inst.eecs.berkeley.edu/~eecs251b

EECS251B : Advanced Digital Circuits and Systems

Lecture 15 – Latches

Borivoje Nikolić

The Startup Flipping Nvidia's Playbook on its Head

March 5, 2024, The Information. Taalas is developing the exact opposite of customizable chips: rigid chips which are each specialized for a different AI model, whether it's Meta Platforms' Llama models or Stable Diffusion. The Toronto-based startup, which was founded in August last year by former Nvidia and AMD veteran Ljubisa Bajic, raised $12 million in September and $38 million in February from Quiet Capital and Pierre Lamond, an advisor at Eclipse Ventures who was previously a general partner at Khosla Ventures and Sequoia Capital.

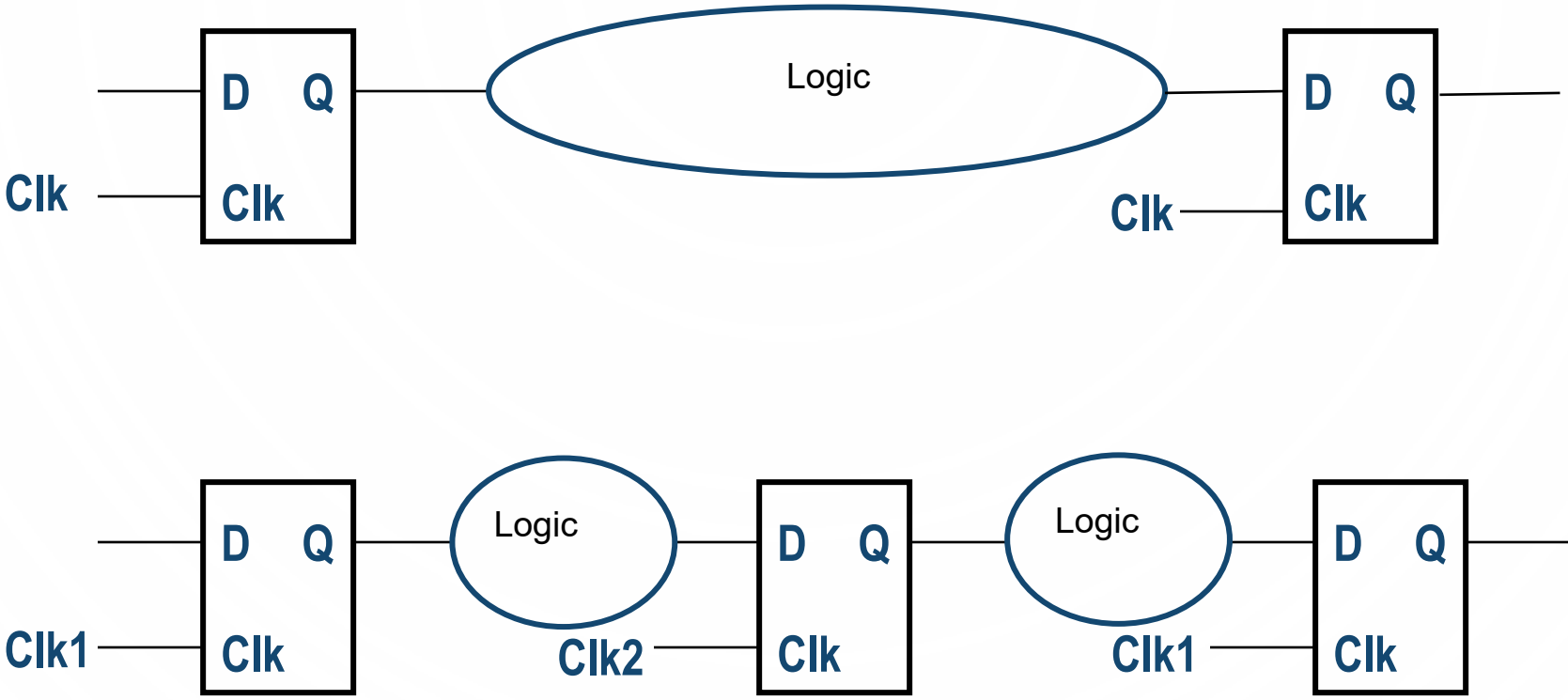
Taalas cofounder and CEO Ljubisa Bajic. Courtesy of Taalas.

Announcements

- Lab 5 still waiting on PDK correction
  - The newest fix brings it very close
- Start project phase 1
  - Spec doc due tomorrow
- Homework 2 due tomorrow
  - Quiz 2 on March 12
  - Homework 3 posted this week

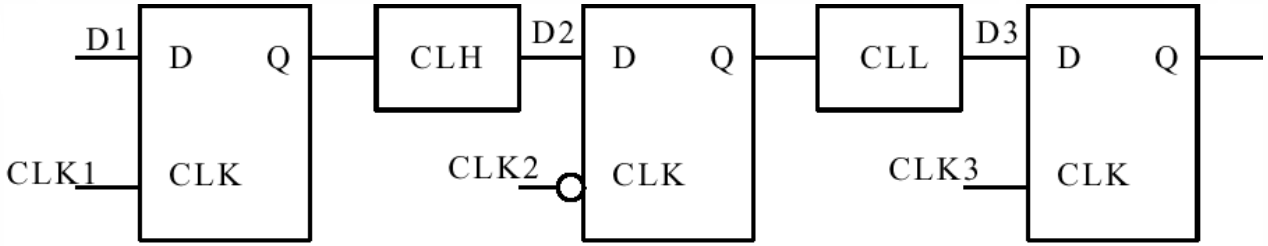
Latch Timing

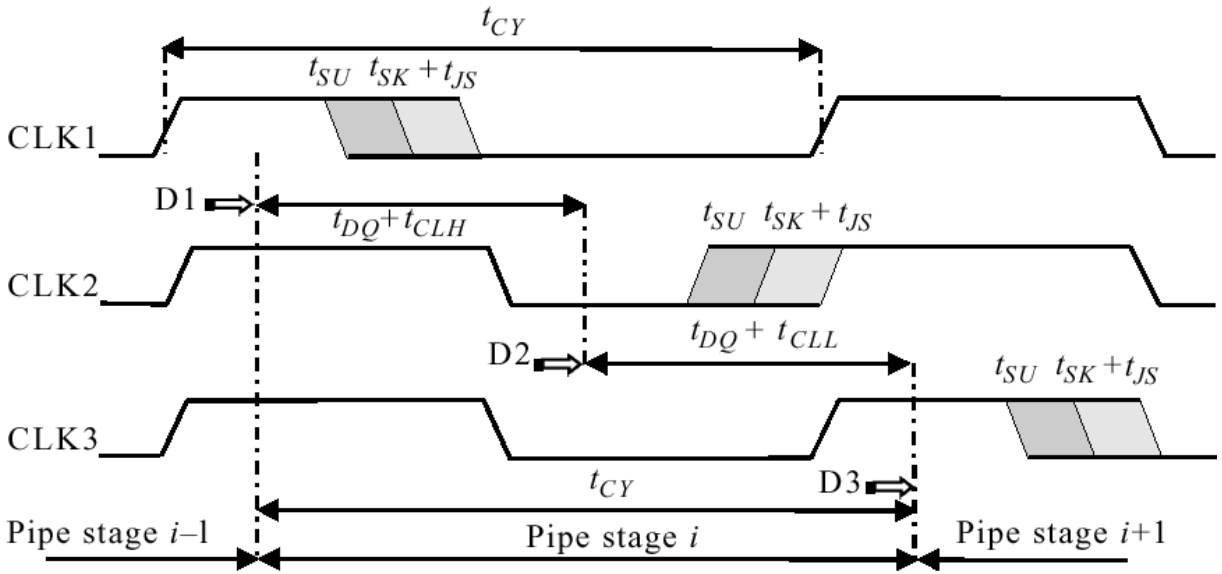
Latch Sequencing



Latch-Based Timing

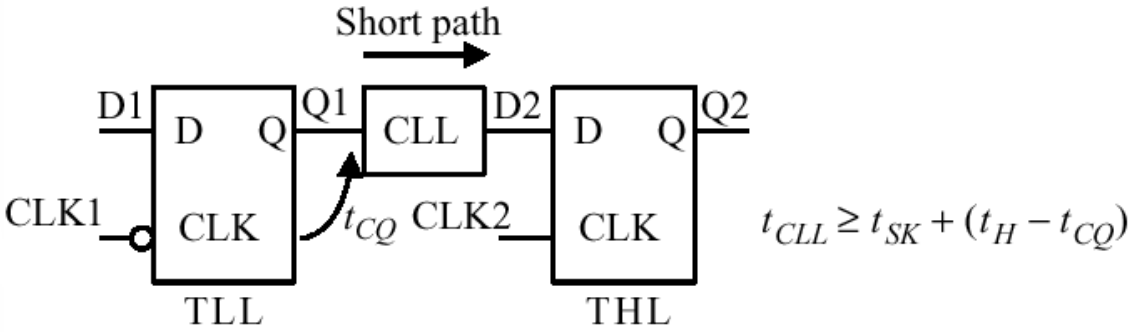
- Single-phase, two-latch



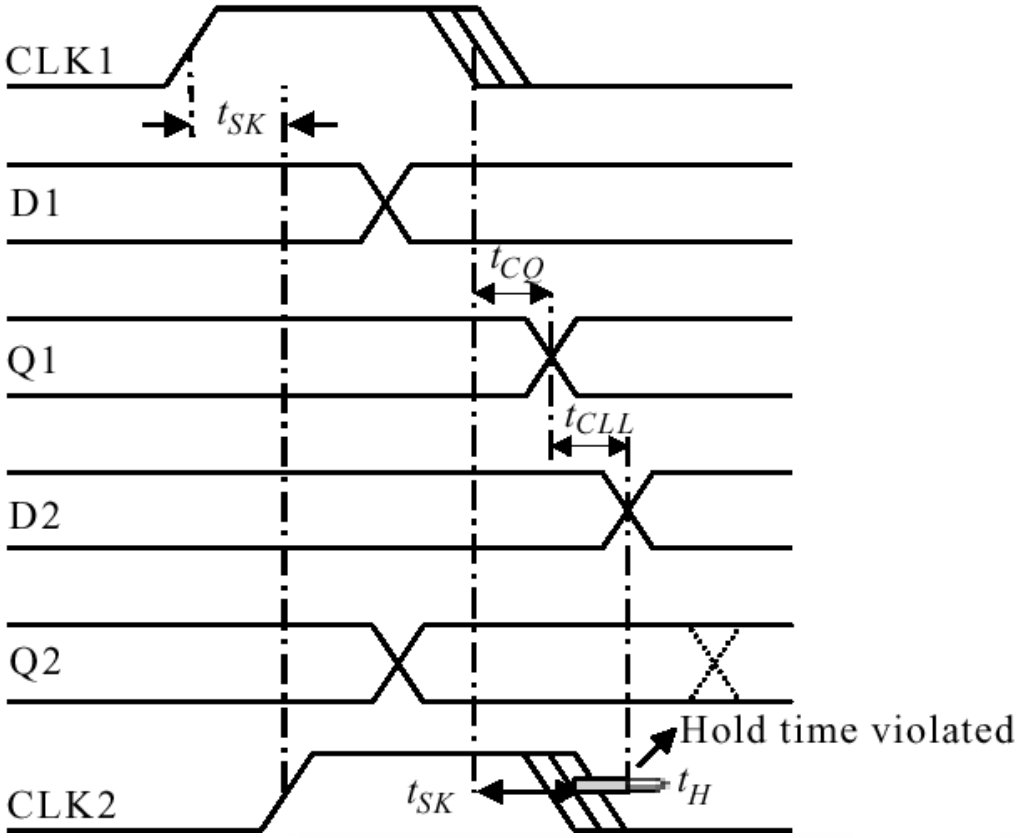


As long as transitions are within the assertion period of the latch, no impact of position of clock edges

Latch Design and Hold Times



$$t_{CLL} \geq t_{SK} + (t_H - t_{CQ})$$

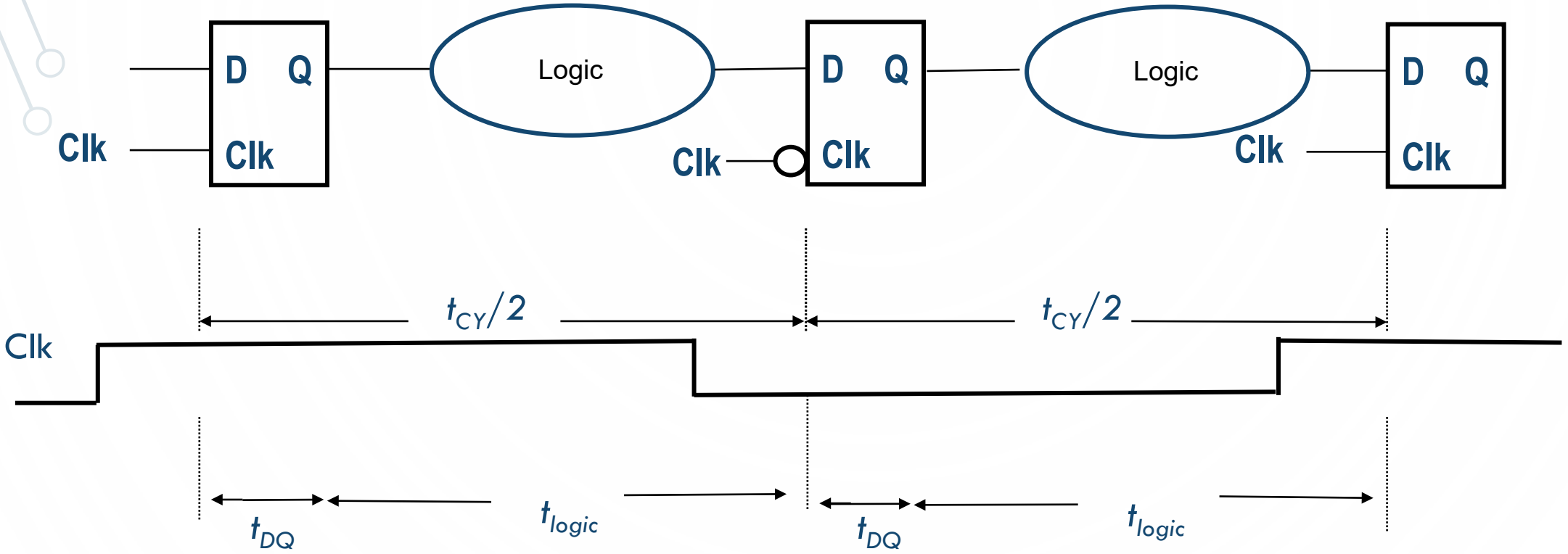


Soft-Edge Properties of Latches

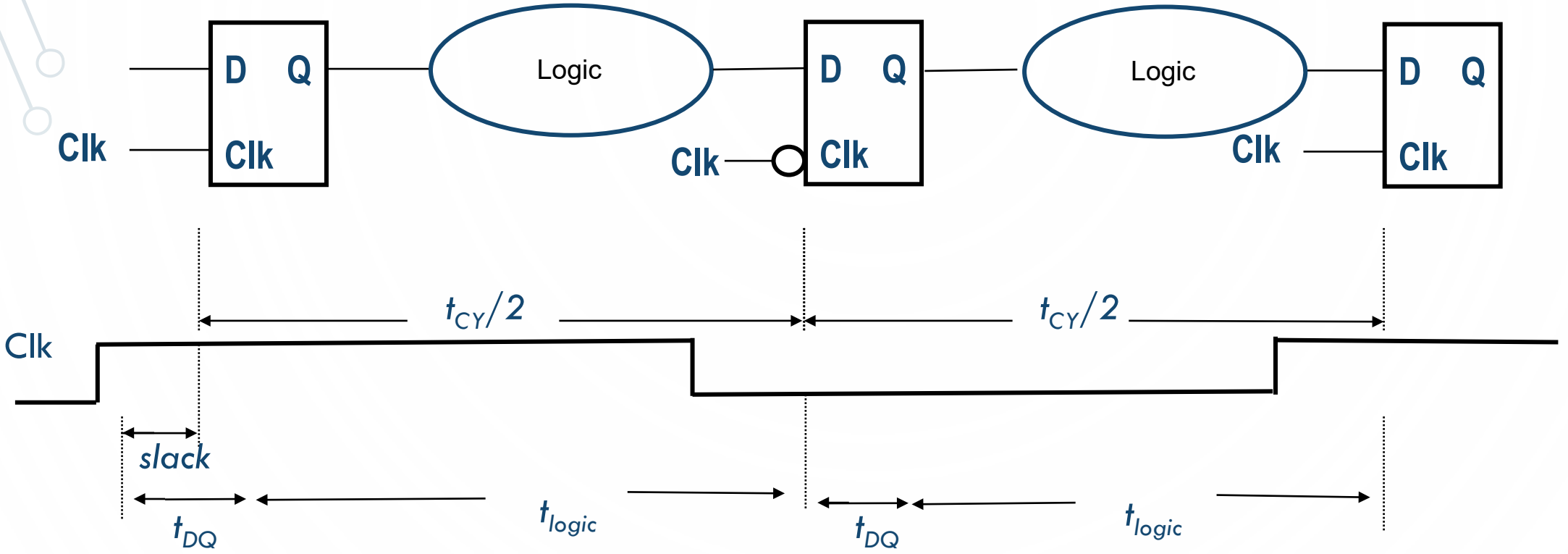
- Slack passing – logical partition uses left over time (slack) from the *previous* partition
- Time borrowing – logical partition utilizes a portion of time allotted to the *next* partition
- Makes most impact in unbalanced pipelines

Bernstein et al, Chapter 8, Chandrakasan, Chap 11 (by Partovi)

Slack Passing and Time Borrowing

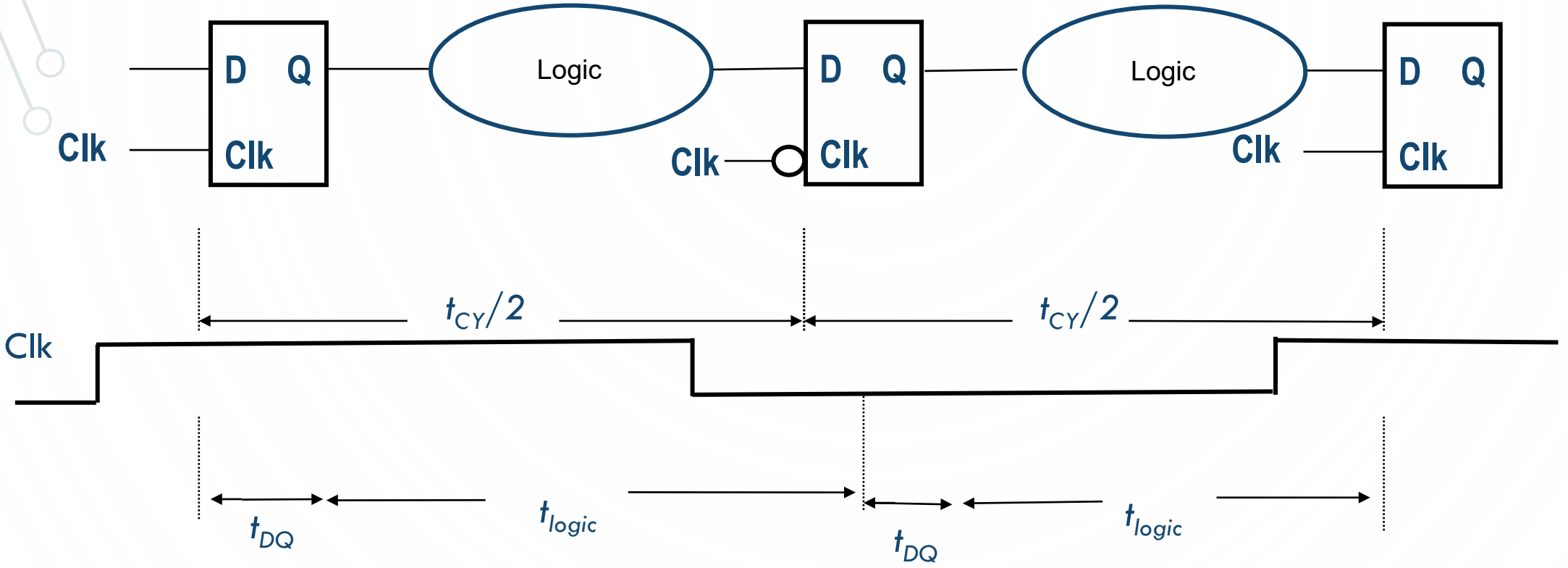


Slack Passing and Time Borrowing



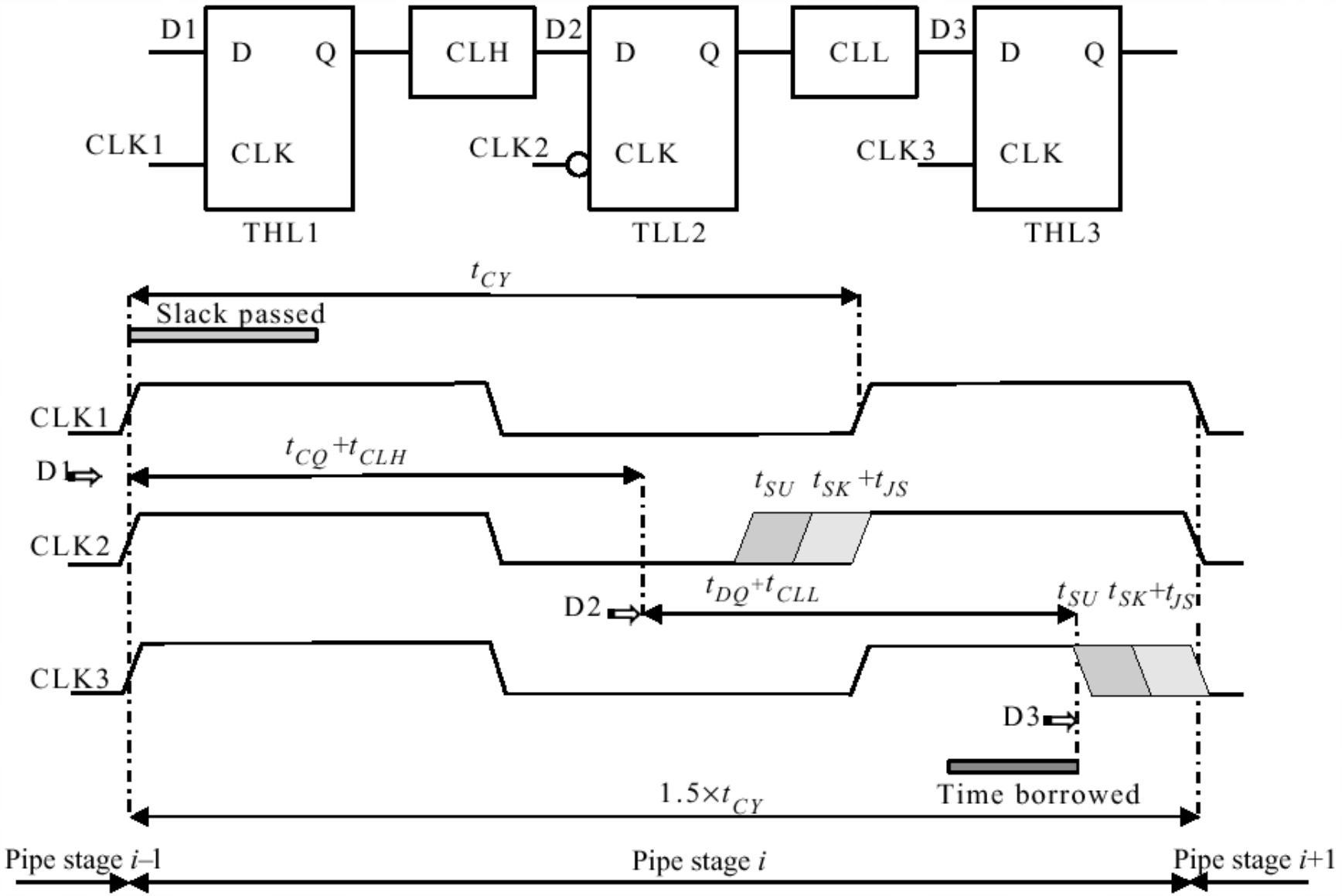
- Slack passed

Slack Passing and Time Borrowing



- Time borrowed

Slack-Passing and Cycle Borrowing



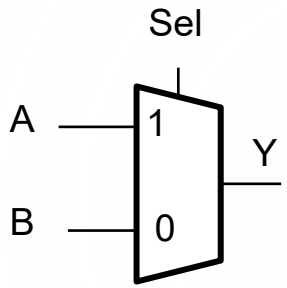
For N stage pipeline, overall logic delay should be < N Tcl

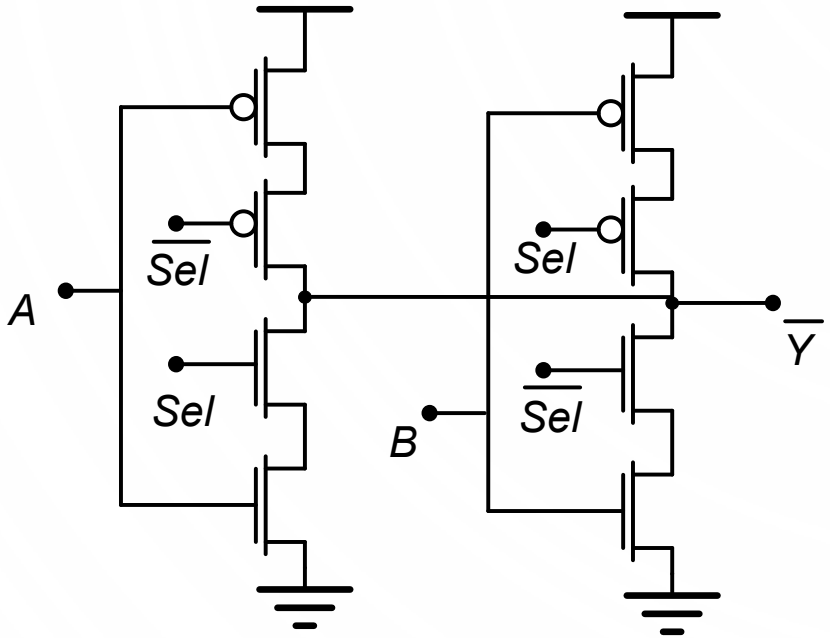
Design for Performance

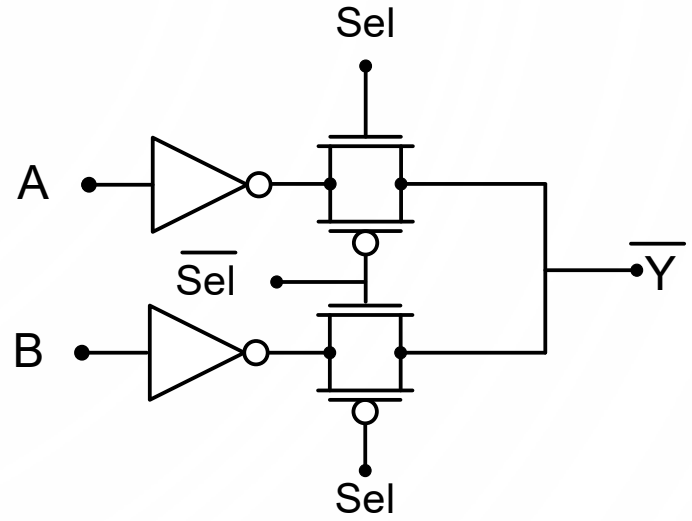
Latch Design

Review: MUX

- 2-input MUX

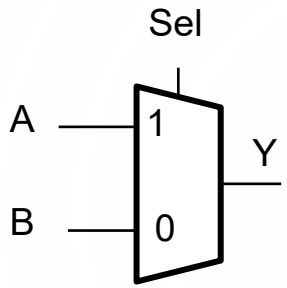


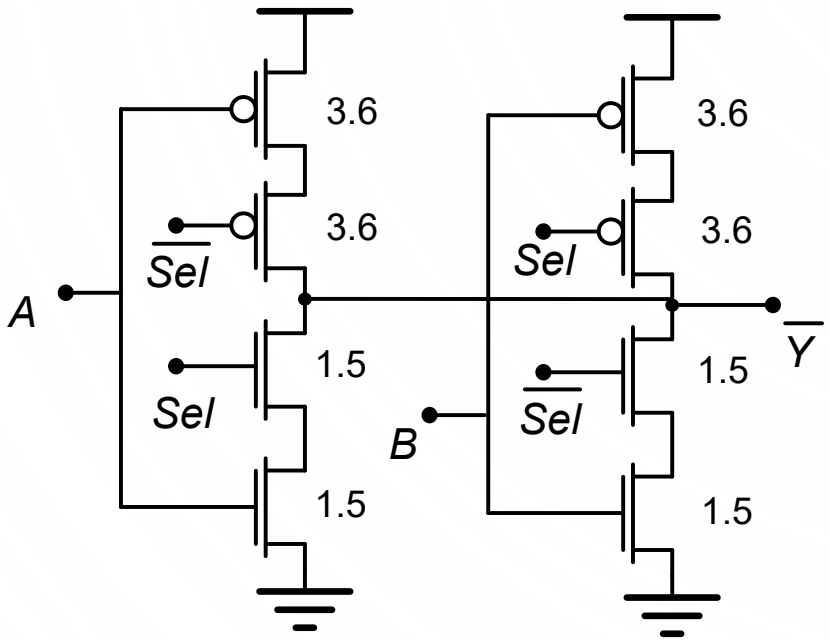


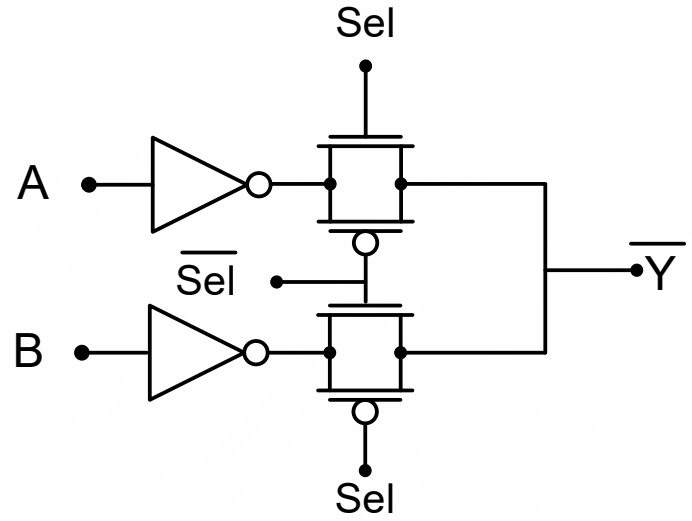


Review: MUX

- 2-input MUX





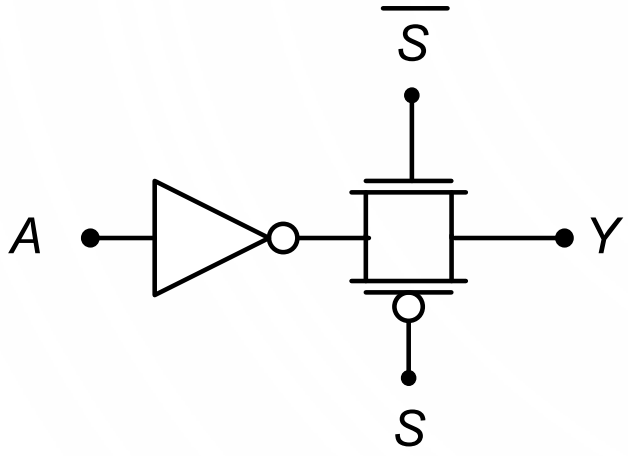


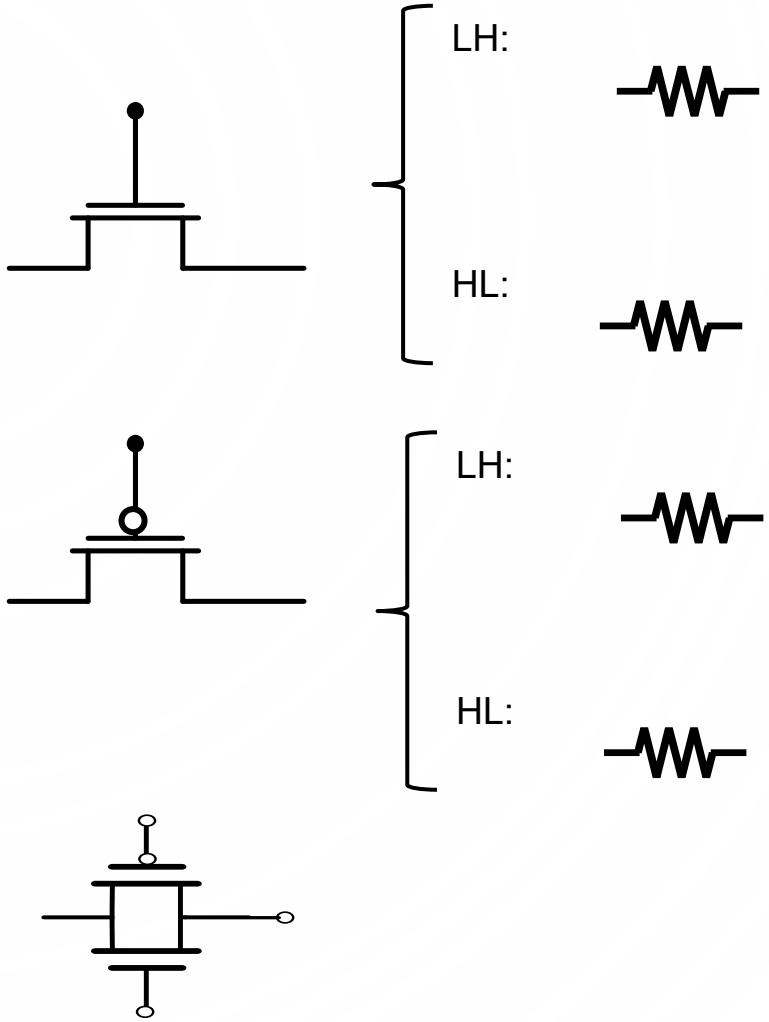
$g_A = 1.7$
$g_B = 1.7$
$g_{Sel} = 3.4$

$p_{C2MOS} = 1.7$
$p_{MUX} = 3.4$

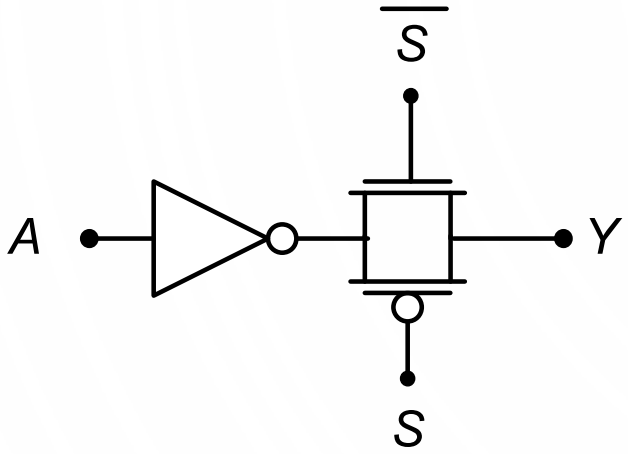
Changing Sel sizing compromises layout

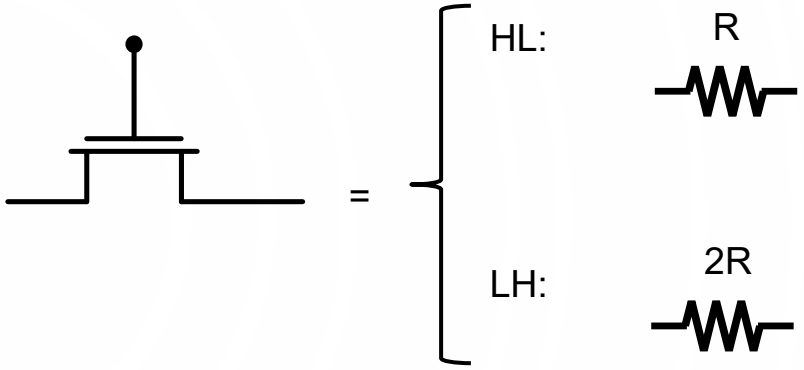
Review: Transmission Gates

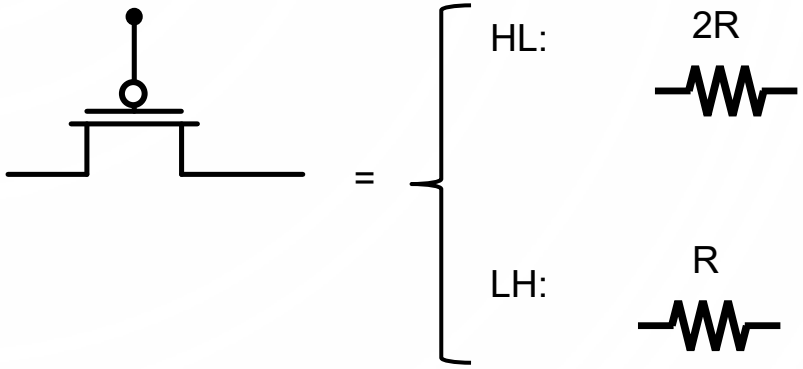


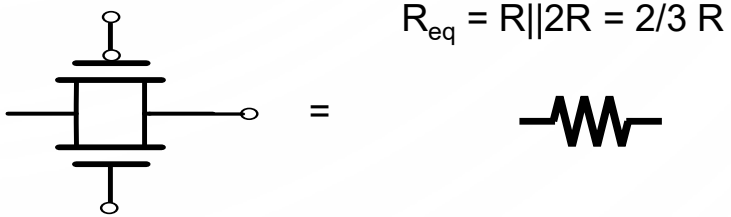


Review: Transmission Gates

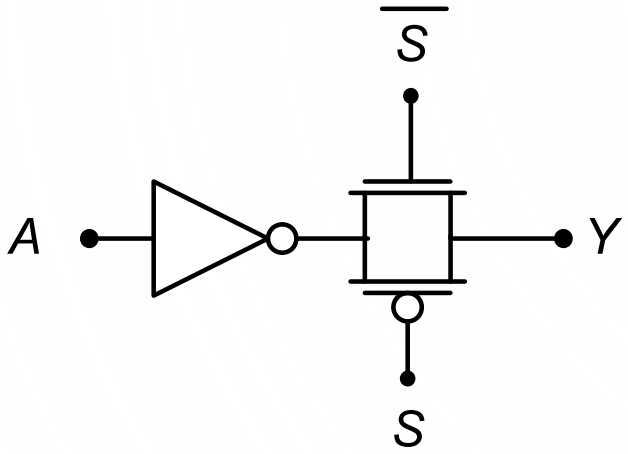


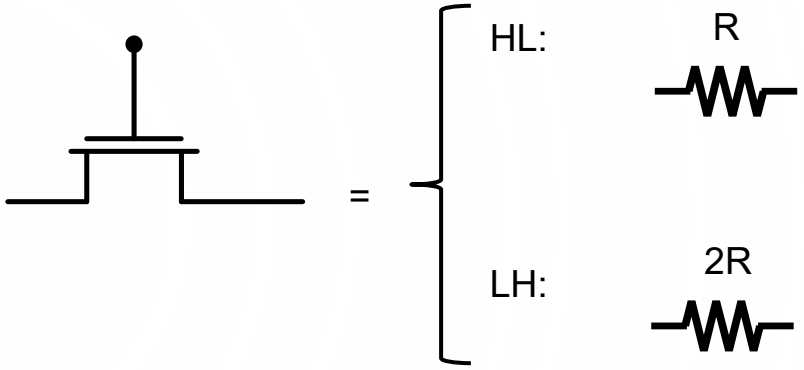


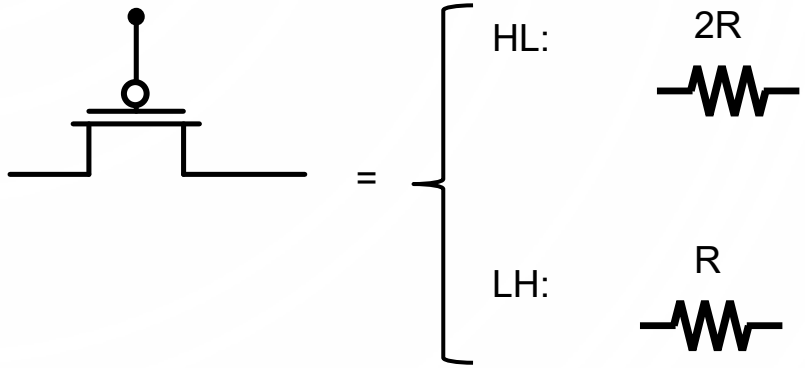




Review: Transmission Gates

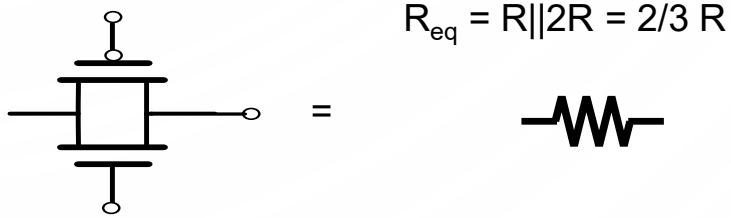






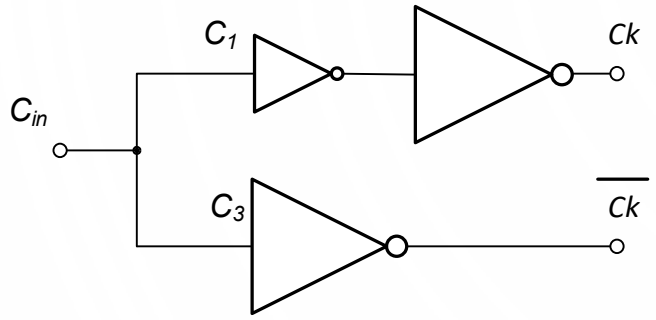
$g_A \sim 1.7$
$g_S = 1.7$

p = 1.7 (but larger in practice because of layout)



Generating Complementary Clocks

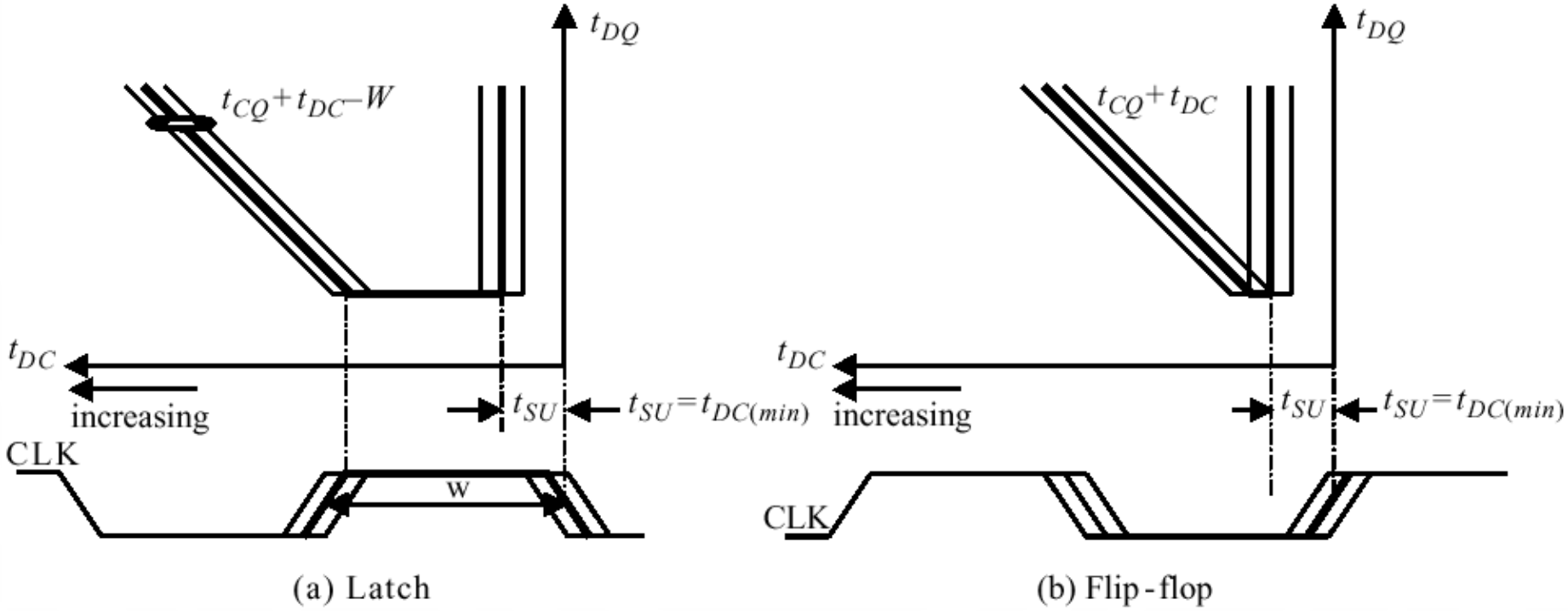
- Inverter fork



$g_{fork} \sim 1$

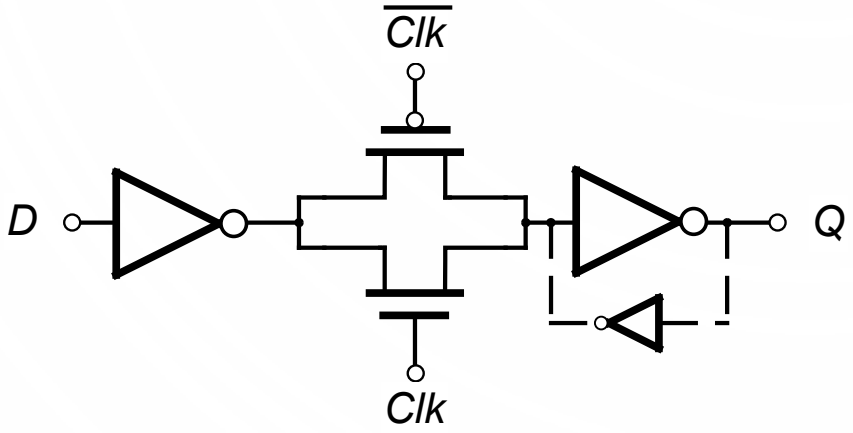
$p_{fork} \sim 3$

Latch vs. Flip-Flop



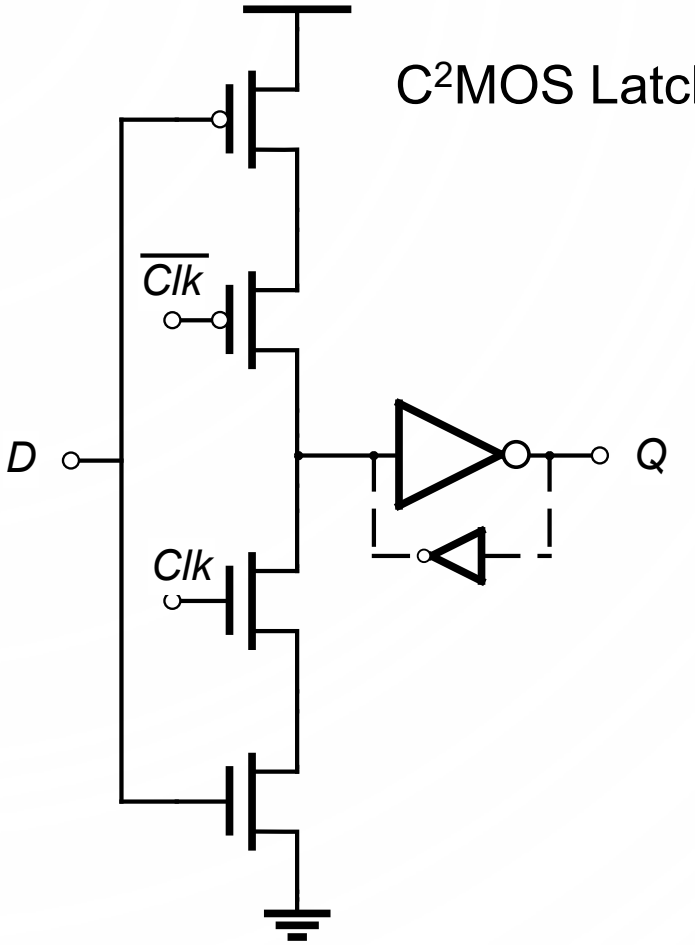
Latches

Transmission-Gate Latch

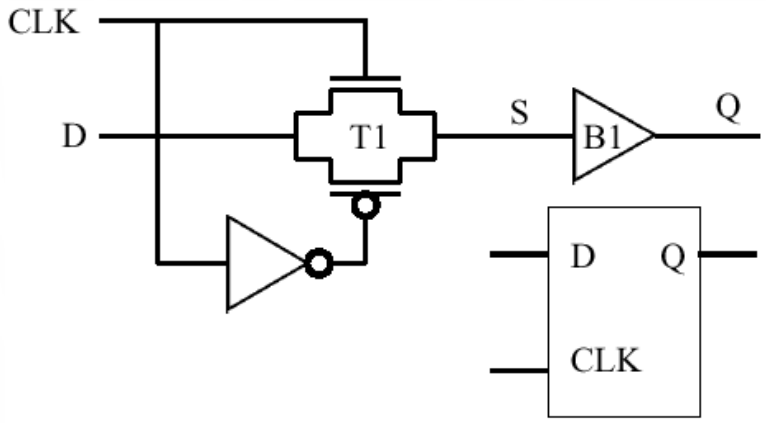


Usually without contention

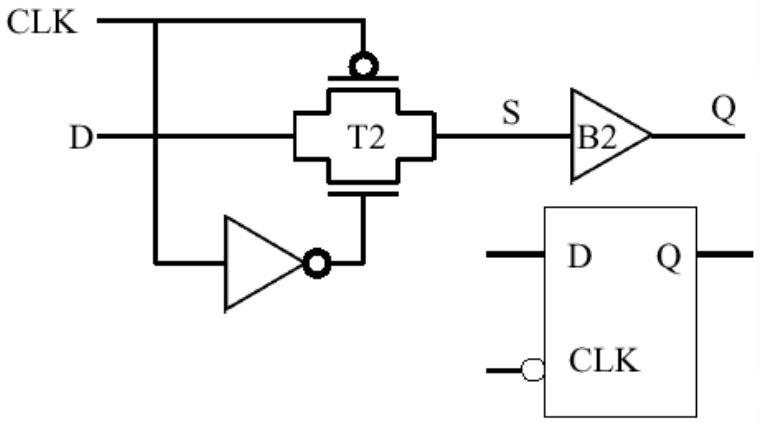
C^2MOS Latch



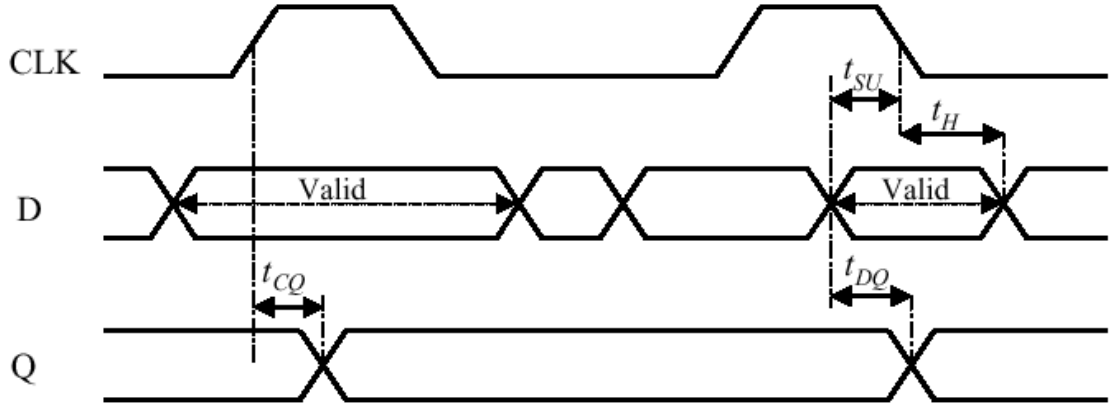
Latches



(a) The transparent high latch (THL)



(b) The transparent low latch (TLL)



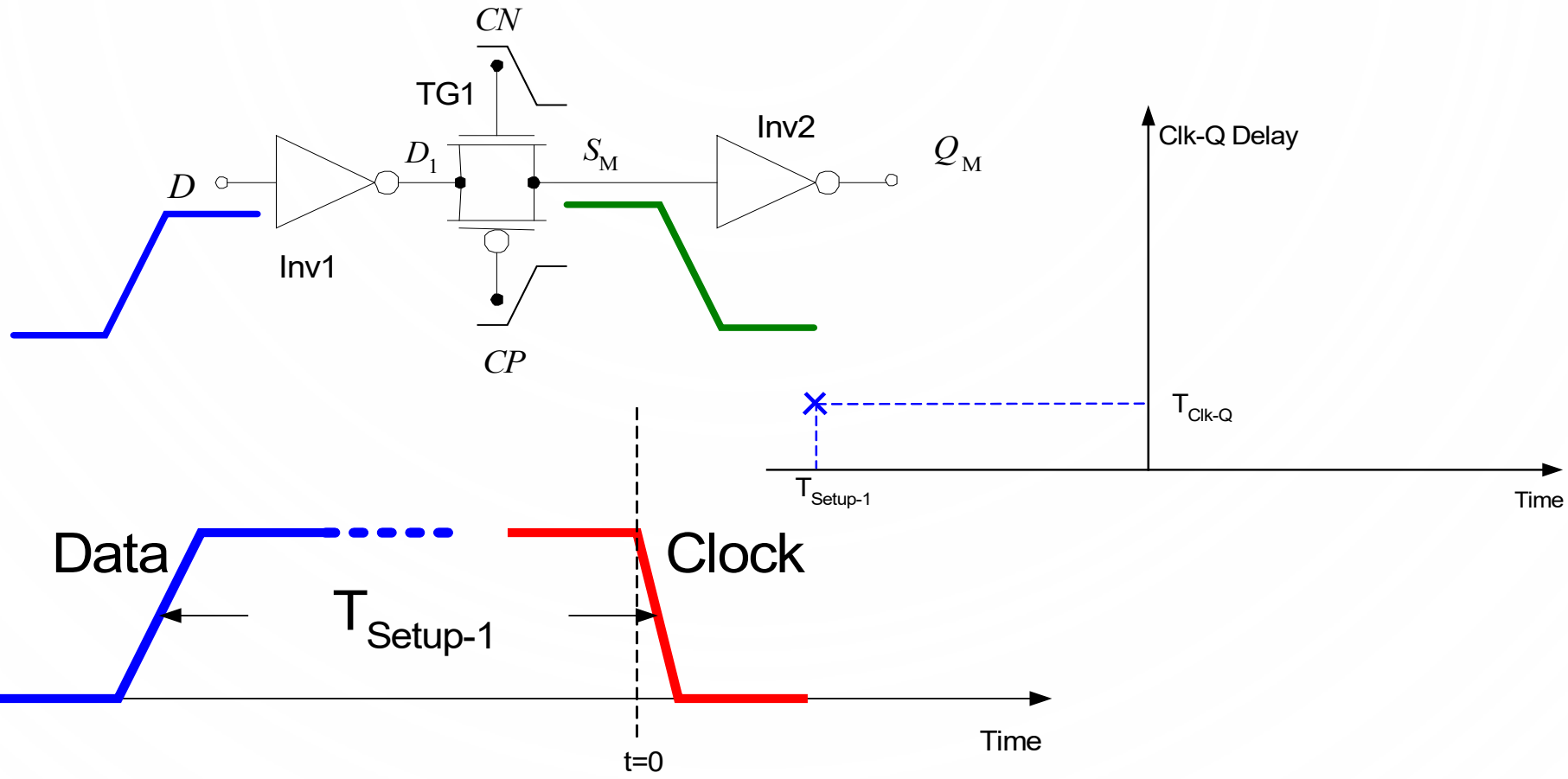
(c) Timing waveforms for the THL

Design for Performance

Delay, Setup, Hold

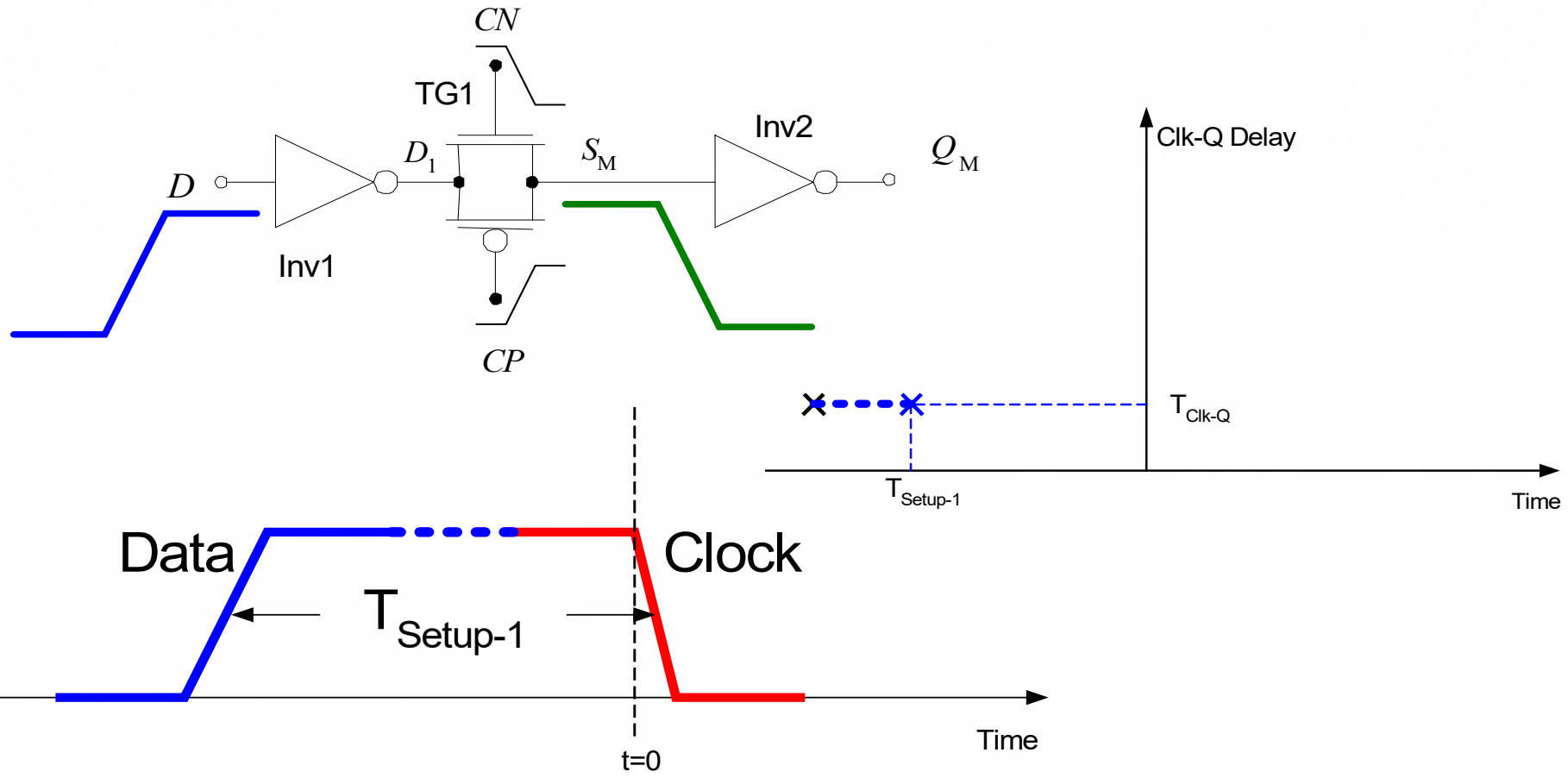
Setup-Hold Time Illustrations

Circuit before clock arrival (Setup-1 case)



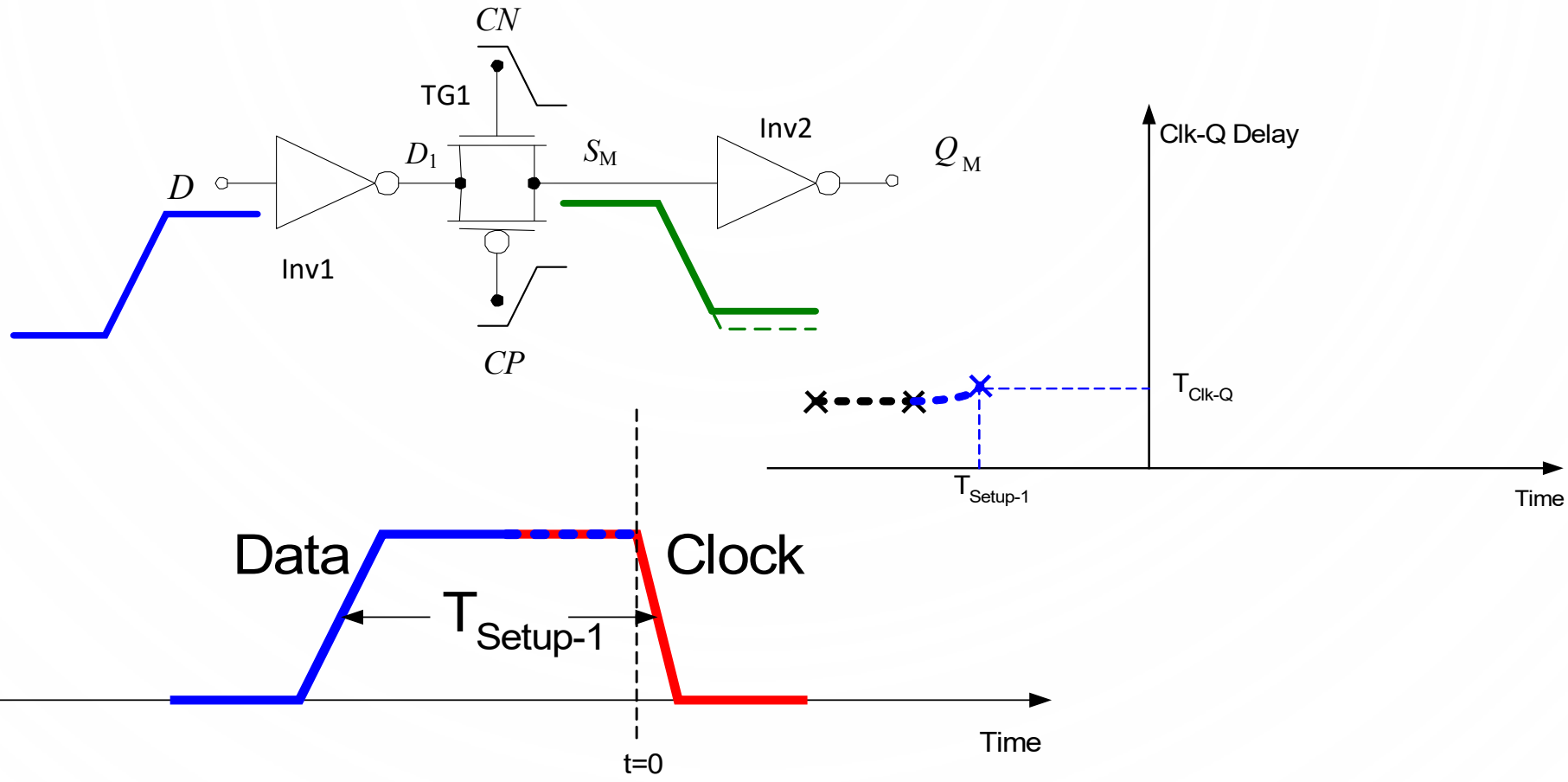
Setup-Hold Time Illustrations

Circuit before clock arrival (Setup-1 case)



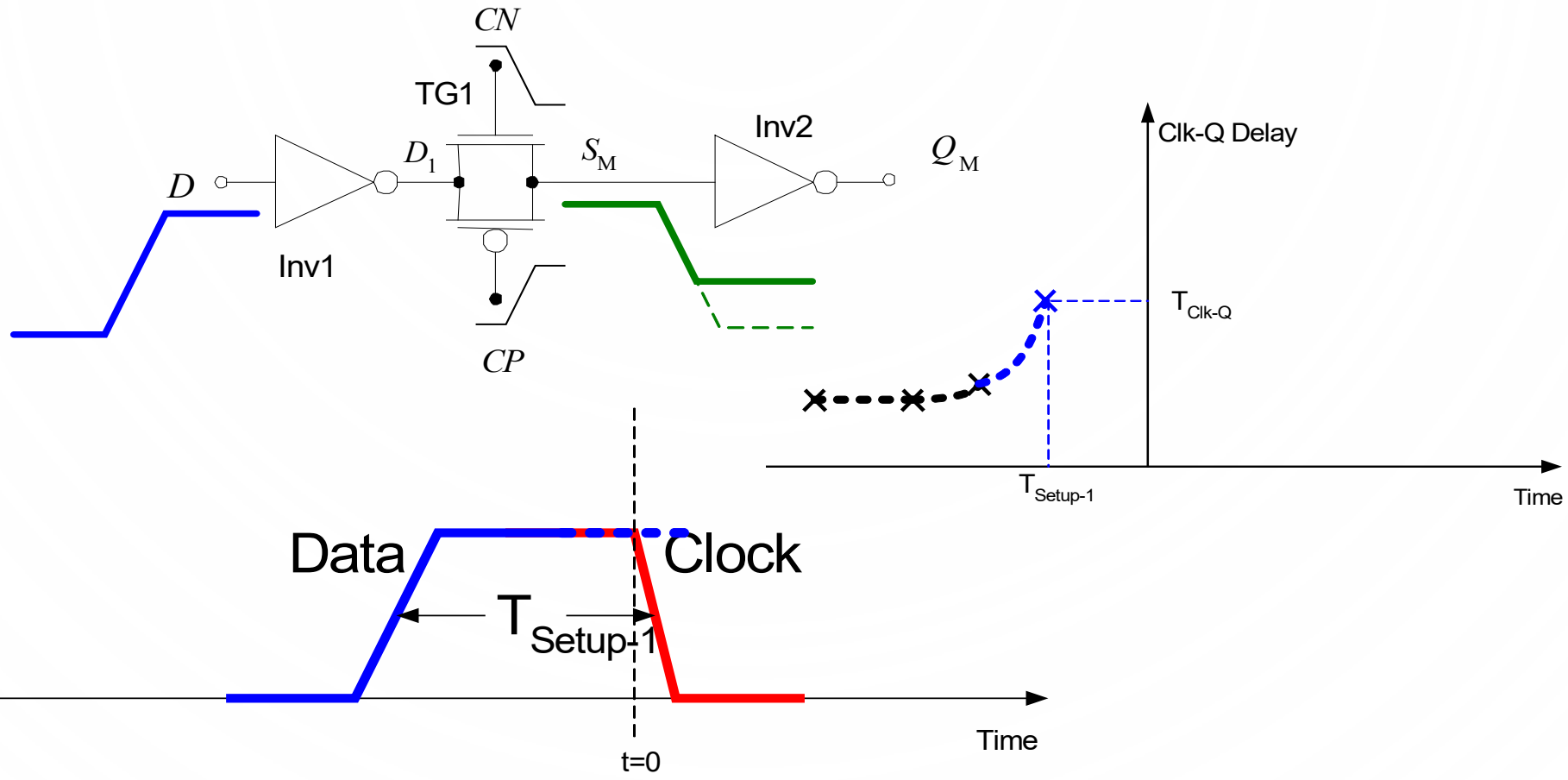
Setup-Hold Time Illustrations

Circuit before clock arrival (Setup-1 case)



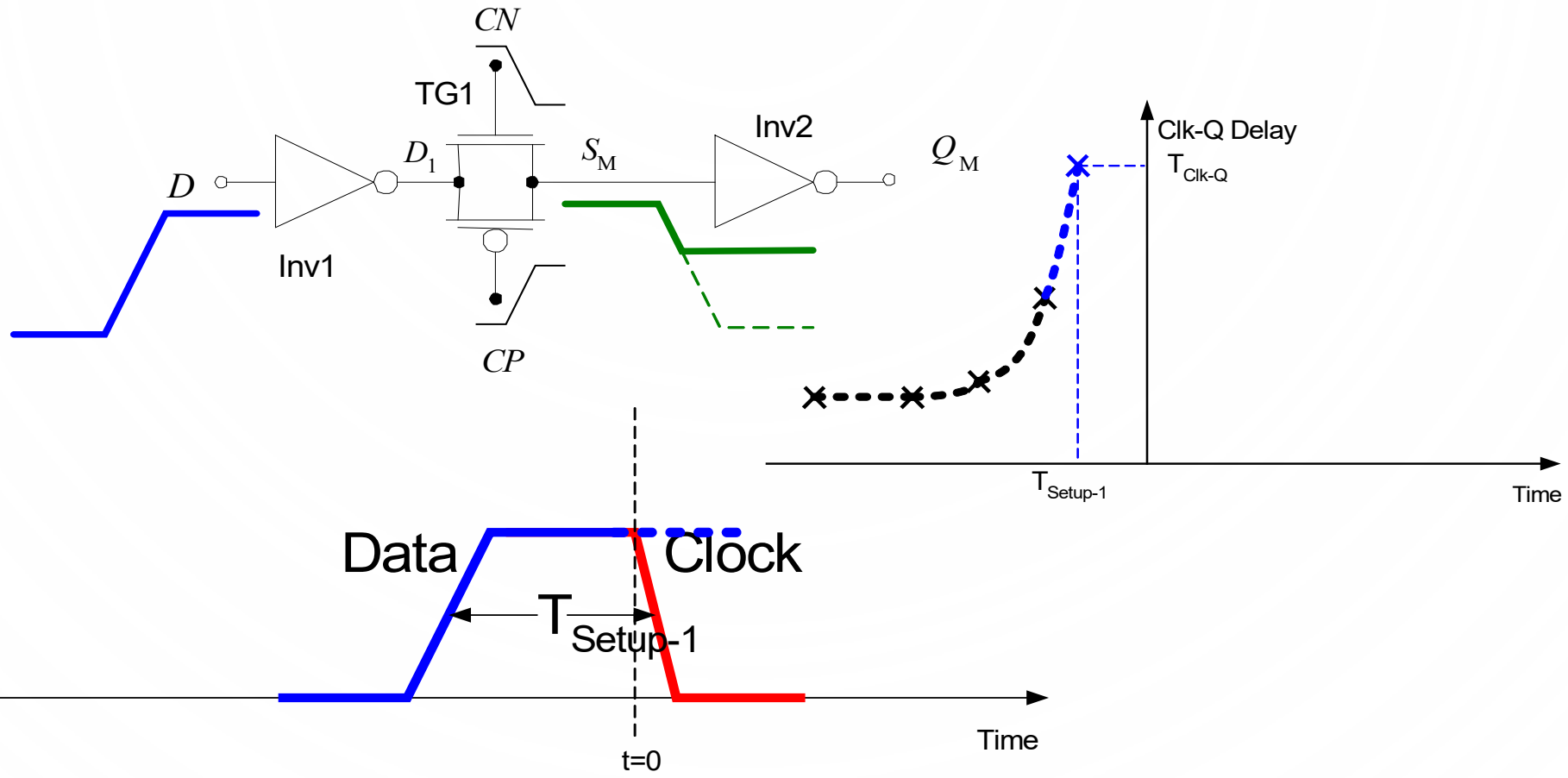
Setup-Hold Time Illustrations

Circuit before clock arrival (Setup-1 case)



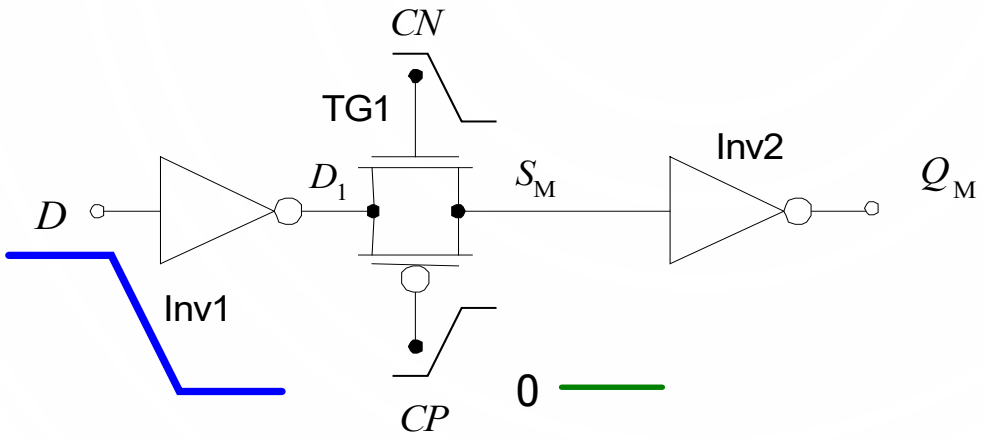
Setup-Hold Time Illustrations

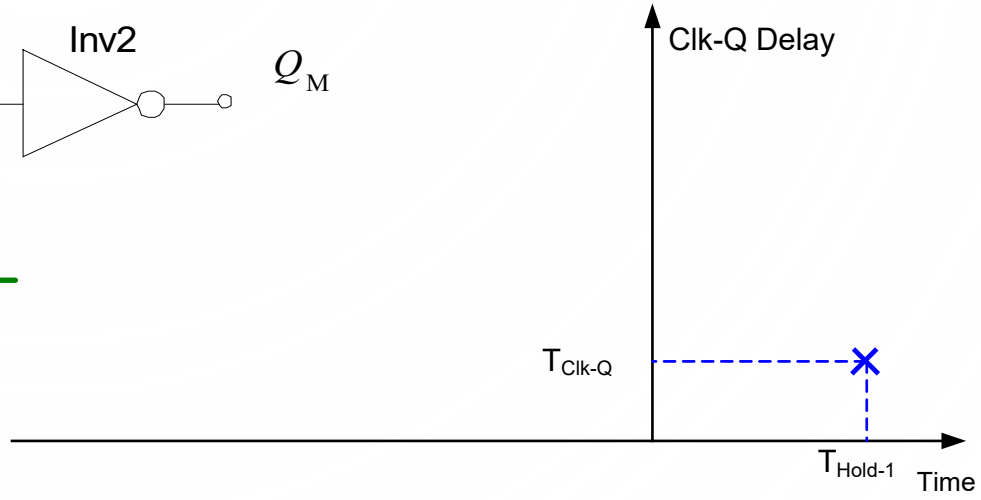
Circuit before clock arrival (Setup-1 case)

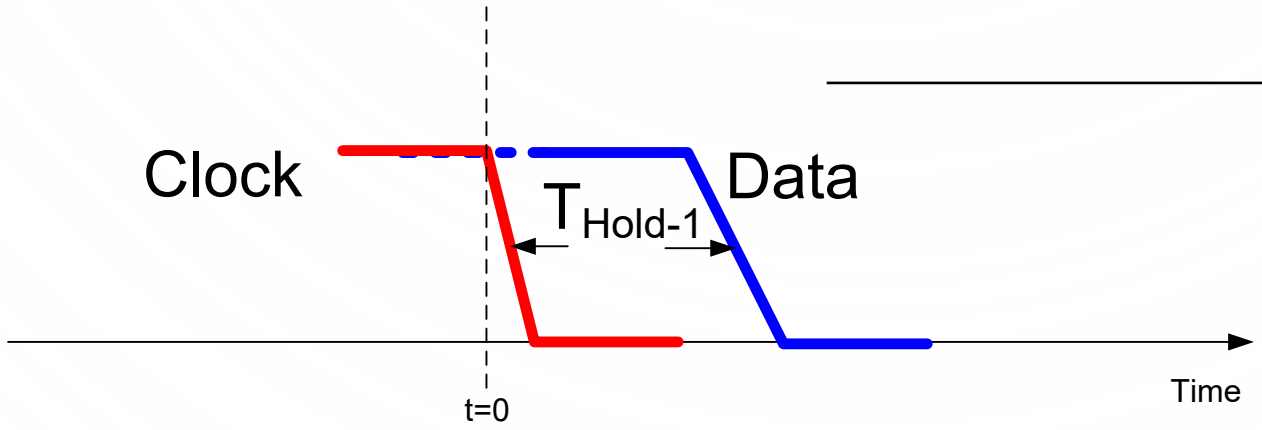


Setup-Hold Time Illustrations

Hold-1 case

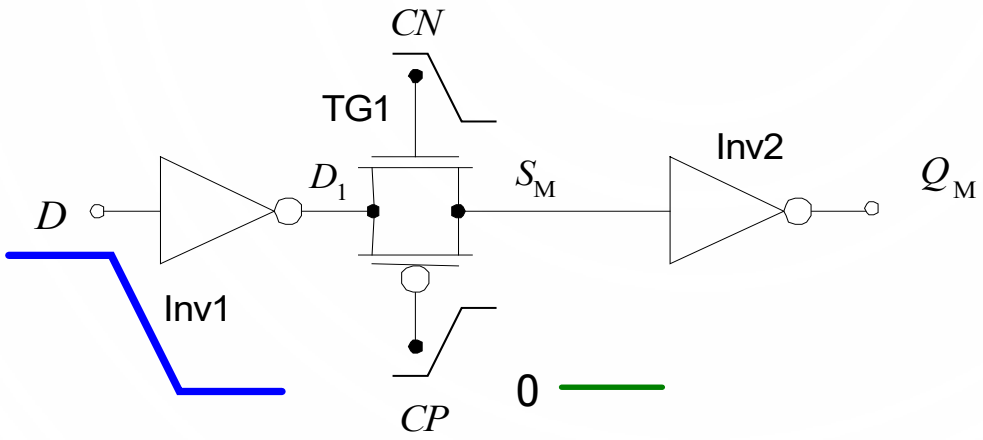


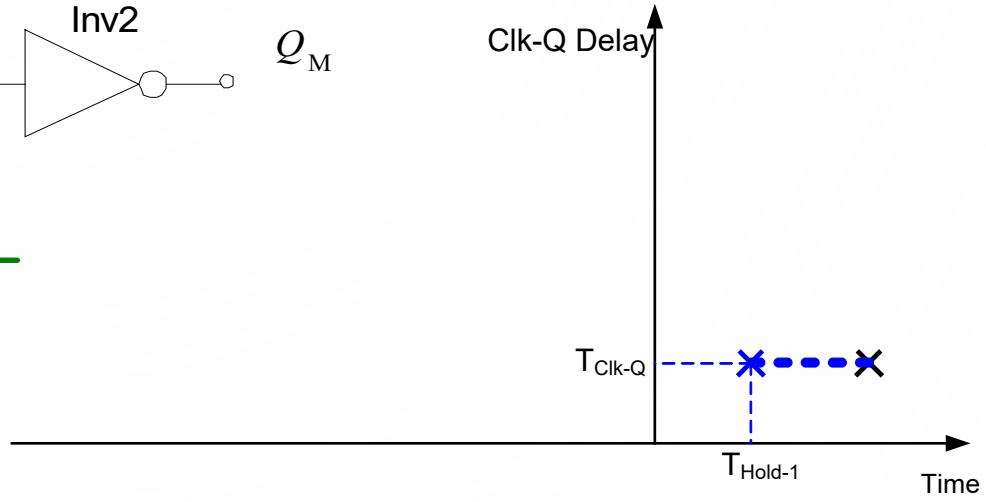


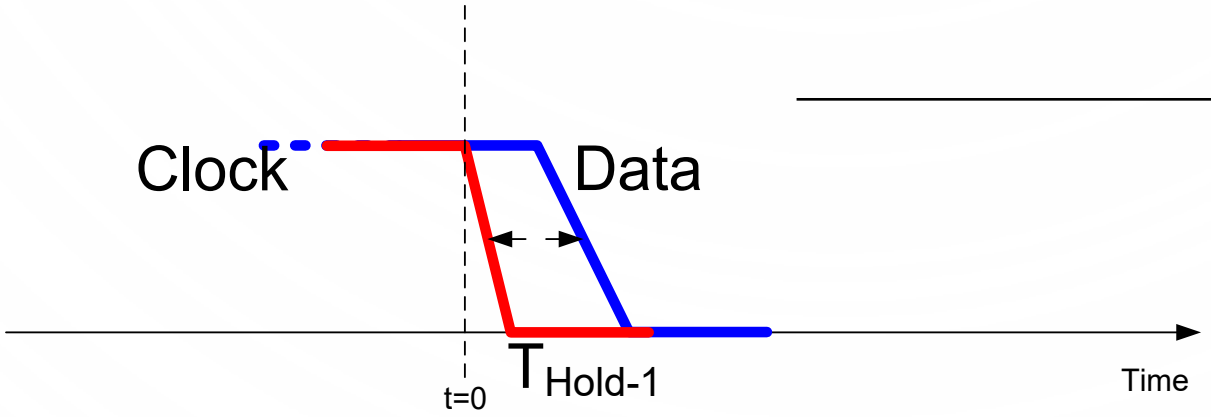


Setup-Hold Time Illustrations

Hold-1 case

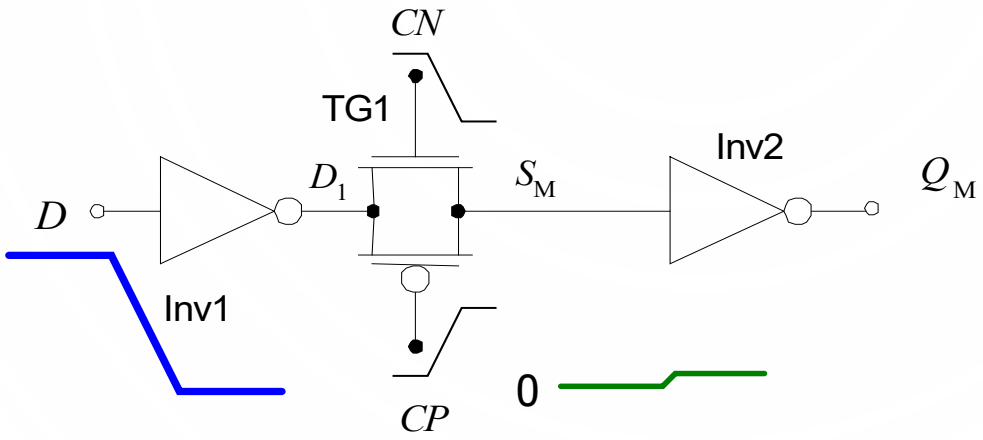


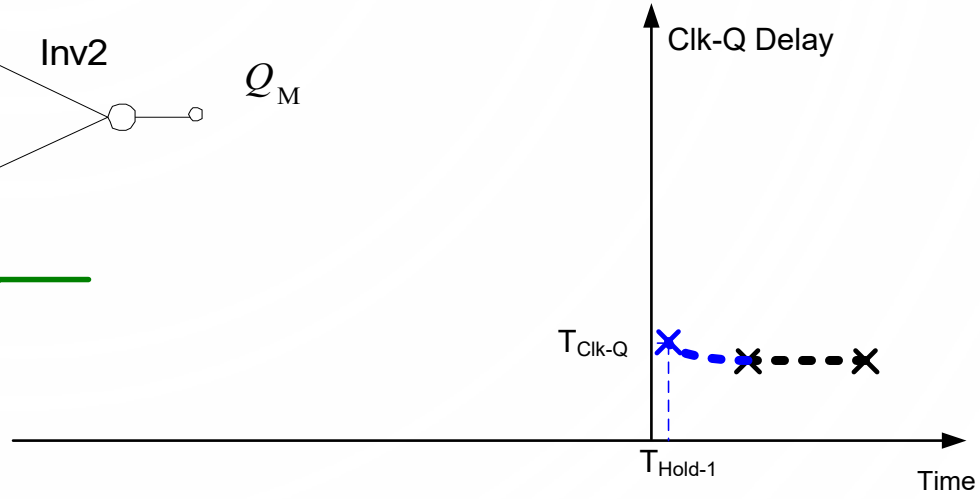


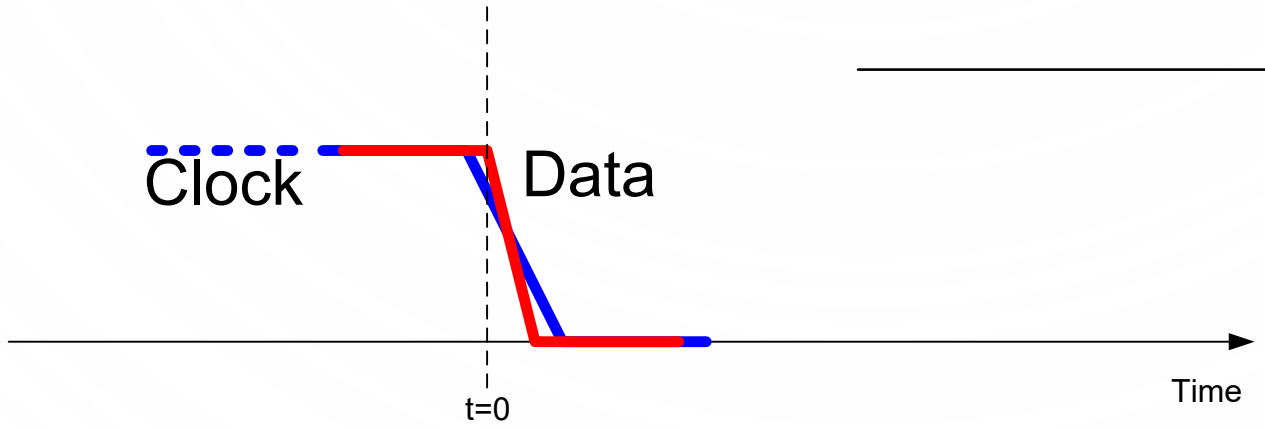


Setup-Hold Time Illustrations

Hold-1 case

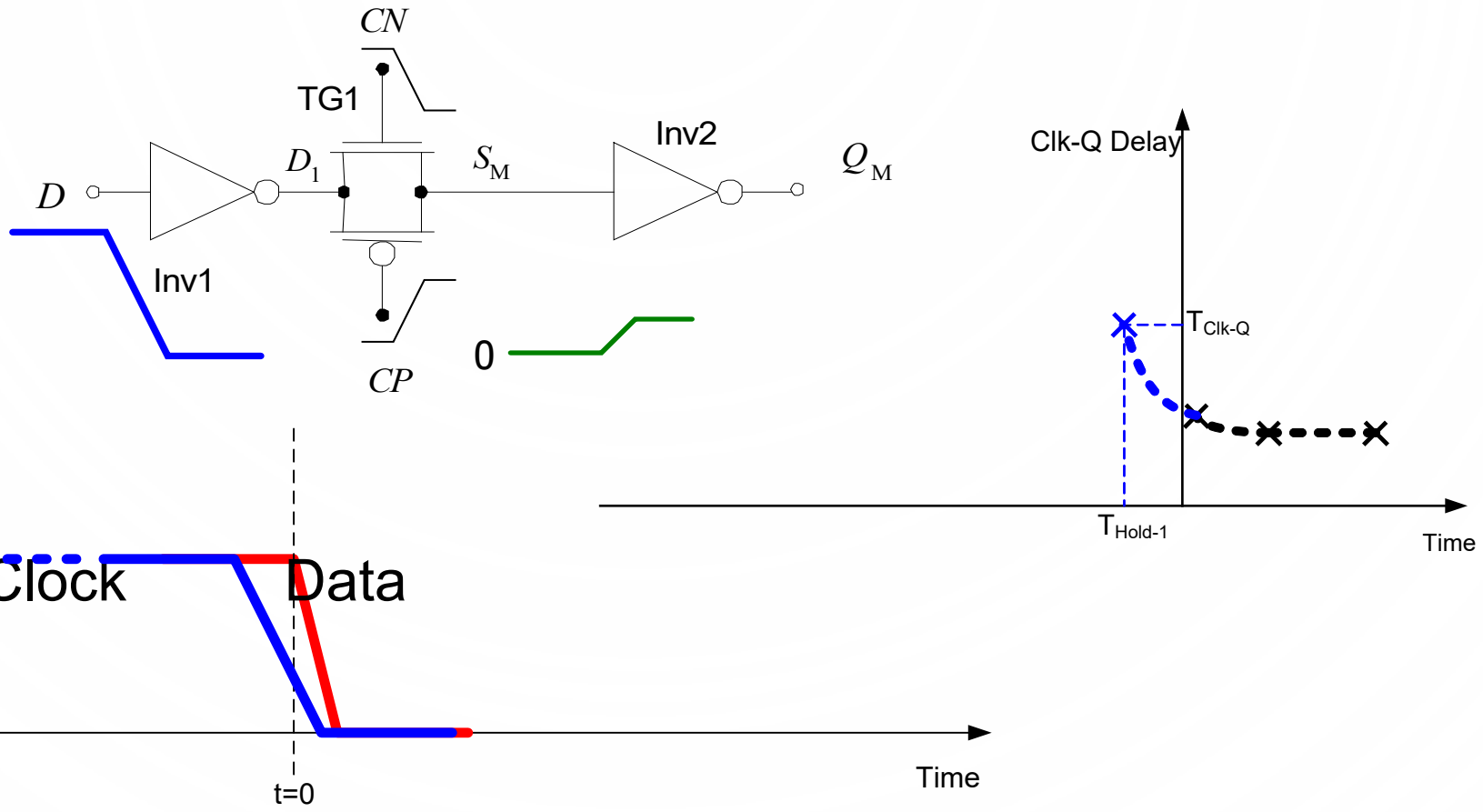




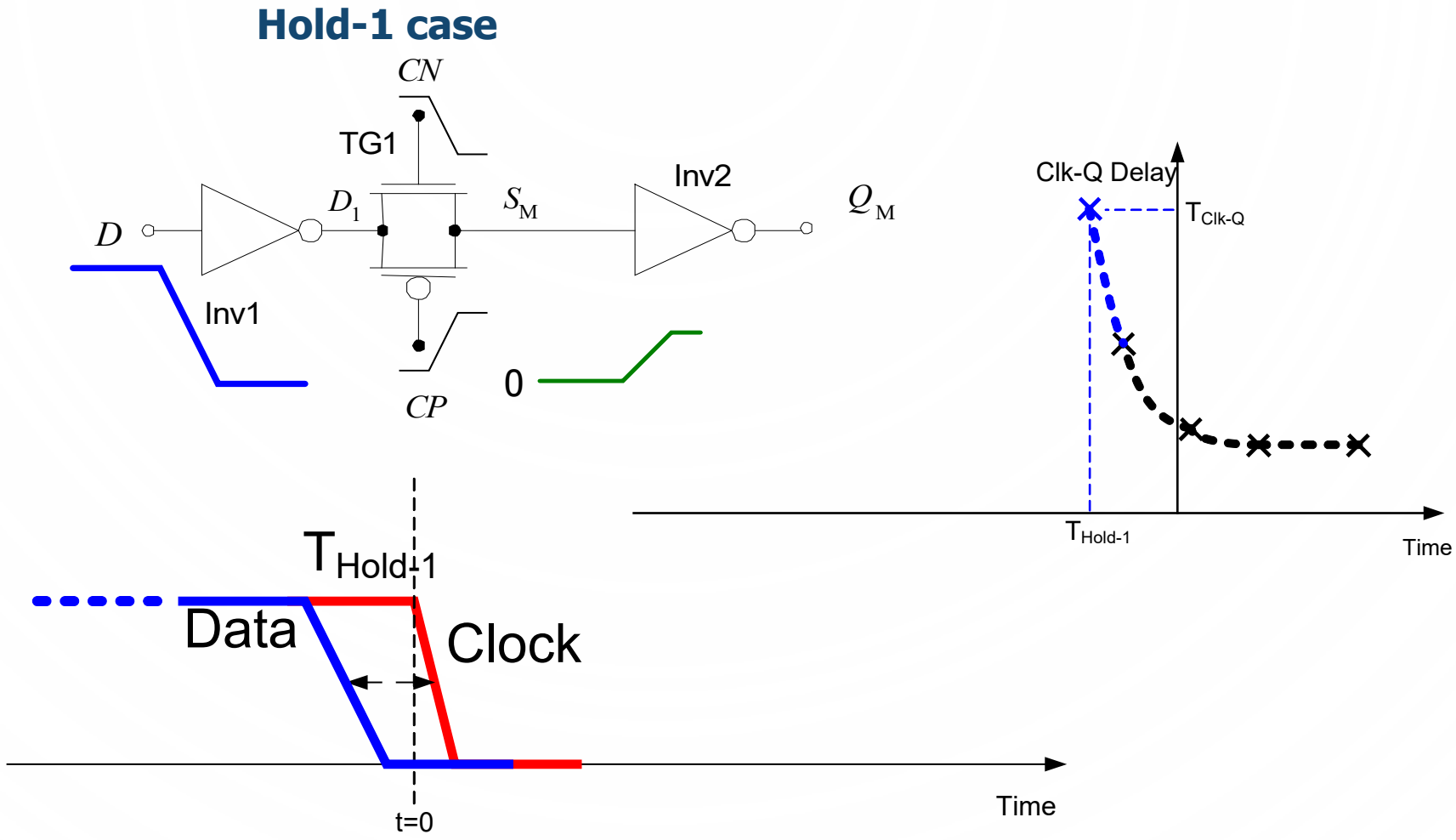


Setup-Hold Time Illustrations

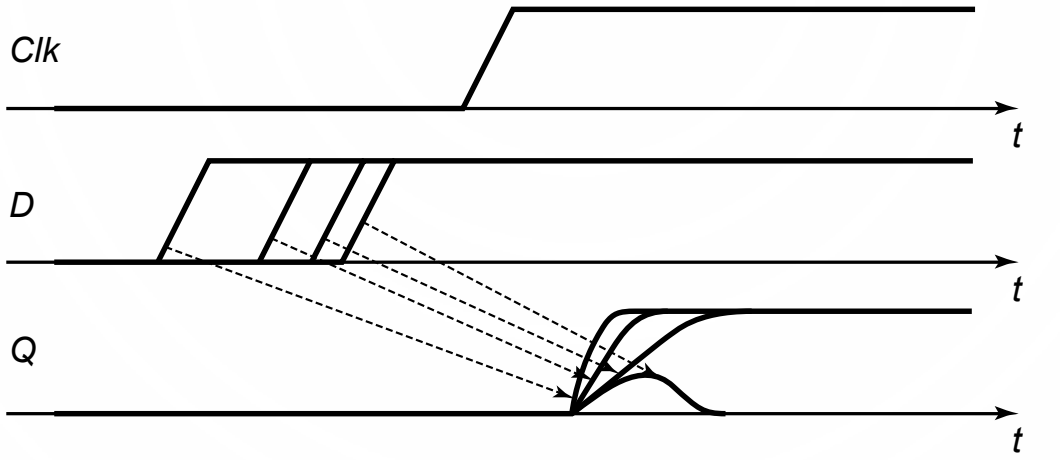
Hold-1 case



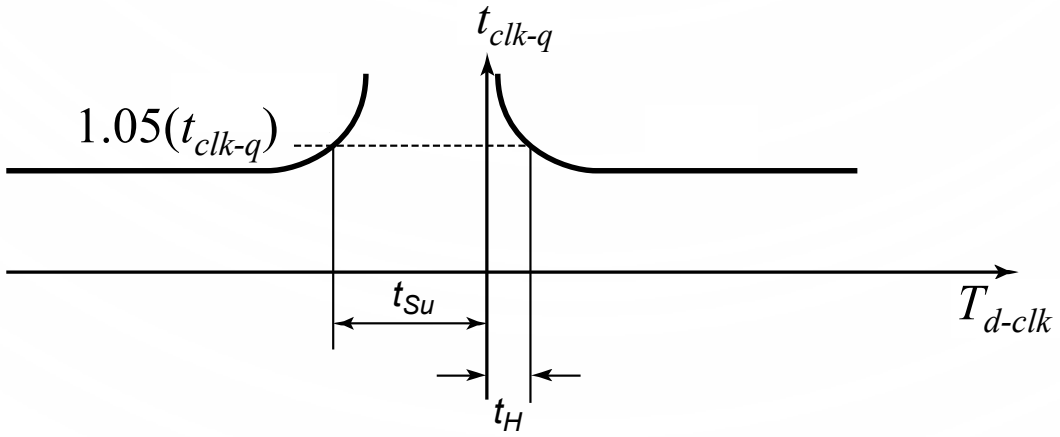
Setup-Hold Time Illustrations



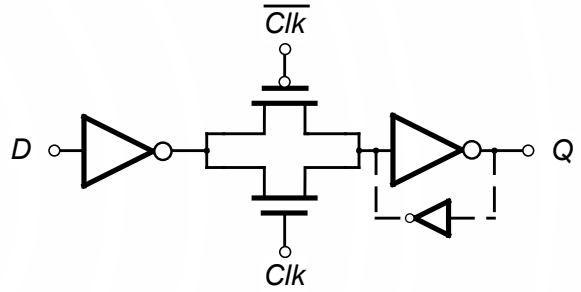
More Precise Setup Time



(a)



Latch t_{D-Q} and t_{Clk-Q}



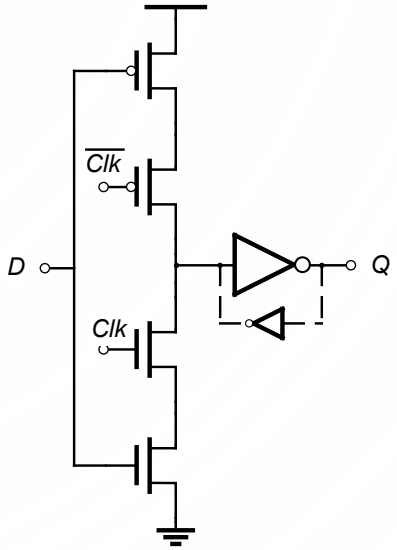
(ignore feedback inverters, assume g_{fork}=1)

$t_{Clk-Q} = g_1f_1 + p_1 + g_2f_2+p_2$; $g_1f_1 = g_2f_2 = \sqrt{(GF)}$

Assume F = 1, for simplicity, (although a latch should drive F>4)

$\sqrt{GF} = 1.3$

$t_{Clk-Q} = (1.3 + 1.7) + (1.3 + 1) = 5.3$ (=1.06 FO4)

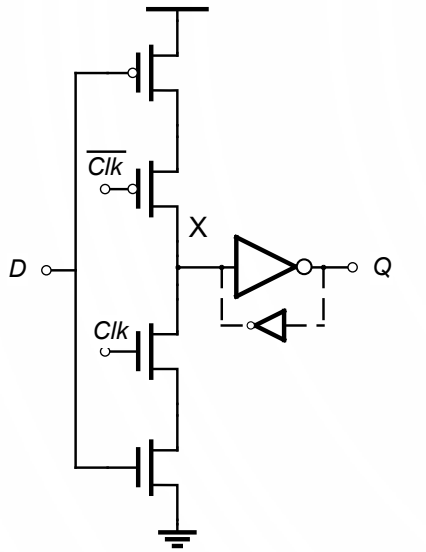


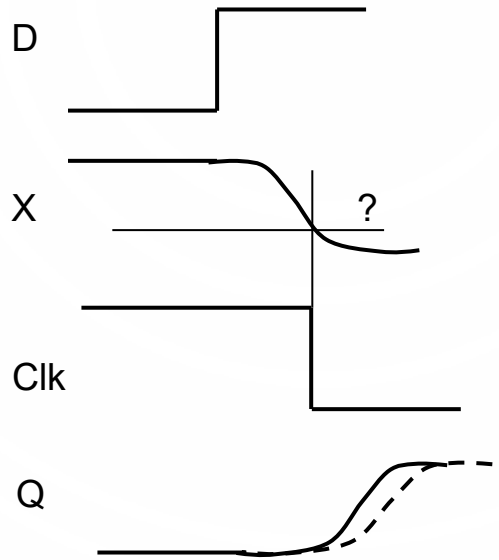
$g_1 = 1.7$ $g_2 = 1$

$p_1 = 1.7$ $p_2 = 1$

(FO4 inverter delay = 1 + 4 = 5 unit delays)

t_{setup}





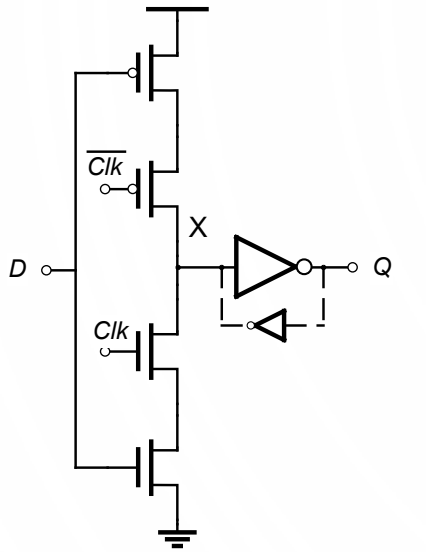
$3t_u$ $2.3t_u \rightarrow 2.58t_u$ For 1.05 $t_{clk\text{-}Q}$

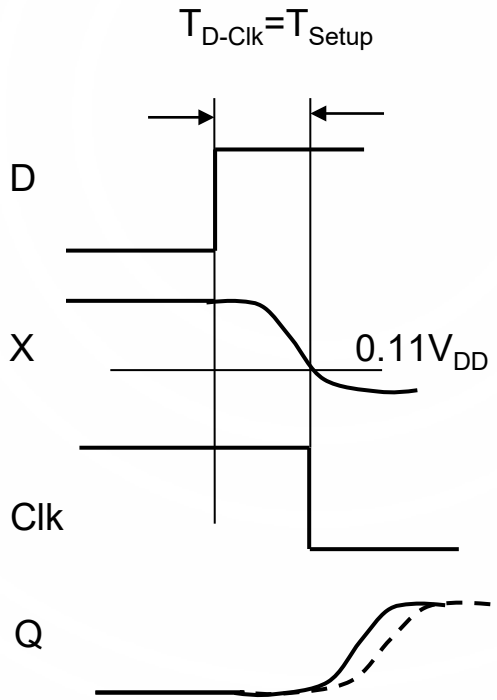
12% delay increase

To find the setup time, we will find the D-Clk offset that increases $t_{clk\text{-}Q}$ by 5%
Overall delay is $5.8t_u$, 5% increase is $0.28t_u$

Note: Voltage level of $.12V_{DD}$ at X causes ~12% delay increase

t_{setup}





$3t_u$ $2.3t_u \rightarrow 2.58t_u$ For 1.05 $t_{clk\text{-}Q}$

11% delay increase

Voltage level of $.12V_{DD}$ at X causes ~12% delay increase

Assuming exponential response

$T_{setup} = -\ln(0.12)\ \tau = 2.1\ \tau$

$T_{setup} = 2.1\ \tau = 3t_{prop}$ ($t_{prop} = 0.7\tau$)

$T_{setup} = 3 \times 3t_u = 9t_u$

$T_{setup} = 1.8$ FO4

$T_{setup} + T_{Clk\text{-}Q} = 2.8$ FO4

$T_{setup} + T_{Clk\text{-}Q}$ is typically 2.5-3.5 FO4 for fast latches (with low fanout)

Flip-Flops

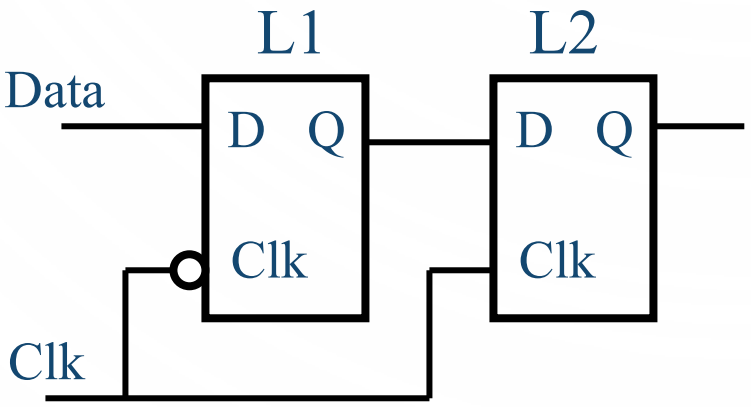
Flip-Flops

- Performance metrics
- Delay metrics
  - Insertion delay
  - Inherent race immunity
  - 'Softness' (Clock skew absorption)
  - Inclusion of logic
  - Small (+constant) clock load
- Power/Energy Metrics
  - Power/energy
- Design robustness
  - Noise immunity

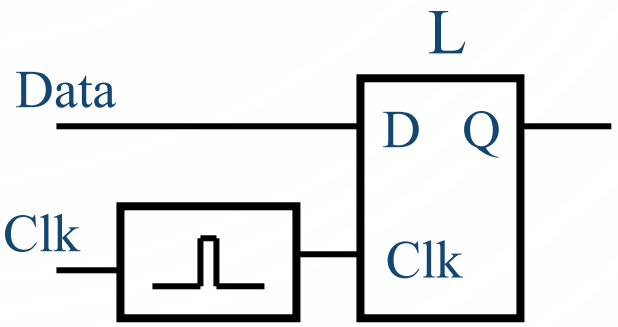
Types of Flip-Flops

- Two ways to design a flip-flop
  - Latch pair (large majority)
  - Pulsed latch

Latch Pair

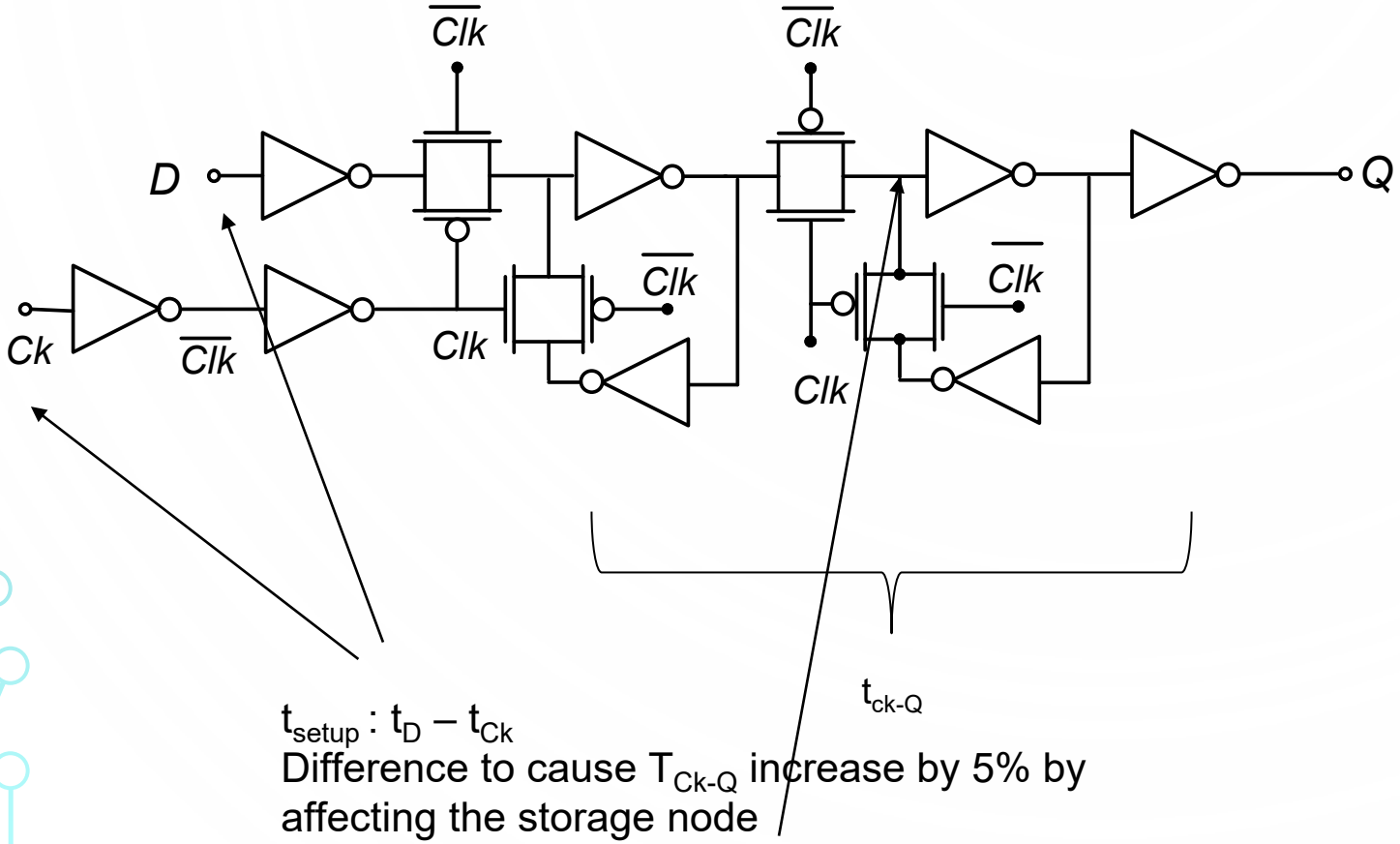


Pulse-Triggered Latch



Flip-Flop (Latch Pair) Clk-Q, setup, hold

Calculation is nearly identical to that of a latch (ignore feedback inverters).
t_{Clk-Q} is the delay of the second latch, which is about 1FO4;
note that t_{Ck-Q} should include the delay of the inverter fork



t_{setup} : $t_D - t_{Ck}$
Difference to cause T_{Ck-Q} increase by 5% by affecting the storage node

Setup time calculation goes the same way!

Flip-Flop Library Timing Characterization

- Combinational logic delay is a function of output load and input slope
- Sequential timing (flip-flop):
  - t_{clk-q} is function of output load and clock rise time
  - t_{Su}, t_H are functions of D and Clk rise/fall times
  - Flip-flop has multiple stages, so the delay is less sensitive to input slope

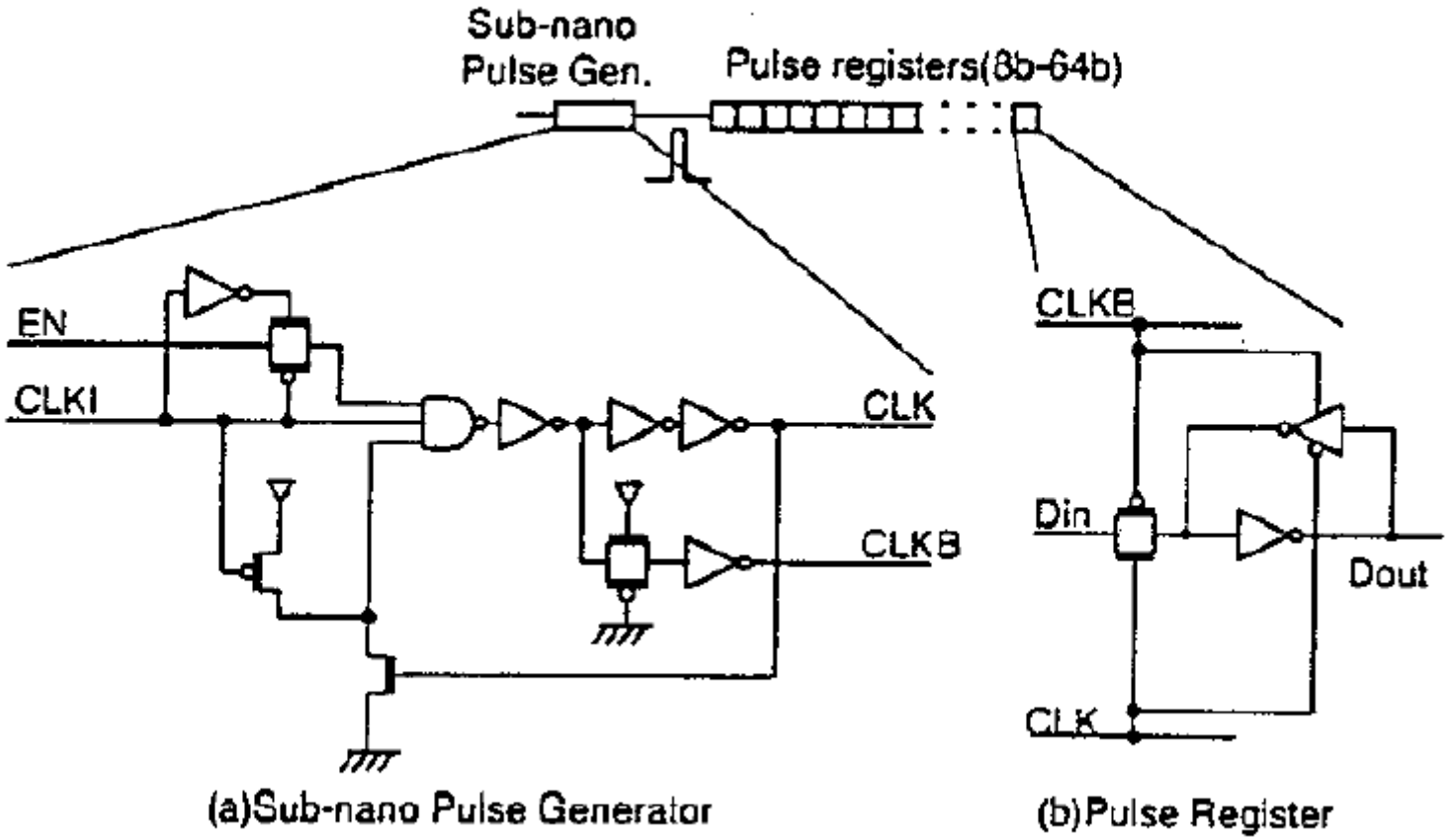
Pulse-Triggered Latches

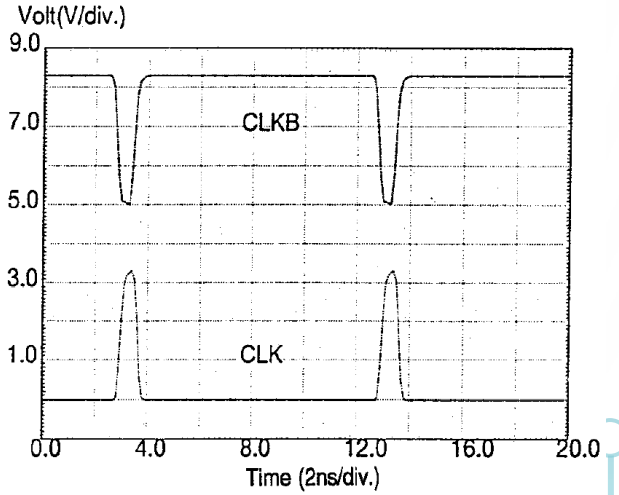
- First stage is a pulse generator
  - generates a pulse (glitch) on a rising edge of the clock
- Second stage is a latch
  - captures the pulse generated in the first stage
- Pulse generation results in a negative setup time
- Frequently exhibit a soft edge property

- Note: power is always consumed in the pulse generator
  - Often shared by a group (register)

Pulsed Latch

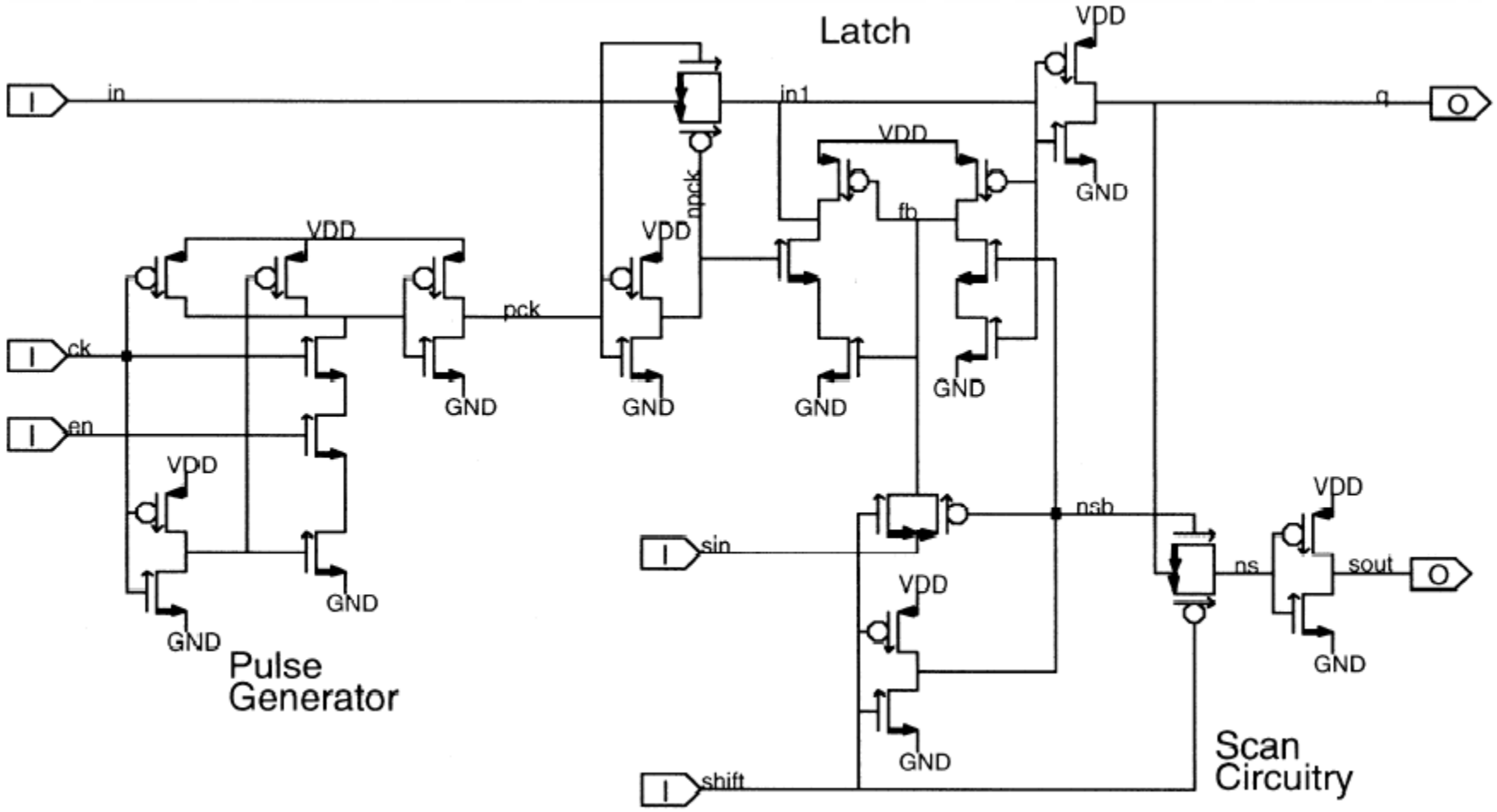
Simple pulsed latch





Kozu, ISSCC'96

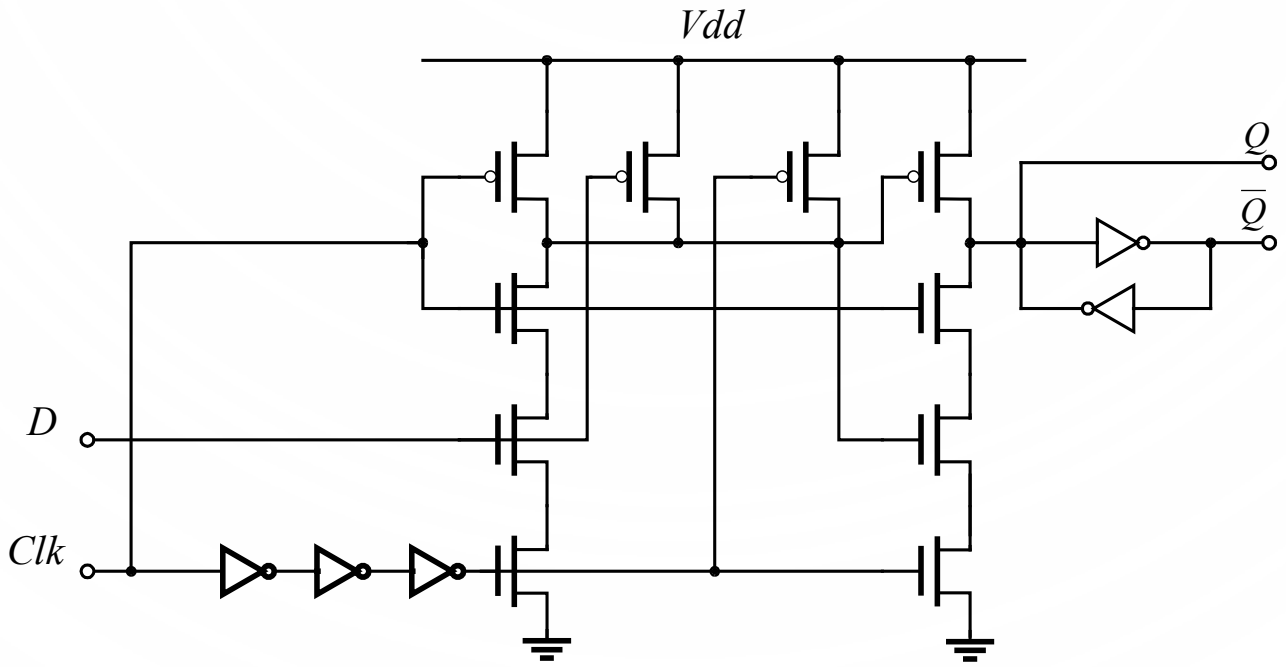
Intel/HP Itanium 2



Naffziger, ISSCC'02

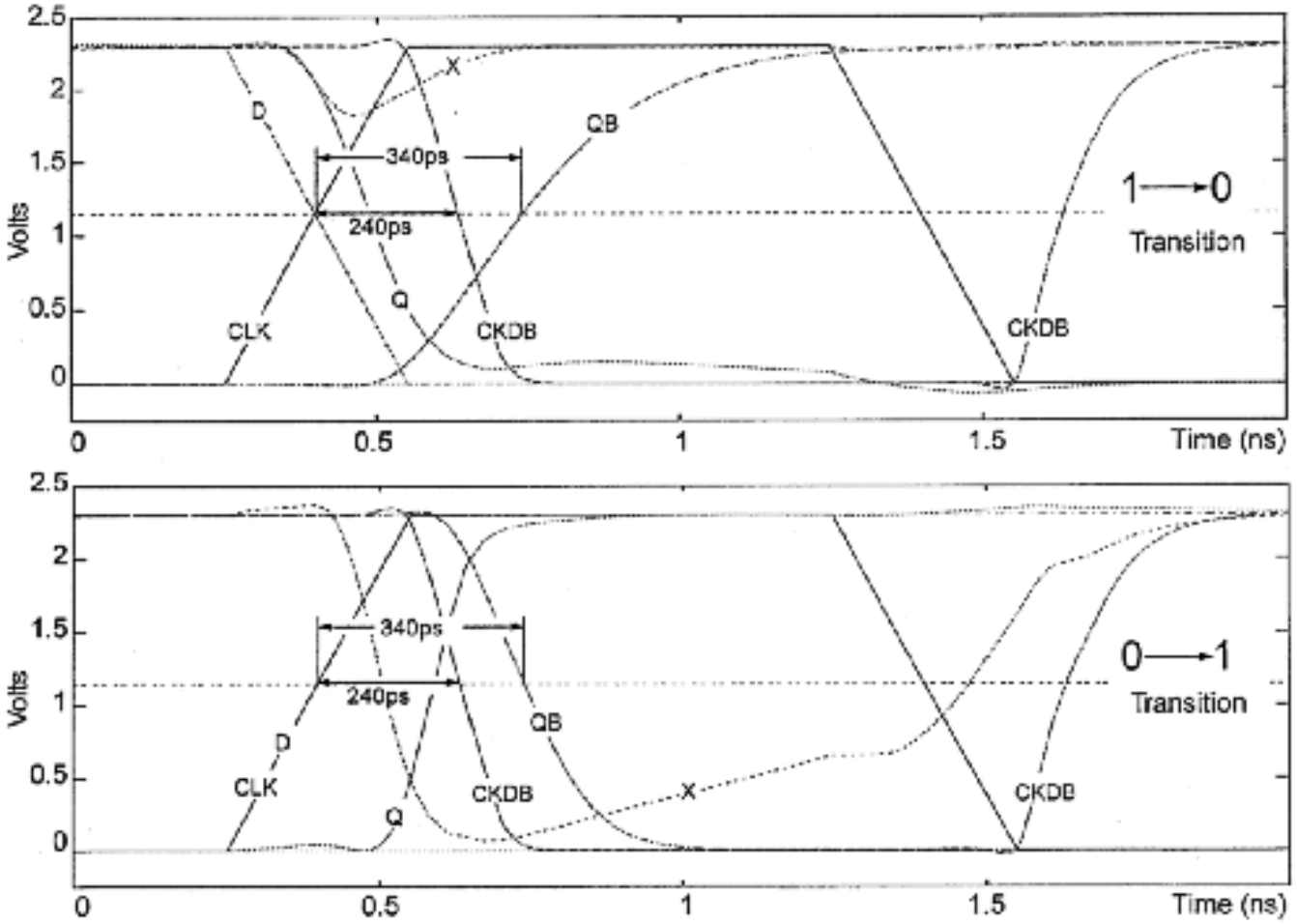
Pulsed Latches

Hybrid Latch Flip-Flop, AMD K-6
Partovi, ISSCC'96



HLFF Operation

1-0 and 0-1 transitions at the input with 0ps setup time

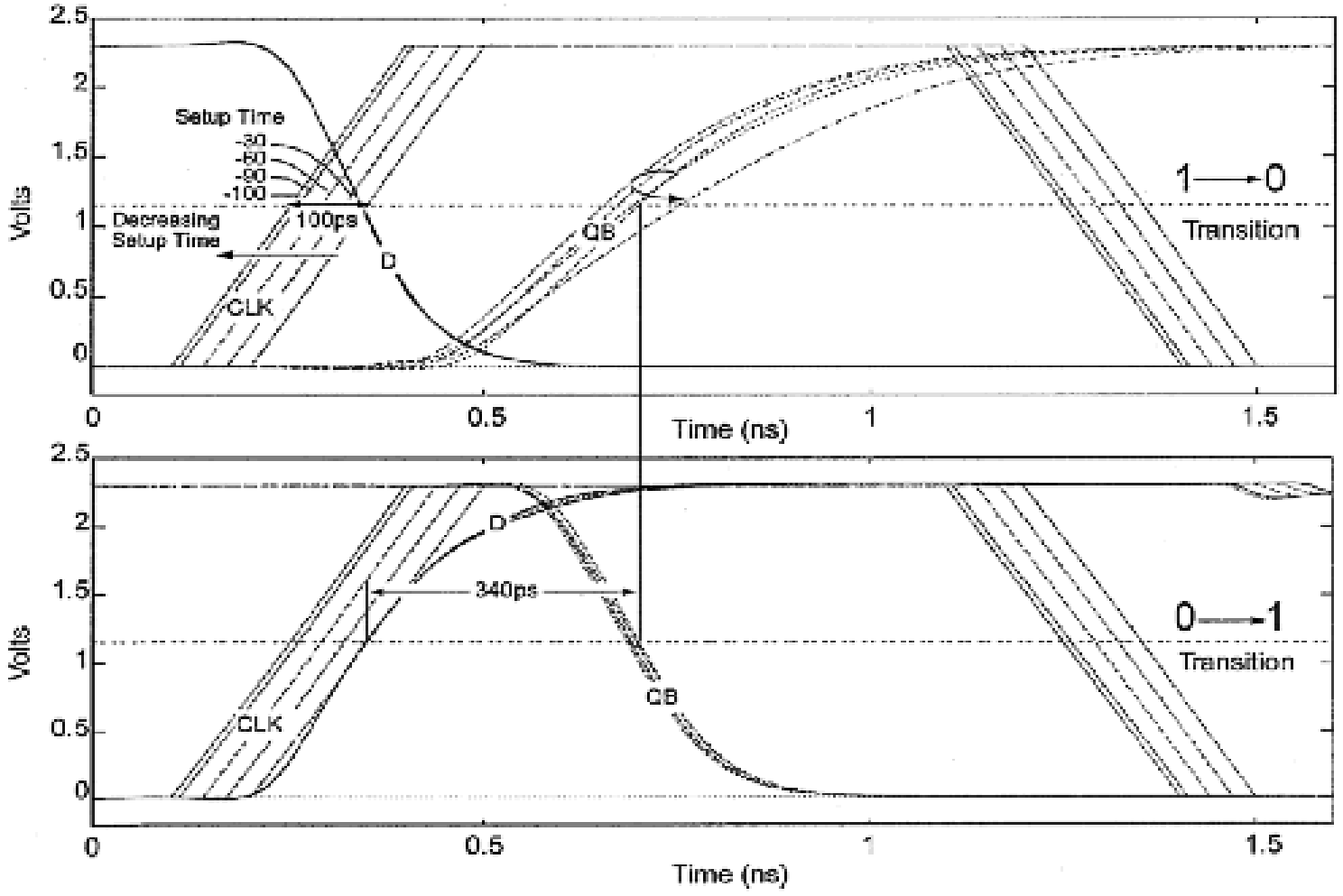


VDD=2.3v, 85°C, Typical Devices
C_{clk} = 60ff, C_{qb} = 300ff

$T_{cq} = T_{dq}$ = 340ps
T_{hold} = 180ps
$T_{su(min)}$ = -90ps

Hybrid Latch Flip-Flop

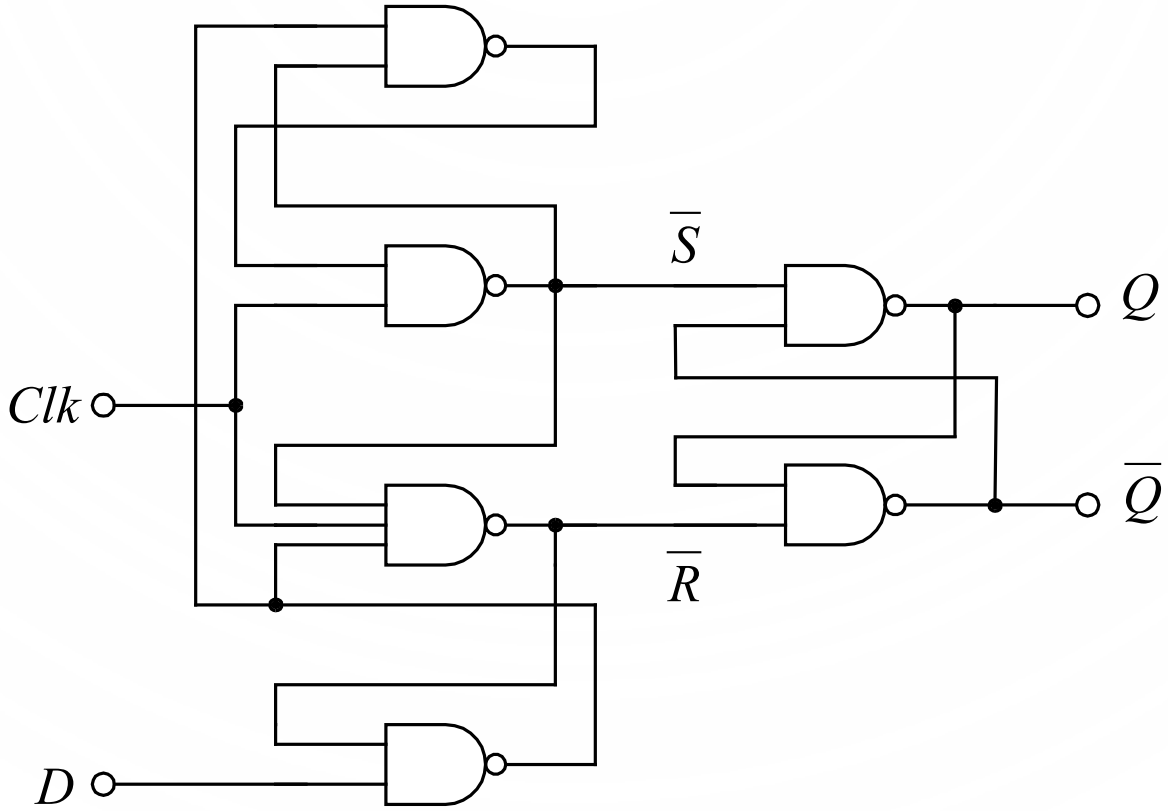
Skew absorption



Partovi *et al*, ISSCC'96

Pulsed Latches

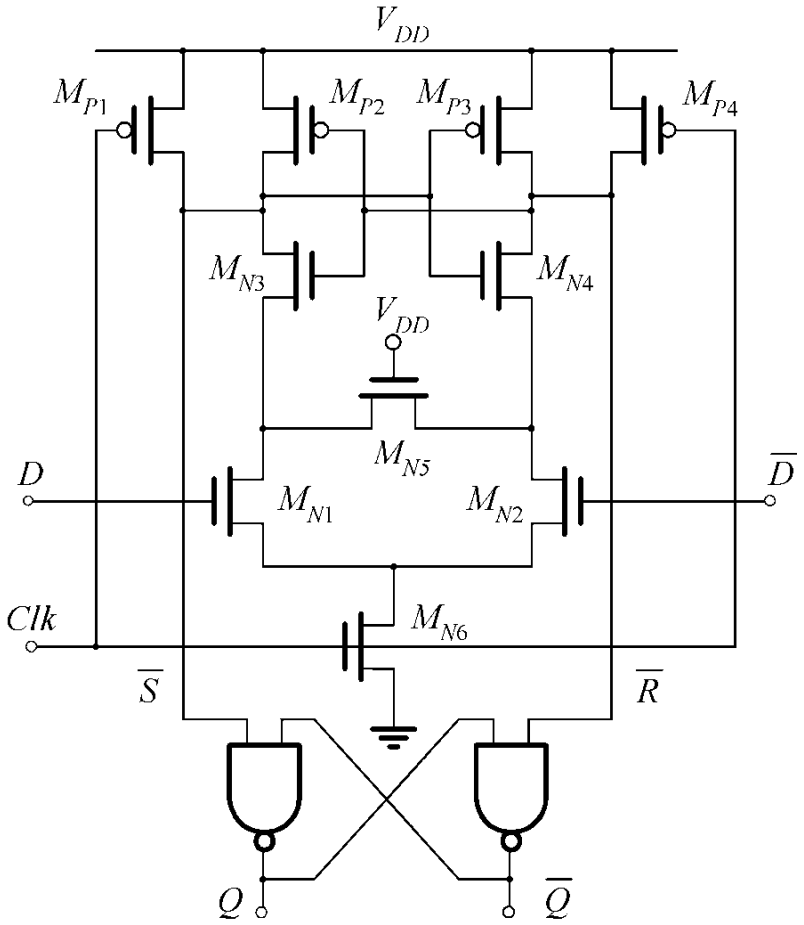
7474, from mid-1960's



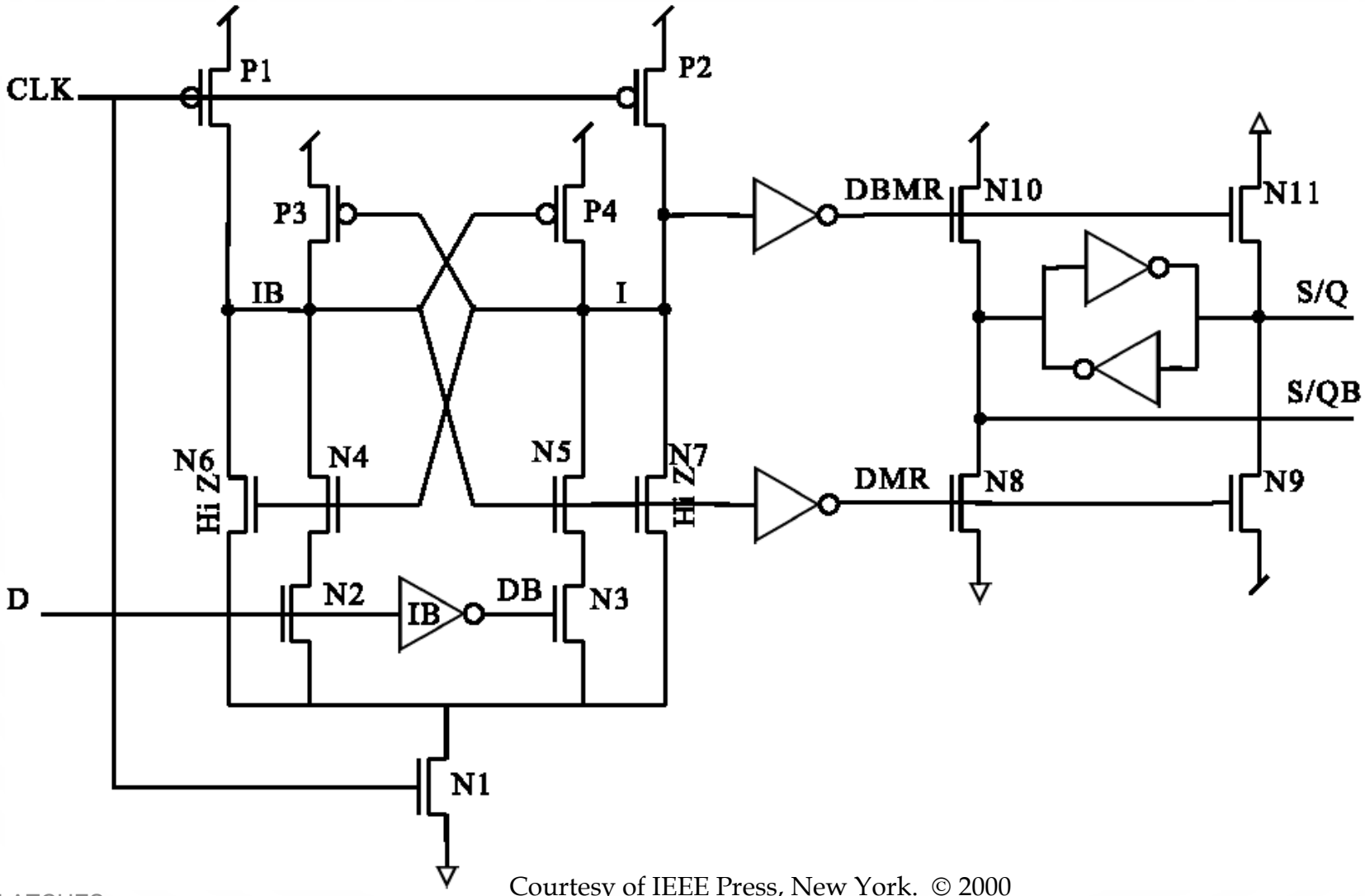
Pulsed Latches

Sense-amplifier-based flip-flop, Matsui 1992.
DEC Alpha 21264, StrongARM 110

First stage is a sense amplifier, precharged to high, when *Clk* = 0
After rising edge of the clock sense amplifier generates the pulse on
S or *R*
The pulse is captured in
S-R latch
Cross-coupled NAND has different propagation delays of rising and falling edges

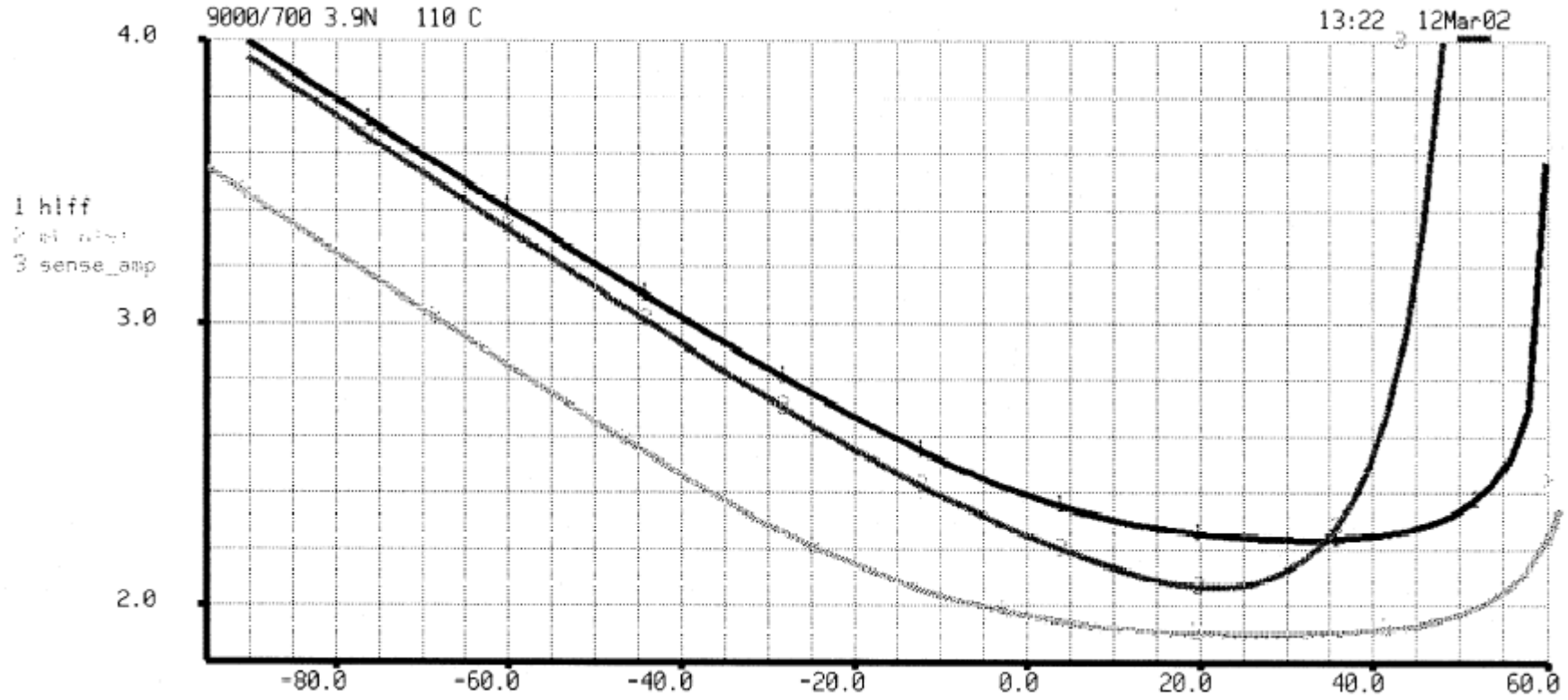


Sense Amplifier-Based Flip-Flop



Courtesy of IEEE Press, New York. © 2000

Sampling Window Comparison



Naffziger, JSSC 11/02

Summary

- Flip-flop-based (edge-triggered) timing dominates today
- Latch-based timing can increase performance, but needs extra care
- There is also asynchronous design
- Logical effort can be used to analyze latch timing

Next Lecture

- Variability