EECS 427
Lecture 12: Low-power ALUs
Reading: 11.7.1

Lecture Overview

- Low power design
  - Parallel vs. pipelined datapaths for lower power
  - Multi-Vdd design
- Quiz in class on Thursday
  - Open notes/book

Last Time

- Barrel shifters are area-intensive but have only 1 pass transistor per path
  - Lots of junction capacitance though
- Log shifters are more versatile for wider data
  - Various choices: log base, reverse order, pass transistor vs. T-gate, buffering
- Low-power design
  - Can focus on reducing any # of things from switching activity to cap to voltage
  - Extremely important today; everyone cares about power

Active Power Reduction

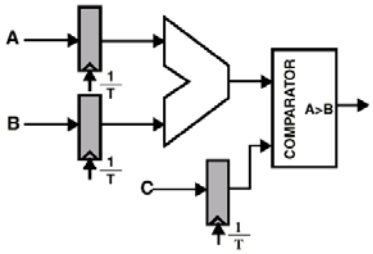
$$P \sim \alpha \cdot C_L \cdot V_{swing} \cdot V_{DD} \cdot f \qquad E \sim \alpha \cdot C_L \cdot V_{swing} \cdot V_{DD}$$

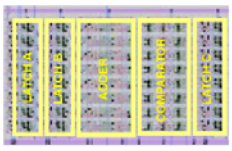
- Reducing load capacitance
  - Technology scaling
  - Gate sizing, logic minimization, better placement tools
  - Logic families (pass transistor logic, …)
- Reducing supply voltage
  - Quadratic impact on power
  - Impact on delay – how to maintain throughput?
- Reducing frequency – performance penalty
- Reducing switching probability (α)
  - Architecture
  - Glitching power reduction (15-20%)

Two types of processing

- Fixed-rate processing (signal processing for multimedia or communications)
  - Stream-based computation
  - No advantage in obtaining throughput in excess of the real-time constraint
- Variable-rate or burst-mode computation (general purpose computation)
  - Mostly idle (or low-load) with bursts of computation
  - Faster is better

Architecture Tradeoff for Fixed-rate Processing Reference Datapath



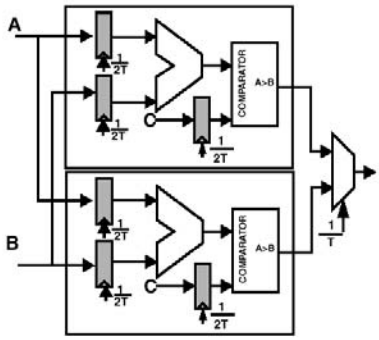


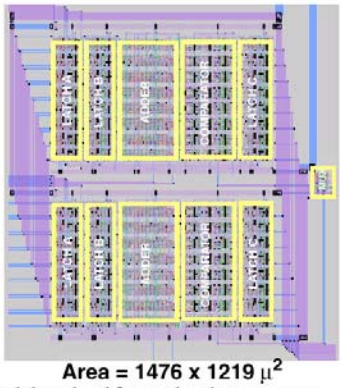
Area = 636 x 833 μ^2

- Critical path delay $\Rightarrow T_{adder} + T_{comparator}$ (= 25ns)
  $\Rightarrow f_{ref}$ = 40Mhz
- Total capacitance being switched = C_{ref}
- $V_{dd} = V_{ref} = 5V$
- Power for reference datapath = $P_{ref} = C_{ref} V_{ref}^2 f_{ref}$

from [Chandrakasan92] (*IEEE JSSC*)

Parallel Datapath

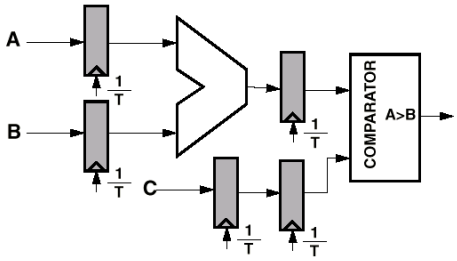


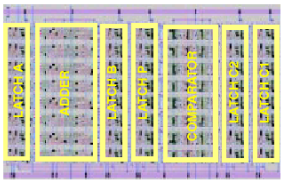


Area = 1476 x 1219 μ^2

- The clock rate can be reduced by half with the same throughput $\Rightarrow f_{par} = f_{ref} / 2$
- $V_{par} = V_{ref} / 1.7$, $C_{par} = 2.15C_{ref}$
- $P_{par} = (2.15C_{ref})\ (V_{ref}/1.7)^2\ (f_{ref}/2) \approx 0.36\ P_{ref}$

Pipelined Datapath





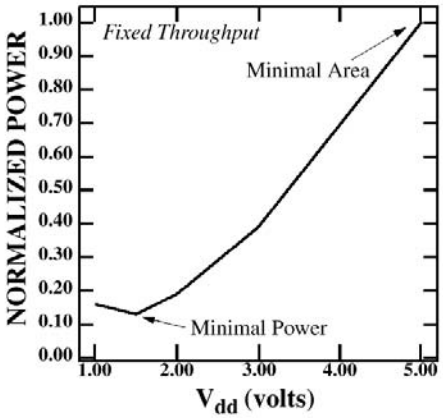
Area = 640 x 1081 μ^2

- Critical path delay is less $\Rightarrow$ max [T_{adder} , $T_{comparator}$]
- Keeping clock rate constant: $f_{pipe} = f_{ref}$
  Voltage can be dropped $\Rightarrow V_{pipe} = V_{ref} / 1.7$
- Capacitance slightly higher: $C_{pipe} = 1.15C_{ref}$
- $P_{pipe} = (1.15C_{ref})\ (V_{ref}/1.7)^2\ f_{ref} \approx 0.39\ P_{ref}$

A Simple Datapath: Summary

Architecture type	Voltage	Area	Power
Simple datapath (no pipelining or parallelism)	5V	1	1
Pipelined datapath	2.9V	1.3	0.39
Parallel datapath	2.9V	3.4	0.36
Pipeline-Parallel	2.0V	3.7	0.2

How Low a Voltage can be Used?



- Capacitance overhead starts to dominate at "high" levels of parallelism and results in an optimum voltage

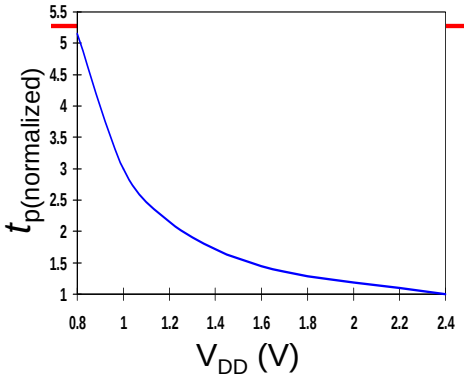
Power and Energy Design Space

	Constant Throughput/Latency		Variable Throughput/Latency
Energy	Design Time	Non-active Modules	Run Time
Active	Logic Design Reduced V_{dd} Sizing Multi-V_{dd}	Clock Gating	DFS, DVS (Dynamic Freq, Voltage Scaling)
Leakage	+ Multi-V_T	Sleep Transistors Variable V_T	+ Variable V_T

Supply Voltage Scaling

- How to maintain throughput under reduced supply?
- Introducing more parallelism/pipelining
  - Area increase – cost increases
  - Cost/power tradeoff
- **Multiple voltage domains**
  - Separate supply voltages for different blocks
  - Lower VDD for slower blocks
  - Cost of DC-DC converters or additional off-chip supplies, distributing multiple power supplies on-chip
- Dynamic voltage scaling – with variable throughput
- Reduce V_{th} to improve speed
  - Exponentially increased leakage eventually dominates

Delay as a Function of V_{DD}



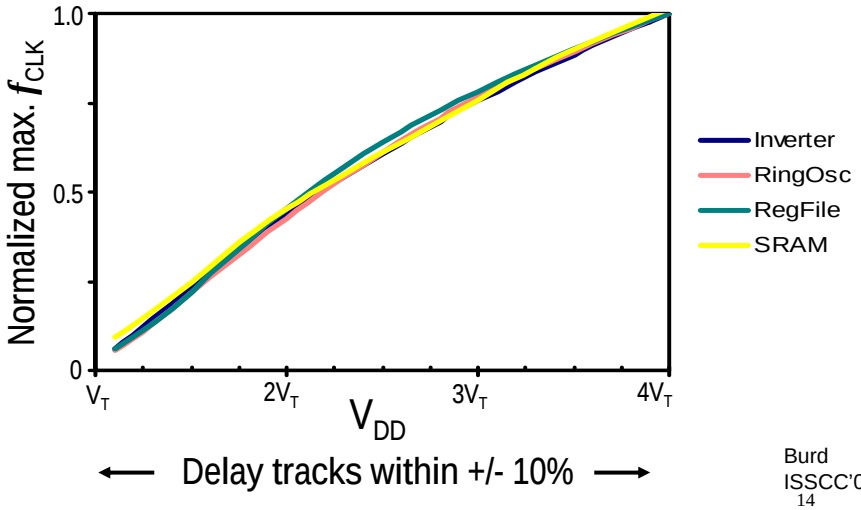
$$T_d = \frac{C_L * V_{dd}}{I}$$

$$I \sim (V_{dd} - V_t)^{1.3}$$

$$\frac{T_{d(Vdd=1.5)}}{T_{d(Vdd=2.5)}} = \frac{(1.5) * (2.5 - 0.4)^{1.3}}{(2.5) * (1.5 - 0.4)^{1.3}} \approx 1.4$$

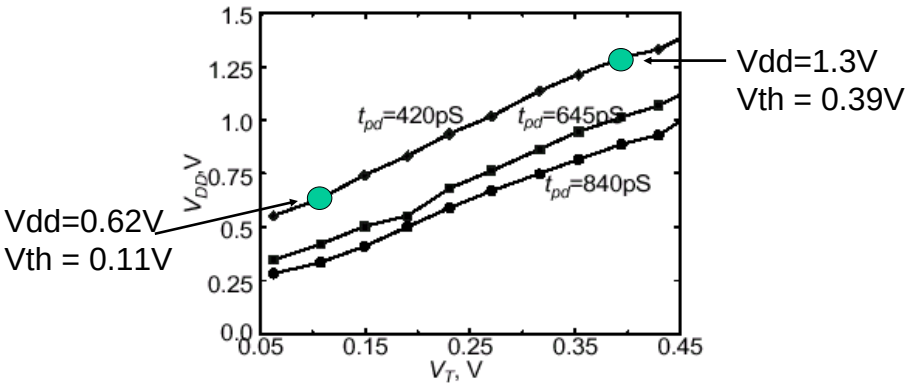
- Decreasing V_{DD} reduces dynamic energy consumption quadratically
- But increases gate delay (decreases performance)
- Determine critical path(s) at **design time** & use high V_{DD} for transistors on those paths for speed. Use lower V_{DD} on other gates

CMOS Circuits Track Over V_{DD}



← Delay tracks within +/- 10% →

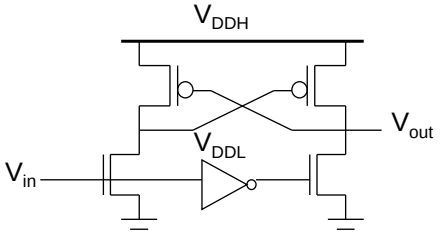
Changing V_{dd} and V_{th} Together



Contours of constant delay show that reductions in V_{th} must accompany smaller V_{dd}'s to maintain speed

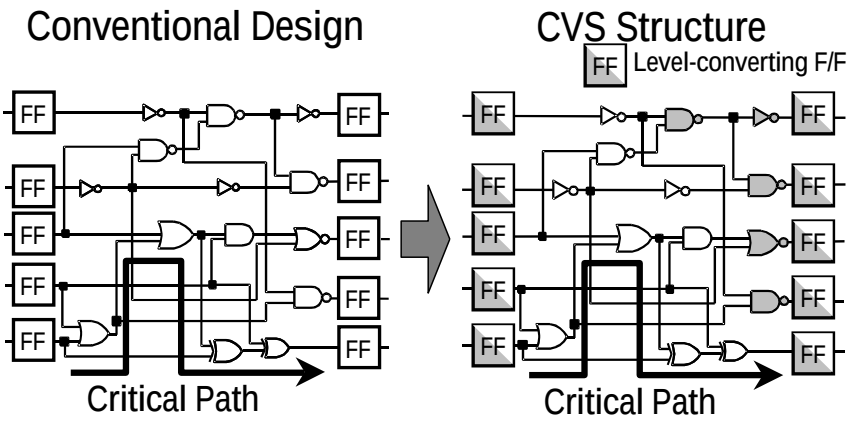
Multiple V_{DD} Considerations

- How many V_{DD}? – 2 is becoming more popular
  - Many chips already have 2 supplies (1 for core and 1 for I/O)
- When combining multiple supplies, **level converters** are required when a module at lower supply drives gate at higher supply (step-up)
  - If a gate supplied with V_{DDL} drives a gate at V_{DDH}, PMOS never turns off
    - Cross-coupled PMOS transistors perform the level conversion
    - NMOS transistors operate at reduced supply
  - Level converters are **not** needed for step-down changes in voltage



  - Overhead of level converters can be reduced by converting at register boundaries & embedding level conversion inside the flop

Multiple Vdd Design



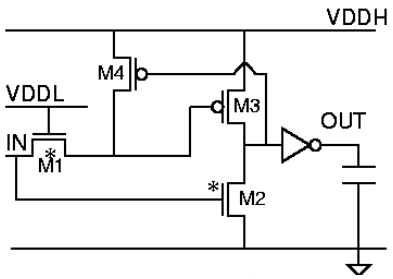
Lower V_{DD} portion is shaded

"Clustered voltage scaling"

M.Takahashi, ISSCC'98.

Existing Level Converters

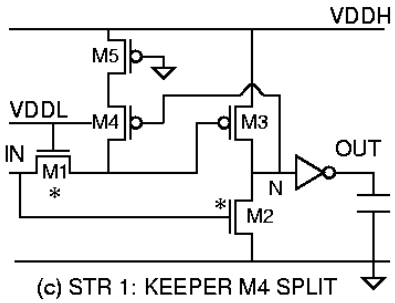
- DCVS (2 slides ago)
- Pass gate (PG)



* = low-Vth candidate

- DCVS – Higher power dissipation due to greater contention and higher transistor count
- PG – Simpler design, faster, lower power than DCVS, critical path is falling input (and output)
  - Key: Purpose of M1

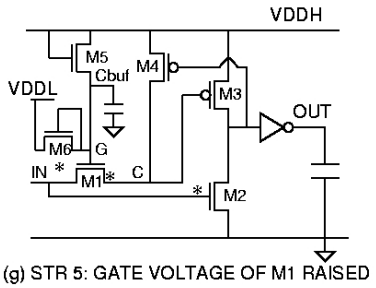
Alternate LC: STR1



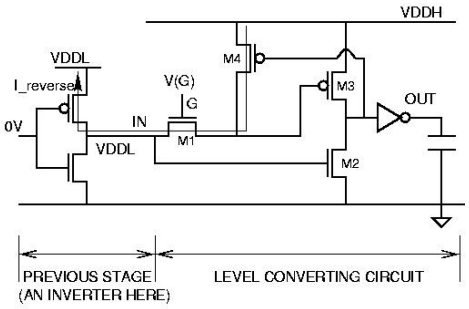
(c) STR 1: KEEPER M4 SPLIT

- Known high-performance design technique, with much improved results in this application space
- Keeper M4 from PG split into M4 and M5
- Reduced loading on node N and reduced contention

Alternate LC: STR5



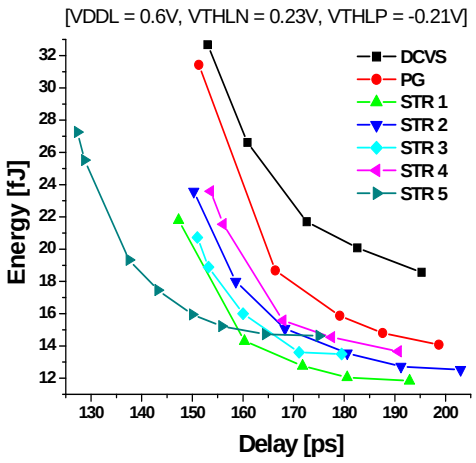
(g) STR 5: GATE VOLTAGE OF M1 RAISED



- Raised gate voltage on pass transistor boosts performance
- Leakage current I_reverse creates tradeoff between power and speed

Simulation Results

- STR1 consumes about 40-50% less energy
- STR1 about 3-4% faster than DCVS and PG
- STR5 best overall, especially at fast speeds
  - Gains diminish at low switching activities due to reverse leakage
  - Overall: Delay of LCs is <2 F04 delays



Summary

- If voltage is scalable, parallelism and pipelining are good levers to reduce dynamic power
  - For a fixed voltage, pipelining is *bad* for power though
- Multi-Vdd is a powerful knob to exploit path imbalances
  - Complicated: level conversion, layout, power distribution issues