EECS 427
Lecture 22: Advanced interconnect techniques

No dedicated reading

Last Time

- Clock distribution styles/goals
  - Reduce skew, area, power
  - H-trees are more common than grids today; grids consume too much power/area
  - On your project, try to pay at least some attention to how the clock is routed

Lecture Overview

- Why intra-chip communication matters
- How to make it more efficient
- Exam 2: Thurs 4/16
- HW6 posted shortly: Project presentations on Tues 4/21
  - Keys: 1-2 presenters per group, don't go over time, must be present to receive credit

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

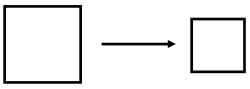
Operation	Power (0.13um)	(0.05um)
32b ALU Operation	5pJ	0.3pJ
32b Register Read	10pJ	0.6pJ
Read 32b from 8KB RAM	50pJ	3pJ
Transfer 32b across chip (10mm)	100pJ	17pJ
Execute a uP instruction (SB-1)	1.1nJ	130pJ
Transfer 32b off chip (2.5G CML)	1.3nJ	400pJ
Transfer 32b off chip (200M HSTL)	1.9nJ	1.9nJ

300:20:1 off-chip to global to local communication/computation energy
1300:56:1 in 2010

From: Dally, HPCA02

Interconnect Scaling Review

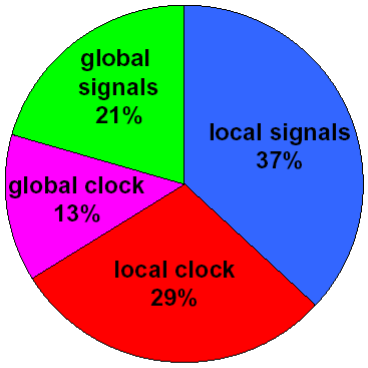
- Scale factor S ~ 1.4 per 2-3 years
  - Shrink linewidth, space, thickness, and dielectric thickness



- Local wires get shorter by S
  - Local wire C is reduced by S
  - Local wire R increases by $S^2/S = S$
  - Devices get faster by S, so constant wire delay appears slower each generation
- Worse for global wires that do not get shorter with scaling
  - Fortunately chip sizes have been fairly constant (not growing)
- Common solution: insert repeaters on global wires to reduce delay dependency on wirelength from L^2 to L

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



Total power

(Gate, Diffusion and Interconnect)

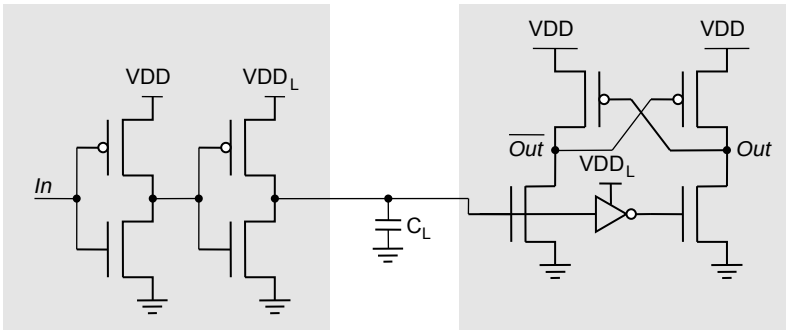
Pentium M power breakdown, [Nagen, SLIP04]

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 some sort of sense amplifier to restore signal level (a la memories)

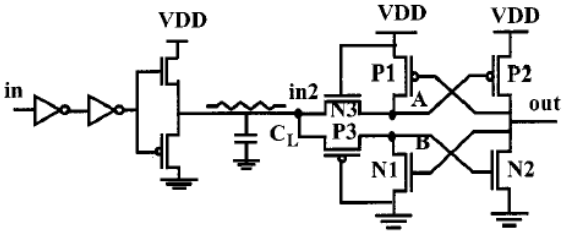
Single-Ended Static Driver and Receiver

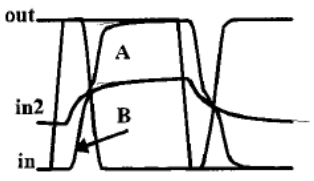


driver receiver

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

Symmetric Source-Follower Driver with Level Converter



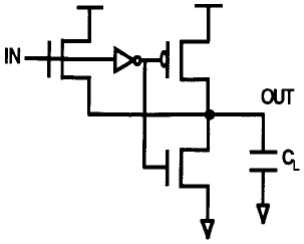


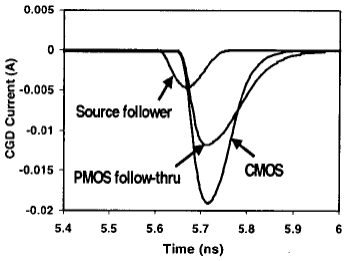
In goes low to high

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

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

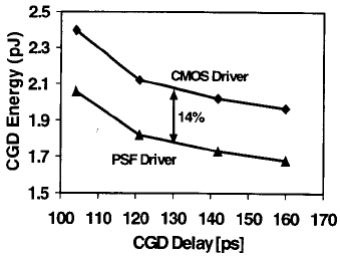
P-boosted source follower





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

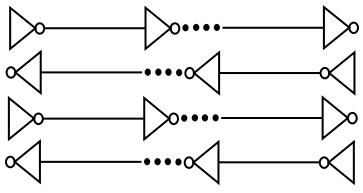
Good for driving large capacitances (clock tree)



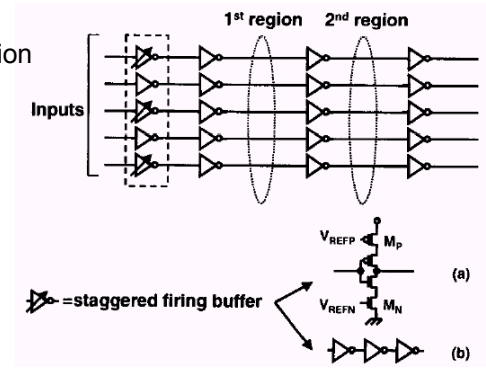
Alternate Signaling Techniques - Performance

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

Interleaved buses

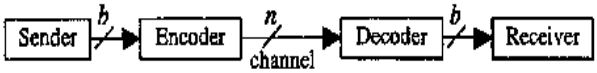


SFB schematic



Bus encoding to mitigate cross-coupling capacitance

- Codes data to be transmitted such that neighboring wires never switch in opposite directions
  - Employs encoder at driving end and decoder at receiving end



- Encoded codeword holds the information
  - Codeword transition controlled to ensure low cross-coupling

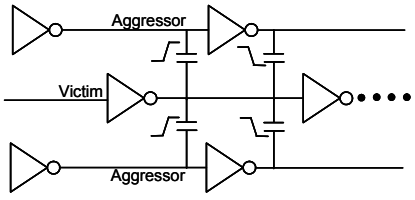
codeword at time 1:	0010	0000	0100	0100
	↓	↓	↓	↓
codeword at time 2:	0110	1111	0001	0010
	valid	*valid*	*valid*	**invalid**

- Provides better routing density than simply inserting power/ground shields

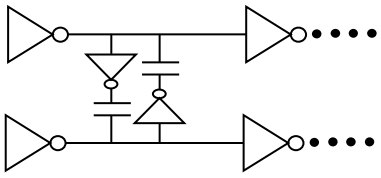
Alternate Signaling Techniques - Robustness

- 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

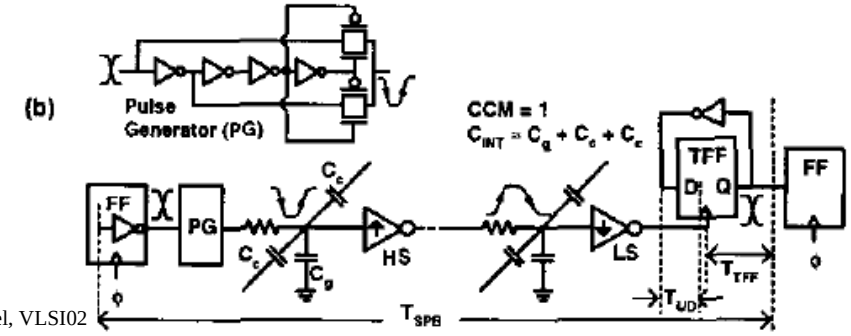


Static Pulsed Buses

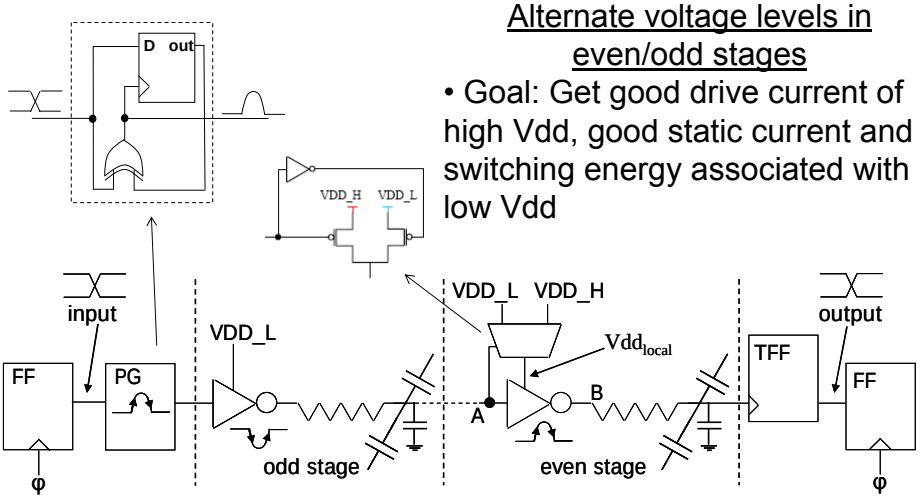
PG creates a low-high-low pulse which propagates through the repeaters

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

No worst-case coupling effects since transitions are monotonic



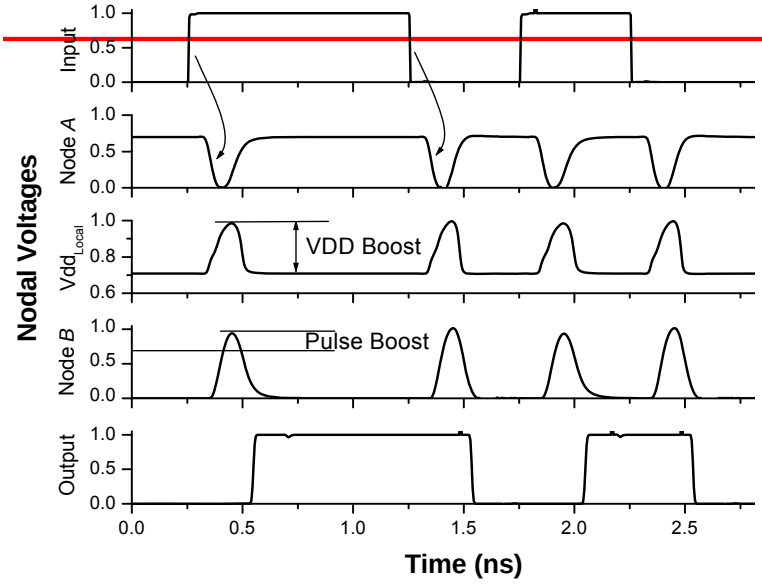
Dual-V_{DD} Boosted Pulsed Bus



Alternate voltage levels in even/odd stages

• Goal: Get good drive current of high Vdd, good static current and switching energy associated with low Vdd

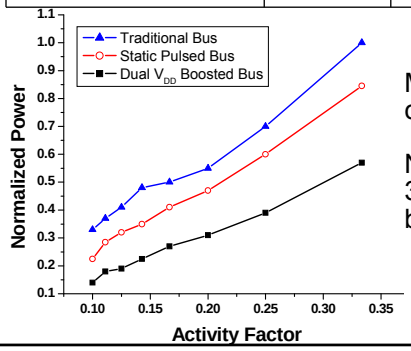
Boosting and Pulse Operation



Power/Performance Results

- Simulated results with VDDL = 0.7V

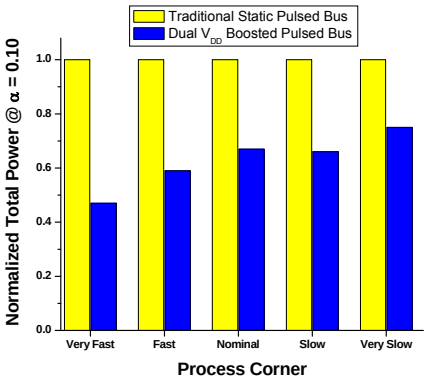
Bus Design Technique	Dynamic Power	Static Power	Delay
Traditional Static Bus	1.00	1.00	1.00
Static Pulsed Bus	0.76	0.78	0.97
Dual-VDD Boosted Pulsed Bus	**0.51**	**0.28**	**0.85**

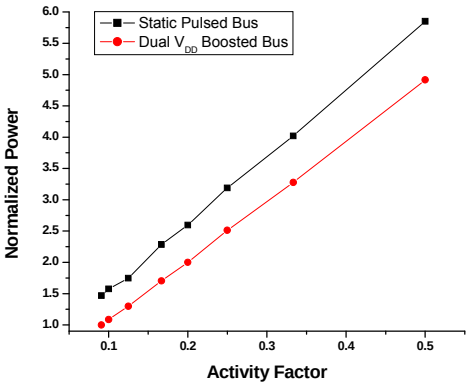


Maximum instantaneous current reduced by 60-64%

Noise margins reduced by 20-30mV vs. traditional repeater-based bus

Hardware Results



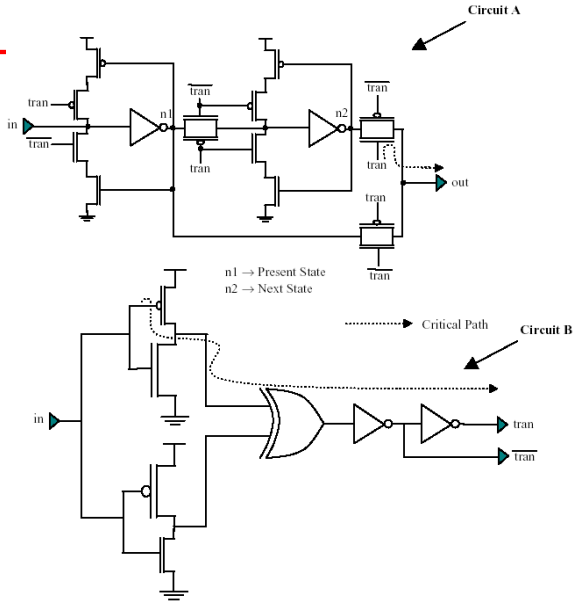


	Normalized Average Leakage	Standard Deviation of Leakage
Traditional Static Pulsed Bus	35.1	37.9
Dual-VDD Boosted Pulsed Bus Prototype	10.6	12.1
Leakage Savings of Dual-VDD Bus	**3.3X**	

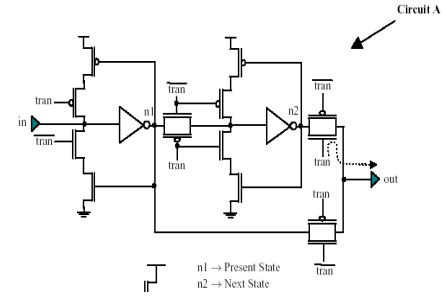
TAGS Concept

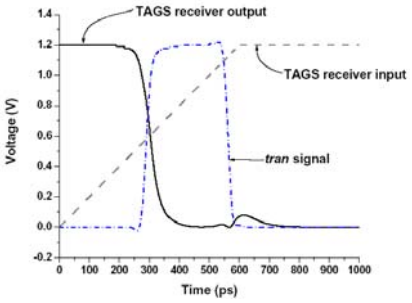
- Store next and present states of receiver output
- When line is quiet, connect output to present state
- Let transitions on line be slow
- On detection of transition, drive output to stored next state
- On completion of transition stored states flip and output connected back to present state
- Early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

TAGS receiver



Typical Waveforms

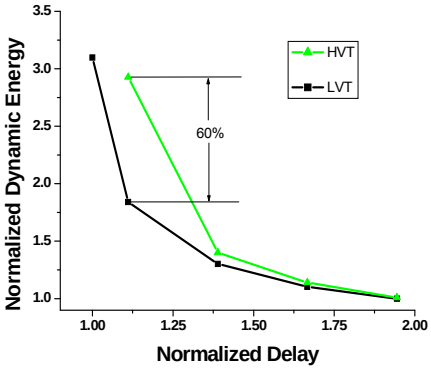




- Pulse generated at *tran*
  - Connects *out* to next state (*n2*)
  - Disconnects receiver from line
- Transition on line nears completion
  - *n1* is allowed to propagate through to *n2* (inverted)
  - Next and present states reset
- Slow transitions at *in* are allowable since *out* is driven by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

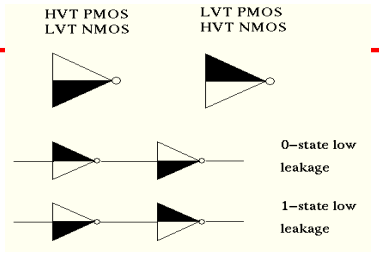
Leakage-Aware Bus Encoding

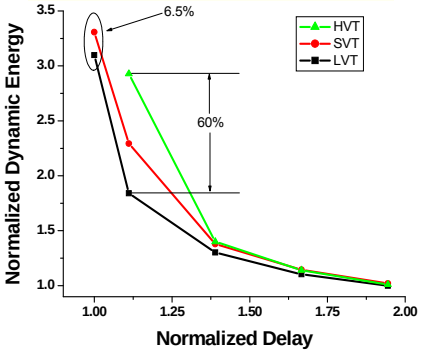
- Leakage in repeaters is significant as seen earlier
- Two device types commonly available – Low (High) Threshold Voltage L(H)VT
  - LVT → Good performance, much higher leakage
  - HVT → Low leakage, worse performance
  - Device sizing needed to achieve comparable performance
- For HVT vs. LVT, performance degradation measured against power reduction
  - HVT cannot meet best delay point of LVT
  - At 11% delay penalty, E_{dyn}(HVT) is 60% > E_{dyn}(LVT)
  - HVT unsuitable for high performance



Staggered V_t configuration

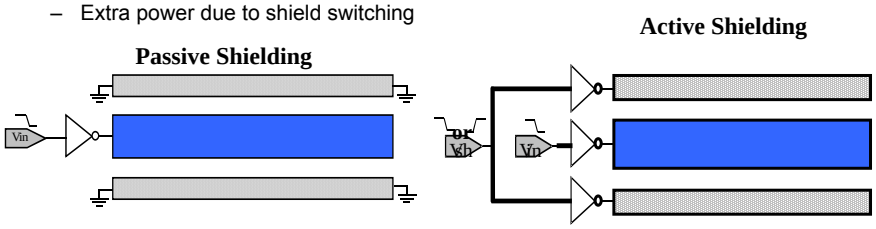
- Staggered Threshold Voltage (SVT) buffers
  - Alternating low/high V_t NMOS/PMOS devices
  - Separate 0/1 low-leakage states
- LVT devices provide higher speed and HVT devices have better leakage
  - LVT reduces sizing requirements on HVT
- Optimal configuration: Highly probable states ↔ Low-leakage configuration
  - Use bus encoding to enforce this condition





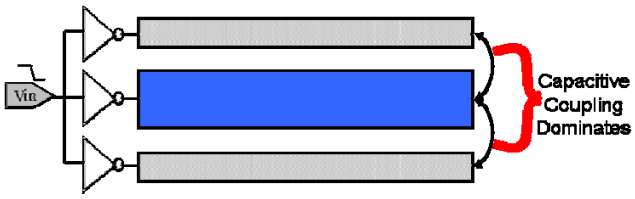
Active Shielding

- **Current paradigm for global wiring**
  - Place GND/VDD shields next to critical signals to limit capacitive and inductive coupling
  - Recommended: 1 shield for every 2 wires (Morton ISSCC'02)
- **Good for**
  - Eliminating capacitive noise
  - Creating stable return paths : Limits self and mutual inductances
- Can we *use* the coupling capacitance and mutual inductance with the shields to optimize delay and/or reduce inductive effects?
  - Use switching activity + layout for optimization
  - Extra power due to shield switching



Active Shielding

- Switch shields in-phase instead!
  - Reduces effective coupling capacitance
  - Speed up transition times and delays
  - Have to watch out for increased ringing (effective inductance rises)



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…