EECS 427
Lecture 23: Advanced interconnect techniques

Lecture Overview

- Reminders/Announcements:
  - CAD8 Due Tomorrow
  - HW5 Due & Course reviews on Tuesday
  - Quiz3 next Thursday
  - Final Demos 6 days from that
  - Presentations 6 days from that
- Why intra-chip communication matters
- How to make it more efficient

Communication vs. Computation: Delay

Operation	Delay (0.13um)	(0.05um)
32b ALU Operation	650ps	250ps
32b Register Read	325ps	125ps
Read 32b from 8KB RAM	780ps	300ps
Transfer 32b across chip (10mm)	1400ps	2300ps
Transfer 32b across chip (20mm)	2800ps	4600ps

2.5:1 global on-chip communication to computation delay
9:1 in 2010

From: Dally, HPCA02

Communication vs. Computation, Energy

Operation	Power (0.13um)	(0.05um)
32b ALU Operation	5pJ	0.3pJ
32b Register Read	10pJ	0.6pJ
Read 32b from 8KB RAM	50pJ	3pJ
Transfer 32b across chip (10mm)	100pJ	17pJ
Execute a uP instruction (SB-1)	1.1nJ	130pJ
Transfer 32b off chip (2.5G CML)	1.3nJ	400pJ
Transfer 32b off chip (200M HSTL)	1.9nJ	1.9nJ

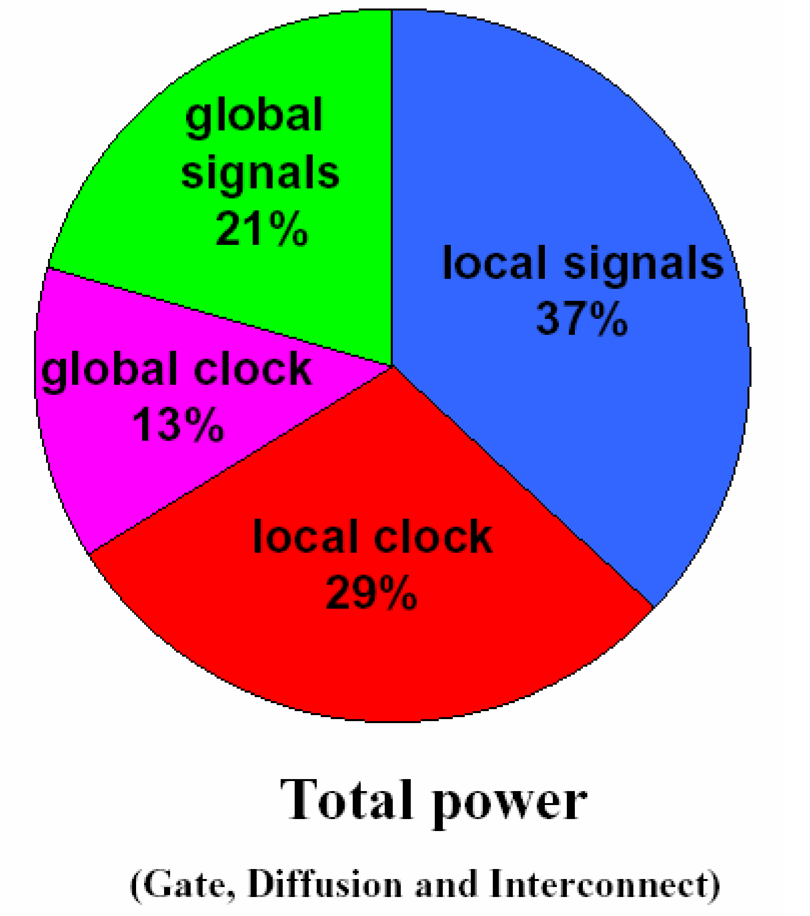
300:20:1 off-chip to global to local communication/computation energy

1300:56:1 in 2010

From: Dally, HPCA02

Status Today

- Repeater count has grown dramatically
- Repeaters are very wide with tight timing constraints
  - Lots of leakage
  - IBM: 50% of leakage in inverters/buffers
- Switching activities are typically low
  - Intel data from Pentium M: 0.05 average activity factor
- Both static and dynamic power are important for global signals



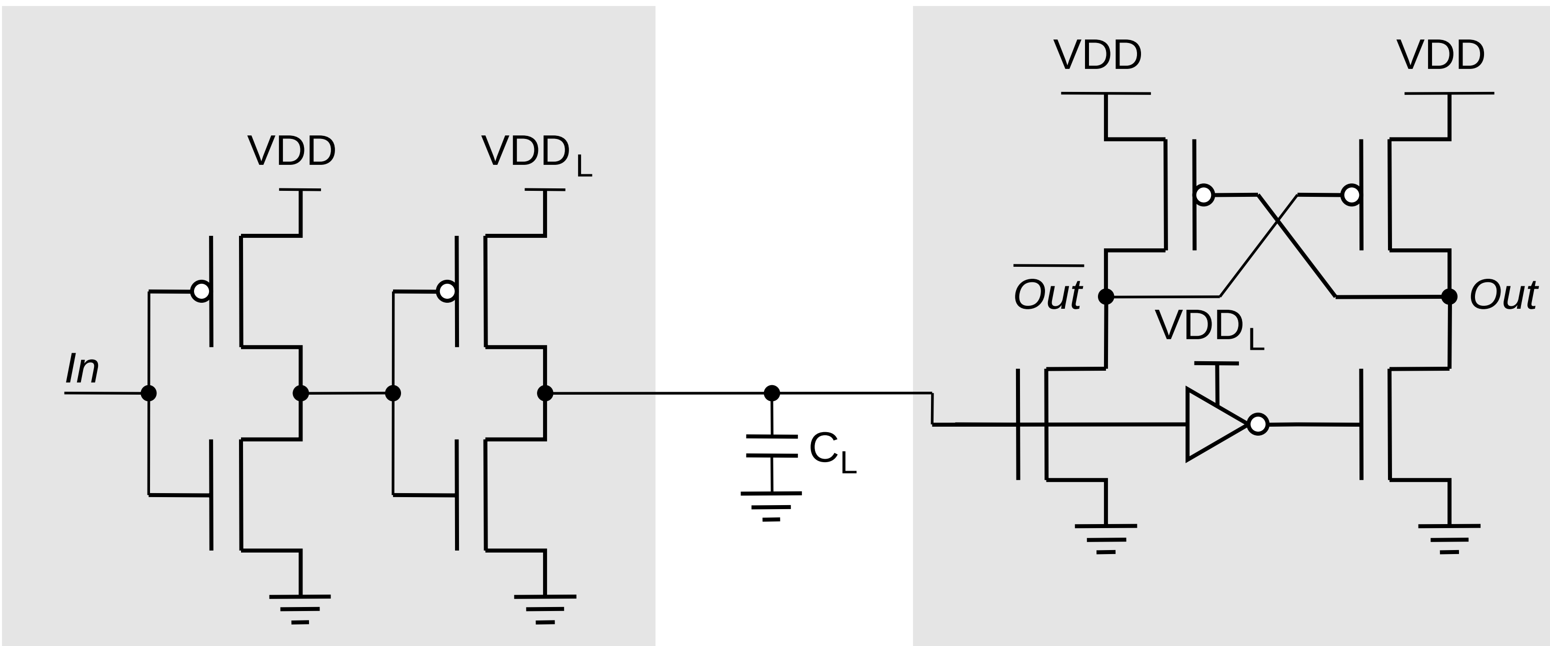
Pentium M power breakdown, [Nagen, SLIP04]

Reducing the swing

$$t_{pHL} = \frac{C_L\, V_{swing}/2}{I_{av}}$$

- Reducing the swing potentially yields linear reduction in delay
- Also results in reduction in power dissipation (from linear to quadratic depending on implementation)
- Delay penalty is paid by the receiver
  - Requires use of “sense amplifier” to restore signal level

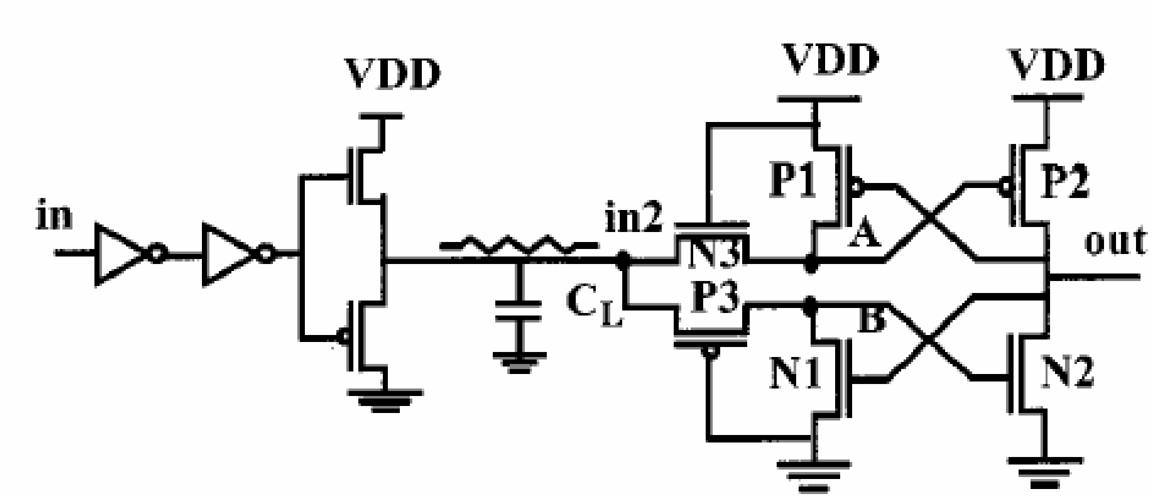
Single-Ended Static Driver and Receiver

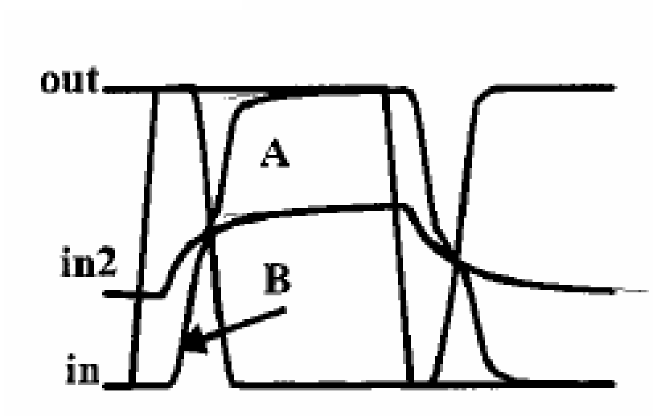


driver receiver

Can be expensive to have an extra voltage supply (becoming less difficult)

Symmetric Source-Follower Driver with Level Converter



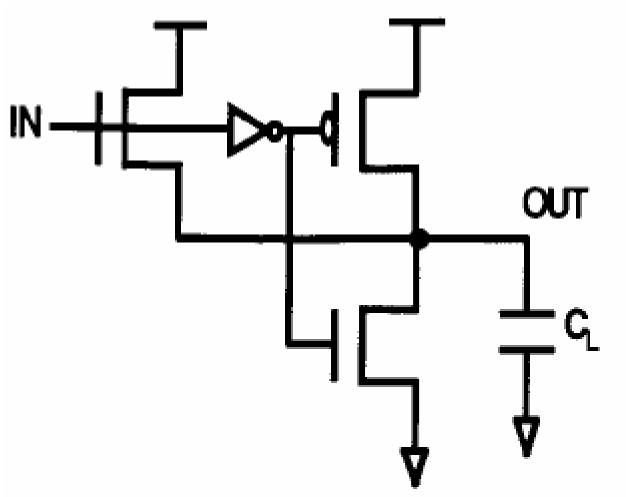


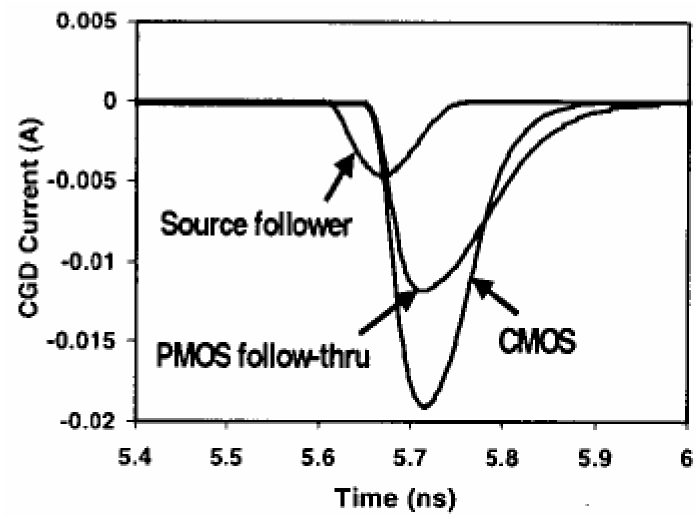
In goes low to high

In2 goes from Vth to Vdd-Vth (with body effect)

B goes to Vdd-Vth(body), turns on N2, pulls OUT low

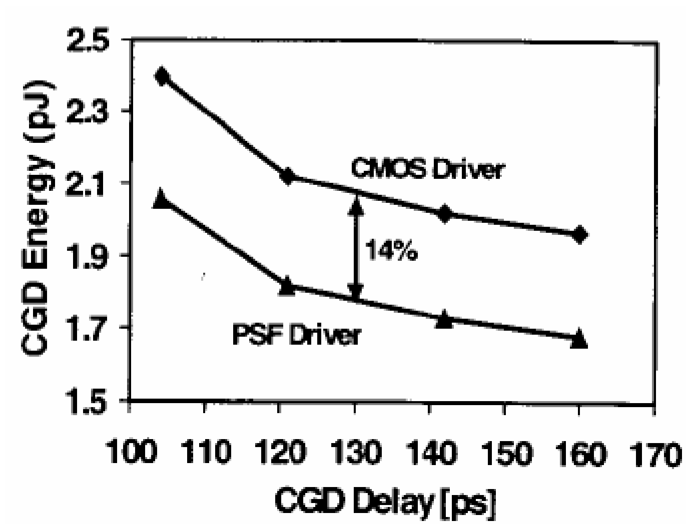
P-boosted source follower





Can make the PMOS pull-up fairly small, rely mainly on better drive of NMOS

Good for driving large capacitances (clock tree)

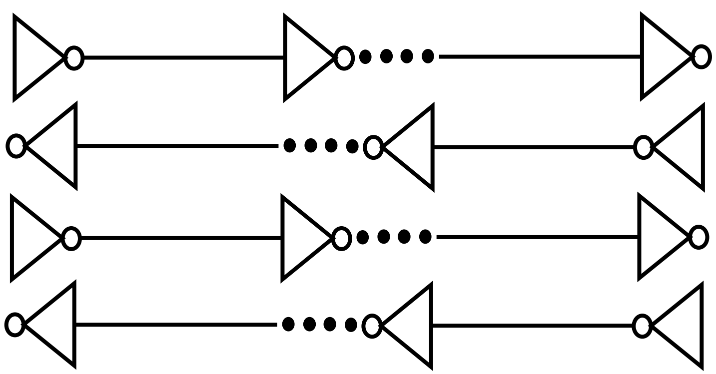


Intel, VLSI02

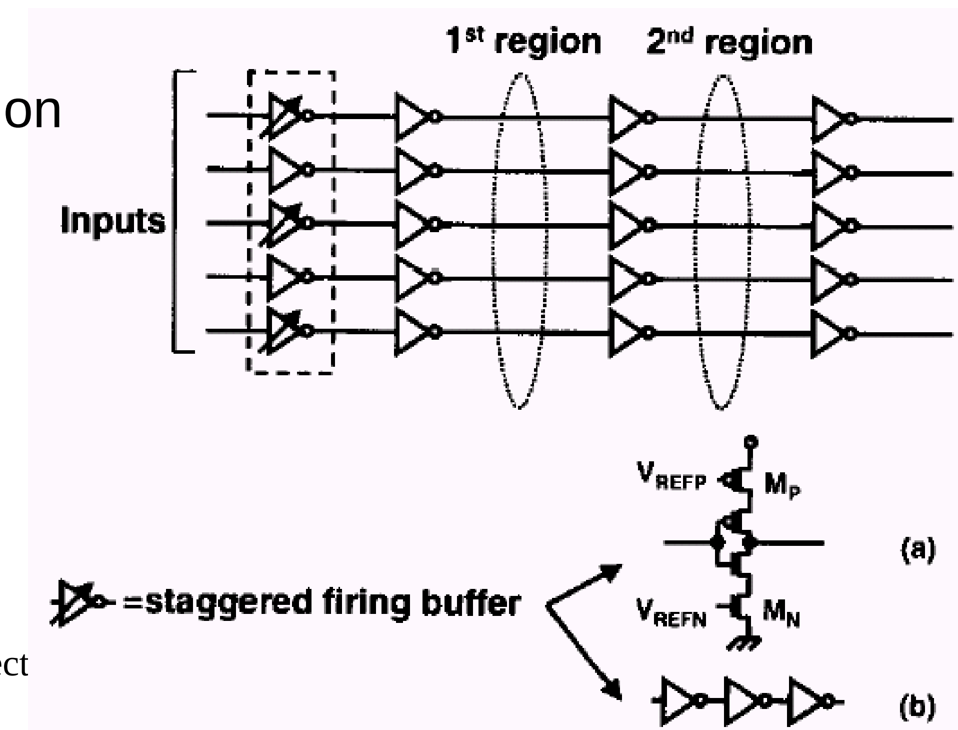
Alternate Signaling Techniques - Performance

- Reduce effective coupling capacitance
  - Insert shield wires
    - Impact on routing density
  - Interleave bidirectional buses
  - Staggered Firing Bus
    - Not feasible; process variation

Interleaved buses



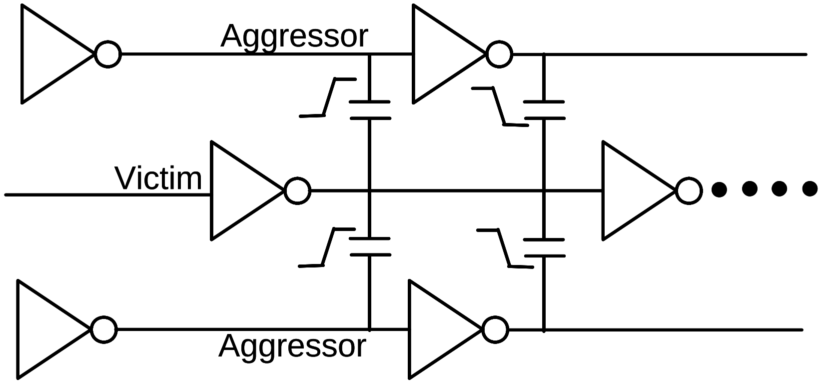
SFB schematic



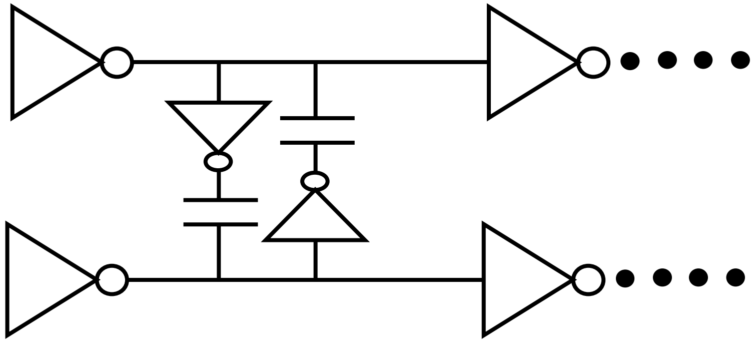
Alternate Signaling Techniques - Robustness

- Other techniques to reduce noise or increased delay arising from coupling capacitance:
  - Staggered repeaters to partition the line
  - Active noise cancellation; dump opposite polarity charge onto adjacent line to compensate

Staggered Repeaters



Active Noise Cancellation

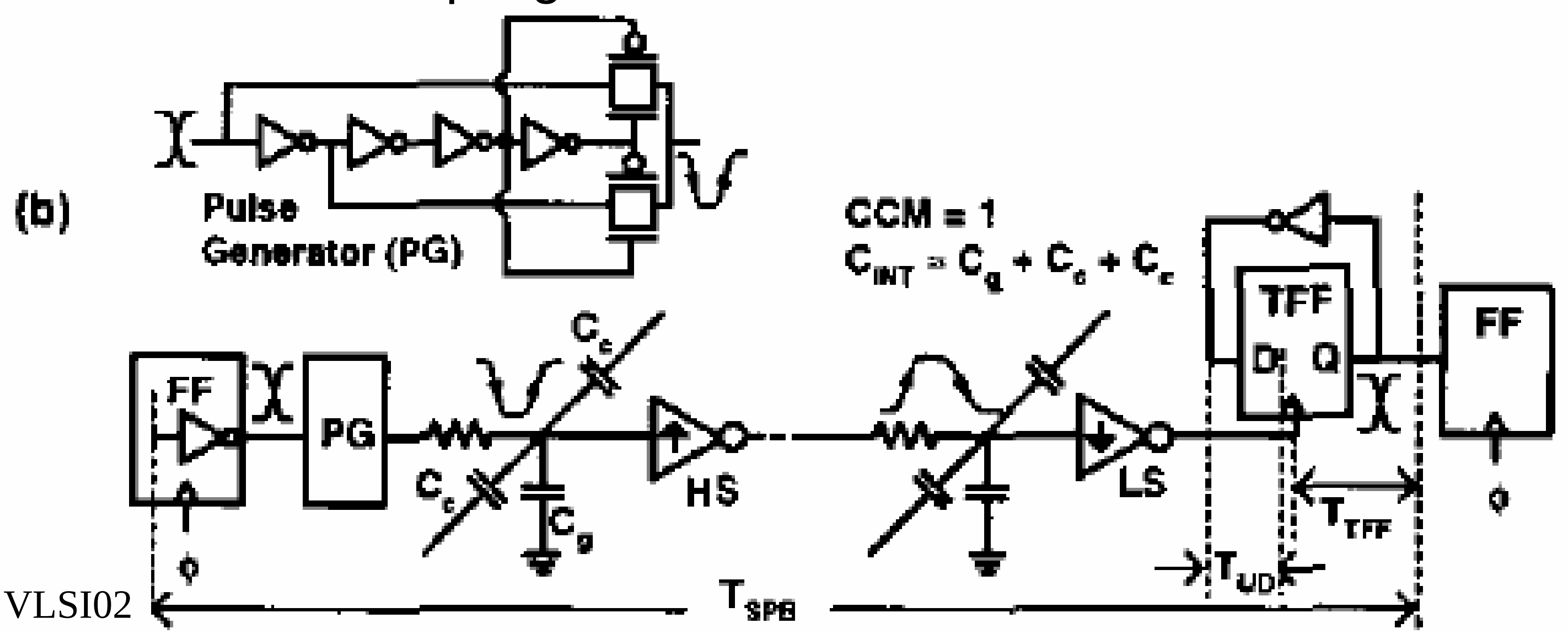


Static Pulsed Buses

PG creates a low-high-low pulse which propagates through the repeaters

Repeaters are skewed to create fast transitions on leading edge only (saving power)

No worst-case coupling effects since transitions are monotonic

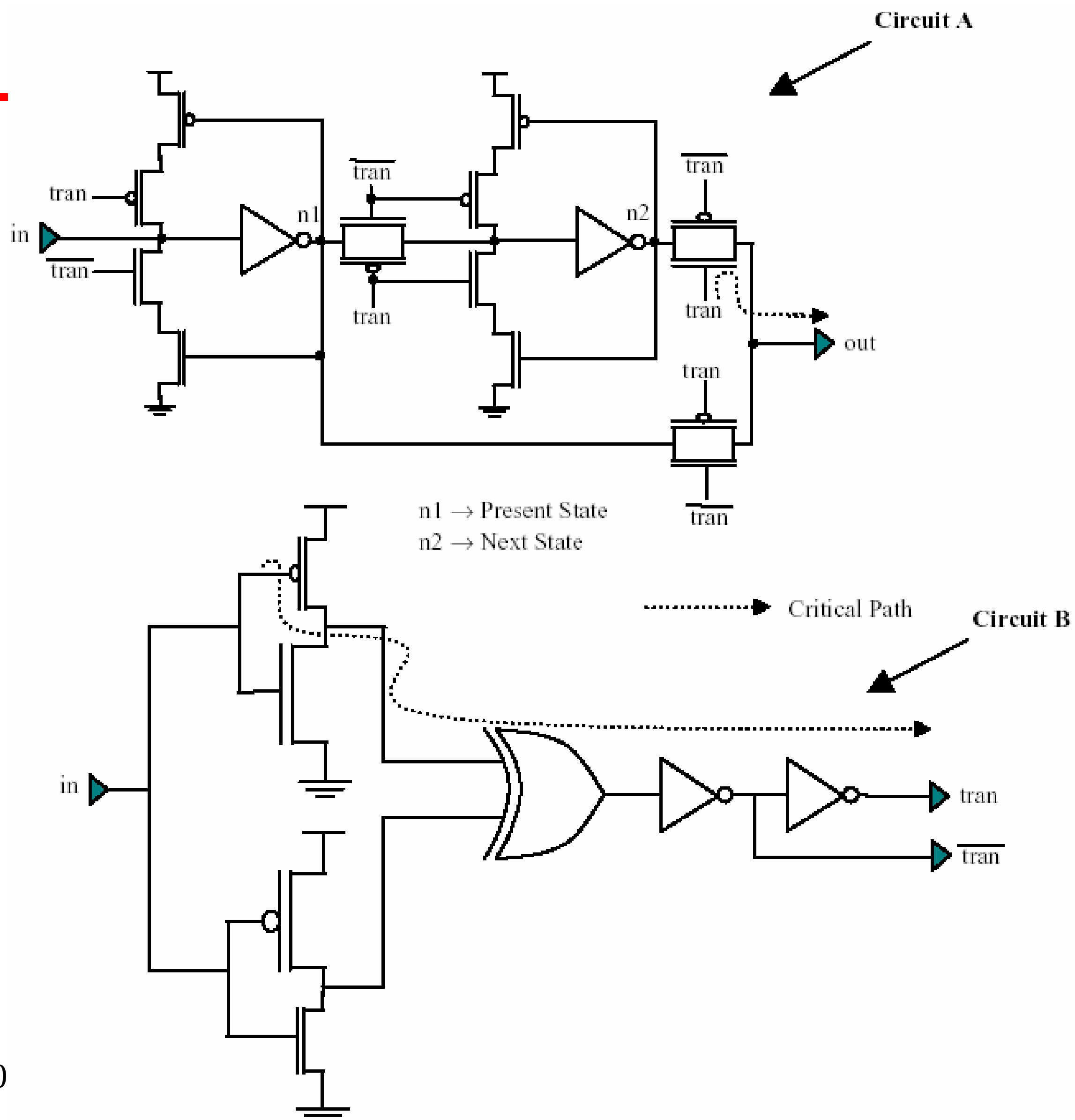


Intel, VLSI02

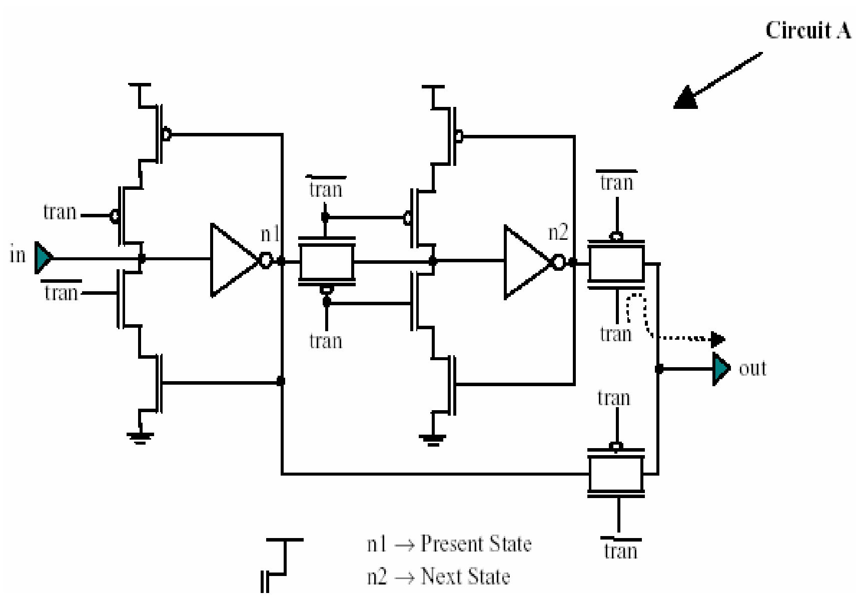
TAGS Concept

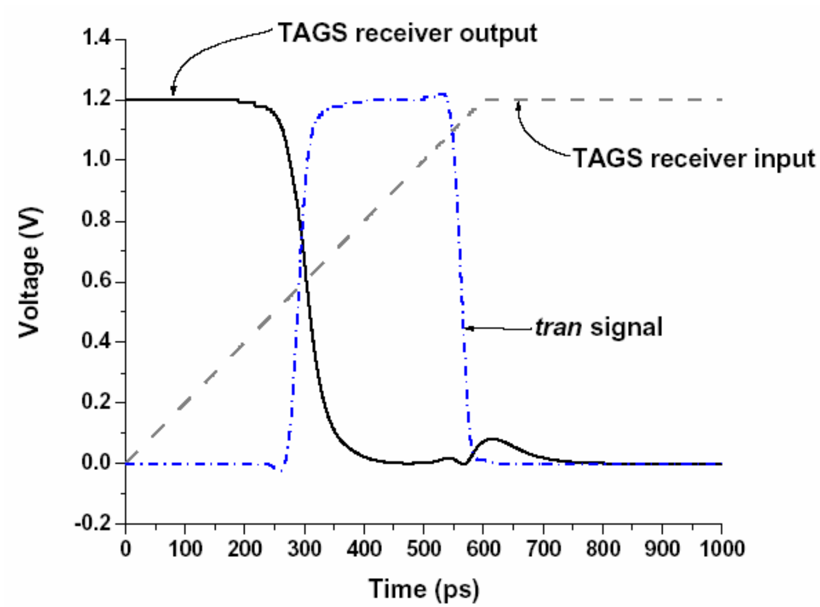
- Store next and present states of receiver output
- When line is quiet, connect output to present state
- Let transitions on line be slow
- On detection of transition, drive output to stored next state
- On completion of transition stored states flip and output connected back to present state
- Early detection of transition can improve delays (or increase unbuffered wire length)
- **T**ransition **A**ware **G**lobal **S**ignaling

TAGS receiver



Typical Waveforms





- Pulse generated at *tran*
  - Connects *out* to next state (*n2*)
  - Disconnects receiver from line
- Transition on line nears completion
  - *n1* is allowed to propagate through to *n2* (inverted)
  - Next and present states reset
- Slow transitions at *in* are allowable since *out* is driven by stored internal state

Summary

- Much of the delay and energy is going to signaling/communication
  - Lots of neat circuit tricks out there to help combat this, but it's still not enough…