Lecture

3

ECE216B

Circuit Optimization

Prof. Dejan Marković

ee216b@gmail.com

Agenda

- **Understand E-D tradeoffs**
  - Sensitivity-based opt.
- **Formulate optimization**
  - Energy & delay models
  - Sensitivity analysis
- **Insights & further steps**
  - Which variables to use
  - Extension to arch. space

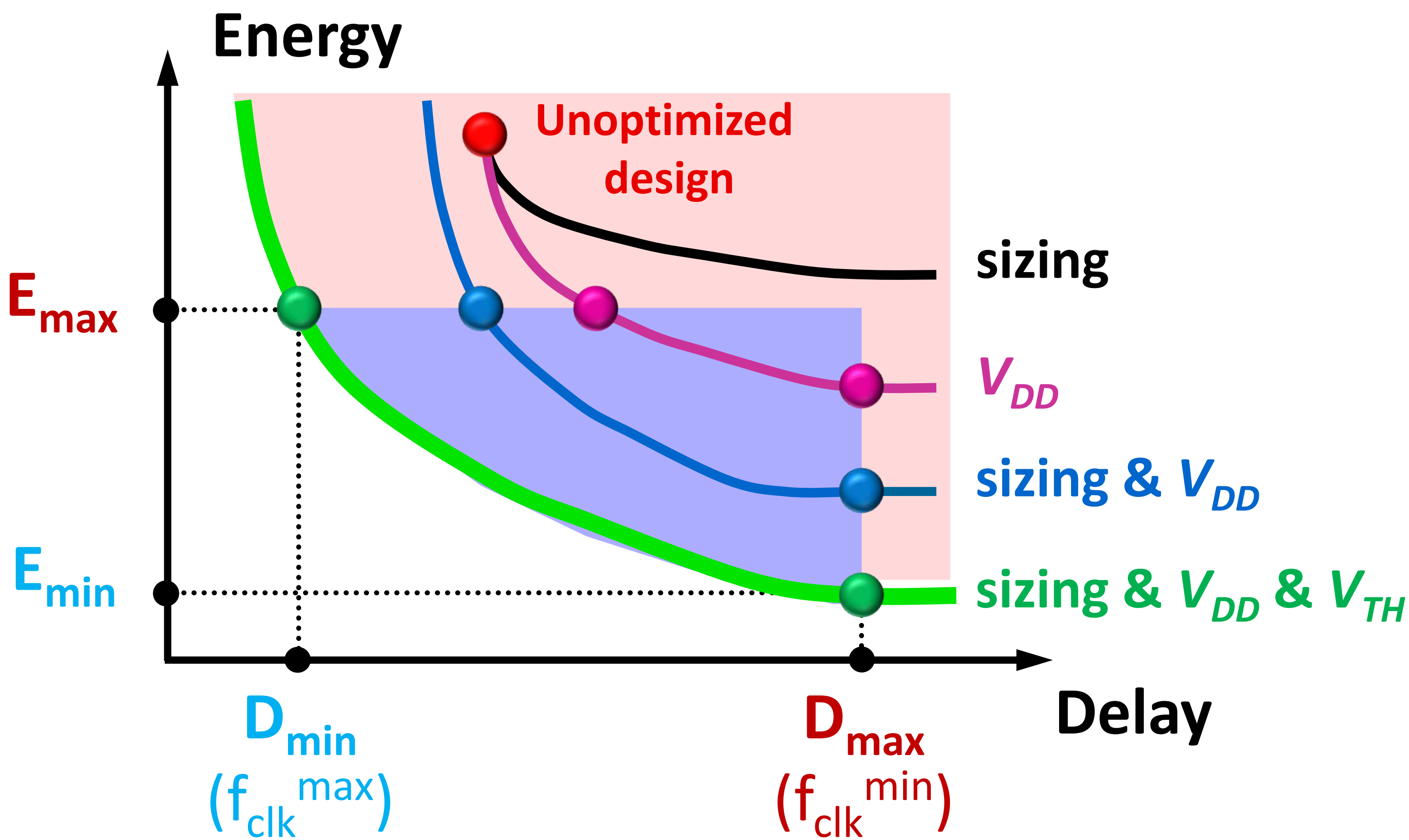
Some Common Questions

- Is sizing better than V_{DD} for energy reduction?
- Optimal values of gate size and V_{DD}?
- Increase or decrease V_{DD} for energy reduction?
- Optimal ratio of leakage / switching for min E?
- Optimal circuit topology?
- Etc.

Energy-Delay Optimization

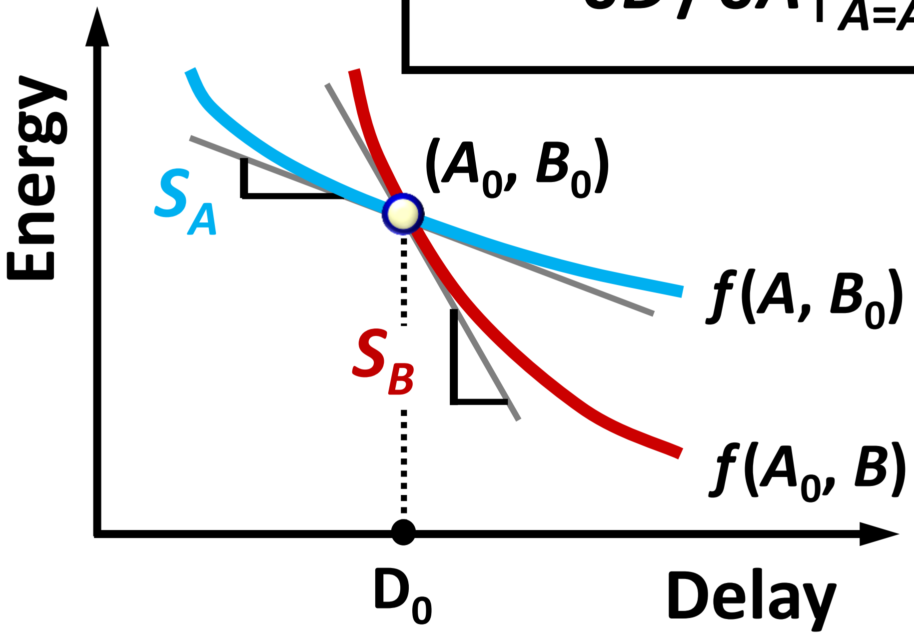
Equivalent formulations

- **Lowest E under D constraint?**
- **Best D under E constraint?**



Energy-Delay Sensitivity

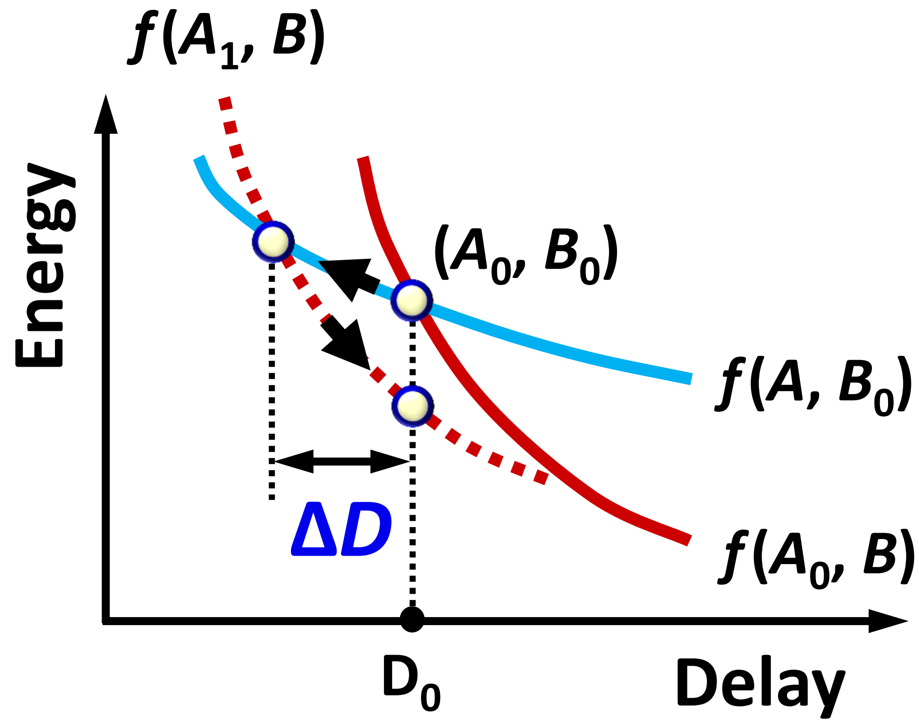
$$S_A = \left.\frac{\partial E/\partial A}{\partial D/\partial A}\right|_{A=A_0}$$



Slope of E-D curve around a design point (e.g. (A_0,B_0))

Solution: Equal Sensitivities

$$\Delta E = S_A \cdot (-\Delta D) + S_B \cdot \Delta D$$



A fixed point is reached when all sensitivities are equal

Circuit-Level Optimization

Objective: minimize E

$E = E(V_{DD}, V_{TH}, W)$

Constraint: Delay

$D = D(V_{DD}, V_{TH}, W)$

Tuning variables

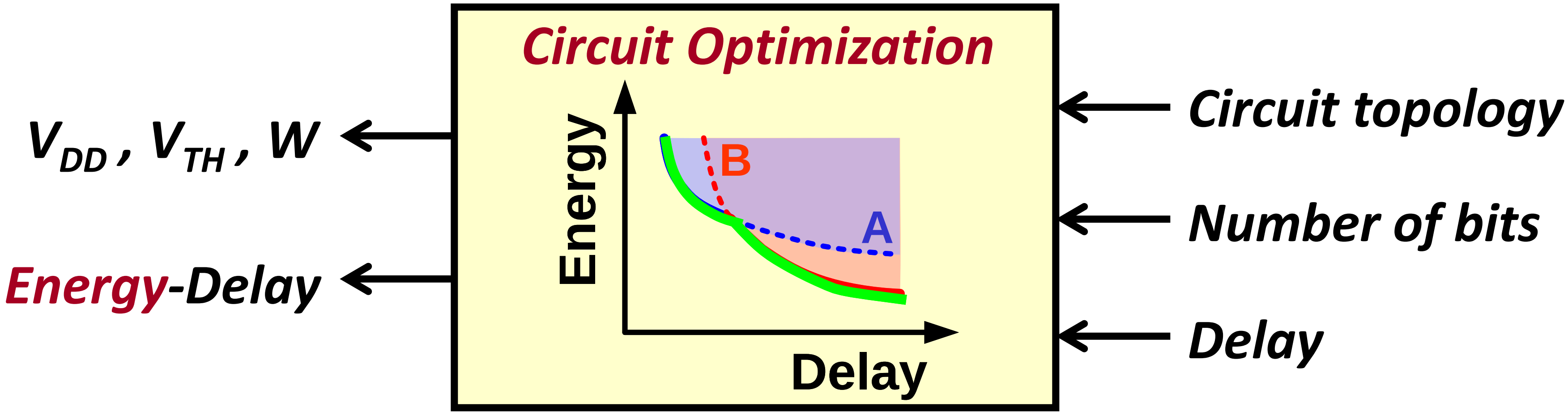
V_{DD}, V_{TH}, W

Constraints

$V_{DD}^{min} < V_{DD} < V_{DD}^{max}$

$V_{TH}^{min} < V_{TH} < V_{TH}^{max}$

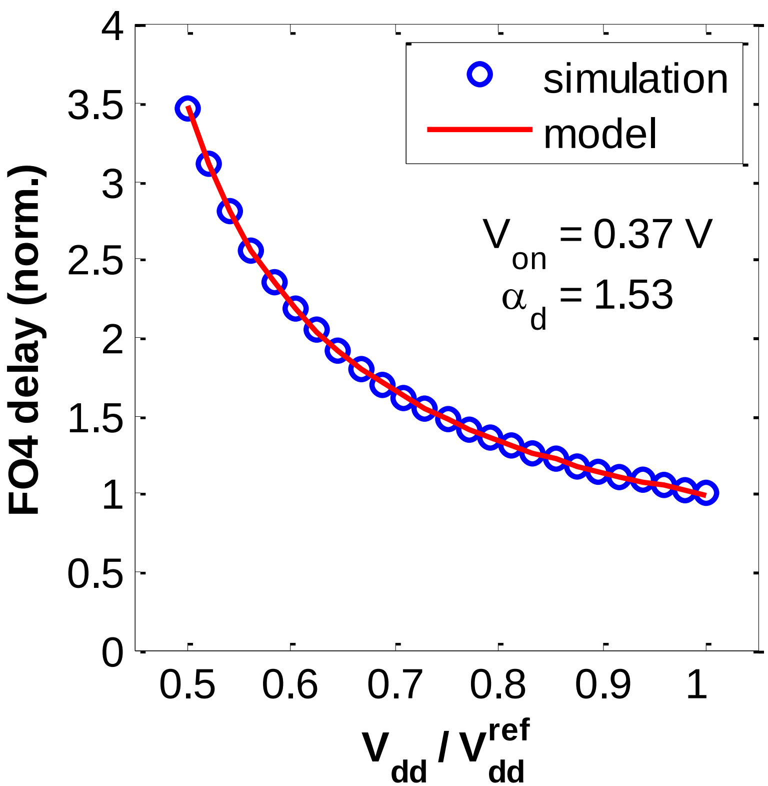
$W^{min} < W$



Energy & Delay Models

Alpha-power based Delay Model

$$Delay = \frac{K_d \cdot V_{DD}}{(V_{DD} - V_{on} - \Delta V_T)^{\alpha_d}} \cdot \left(\frac{W_{out}}{W_{in}} + \frac{W_{par}}{W_{in}} \right)$$



Fitting parameters:
V_{on}, α_d, K_d

V_{DD}^{ref} = 1.2V

FO4 (V_{DD}^{ref}) = 25ps

(*) [Sutherland et al., Logical Effort, 1999]

Energy Model for Circuit Optimization

- **Switching Energy**

$$E_{sw} = \alpha_{0\to 1} \cdot \left(C(W_{out}) + C\left(W_{par}\right)\right) \cdot V_{DD}{}^{2}$$

- **Leakage Energy**

$$E_{lk} = \frac{W_{in}}{W_0} \cdot I_0(S_{in}) \cdot 10^{-\frac{V_T - \gamma_D V_{DD}}{S}} \cdot V_{DD} \cdot D$$

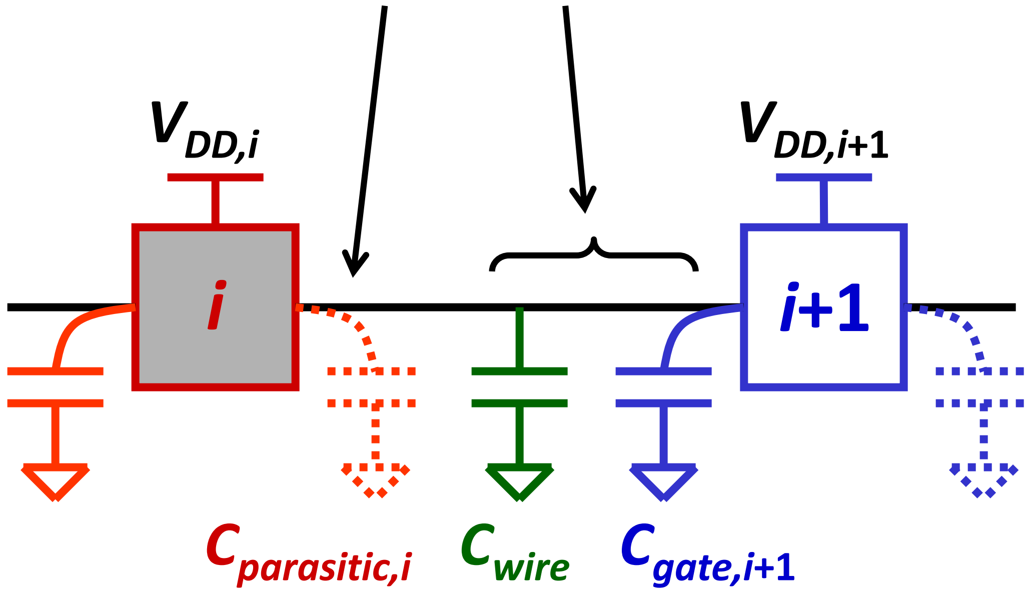
with:

D: the cycle time

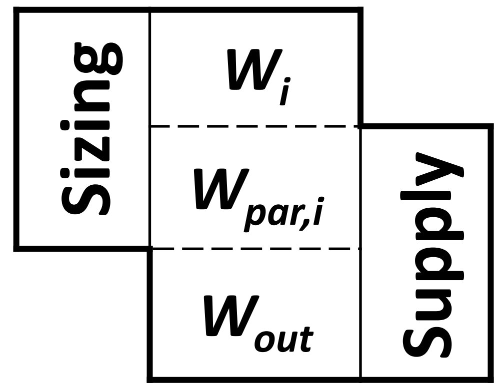
$I_0(S_{in})$: normalized leakage current with inputs in state S_{in}

Switching Component of Energy

$$E = K_e \cdot (W_{par} + W_{out}) \cdot V_{DD}^2$$



Impact of opt variables



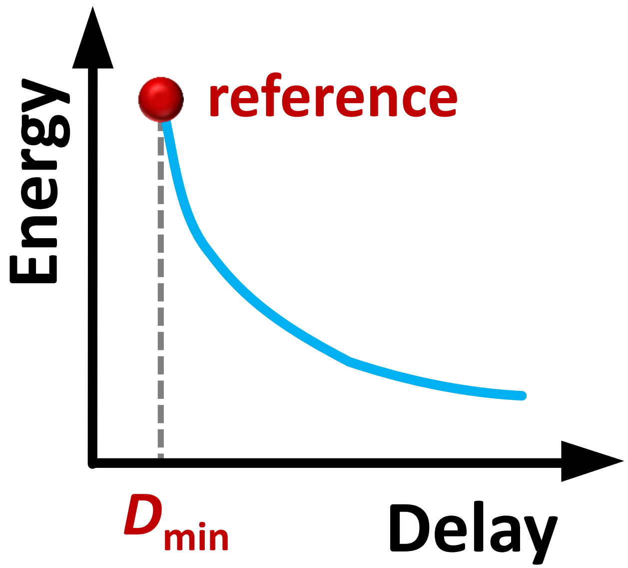
$$e_{ci} = K_e \cdot W_i \cdot (V_{DD,i-1}^2 + \gamma_i \cdot V_{DD,i}^2)$$

(energy stored on the logic gate *i*)

Optimization

Optimization Setup

- **Reference/nominal circuit**
  - Sized for D_{min} @ V_{DD}^{max}, V_T^{ref}
  - Known average activity
- **Define delay constraint**
  - $D_{con} = D_{min}\,(1 + d_{inc}/100)$
- **Minimize energy under delay constraint**
  - Gate sizing (*W*), optional buffering
  - V_{DD} and V_T scaling



Sensitivity to Sizing and Supply

- **Gate sizing (*W*)**

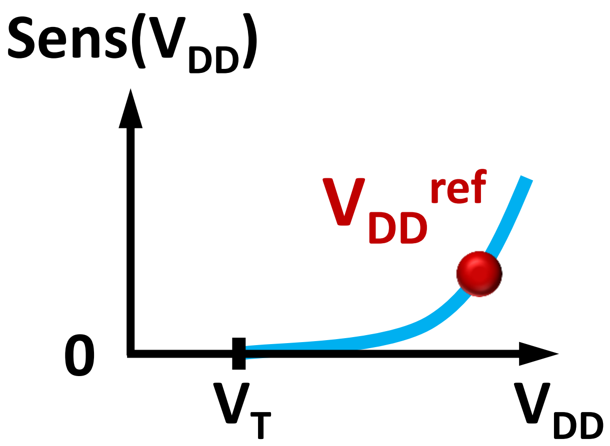
$$-\frac{\partial E_{Sw} / \partial W_i}{\partial D / \partial W_i} = \frac{ec_i}{\tau_{ref} \cdot \left(h_{eff,i} - h_{eff,i-1}\right)}$$

∞ for equal h_{eff} (D_{min})

- **Supply voltage (V_{DD})**

$$-\frac{\partial E_{Sw} / \partial V_{DD}}{\partial D / \partial V_{DD}} = 2 \cdot \frac{E_{Sw}}{D} \cdot \frac{1 - x_v}{\alpha_d - 1 + x_v}$$

$$x_v = \frac{V_{on} + \Delta V_{TH}}{V_{DD}}$$



V. Stojanović *et al.*, ESSCIRC 2002, pp. 211-214. | D. Marković *et al.*, IEEE JSSC, pp. 1282-1293, 8/04.

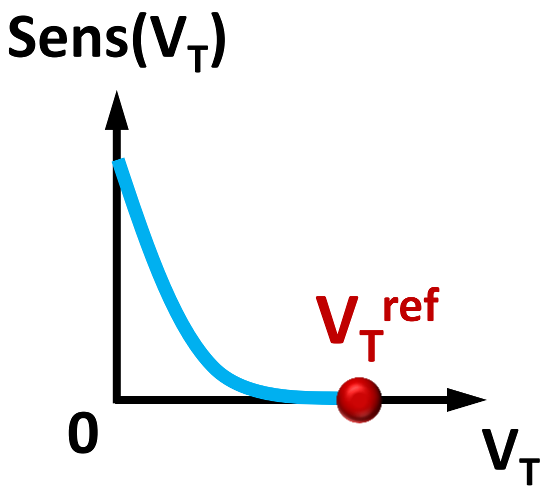
Sensitivity to Threshold Voltage

- **Threshold voltage (V_T)**

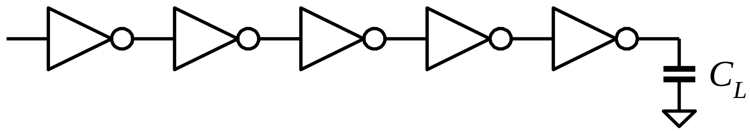
$$-\frac{\partial E / \partial(\Delta V_{TH})}{\partial D / \partial(\Delta V_{TH})} = P_{Lk} \cdot \left(\frac{V_{DD} - V_{on} - \Delta V_{TH}}{\alpha_d \cdot V_0} - 1 \right)$$

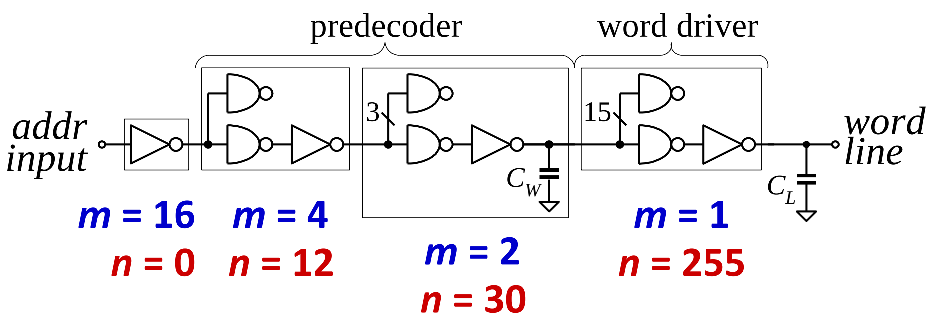
Low initial leakage

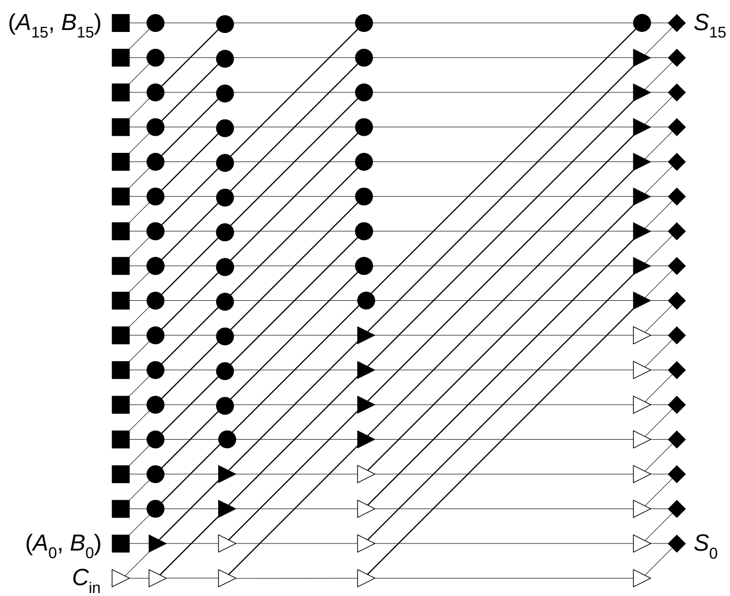
$\Rightarrow$ speedup comes for "free"



Circuit Optimization Examples



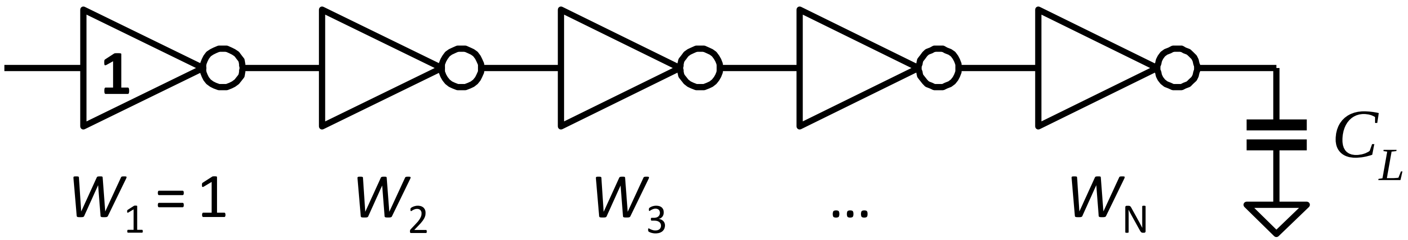




- **Inverter chain**
- **Memory decoder**
  - Branching
  - Inactive gates
- **Tree adder**
  - Long wires
  - Re-convergent paths
  - Multiple active outputs

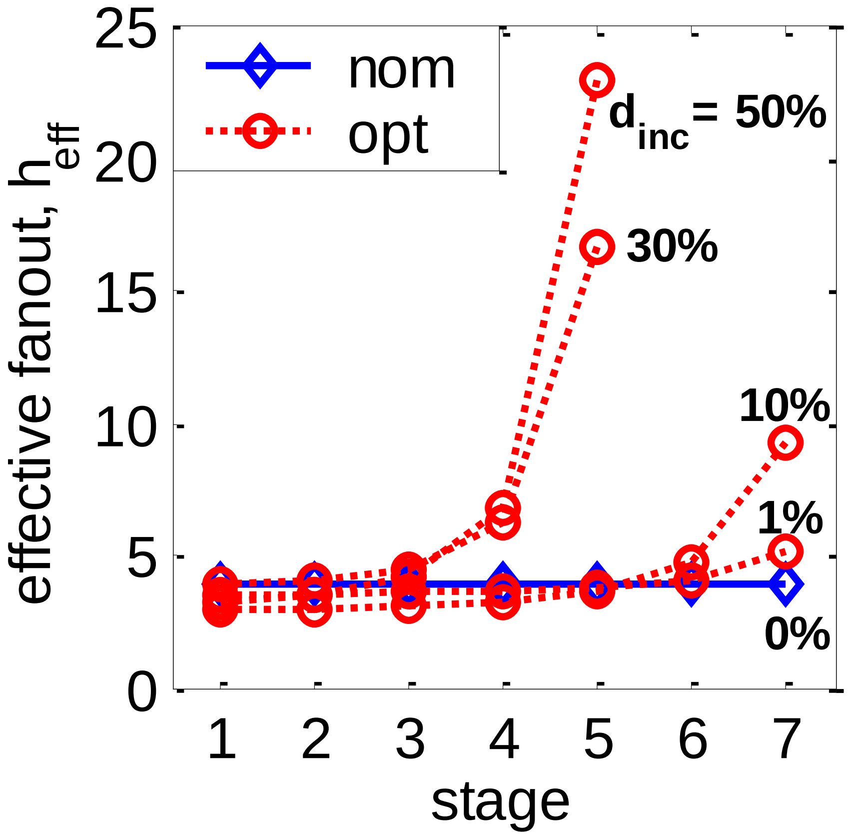
Example 3.1: Inverter Chain

- **Properties of inverter chain**
  - Single path topology
  - Energy increases geometrically from input to output



- **Goal**
  - Find optimal sizing $\boldsymbol{W} = [W_1, W_2, \ldots, W_N]$, supply voltage and buffering strategy to minimize energy

Inverter Chain: Gate Sizing & Buffering



$$W_i^2 = \frac{W_{i-1} \cdot W_{i+1}}{1 + \mu \cdot W_{i-1}}$$

[Ma, Franzon, IEEE JSSC, 9/94]

$$\mu = -\frac{2 \cdot K_e \cdot V_{DD}^2}{\tau_{ref} \cdot S_W}$$

$$S_W \propto \frac{ec_i}{h_{eff,i} - h_{eff,i-1}}$$

- **Variable taper achieves minimum energy**
- **Reduce number of stages at large d_{inc}**

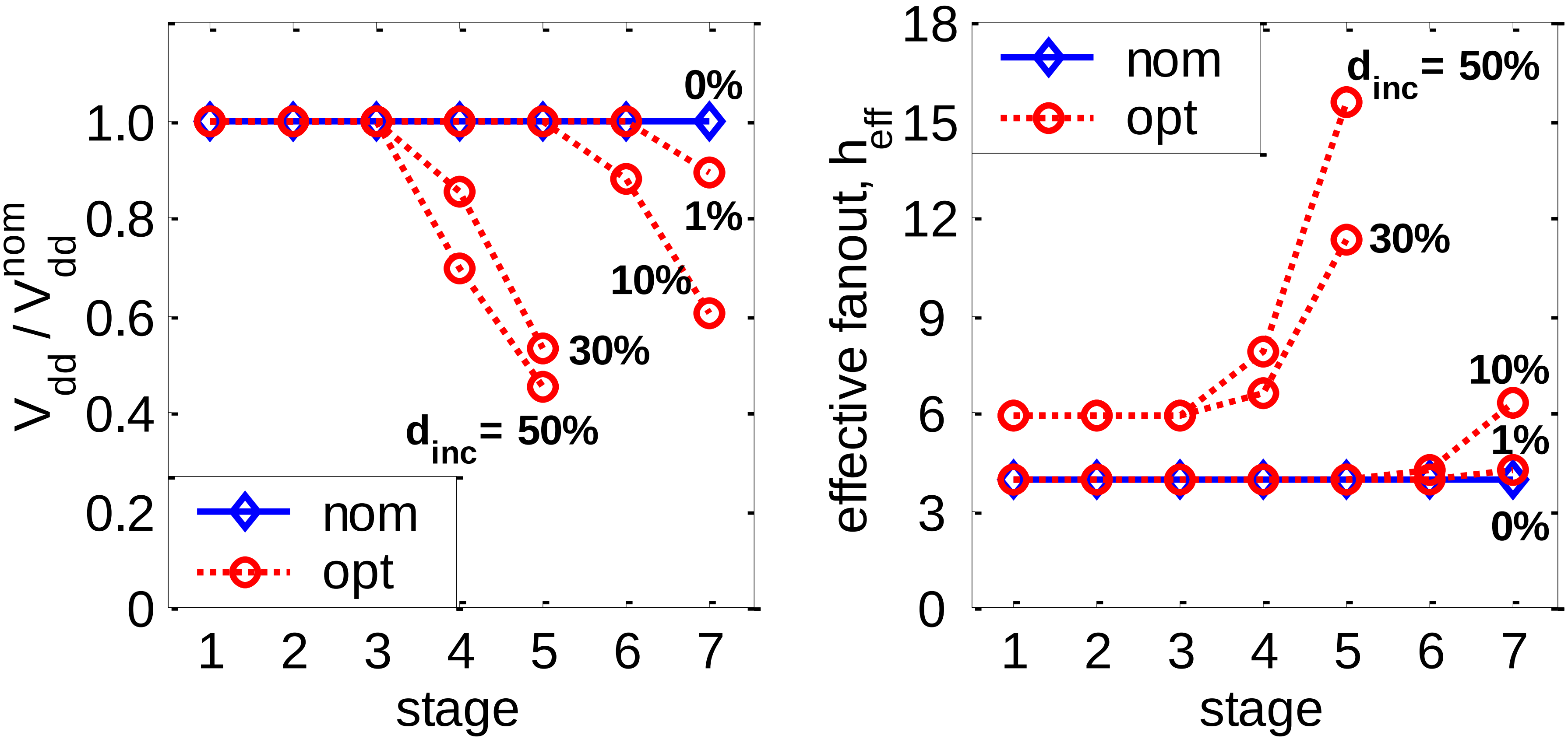
MS Excel Optimization: Inverter Chain

- **5-stage inverter chain loaded with C_L = 1024**
  - Sizing optimization for D_{min} + 10% constraint

Inv chain										
stage	1	2	3	4	5					
size	1	2.74	7.83	25.20	109.79	1024				
Energy	3.44	9.75	30.68	127.42	1100.85		Etot	1272.14		objective
Delay	3.44	3.56	3.92	5.06	10.03		Dtot	26.00		constraint
										variable
fanout	2.74	2.86	3.22	4.36	9.33		Emax	1602.7		
							Dmin	23.5		

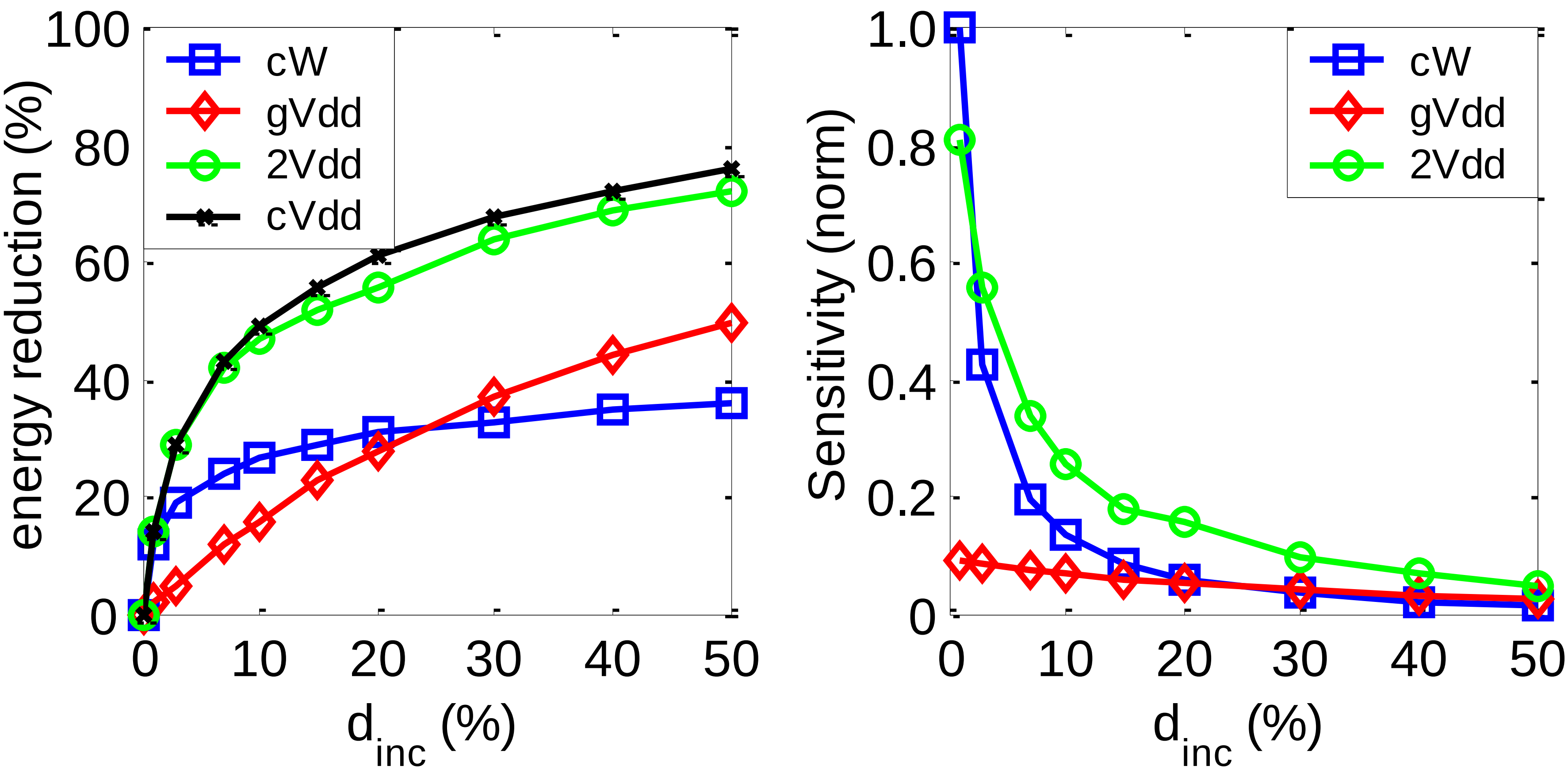
- **Method: use Solver Add-In** (under File / Options)
  - A push-button optimization…

Inverter Chain: V_{DD} Optimization



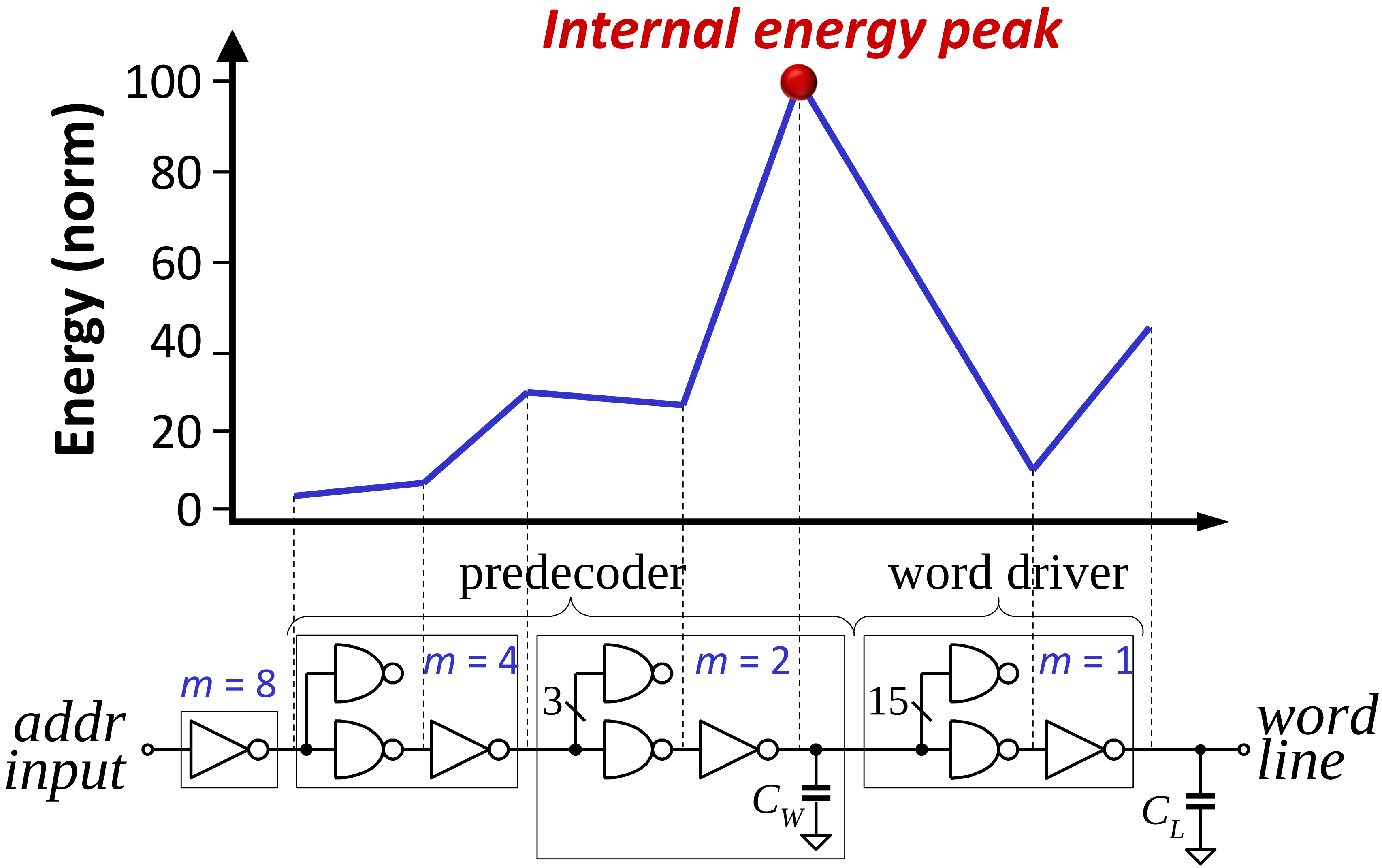
- **Variable taper achieved by voltage scaling**
- **V_{DD} reduces energy of the final load first**

Inverter Chain: Optimization Results

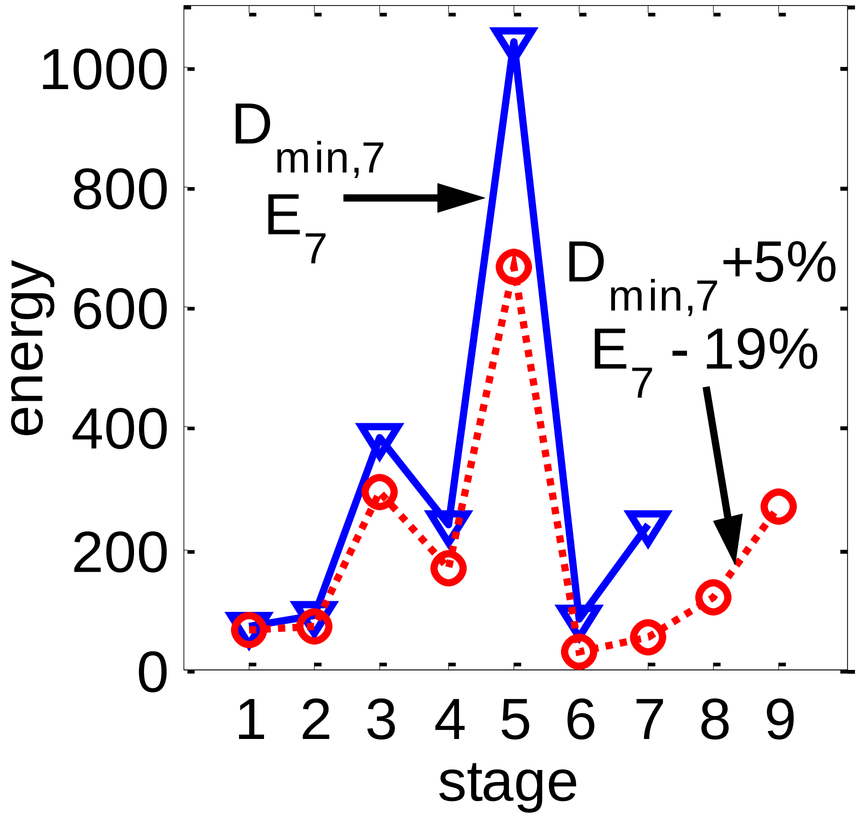


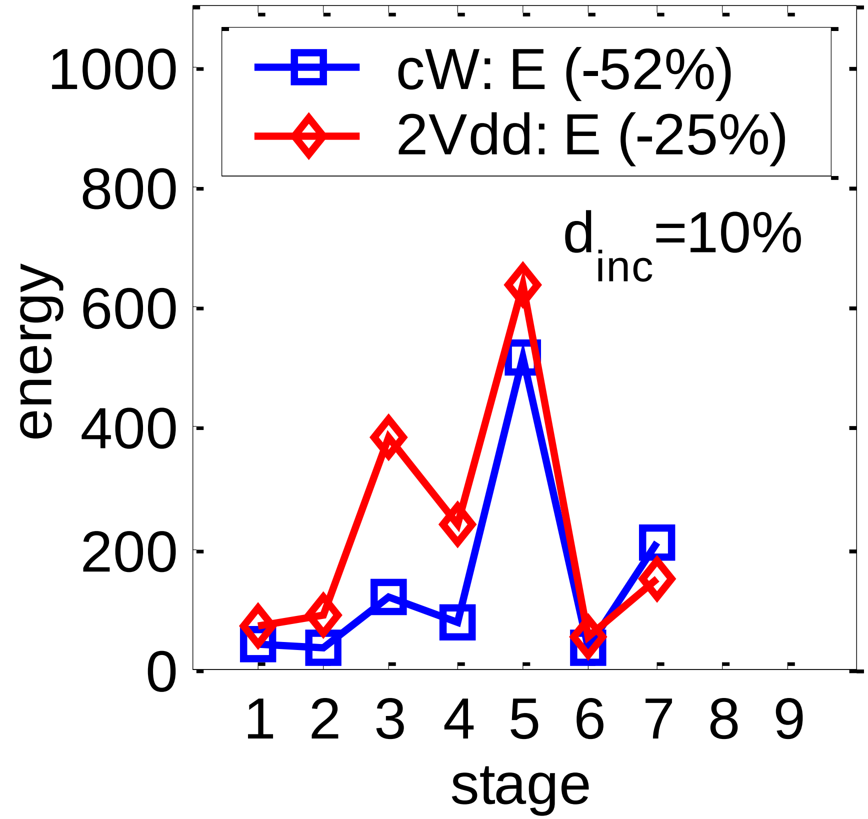
- **Parameter with the largest sensitivity has the largest potential for energy reduction**
- **Two discrete supplies mimic per-stage V_{DD}**

Example 3.2: SRAM Decoder



W vs. V_{DD} for Reducing Energy Peak

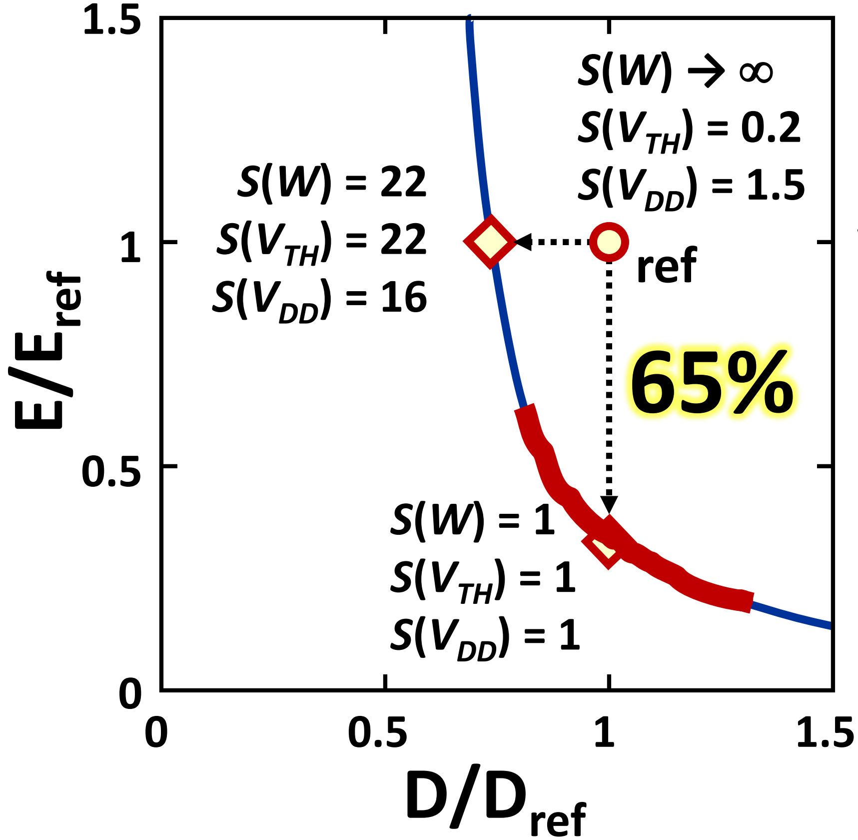


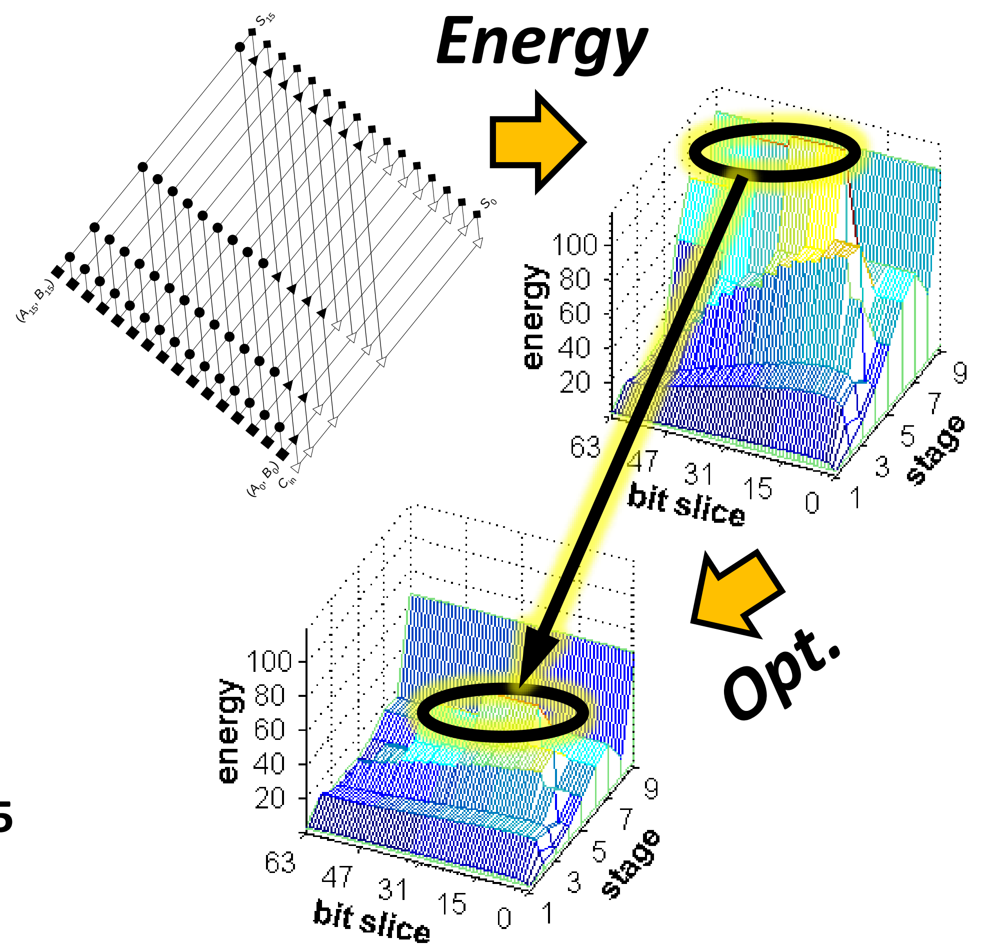


- **V_{DD} less effective than *W* optimization**
- **Buffering also reduces energy peak**

[B. Amrutur, Ph.D. Thesis, Stanford, 8/99]

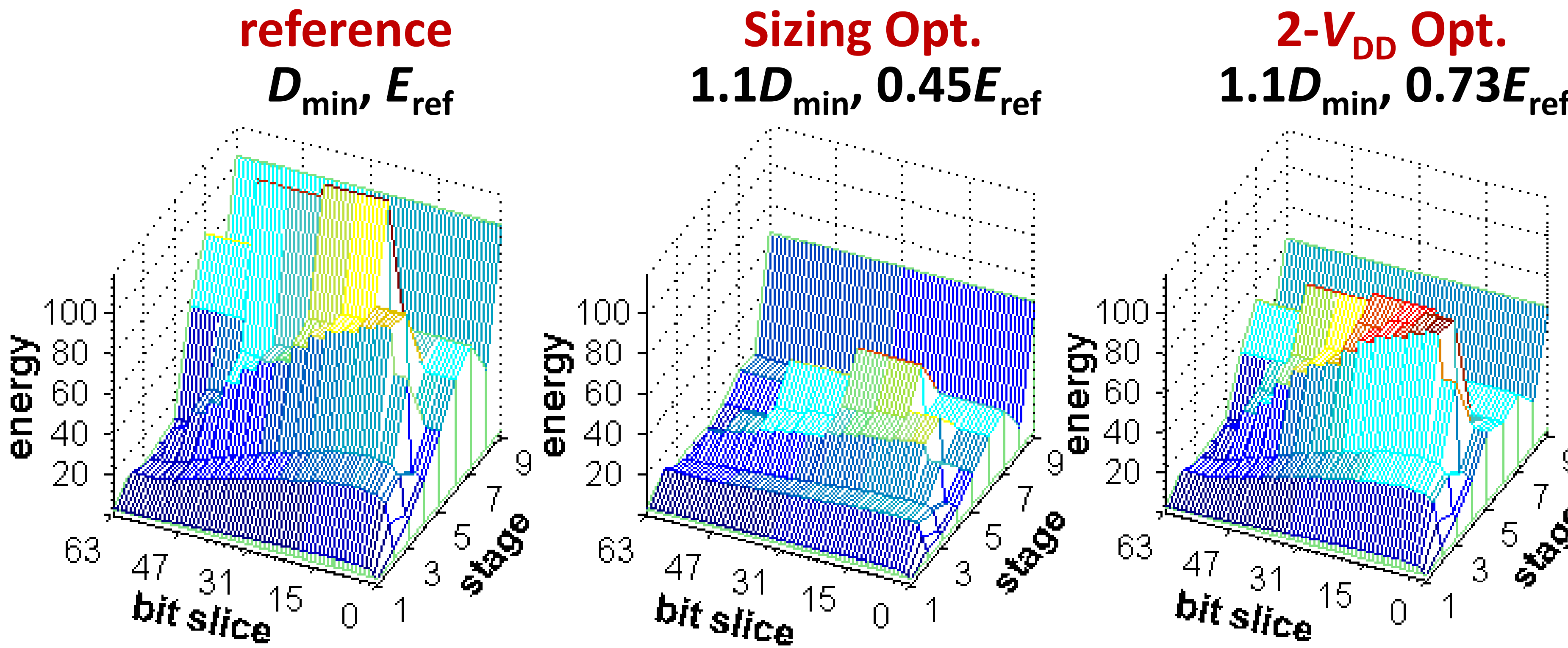
Example 3.3: Tree Adder





Tree Adder: Optimization Results

Reference: all paths are critical

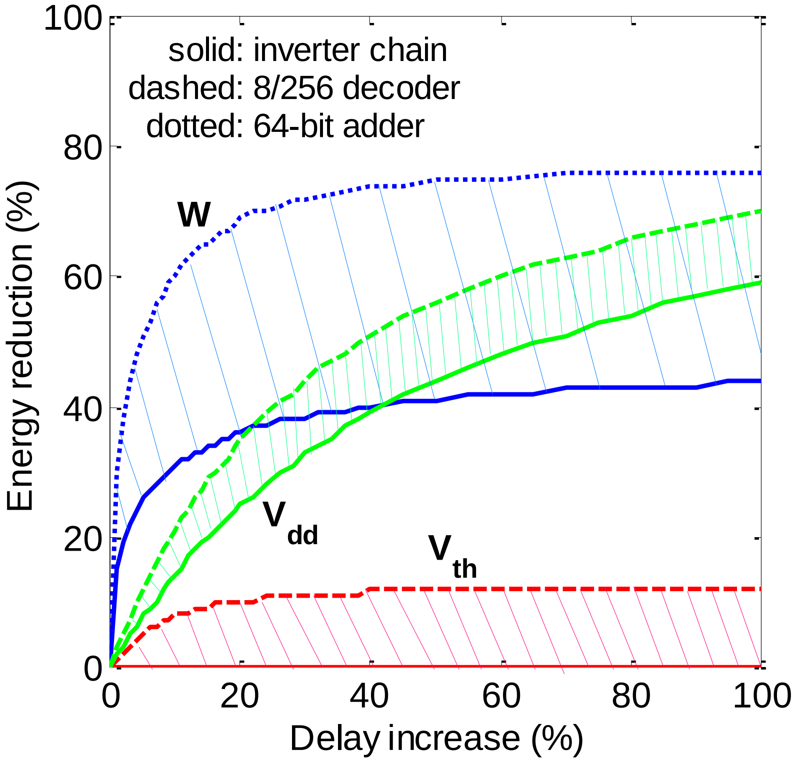


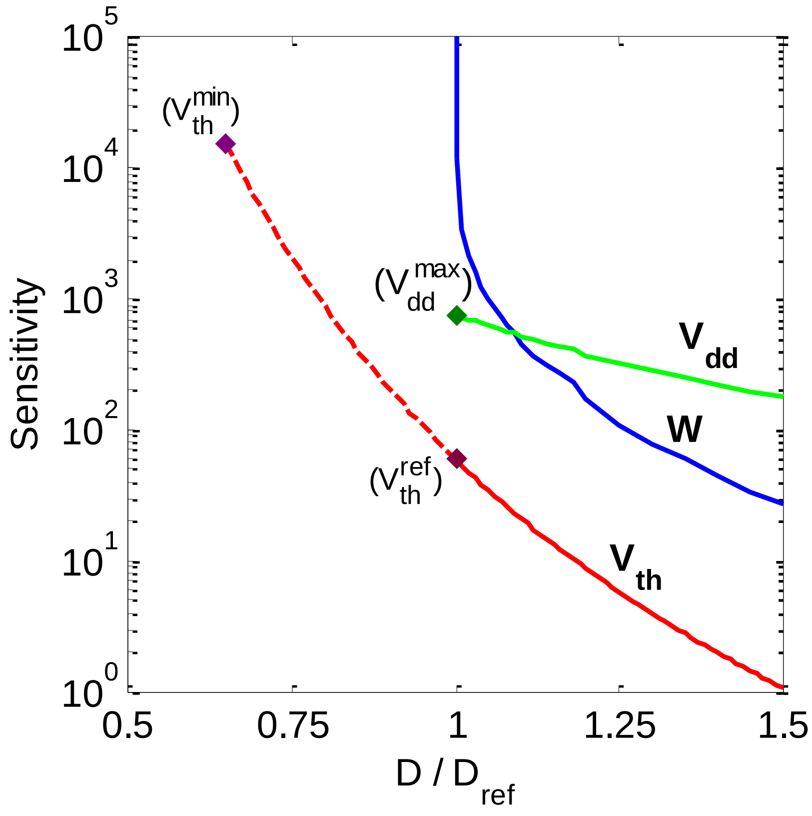
- **Internal energy: *W* more effective than V_{DD}**
  - For d_{inc} = 10%: $\Delta E_W = -55\%$, $\Delta E_{2Vdd} = -27\%$

Few Insights

A Look at Tuning Variables…

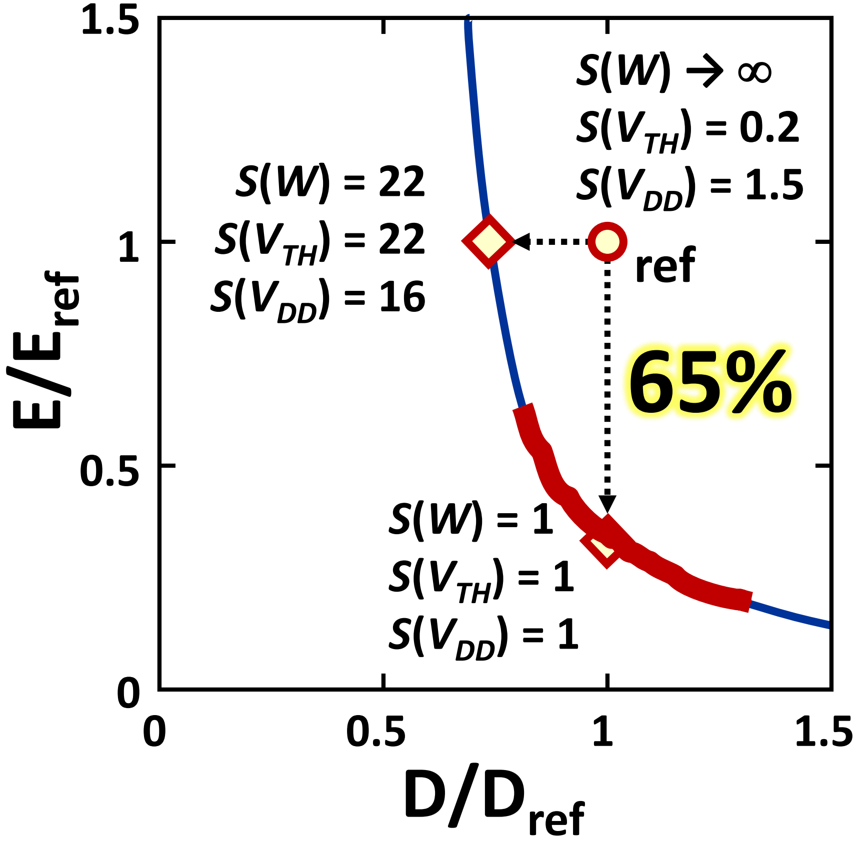
10% excess delay → 30-70% energy reduction





Peak performance is very power inefficient!

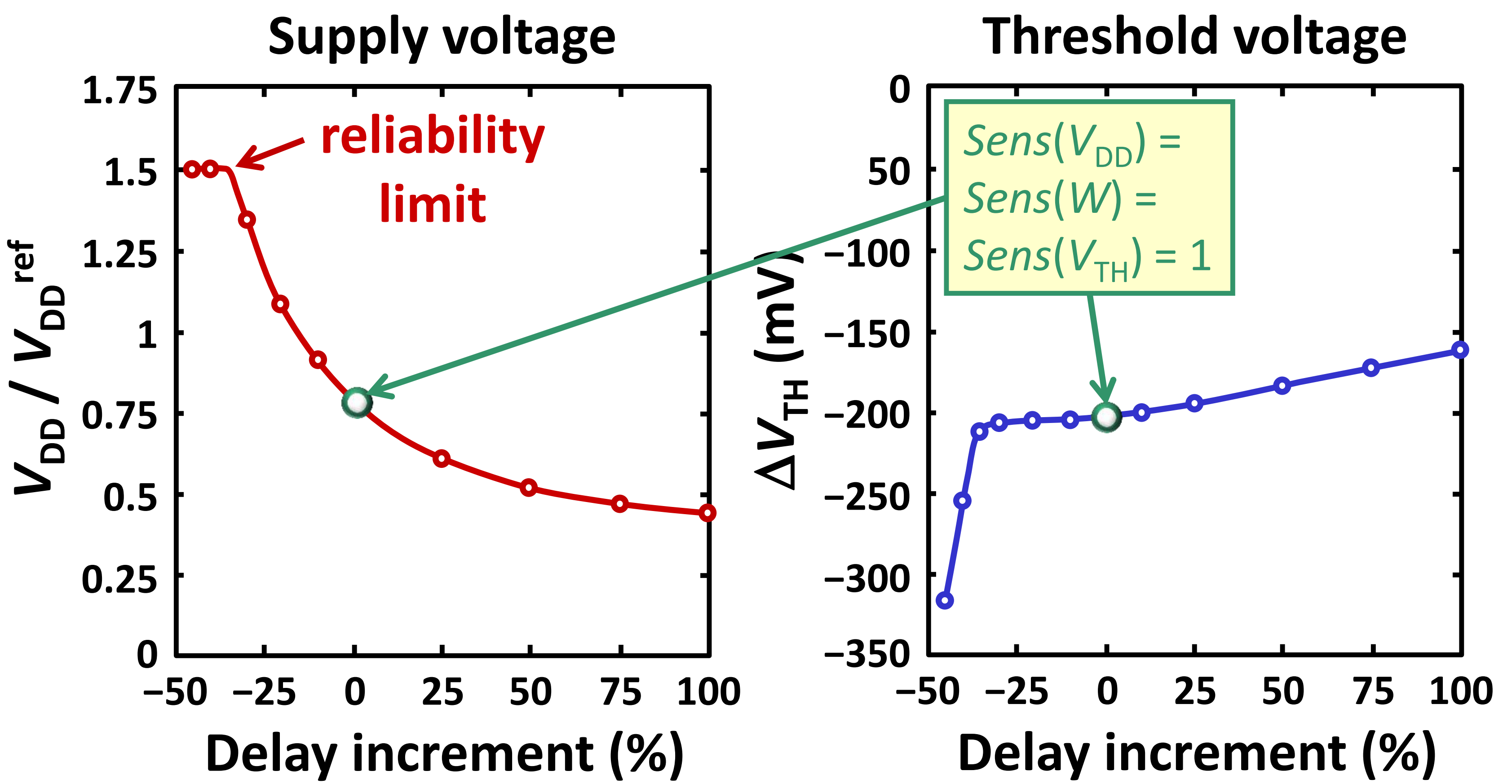
Limited Range of Circuit Optimization



- ±30% around D_{ref}
- Else, too much E or D
- Need for arch. opt.

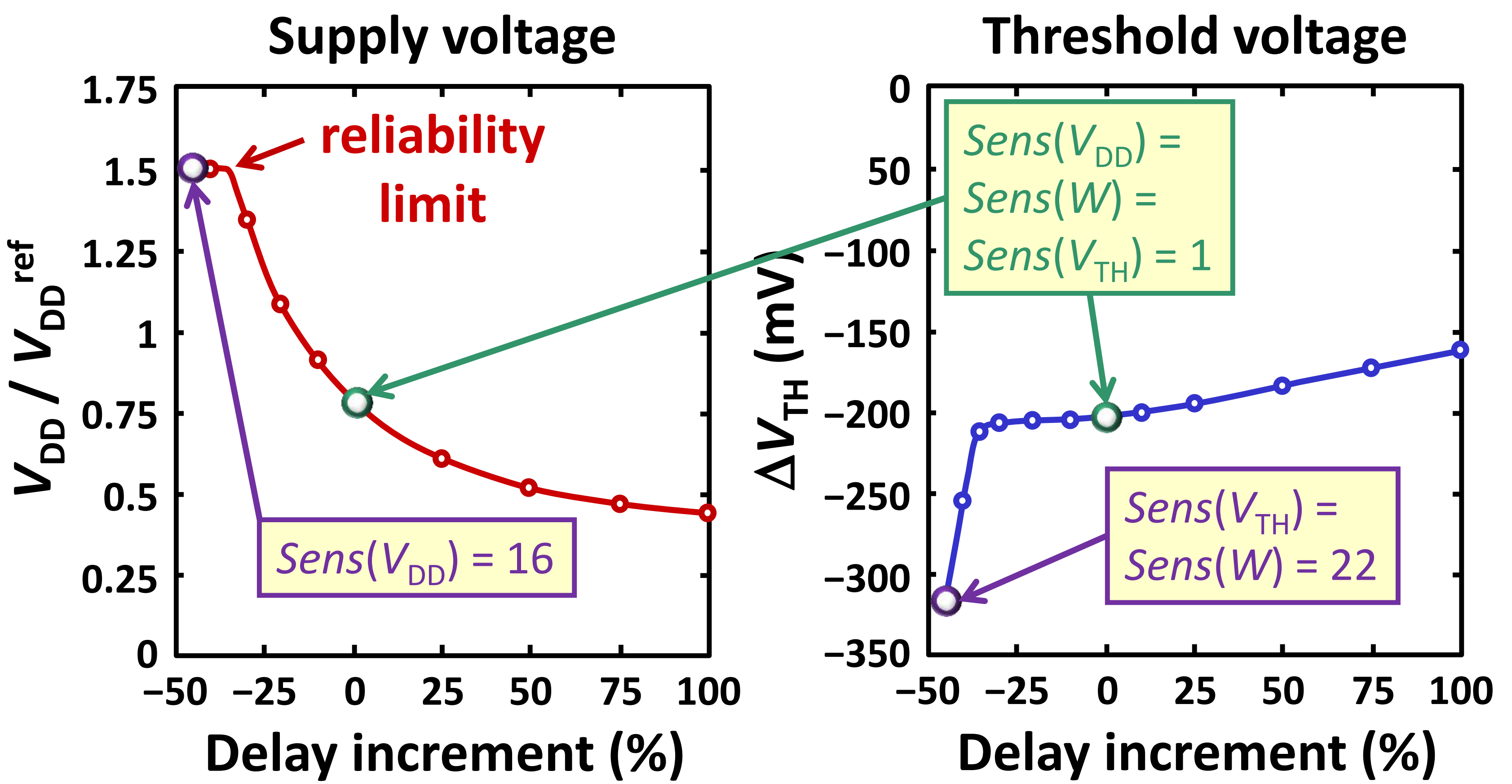
A Look at Tuning Variables

Equal sensitivity unless variables reach their bounds



A Look at Tuning Variables

Bound reached = fewer variables to work with

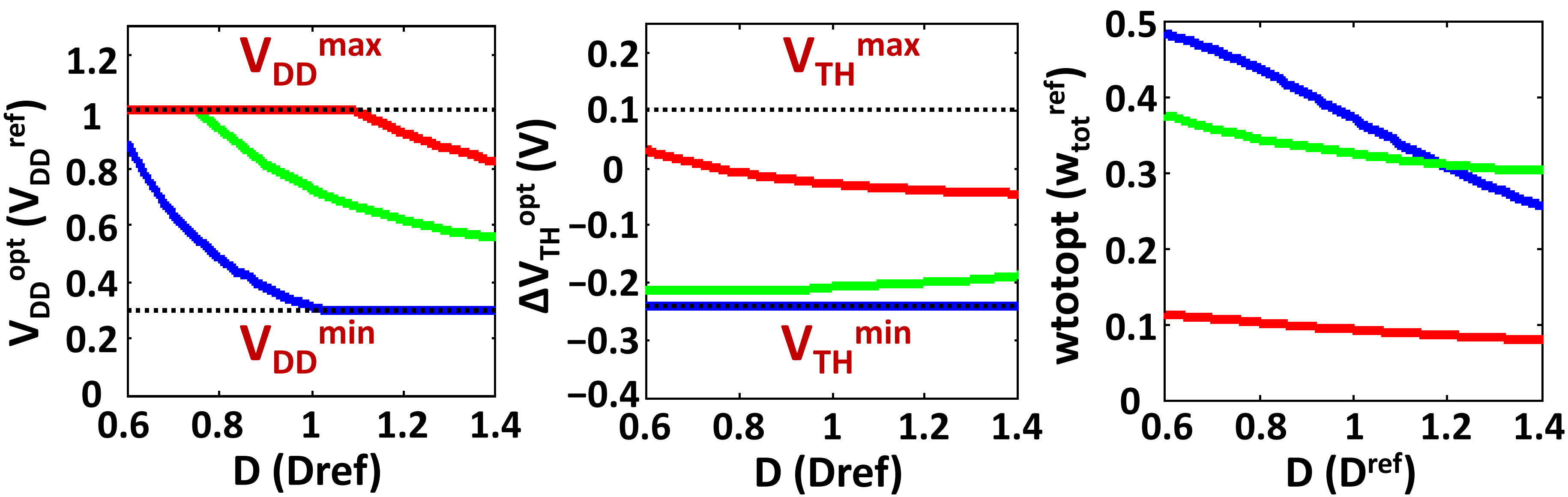


Optimal Circuit Parameters

Reference Design:
D^{ref} (V_{DD}^{ref}, V_{TH}^{ref})

Topology	Inv	Add	Dec
$(E_{Lk}/E_{Sw})^{ref}$	0.1%	1%	10%

- **Large variation in optimal parameters V_{DD}^{opt}, V_{TH}^{opt}, W^{opt}**



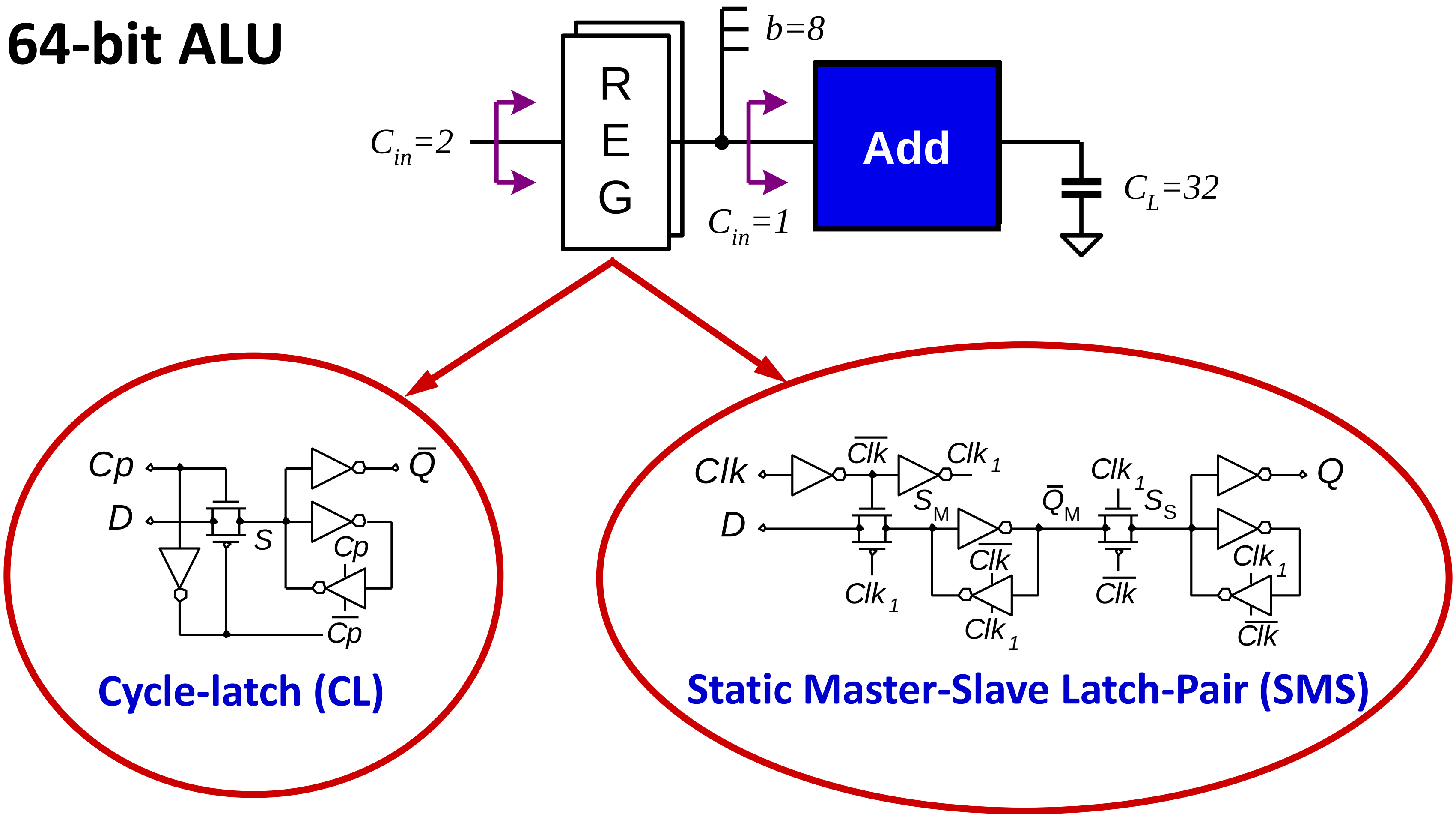
Reference/nominal parameters (V_{DD}^{ref}, V_{TH}^{ref}) are rarely optimal

Lessons from Circuit Optimization

- **Sensitivity-based optimization framework**
  - Equal marginal costs ⇔ Energy-efficient design
- **Effectiveness of tuning variables**
  - Sizing is the most effective for small d_{inc}
  - V_{DD} is better for large delay increments
- **Peak performance is VERY power inefficient**
  - ~70% energy reduction for 20% delay penalty
- **Limited performance range of tuning variables**
  - Additional variables for higher energy-efficiency

Choosing Circuit Topology: Optimal Register?

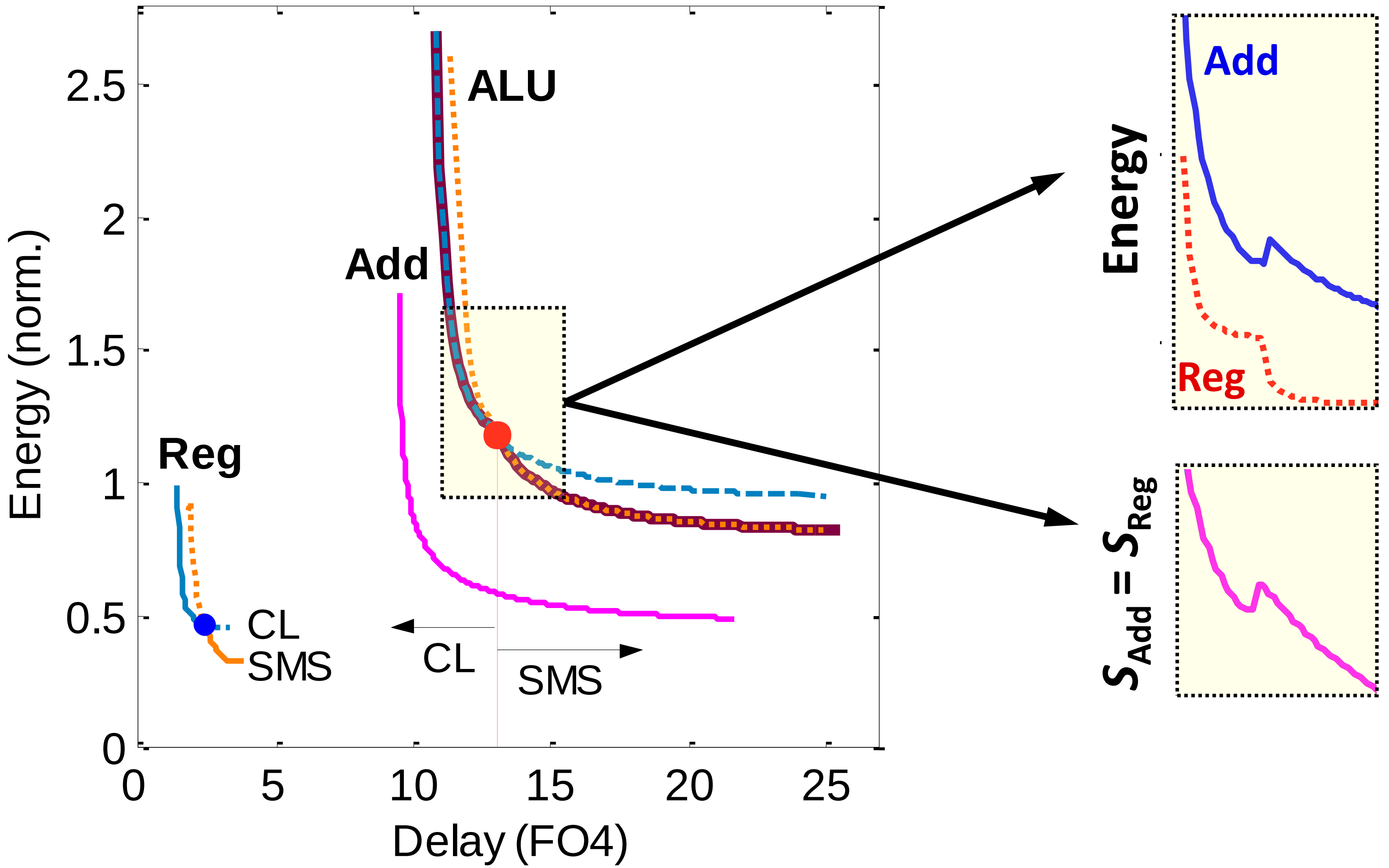
- **64-bit ALU**



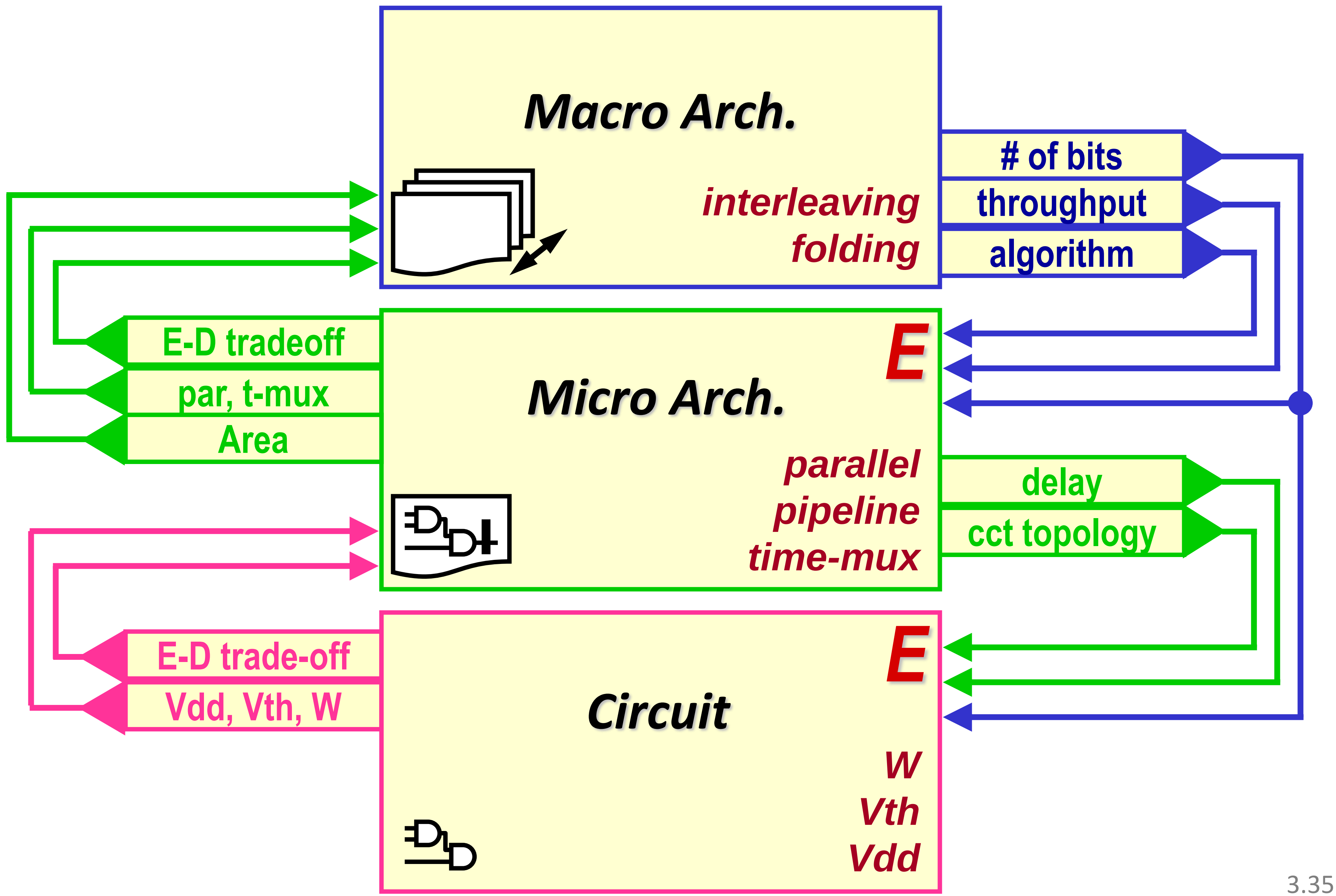
- **Given energy-delay tradeoff for adder and register (two register options), what is the best energy-delay tradeoff in the ALU?**

Balancing Sensitivity Across Circuit Blocks

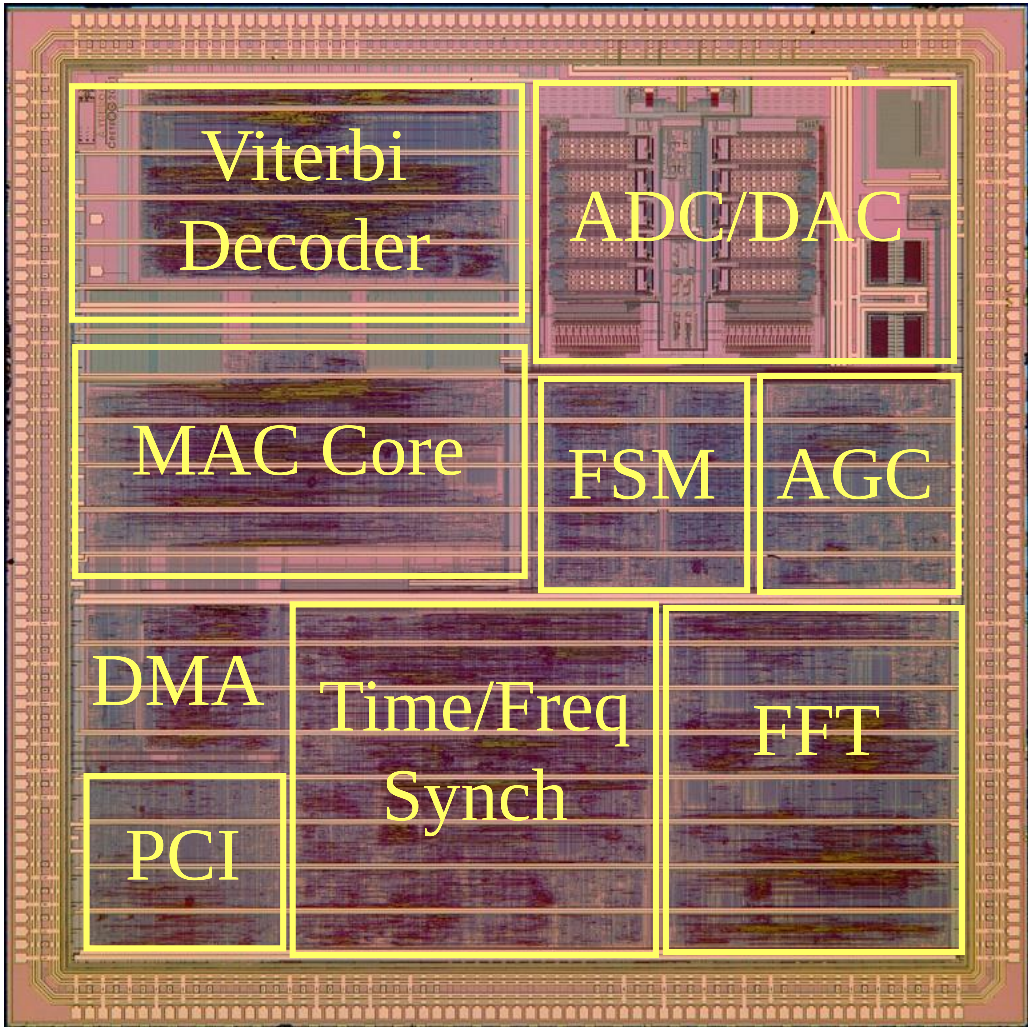
Upsize lower activity blocks (Add) to save energy



Micro-Architectural Optimization



Example 3.4: 802.11a Baseband

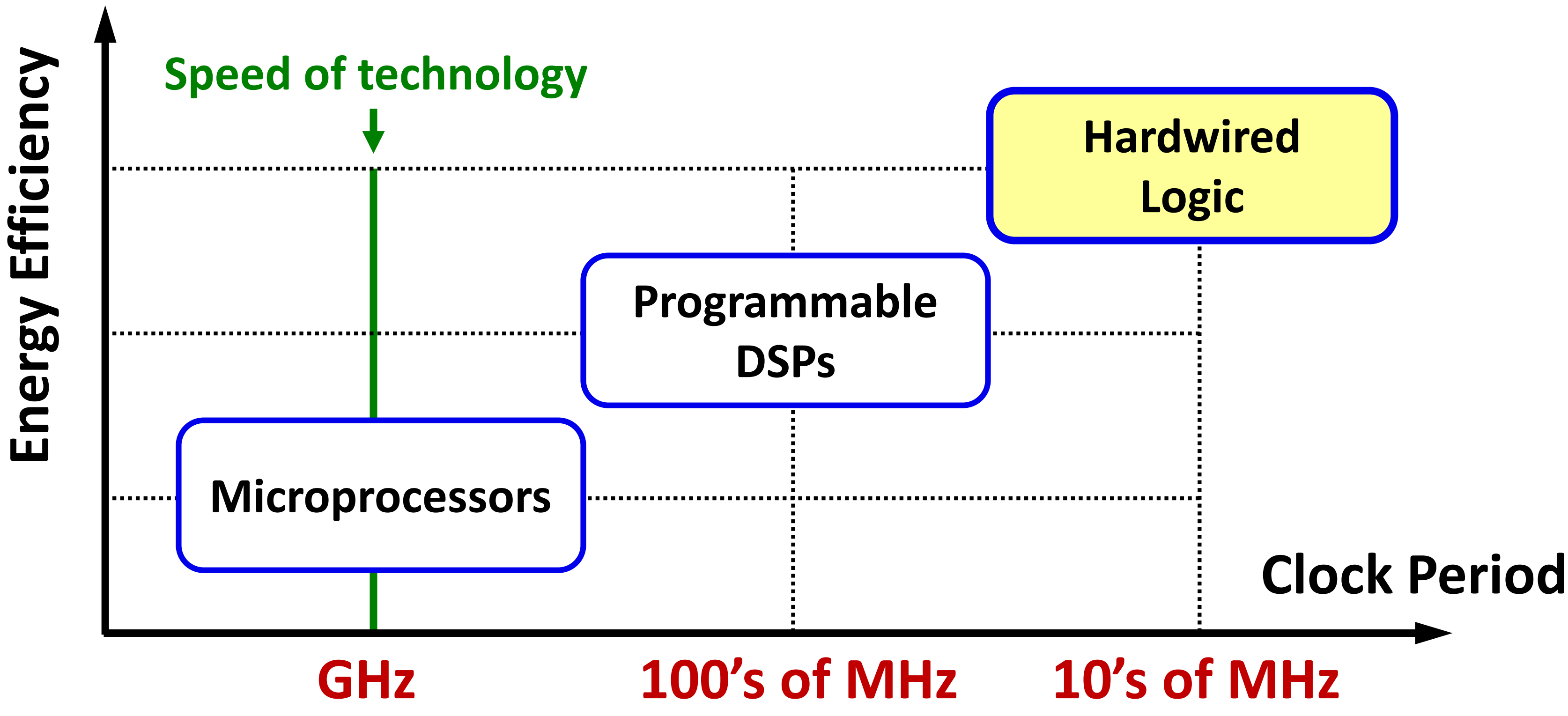


[An 802.11a baseband processor]

- **Direct mapped architecture**
- **200 MOPS/mW**
  - 80 MHz clock!
  - 40 GOPS
  - Power = 200 mW
  - 0.25 µm CMOS
- **The architecture has to track technology**

Wireless Baseband Chip Design

- **Direct mapping is the most energy-efficient**
- **Technology is too fast for dedicated hardware**
  - Opportunity to further reduce energy and area

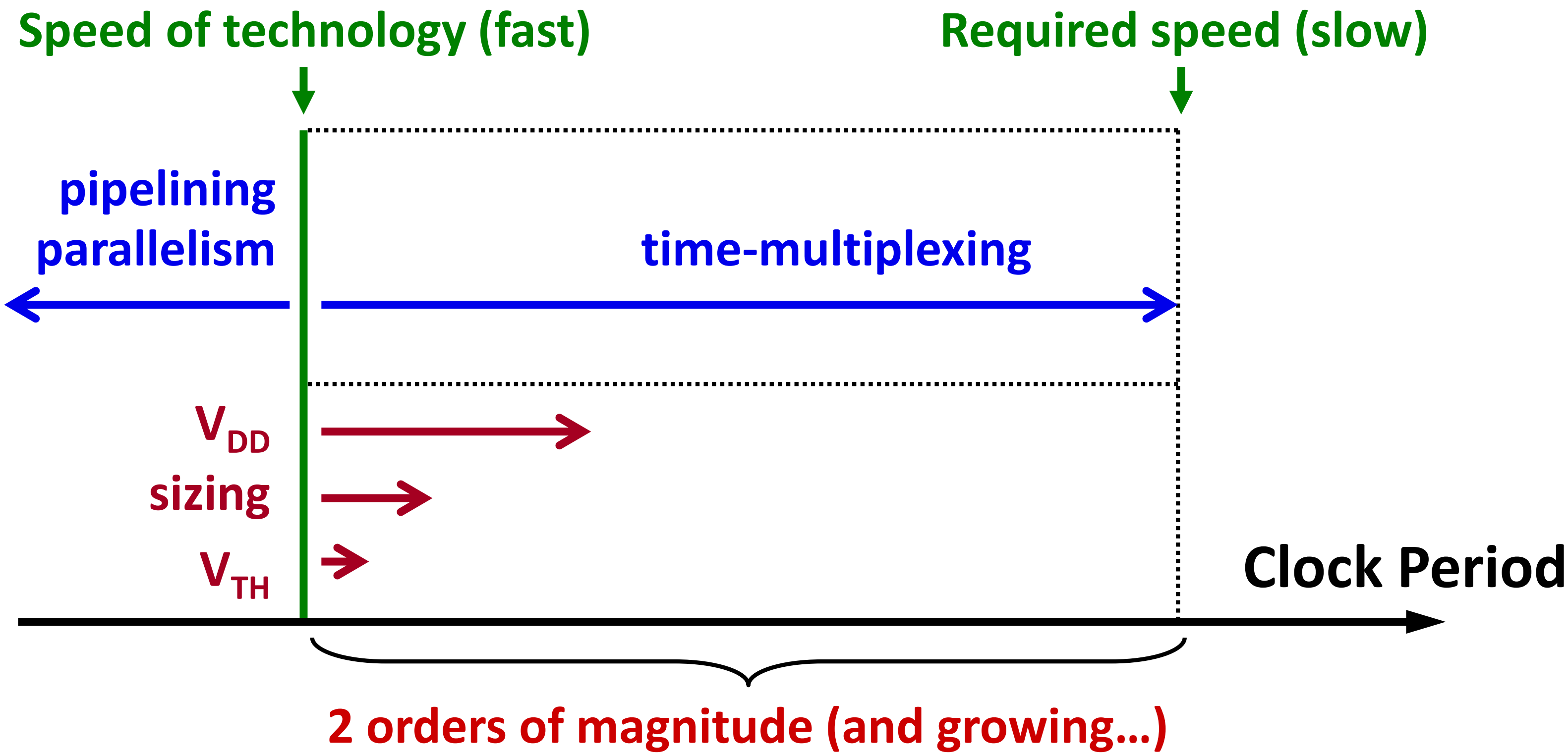


A Number of Variables to Consider

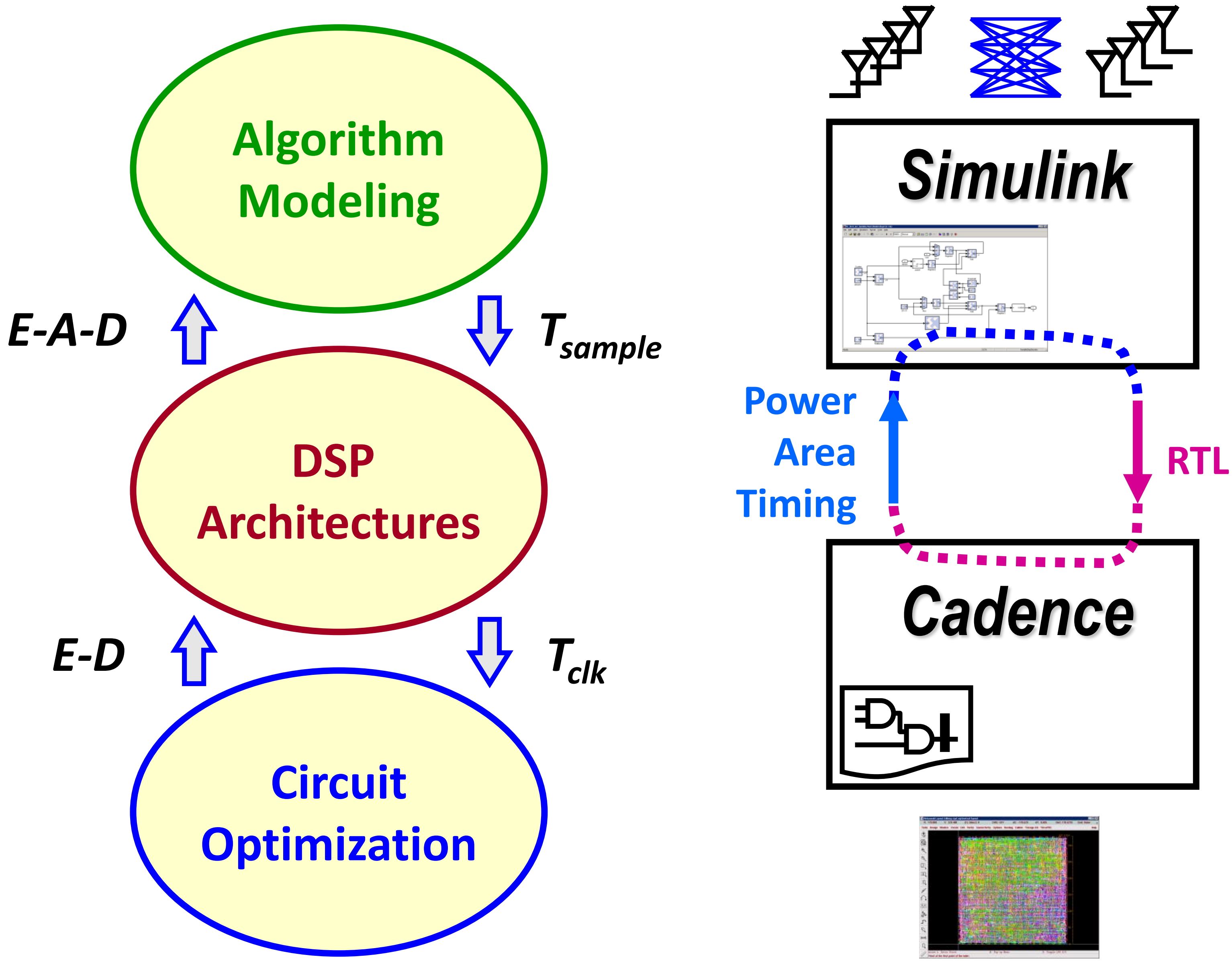
- **How to optimally combine all variables?**

$$D \propto \frac{V_{DD}}{(V_{DD} - V_{on} - \Delta V_{TH})^{\alpha_d}} \cdot \frac{1}{W}$$

$$E_{Lk} \propto W \cdot e^{-V_{TH}/V_0} \cdot V_{DD}^3 \cdot T \qquad E_{Sw} \propto W \cdot V_{DD}^2$$



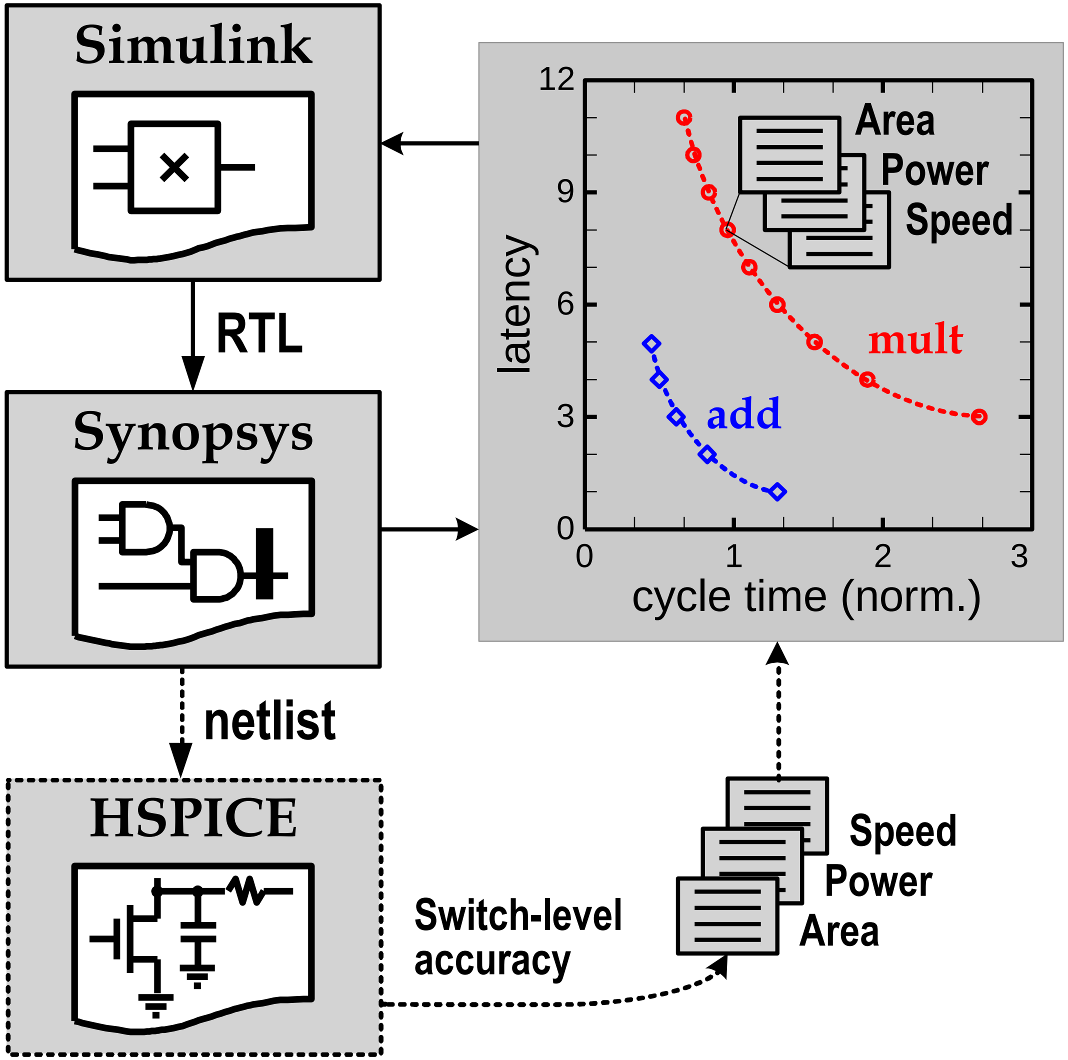
Towards Architecture Optimization



Architectural Feedback from Technology

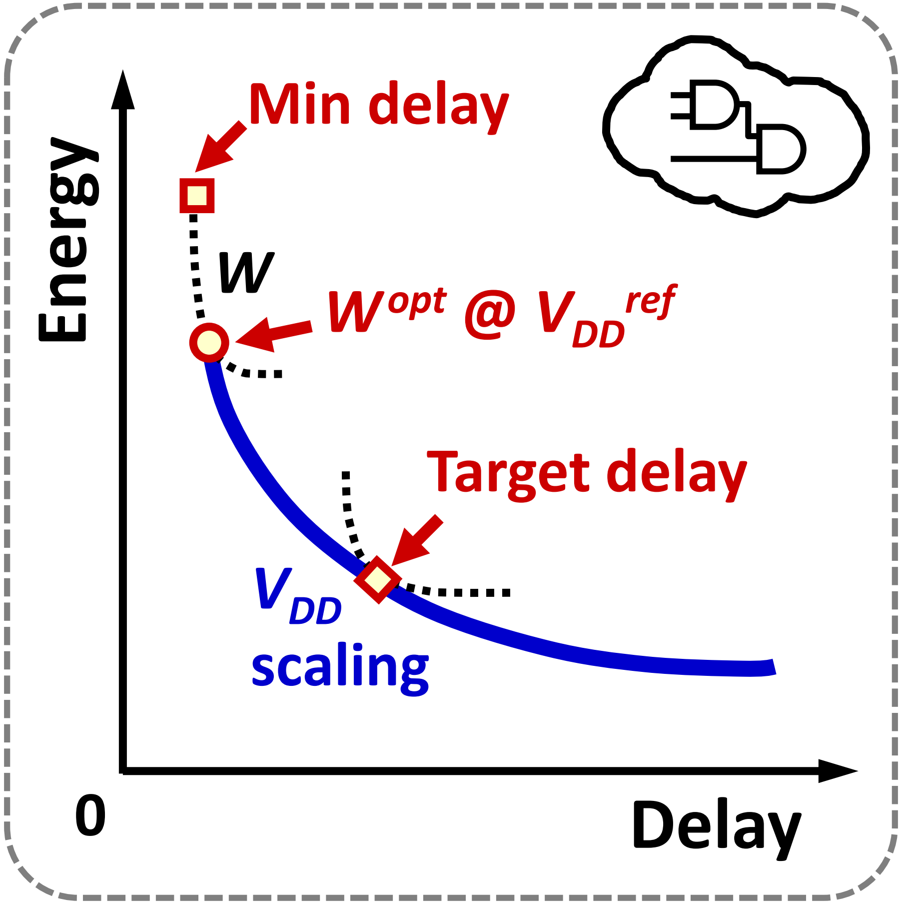
- **Simulink hardware library implicitly carries information only about latency and wordlength (we can later choose sample period when targeting an FPGA)**
- **ASIC flow: block characterization exposes technology features such as speed, power, and area**
- **But, technology parameters scale each generation**
  - Need a general and quick characterization methodology
  - Propagate results back to Simulink to avoid iterations

Cycle Time is Common for All Blocks



Datapath Characterization

- **Balance tradeoffs due to gate size (*W*) and supply voltage (V_{DD})**



Circuit Level

- **Optimal design point**
  - Curves from *W* and V_{DD} are tangent (equal sensitivity)
- **Goal: keep all pipelines at the same E-D point**

Summary | Method & Results

- **Optimal energy-delay tradeoff obtained by tuning gate size, V_{DD} and V_{TH} can be calculated by optimization**
  - Convex formulation of delay-constrained energy min
  - Circuit-level E-D tradeoff allows for quick comparison of multiple circuit topologies

- **Insights from circuit optimization**
  - Min-delay design consumes the most energy
  - **Gate sizing** is the most effective for small delays
  - V_{DD} is the most effective for medium delays
  - V_{TH} is the most effective for large delays

Summary | Limitations

- **Circuit optimization is effective in a narrow band of delay**

- **More degrees of freedom (e.g. architectural) are needed for broader range of performance tuning in an energy-efficient way**
  - Next lecture