of SOP terms to cover the following function?
A. 1
B. 2
C. 3
D. 4
E. 5
Input
A B C
Output
Poll close in 1:13
Minimum number of SOP terms
• Minimum number of SOP terms to cover the following function?
A. 1
B. 2
C. 3
D. 4
E. 5
Top chat
SAMSUNG

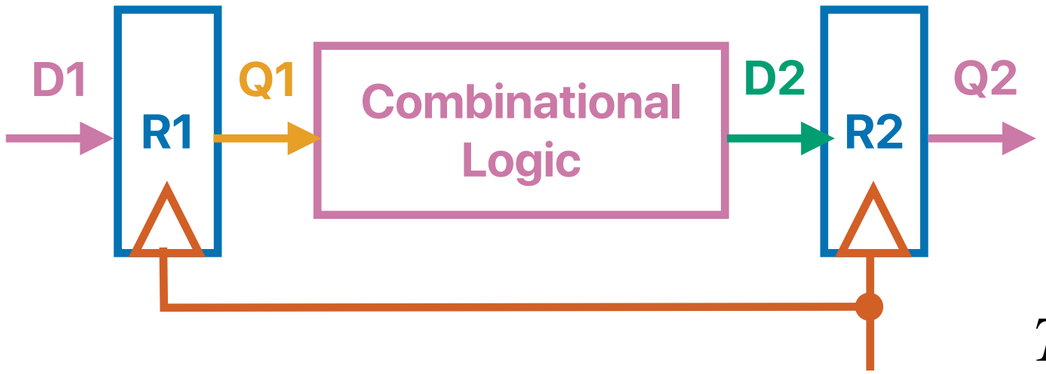
Clock Skew

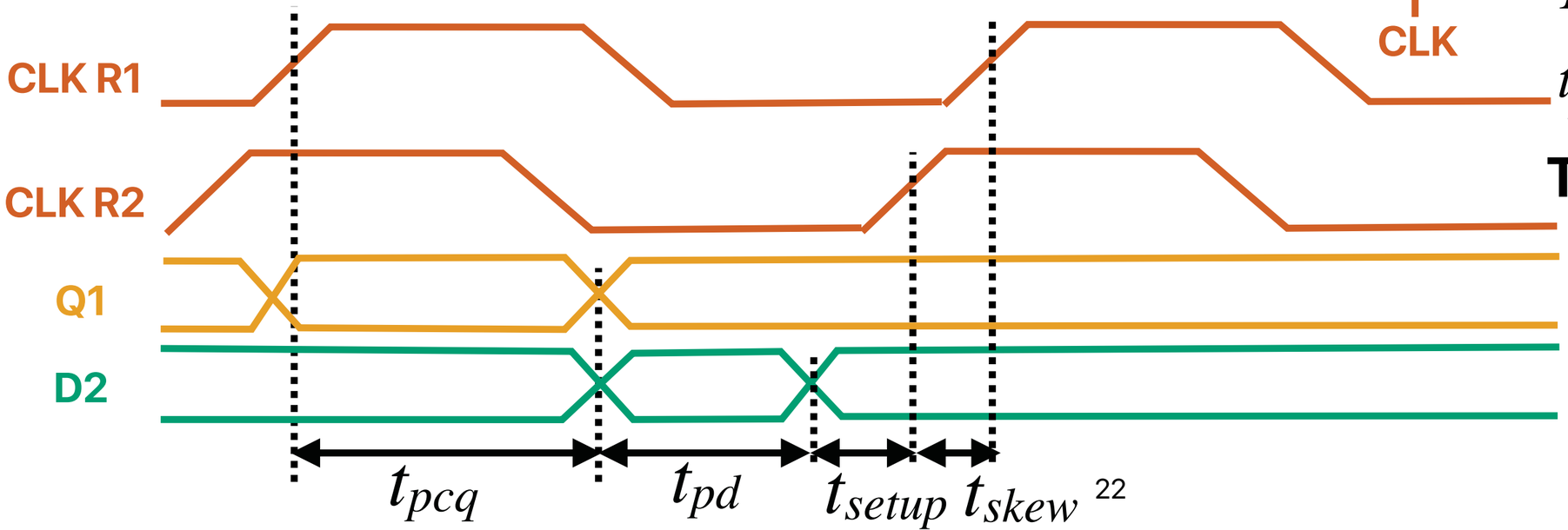
- The clock doesn't arrive at all registers at the same time
- **Skew**: difference between the two clock edges
- Perform the **worst case analysis**



Setup Time Constraint with Skew

- In the worst case, CLK2 is earlier than CLK1
- t_{pcq} is max delay through FF, t_{pd} is max delay through logic





$$T_c \geq t_{pcq} + t_{pd} + t_{setup} + t_{skew}$$

$$t_{pd} \leq T_c - (t_{setup} + t_{pcq} + t_{skew})$$

The larger the design, the longer the t_{skew}

Power consumption

Poll close in 1:30

Power & Energy

- Regarding power and energy, how many of the following statements are correct?
  - ① Lowering the power consumption helps extending the battery life
  - ② Lowering the power consumption helps reducing the heat generation
  - ③ Lowering the energy consumption helps reducing the electricity bill
  - ④ A CPU with 10% utilization can still consume 33% of the peak power

  A. 0
  B. 1
  C. 2
  D. 3
  E. 4

Power v.s. Energy

- Power is the direct contributor of “heat”
  - Packaging of the chip
  - Heat dissipation cost
  - Power = $P_{Dynamic} + P_{static}$
- Energy = P * ET
  - The electricity bill and battery life is related to energy!
  - Lower power does not necessary means better battery life if the processor slow down the application too much

Power & Energy

- Regarding power and energy, how many of the following statements are correct?
  - ① Lowering the power consumption helps extending the battery life
  - ② Lowering the power consumption helps reducing the heat generation
  - ③ Lowering the energy consumption helps reducing the electricity bill
  - ④ A CPU with 10% utilization can still consume 33% of the peak power
  - A. 0
  - B. 1
  - C. 2
  - D. 3
  - E. 4

Dynamic/Active Power

- The power consumption due to the switching of transistor states
- Dynamic power per transistor

  $P_{dynamic} \sim \alpha \times C \times V^2 \times f \times N$

  - α: average switches per cycle
  - C: capacitance
  - V: voltage
  - f: frequency, usually linear with V
  - N: the number of transistors

Static/Leakage Power

- The power consumption due to leakage — transistors do not turn all the way off during no operation
- Becomes the **dominant** factor in the most advanced process technologies.

$$P_{leakage} \sim N \times V \times e^{-V_t}$$

  - N: number of transistors
  - V: voltage
  - V_t: threshold voltage where transistor conducts (begins to switch)

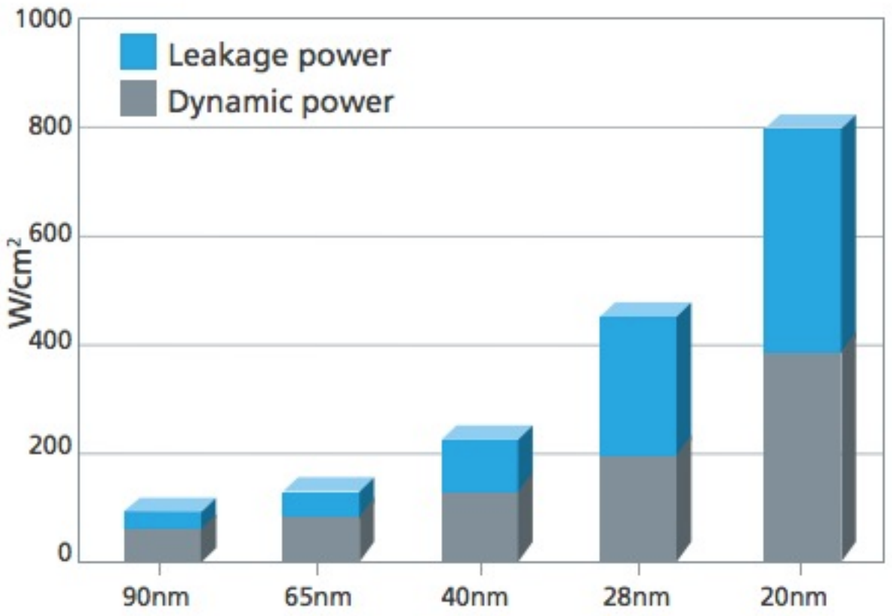


Figure 1: Leakage power becomes a growing problem as demands for more performance and functionality drive chipmakers to nanometer-scale process nodes (Source: IBS).

Power & Energy

- Regarding power and energy, how many of the following statements are correct?
  - ① Lowering the power consumption helps extending the battery life
  - ② Lowering the power consumption helps reducing the heat generation
  - ③ Lowering the energy consumption helps reducing the electricity bill
  - ④ A CPU with 10% utilization can still consume 33% of the peak power
  - A. 0
  - B. 1
  - C. 2
  - D. 3
  - E. 4

Dynamic/Active Power

- The power consumption due to the switching of transistor states
- Dynamic power per transistor

  $P_{dynamic} \sim \alpha \times C \times V^2 \times f \times N$

  - α: average switches per cycle
  - C: capacitance
  - V: voltage
  - f: frequency, usually linear with V
  - N: the number of transistors

Dennardian Broken

- Given a scaling factor S

Parameter	Relation	Classical Scaling	Leakage Limited
Power Budget		1	1
Chip Size		1	1
Vdd (Supply Voltage)		1/S	1
Vt (Threshold Voltage)	1/S	1/S	1
tex (oxide thickness)		1/S	1/S
W, L (transistor dimensions)		1/S	1/S
Cgate (gate capacitance)	WL/tox	1/S	1/S
Isat (saturation current)	WVdd/tox	1/S	1
F (device frequency)	Isat/(CgateVdd)	S	S
D (Device/Area)	1/(WL)	S^2	S^2
p (device power)	IsatVdd	$1/S^2$	1
P (chip power)	Dp	1	S^2
U (utilization)	1/P	1	$1/S^2$

Poll close in 1:30

What happens if power doesn't scale with process technologies?

- If we are able to cram more transistors within the same chip area (Moore's law continues), but the power consumption per transistor remains the same. Right now, if put more transistors in the same area because the technology allows us to. How many of the following statements are true?
  - ① The power consumption per chip will increase
  - ② The power density of the chip will increase
  - ③ Given the same power budget, we may not able to power on all chip area if we maintain the same clock rate
  - ④ Given the same power budget, we may have to lower the clock rate of circuits to power on all chip area

  A. 0
  B. 1
  C. 2
  D. 3
  E. 4

Power consumption

Chip

1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1

=49W

Dennardian Scaling

Chip

0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5

=50W

Dennardian Broken

Chip

1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1

=100W!

Power density

Chip

1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1

$$= \frac{49W}{Chip\ Area}$$

Dennardian Scaling

Chip

0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5

$$= \frac{50W}{Chip\ Area}$$

Dennardian Broken

Chip

1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1

$$= \frac{100W}{Chip\ Area}$$

Power density

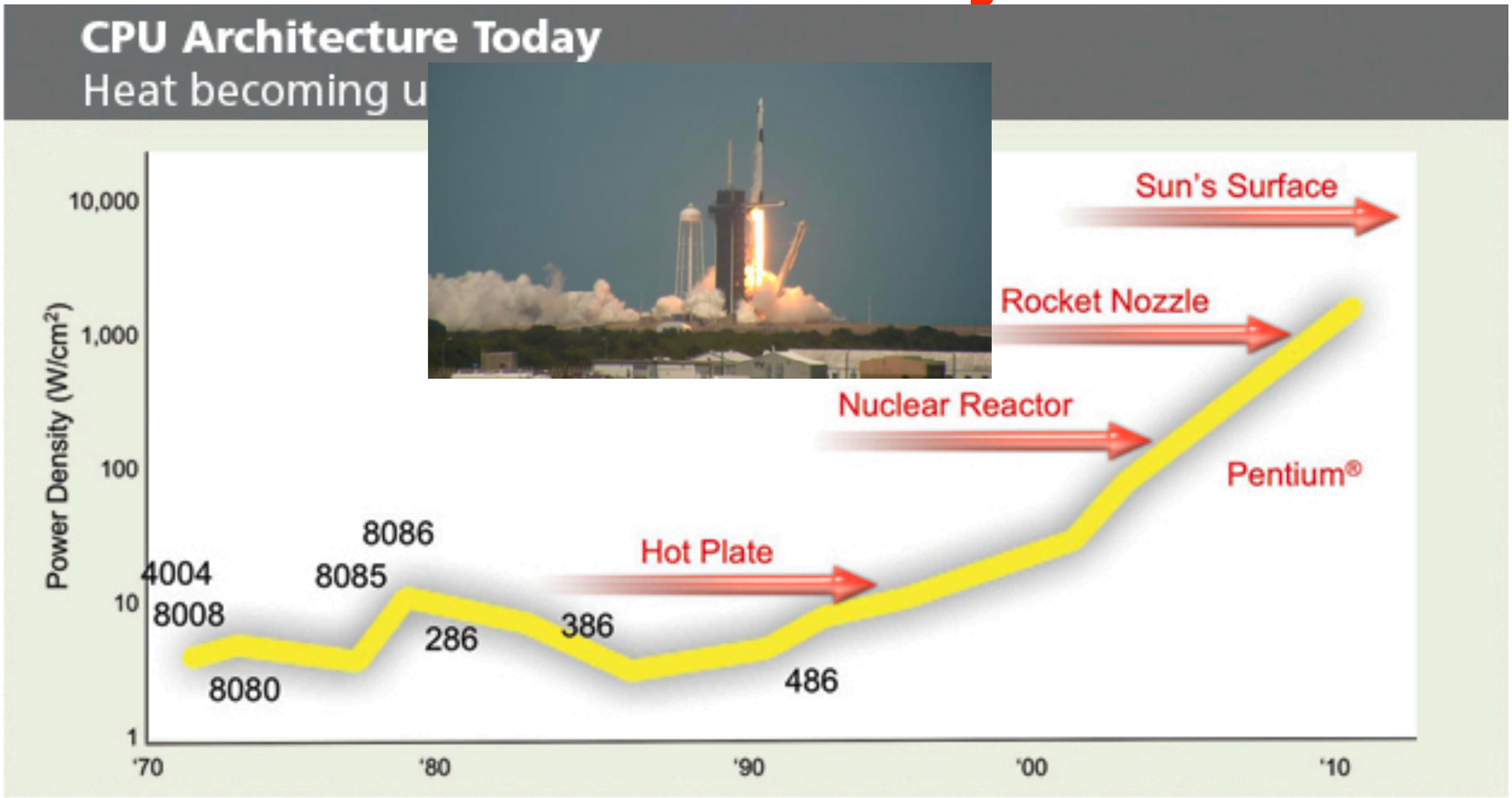


Figure 1. In CPU architecture today, heat is becoming an unmanageable problem. (Courtesy of Pat Gelsinger, Intel Developer Forum, Spring 2004)

https://www.cadalyst.com/hardware/workstation-performance-tomorrow039s-possibilities-viewpoint-column-6351

What happens if power doesn’t scale with process technologies?

- If we are able to cram more transistors within the same chip area (Moore’s law continues), but the power consumption per transistor remains the same. Right now, if put more transistors in the same area because the technology allows us to. How many of the following statements are true?
  - ✓ ① The power consumption per chip will increase
  - ✓ ② The power density of the chip will increase
  - ③ Given the same power budget, we may not able to power on all chip area if we maintain the same clock rate
  - ④ Given the same power budget, we may have to lower the clock rate of circuits to power on all chip area

  A. 0
  B. 1
  C. 2
  D. 3
  E. 4

Power consumption to light on all transistors

If we can only cool down 50W in the same area —

Chip

1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1

=49W

Dennardian Scaling

Chip

0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5

=50W

Dennardian Broken

Chip

1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1

On ~ 50W

Off ~ 0W

Dark!

=100W!

Dark silicon

- Your power consumption goes up as the number of transistors goes up
- Even Moore's Law allows us to put more transistors within the same area —we cannot use them all simultaneously!
- We have no choice to not activate all transistors at the same time!

What happens if power doesn't scale with process technologies?

- If we are able to cram more transistors within the same chip area (Moore's law continues), but the power consumption per transistor remains the same. Right now, if put more transistors in the same area because the technology allows us to. How many of the following statements are true?
  - ① The power consumption per chip will increase
  - ② The power density of the chip will increase
  - ③ Given the same power budget, we may not able to power on all chip area if we maintain the same clock rate
  - ④ Given the same power budget, we may have to lower the clock rate of circuits to power on all chip area

  A. 0
  B. 1
  C. 2
  D. 3
  E. 4

Dynamic/Active Power

- The power consumption due to the switching of transistor states
- Dynamic power per transistor

$$P_{dynamic} \sim \alpha \times C \times V^2 \times f \times N$$

  - α: average switches per cycle
  - C: capacitance
  - V: voltage
  - f: frequency, usually linear with V
  - N: the number of transistors

What happens if power doesn't scale with process technologies?

- If we are able to cram more transistors within the same chip area (Moore's law continues), but the power consumption per transistor remains the same. Right now, if put more transistors in the same area because the technology allows us to. How many of the following statements are true?
  - ① The power consumption per chip will increase
  - ② The power density of the chip will increase
  - ③ Given the same power budget, we may not able to power on all chip area if we maintain the same clock rate
  - ④ Given the same power budget, we may have to lower the clock rate of circuits to power on all chip area

  A. 0
  B. 1
  C. 2
  D. 3
  E. 4

Solutions/trends in dark silicon era

Trends in the Dark Silicon Era

- Aggressive dynamic voltage/frequency scaling
- Throughout oriented — slower, but more
- Just let it dark — activate part of circuits, but not all
- From general-purpose to domain-specific — ASIC

Aggressive dynamic frequency scaling

Dynamic/Active Power

- The power consumption due to the switching of transistor states
- Dynamic power per transistor

$$P_{dynamic} \sim \alpha \times C \times V^2 \times f \times N$$

  - α: average switches per cycle
  - C: capacitance
  - V: voltage
  - f: frequency, usually linear with V
  - N: the number of transistors

Frequency varies per core

Products Solutions Support

intel

	Intel® Xeon® Processor E7-8890 v4	Intel® Xeon® Processor E7-8893 v4	Intel® Xeon® Processor E7-8880 v4
Status	Launched	Launched	Launched
Launch Date	Q2'16	Q2'16	Q2'16
Lithography	14 nm	14 nm	14 nm
Performance			
# of Cores	24	4	22
# of Threads	48	8	44
Processor Base Frequency	2.20 GHz	3.20 GHz	2.20 GHz
Max Turbo Frequency	3.40 GHz	3.50 GHz	3.30 GHz
Cache	60 MB	60 MB	55 MB
Bus Speed	9.6 GT/s	9.6 GT/s	9.6 GT/s
# of QPI Links	3	3	3
TDP	165 W	140 W	150 W

Demo

- You may use cat /proc/cpuinfo to see all the details of your processors
- You may add “| grep MHz” to see the frequencies of your cores
- Only very few of them are on the boosted frequency

Slower, but more

More cores per chip, slower per core

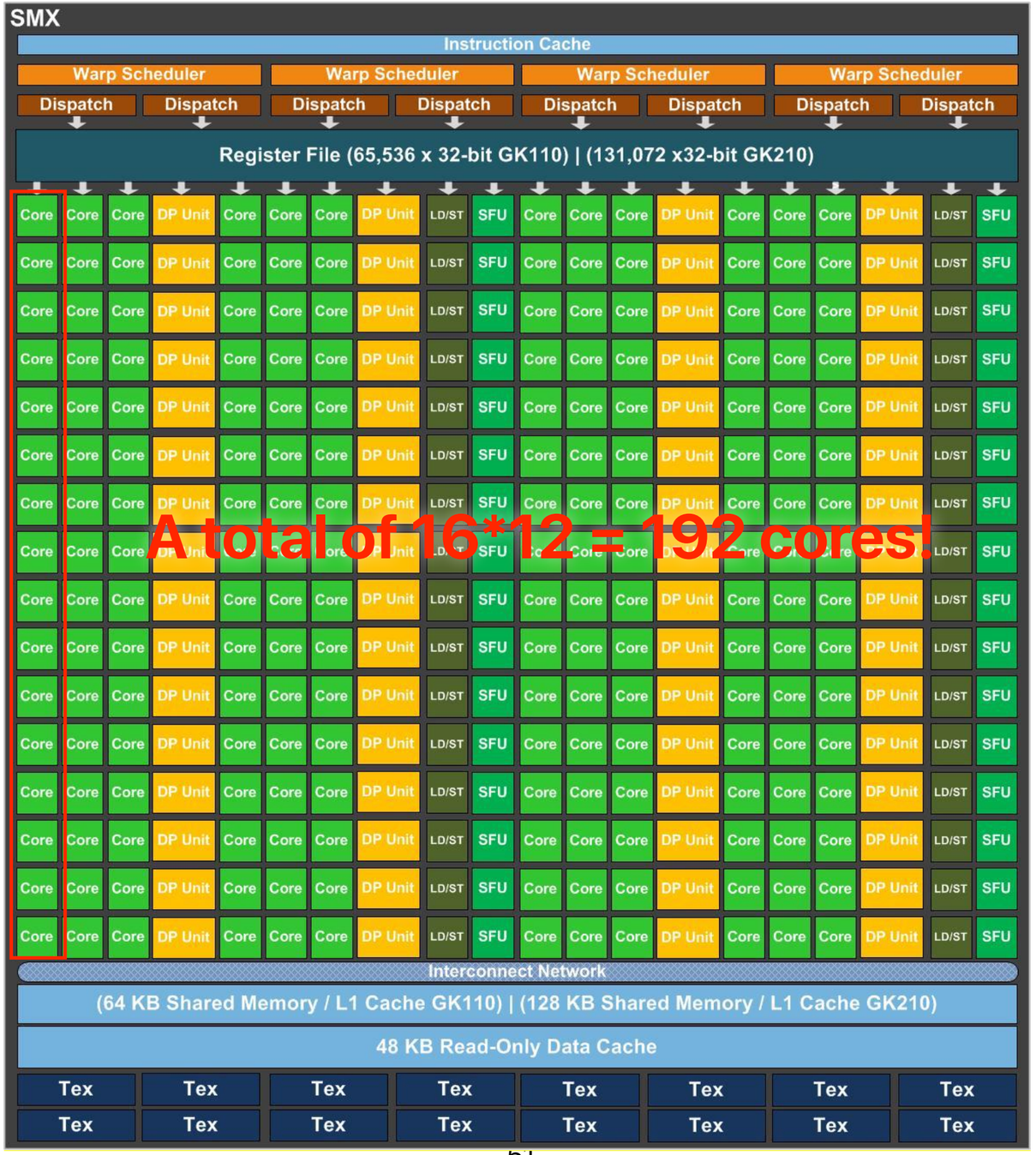
Products Solutions Support

intel

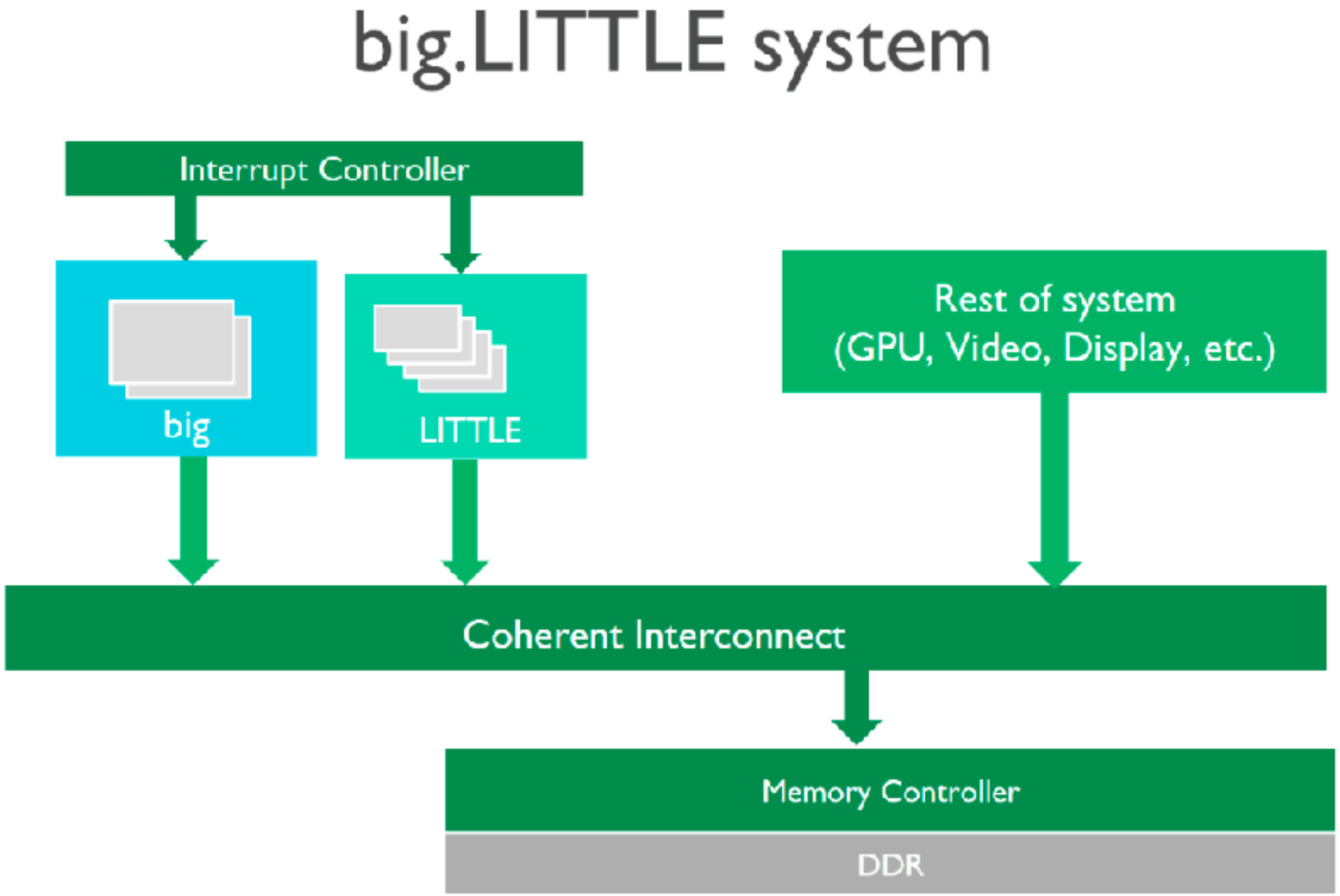
	Intel® Xeon® Processor E7-8890 v4	Intel® Xeon® Processor E7-8893 v4	Intel® Xeon® Processor E7-8880 v4
Status	Launched	Launched	Launched
Launch Date	Q2'16	Q2'16	Q2'16
Lithography	14 nm	14 nm	14 nm
Performance			
# of Cores	24	4	22
# of Threads	48	8	44
Processor Base Frequency	2.20 GHz	3.20 GHz	2.20 GHz
Max Turbo Frequency	3.40 GHz	3.50 GHz	3.30 GHz
Cache	60 MB	60 MB	55 MB
Bus Speed	9.6 GT/s	9.6 GT/s	9.6 GT/s
# of QPI Links	3	3	3
TDP	165 W	140 W	150 W

The rise of GPU

Each of these performs the same operation, but each of these is also a "thread"

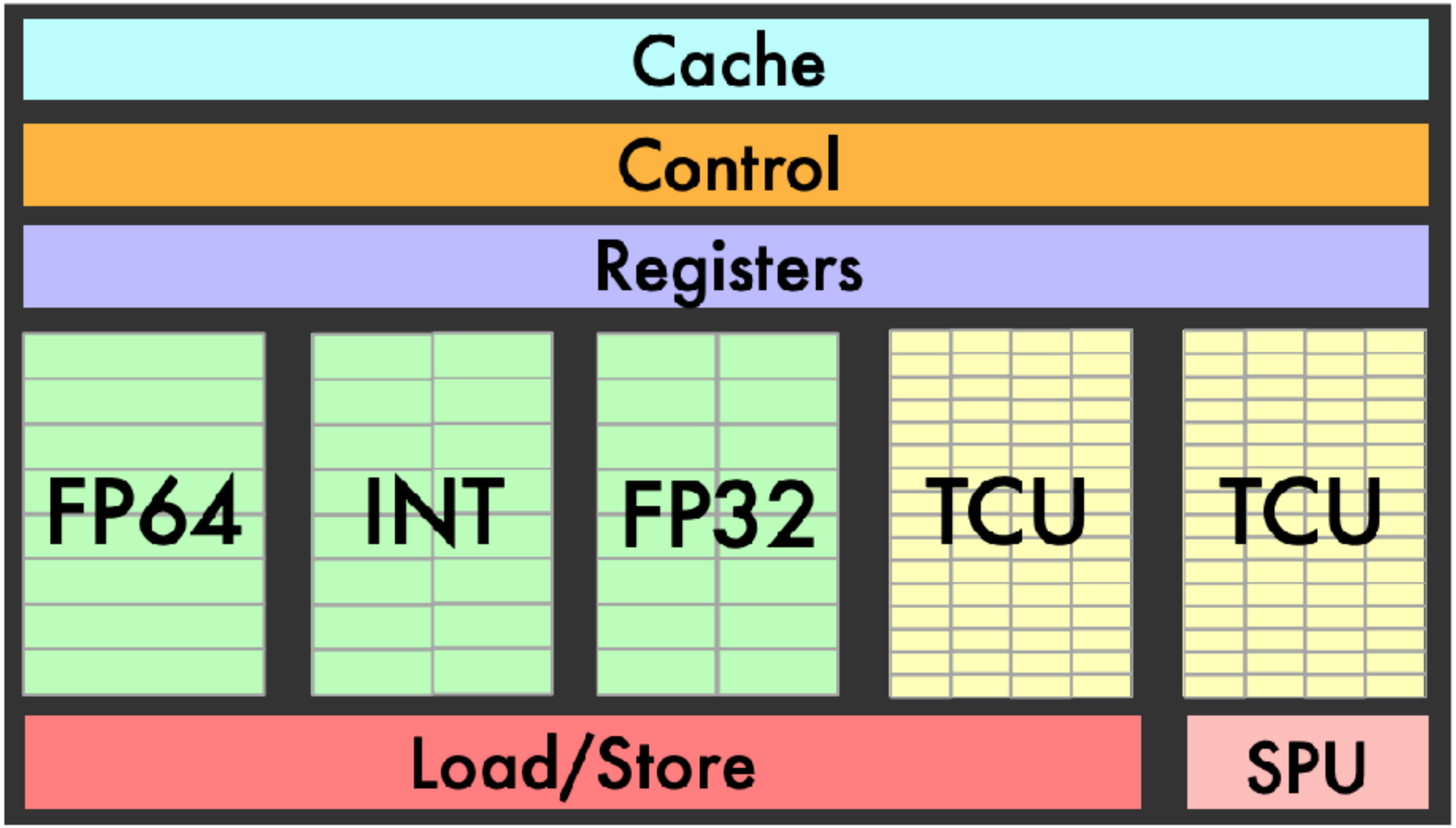


ARM's big.LITTLE architecture



Just let it dark

NVIDIA's Turing Architecture



Programming in Turing Architecture

Use tensor cores

```
cublasErrCheck(cublasSetMathMode(cublasHandle, CUBLAS_TENSOR_OP_MATH));
```

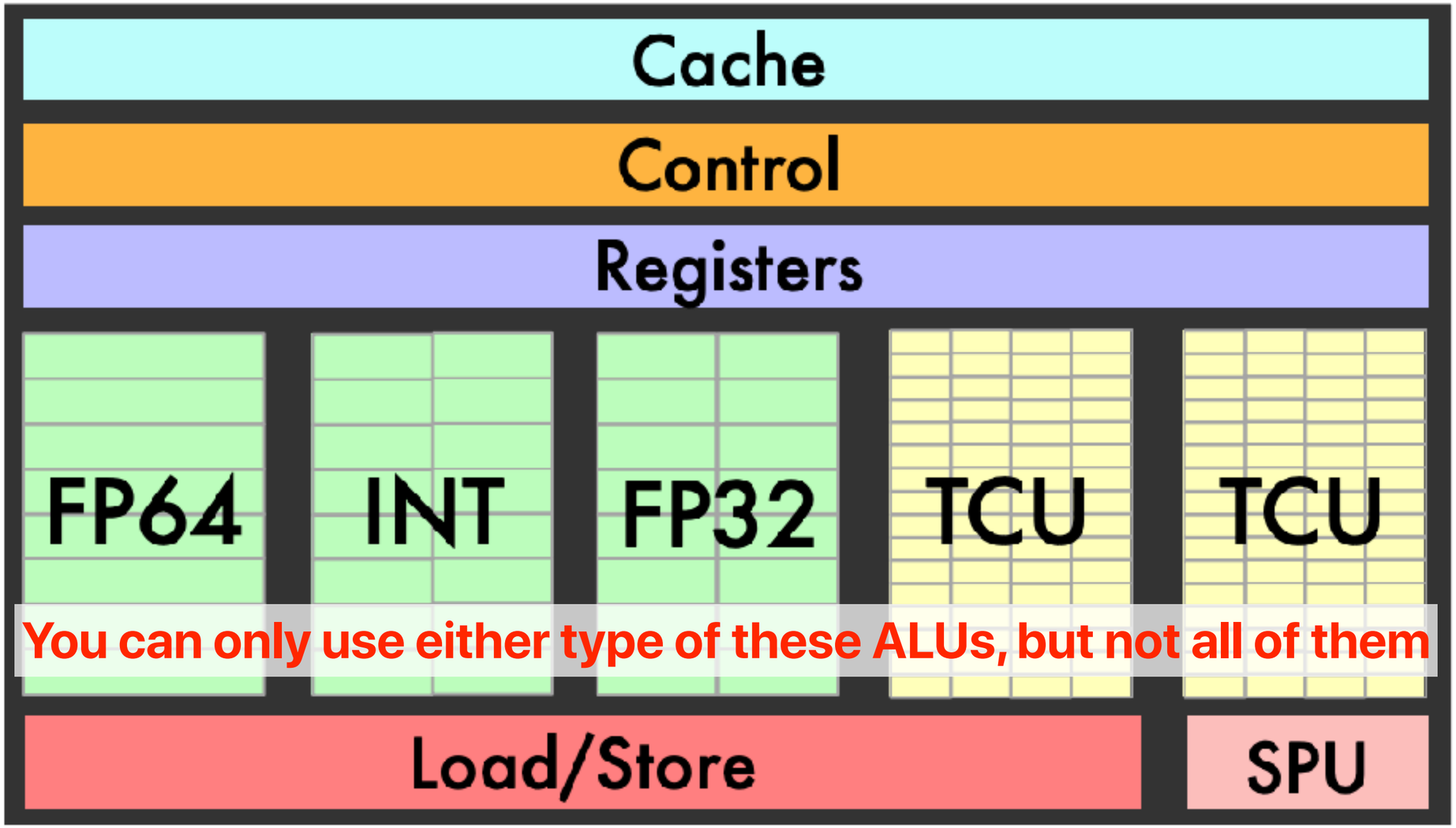
Make them 16-bit

```
convertFp32ToFp16 <<< (MATRIX_M * MATRIX_K + 255) / 256, 256 >>> (a_fp16, a_fp32,
MATRIX_M * MATRIX_K);
   convertFp32ToFp16 <<< (MATRIX_K * MATRIX_N + 255) / 256, 256 >>> (b_fp16, b_fp32,
MATRIX_K * MATRIX_N);
```

```
cublasErrCheck(cublasGemmEx(cublasHandle, CUBLAS_OP_N, CUBLAS_OP_N,
                 MATRIX_M, MATRIX_N, MATRIX_K,
                 &alpha,
                 a_fp16, CUDA_R_16F, MATRIX_M,
                 b_fp16, CUDA_R_16F, MATRIX_K,
                 &beta,
                 c_cublas, CUDA_R_32F, MATRIX_M,
                 CUDA_R_32F, CUBLAS_GEMM_DFALT_TENSOR_OP));
```

call Gemm

NVIDIA's Turing Architecture



The rise of ASICs

Real-time AI: Microsoft announces preview of Project Brainwave

AI & MACHINE LEARNING

An in-depth look at Google's first Tensor Processing Unit (TPU)

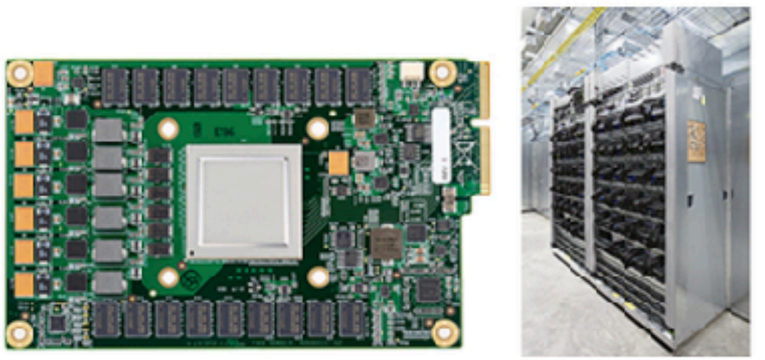
Kaz Sato
Staff Developer Advocate, Google Cloud

Cliff Young
Software Engineer, Google Brain

David Patterson
Distinguished Engineer, Google Brain

May 12, 2017

There's a common thread that connects Google services such as Google Search, Street View, Google Photos and Google Translate: they all use Google's Tensor Processing Unit, or TPU, to accelerate their neural network computations behind the scenes.



Google's first Tensor Processing Unit (TPU) on a printed circuit board (left); TPUs deployed in a Google datacenter (right)

There is no pure “software” or “hardware” design in the dark silicon era.
Everything needs to be hardware/software co-designed.

–Prof. Usagi

Announcement

- iEval — Capture your screenshot, submit through iLearn and you will receive a full credit assignment
- Assignment 6 due 6/4
- Lab 6 due **this Friday**
- Please fill out ABET survey through iLearn
- Final exam will be held during the campus scheduled period to avoid conflicts
  - Final review — 6/4 during the lecture, will also release the sample final
  - 6/11 11:30am — 2:59:59pm
  - About the same format as midterm, but longer
  - Will have a final review on 6/6 to help you prepare

Electrical
Computer Science
Engineering

120A

つづく