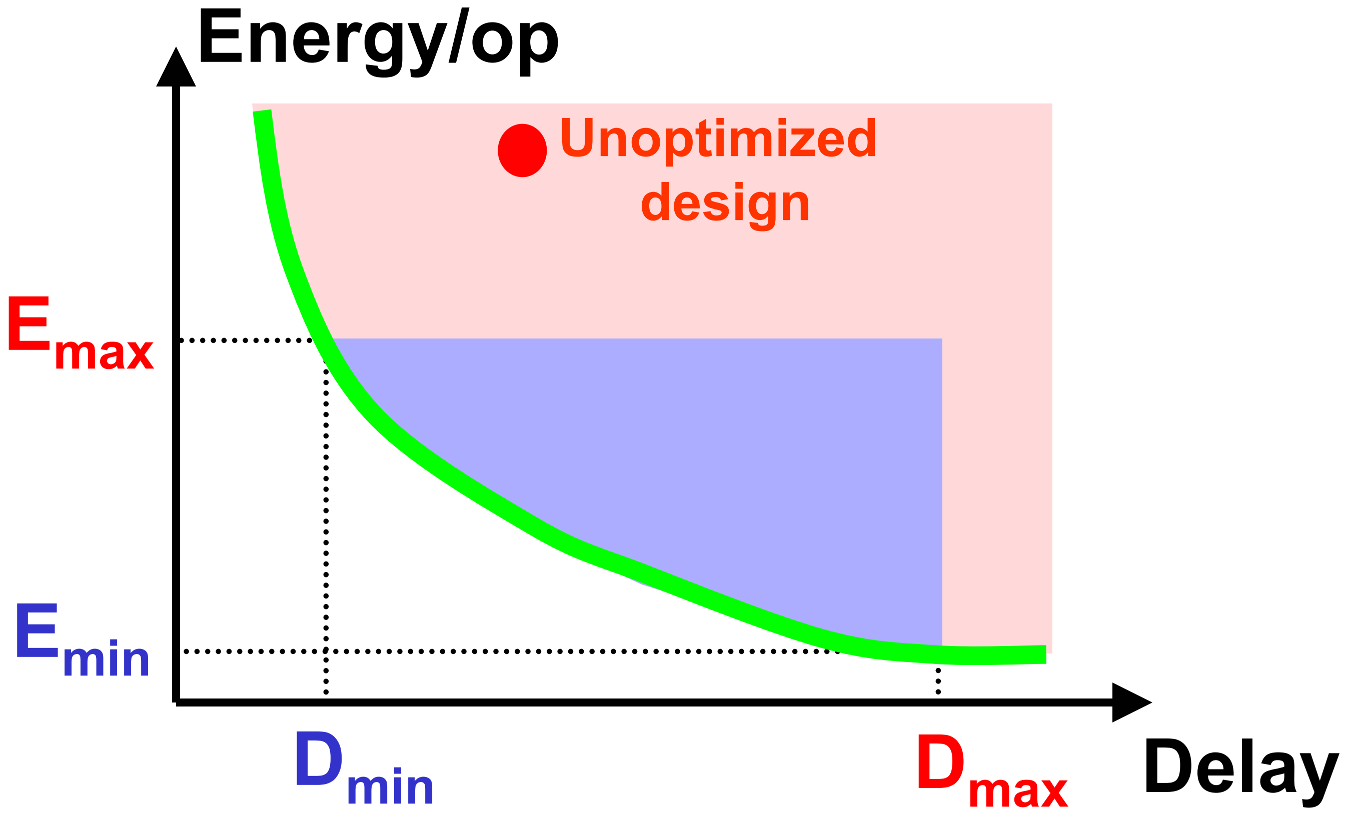
Methods for True Power Minimization

Robert W. Brodersen[1], Mark A. Horowitz[2], Dejan Markovic[1],
Borivoje Nikolic[1] and Vladimir Stojanovic[2]

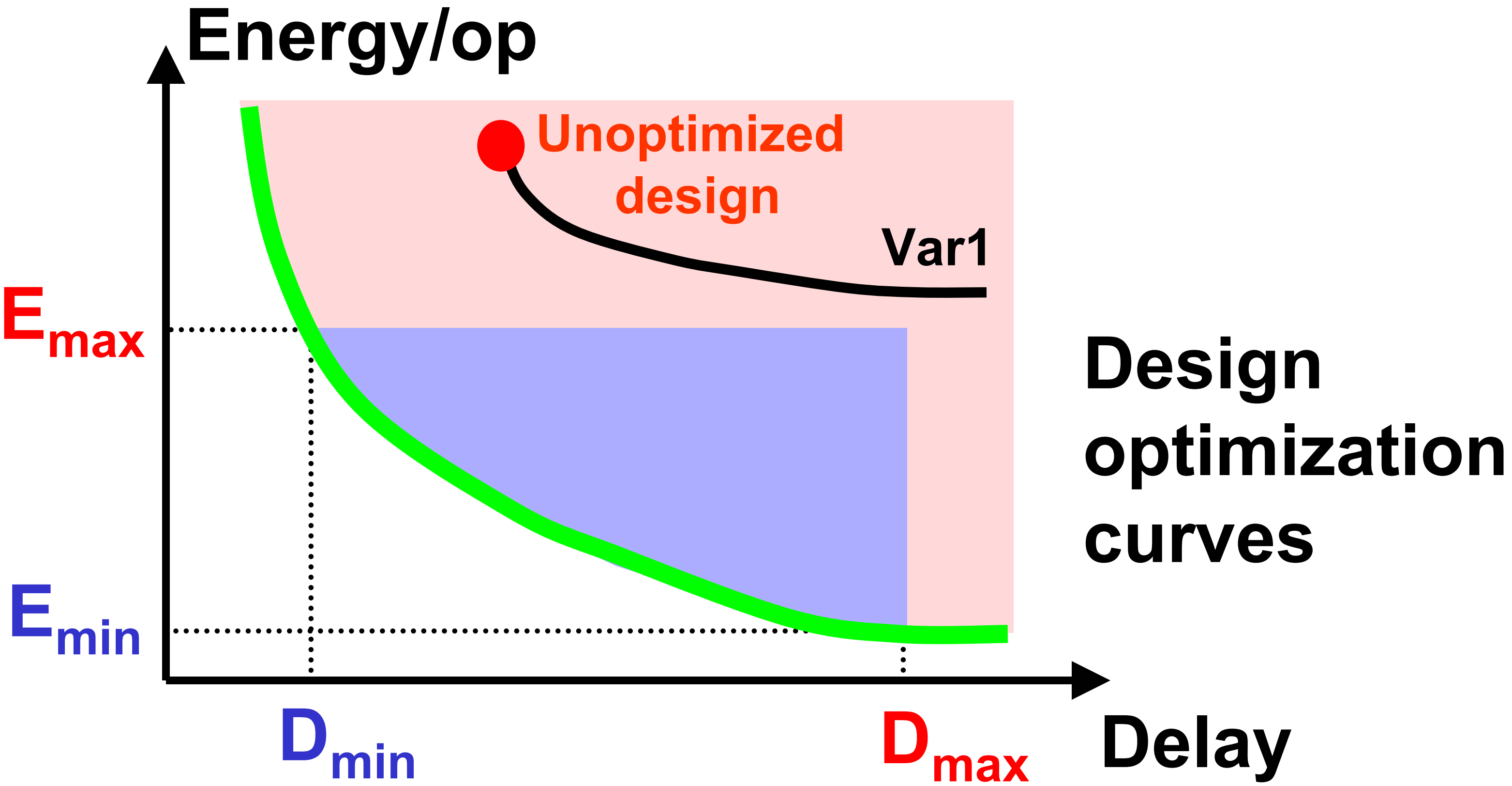
[1]Berkeley Wireless Research Center, UC Berkeley
[2]Stanford University

Power limited operation



Achieve the highest performance under the power cap

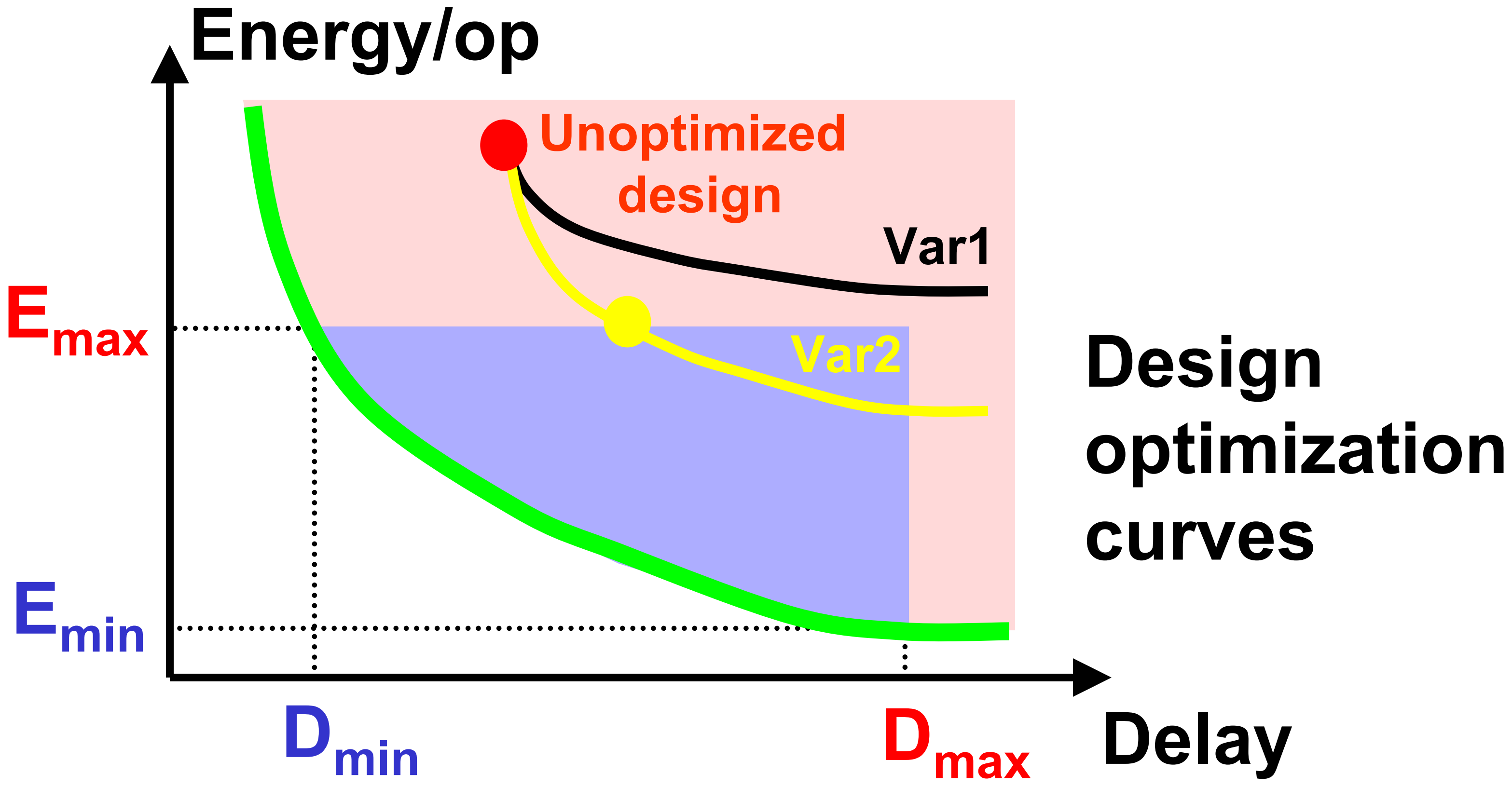
Power limited operation



Design optimization curves

Achieve the highest performance under the power cap

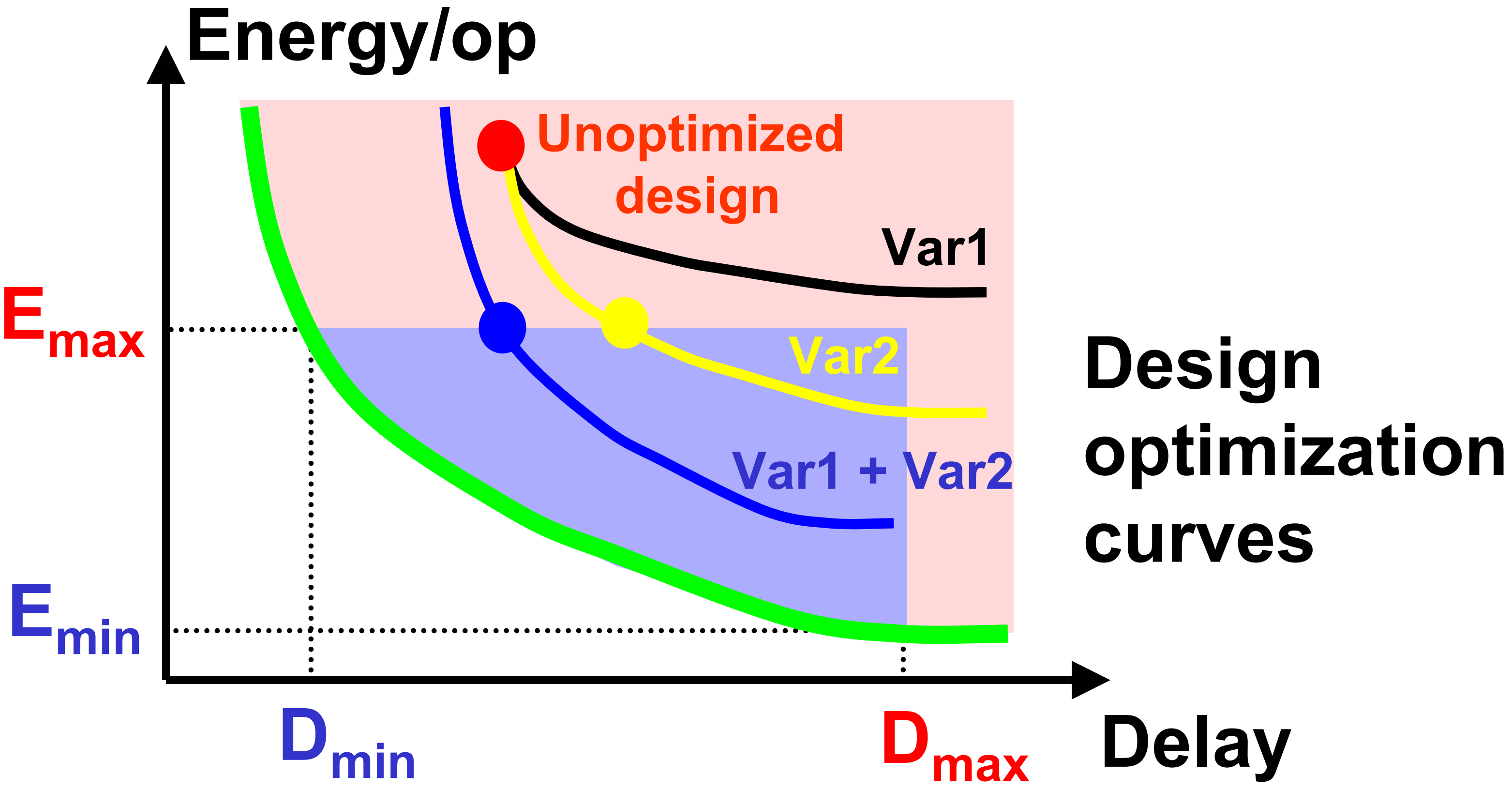
Power limited operation



Design optimization curves

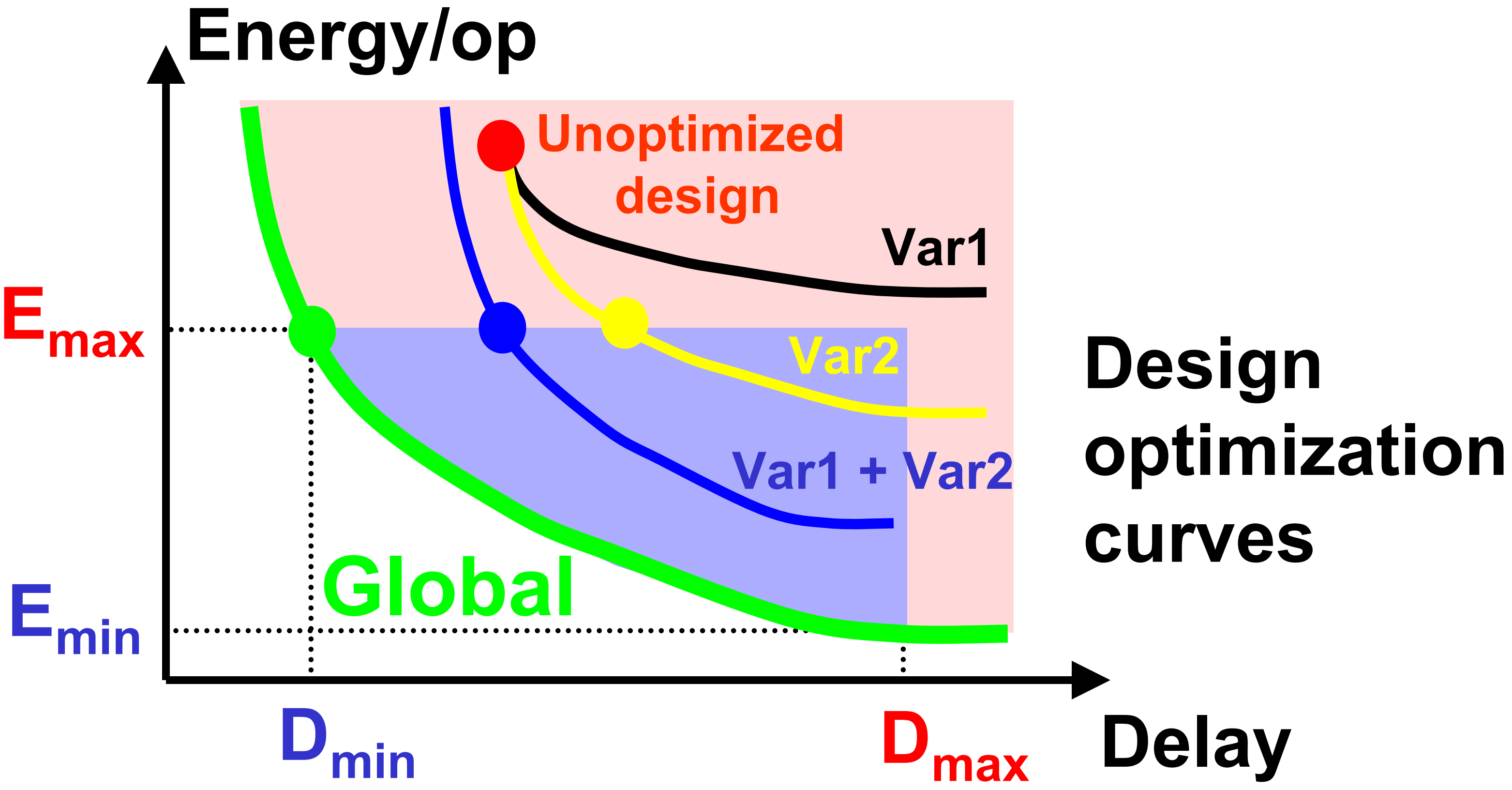
Achieve the highest performance under the power cap

Power limited operation



How far away are we from the optimal solution?

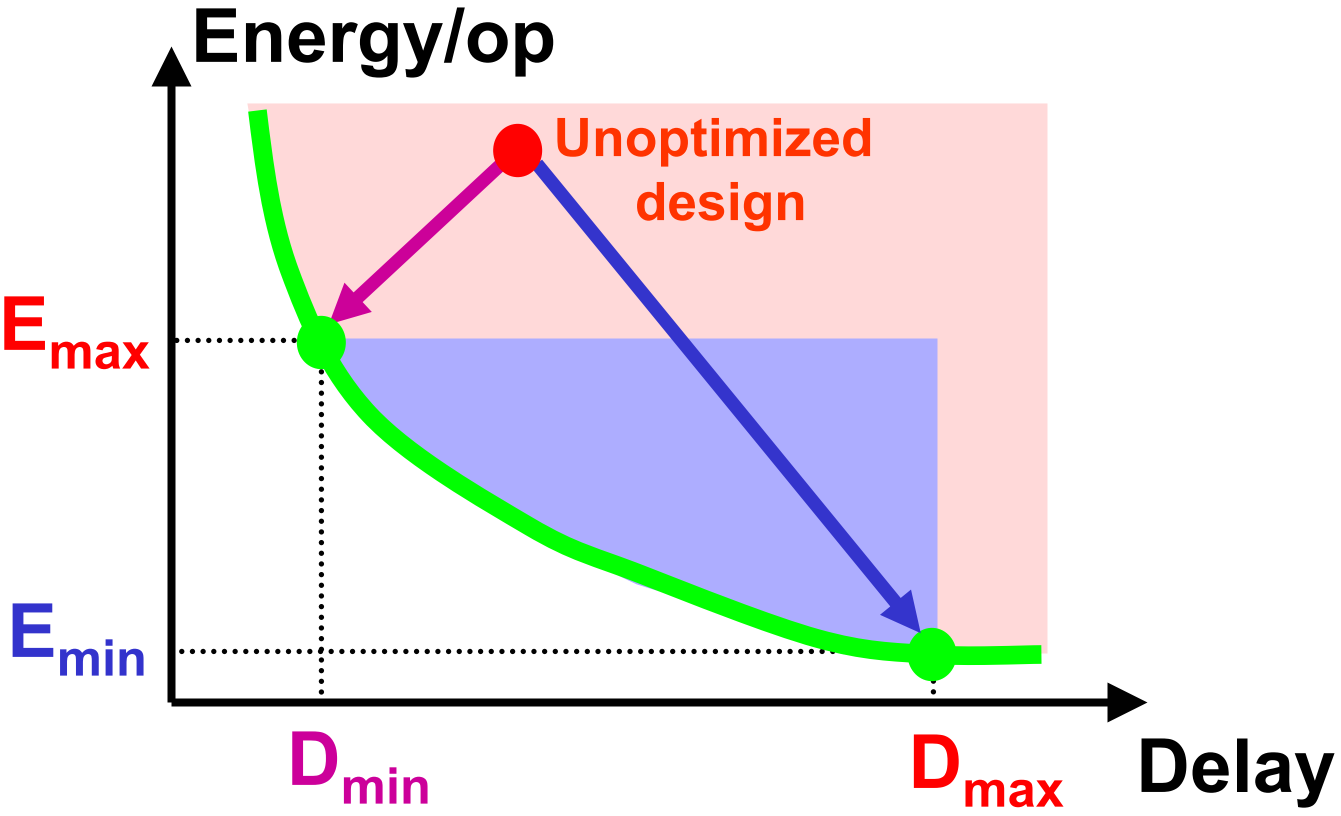
Power limited operation



Design optimization curves

Global optimum – best performance

Power limited operation

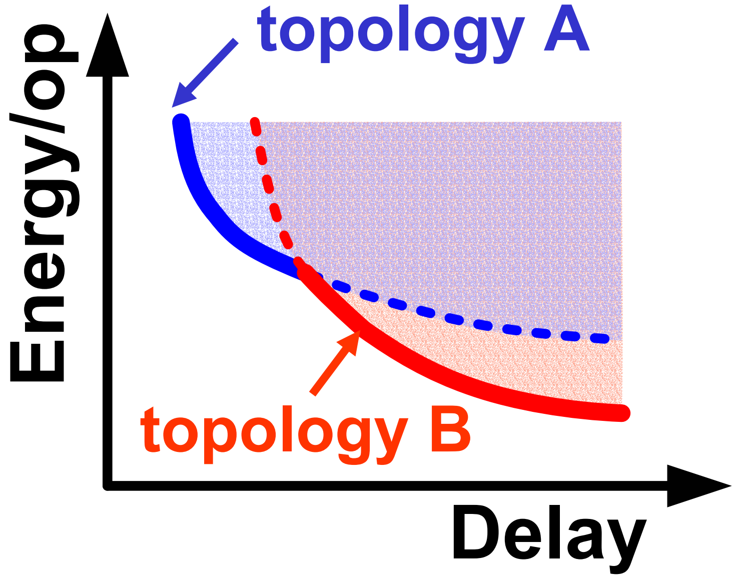


Maximize throughput for given energy or

Minimize energy for given throughput

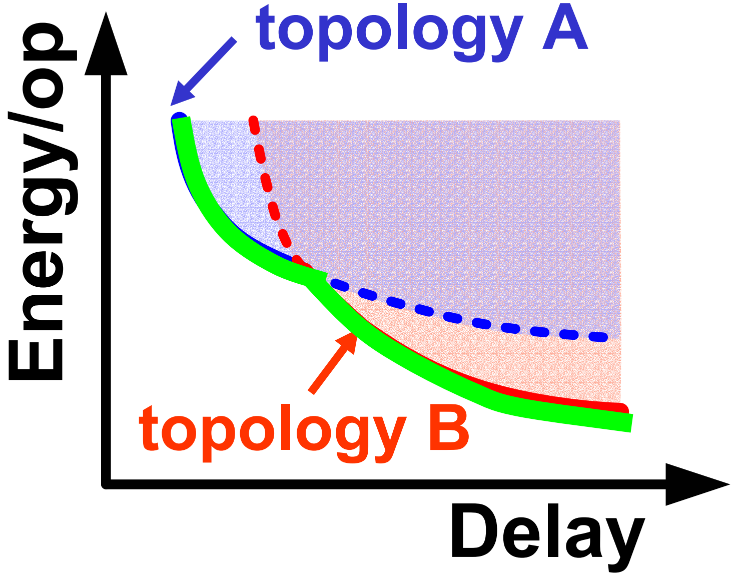
Design optimization

- There are many sets of parameters to adjust
- Tuning variables
  - Circuit
    (sizing, supply, threshold)
  - Logic style
    (domino, pass-gate, …)
  - Block topology
    (adder: CLA, CSA, …)
  - Micro-architecture
    (parallel, pipelined)



Design optimization

- There are many sets of parameters to adjust
- Tuning variables
  - Circuit
    (sizing, supply, threshold)
  - Logic style
    (domino, pass-gate, …)
  - Block topology
    (adder: CLA, CSA, …)
  - Micro-architecture
    (parallel, pipelined)



**Globally optimal boundary curve:
pieces of E-D curves for different topologies**

Outline

- **Circuit optimization**
  - Joint optimization
  - Select the most promising sets of tuning variables
- **Circuit & μArchitecture examples**
  - Adder
  - Add-Compare
- **Conclusions**

Energy-delay sensitivity

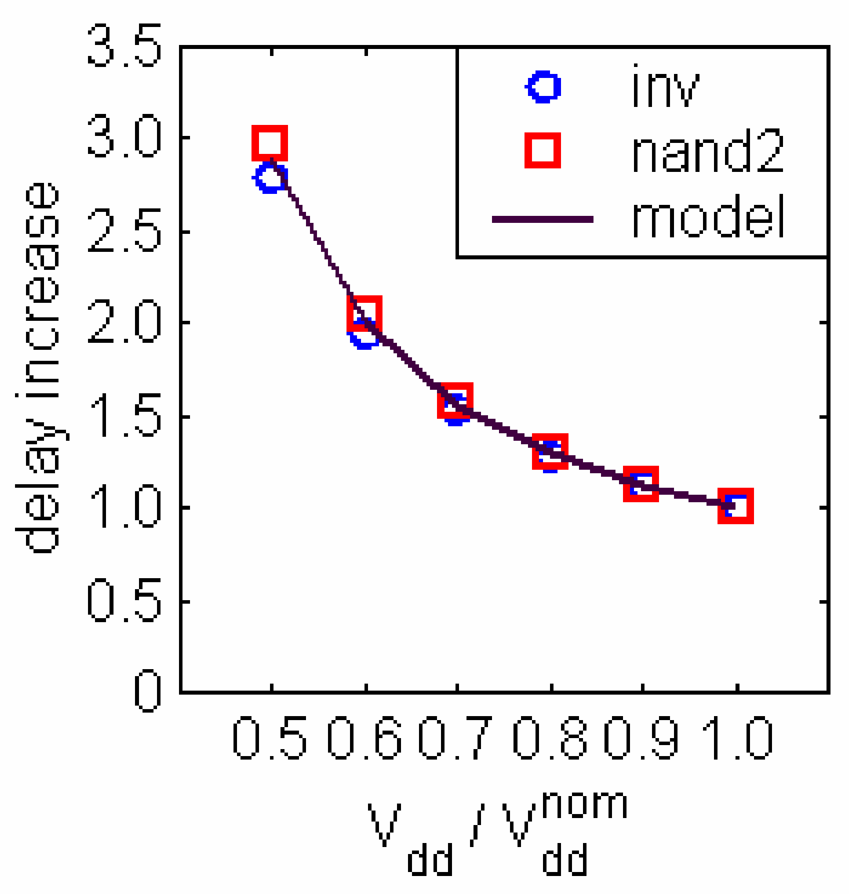
$$Sens\left(V_{dd}\right) = -\left.\frac{\partial E / \partial V_{dd}}{\partial D / \partial V_{dd}}\right|_{V_{dd}=V_{dd}*}$$

- Proposed by Zyban at *ISLPED02*
- ΔE = Sens(*A*)•(-ΔD) + Sens(*B*)•ΔD

At the optimal point,
all sensitivities should be the same

Alpha-power based delay model

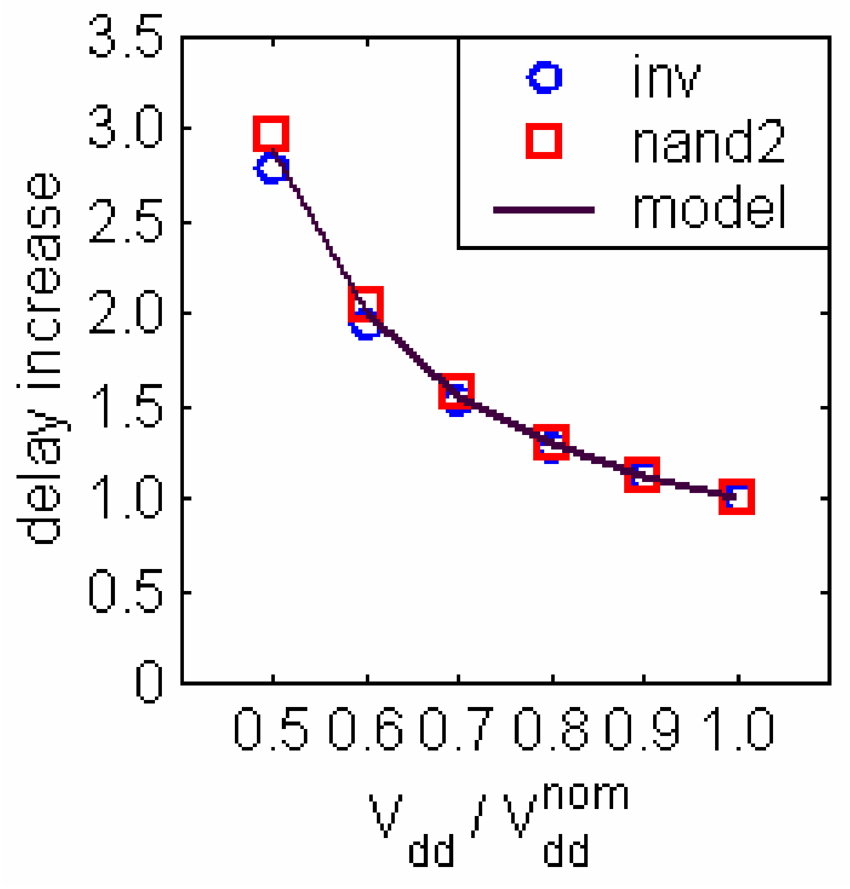
$$t_p = \frac{K_d \cdot V_{dd}}{\left(V_{dd} - V_{on}\right)^{\alpha_d}} \cdot \left(\frac{W_{out}}{W_{in}} + \frac{W_{par}}{W_{in}} \right)$$



- ♦ Fitting parameters V_{on}, α_d, K_d

Alpha-power based delay model

$$t_p = \underset{h_{eff}}{\boxed{\frac{K_d \cdot V_{dd}}{\left(V_{dd} - V_{on}\right)^{\alpha_d}} \cdot \left(\frac{W_{out}}{W_{in}}\right.}} + \left.\frac{W_{par}}{W_{in}}\right)$$



- Fitting parameters V_{on}, α_d, K_d
- Effective fanout, h_{eff}

Energy model

- Switching energy

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

- Leakage energy

$$E_{Lk} = W_{in} \cdot I_0(S_{in}) \cdot e^{\frac{-\left(V_{th} - \gamma V_{dd}\right)}{V_0}} \cdot V_{dd} \cdot D$$

Sensitivity to sizing and supply

- Gate sizing (W_i)

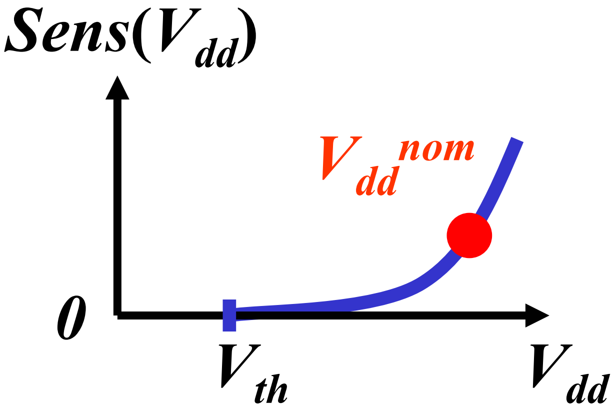
$$-\frac{\partial E_{Sw}/\partial W_i}{\partial D/\partial W_i} = \frac{ec_i}{\tau_{nom}\cdot\left(h_{eff,i}-h_{eff,i-1}\right)}$$

∞ for equal h_{eff} (D_{min})

- Supply voltage (V_{dd})

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

$$x_v = (V_{on}+\Delta V_{th})/V_{dd}$$



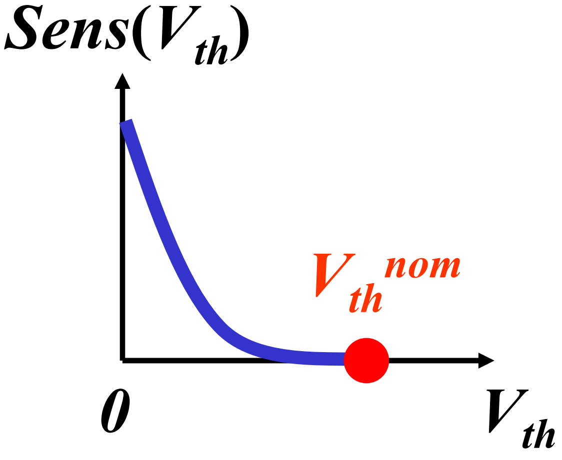
Sensitivity to Vth

- Threshold voltage (V_{th})

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

Low initial leakage

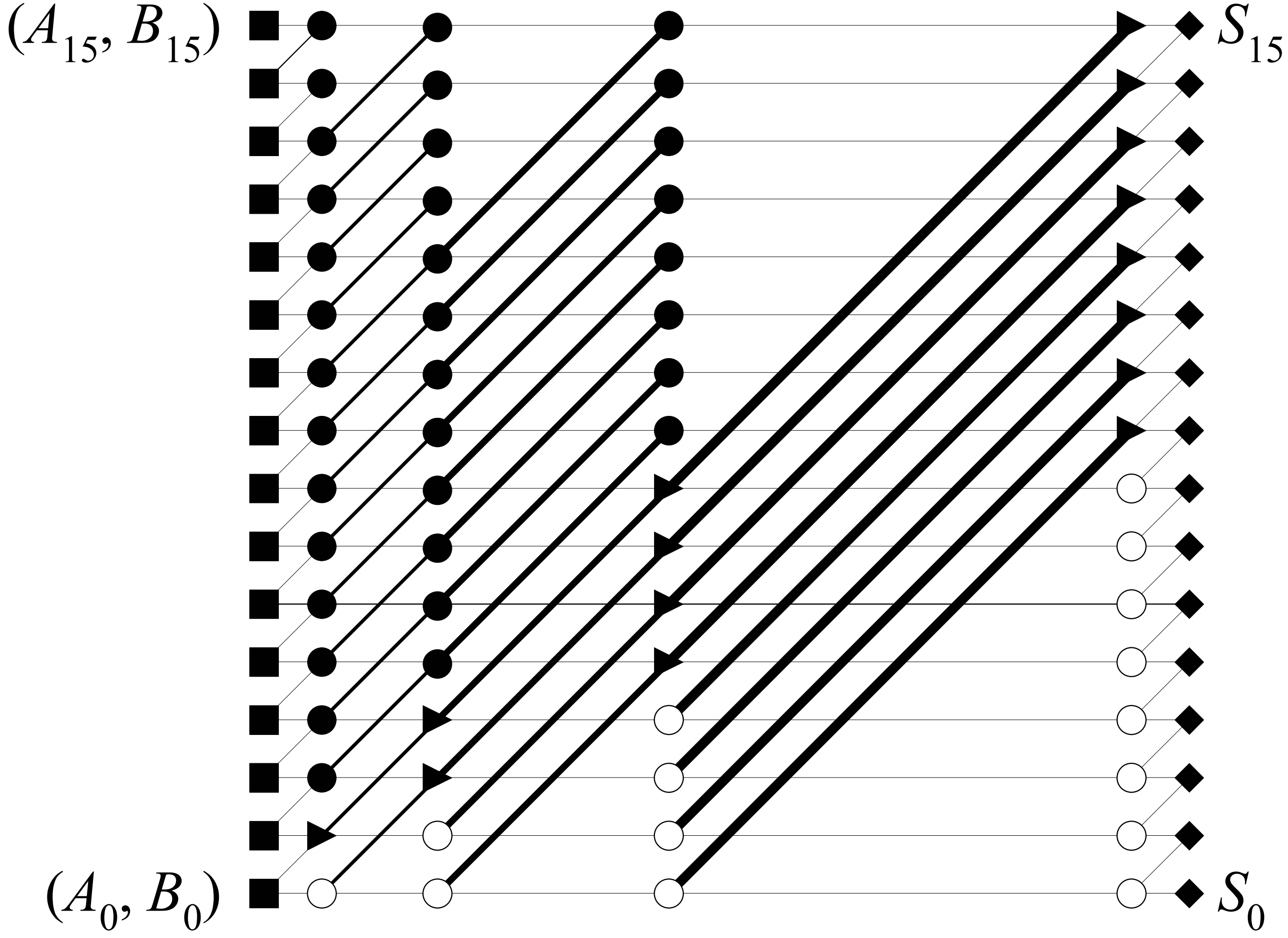
⇒ speedup comes for “free”



Optimization setup

- ♦ Reference/nominal circuit
  - sized for D_{min} @ V_{dd}^{nom}, V_{th}^{nom}
  - known average activity
- ♦ Set delay constraint
- ♦ Minimize energy under delay constraint
  - gate sizing
  - V_{dd} , V_{th} scaling

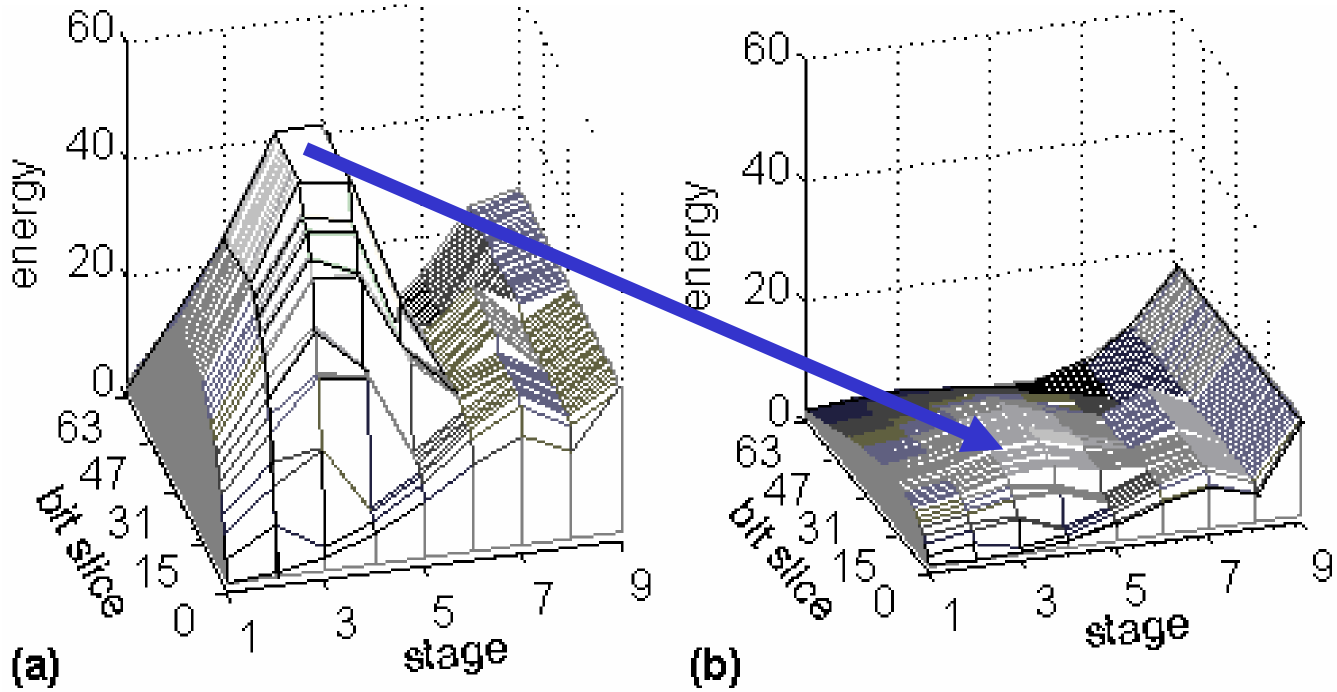
Kogge-Stone tree adder topology



- Off-path load (gates + wires)
- Reconvergence (inside ●-block)

Tree adder: Sizing optimization

- Reference: all paths are critical

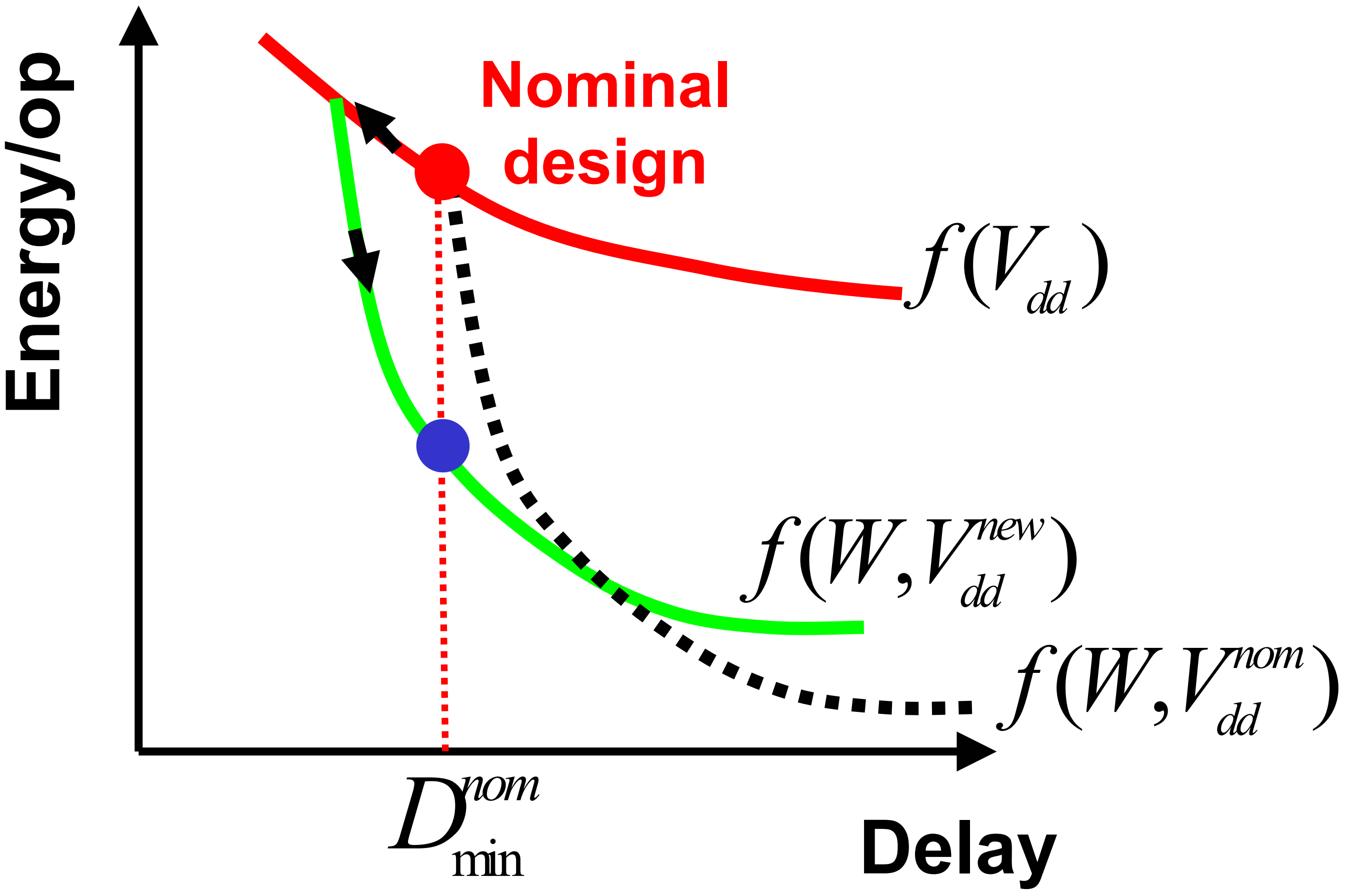


Nominal
(D_{min}, E_{nom})

Sizing opt.
($1.1D_{min}$, $0.3E_{nom}$)

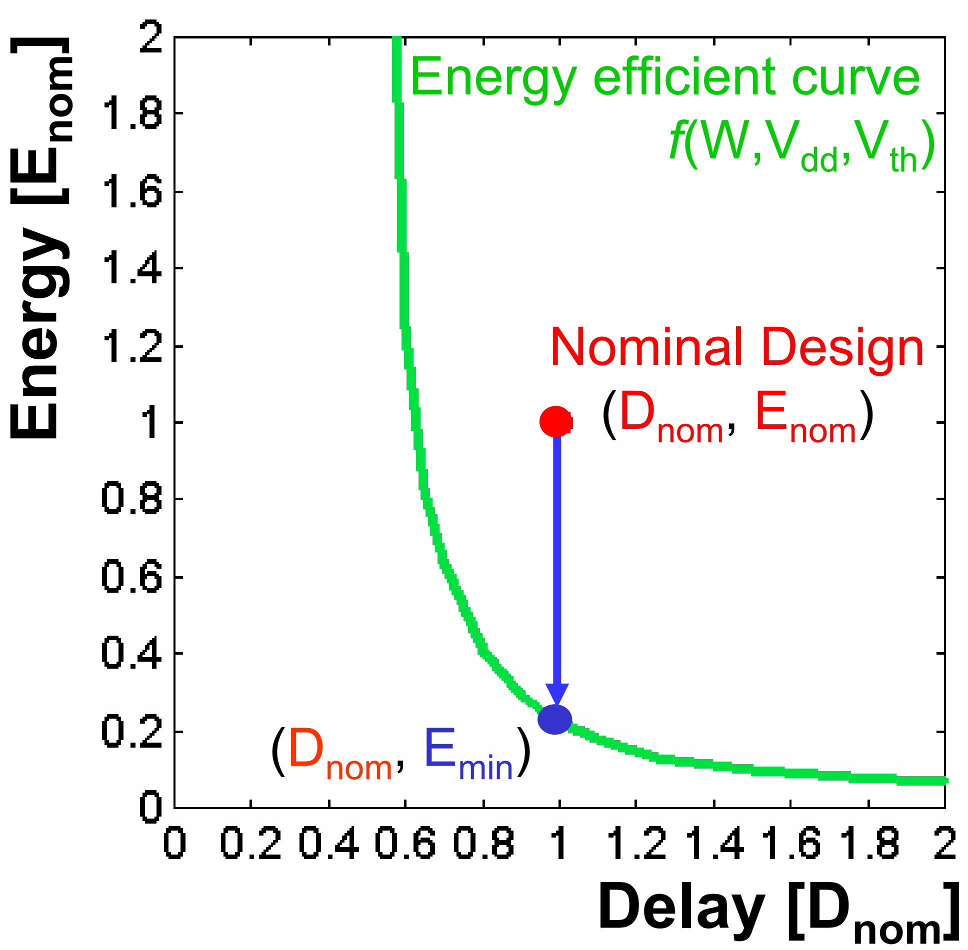
Internal energy peaks ⇒
Big savings for small delay penalty with resizing

Joint optimization: sizing and Vdd



ΔE = Sens(V_{dd})•(-ΔD) + Sens(W)•ΔD

Results of joint optimization

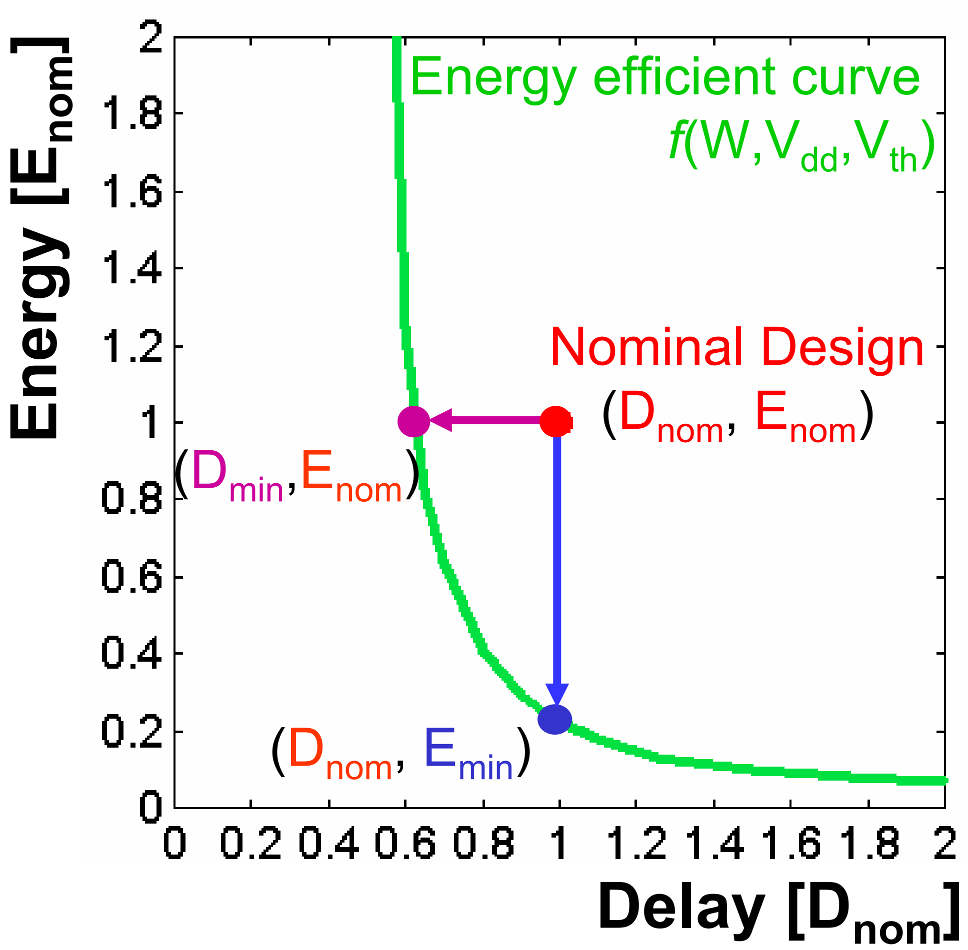


Sensitivity table

Sens	W	Vdd	Vth
(D_{nom}, E_{nom})	∞	**1.5**	**0.2**
(D_{nom}, E_{min})	**1 (reference)**		

80% of energy saved without delay penalty

Results of joint optimization



Sensitivity table

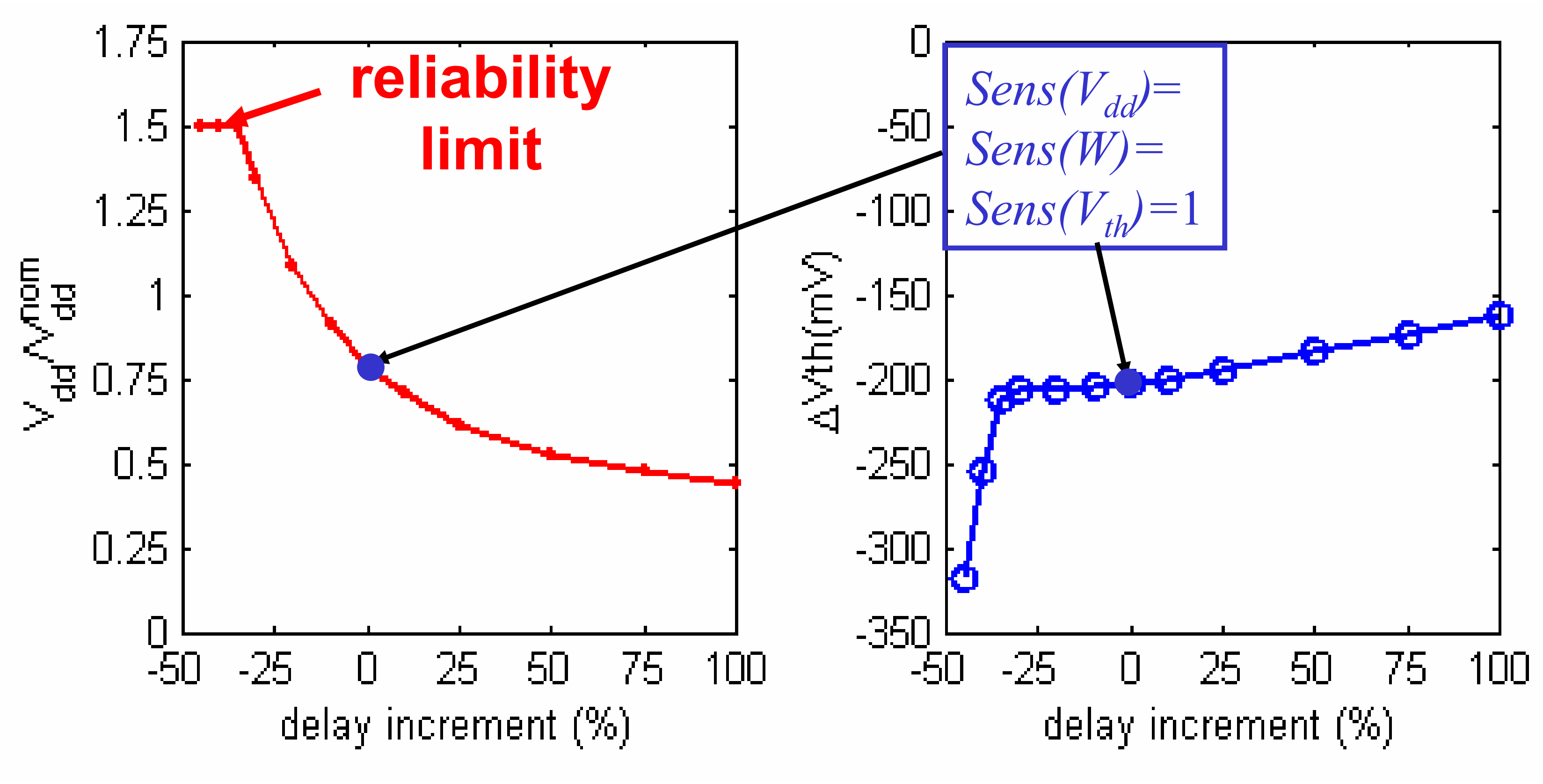
Sens	W	Vdd	Vth
(D_{nom}, E_{nom})	∞	**1.5**	**0.2**
(D_{nom}, E_{min})	**1 (reference)**		
(D_{min}, E_{nom})	**22**	**16**	**22**

80% of energy saved without delay penalty

40% speedup for same energy

A look at tuning variables

Supply **Threshold**

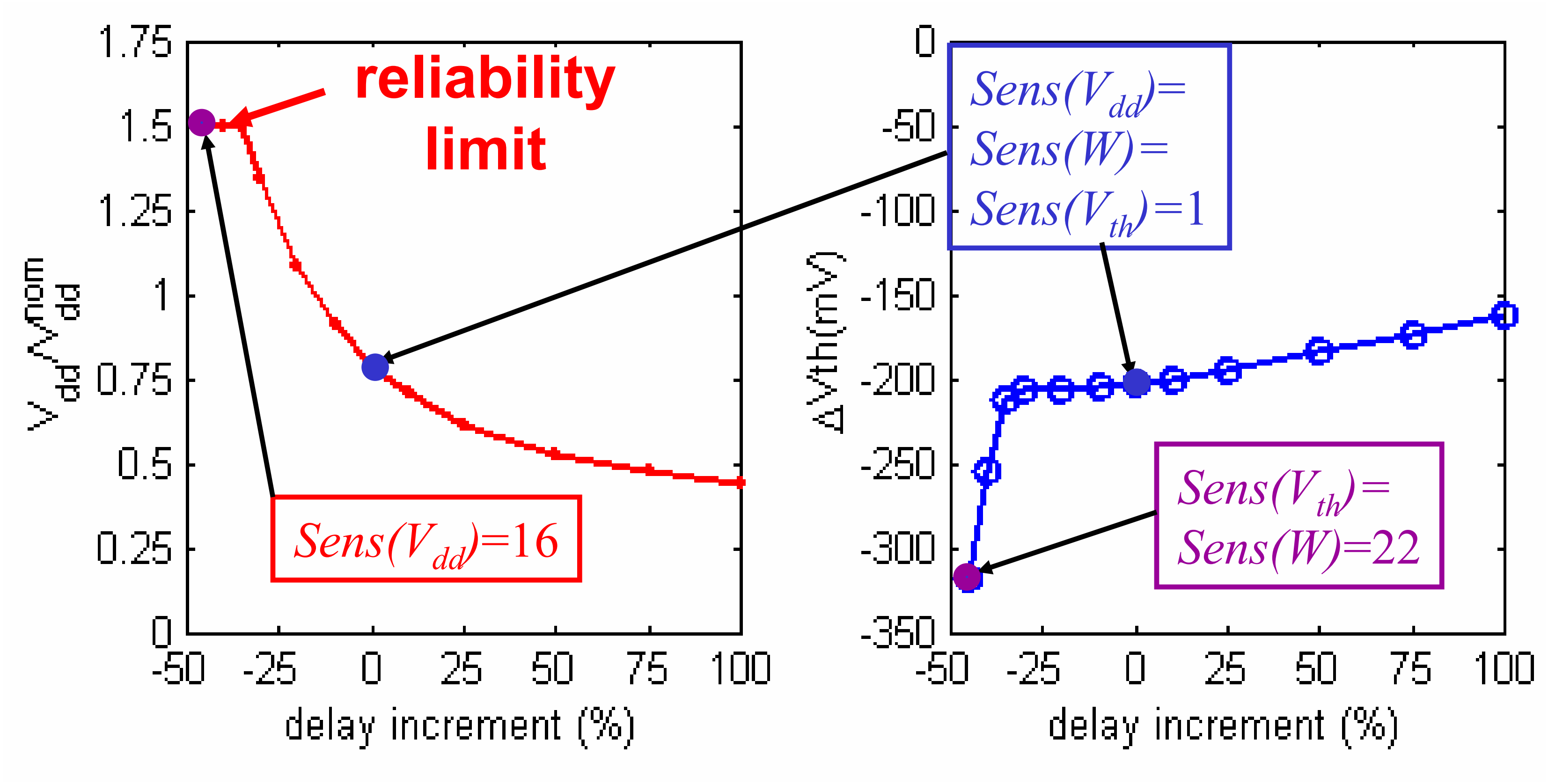


reliability limit

$Sens(V_{dd}) =$
$Sens(W) =$
$Sens(V_{th}) = 1$

Limited range of tuning variables

A look at tuning variables



Supply

Threshold

reliability limit

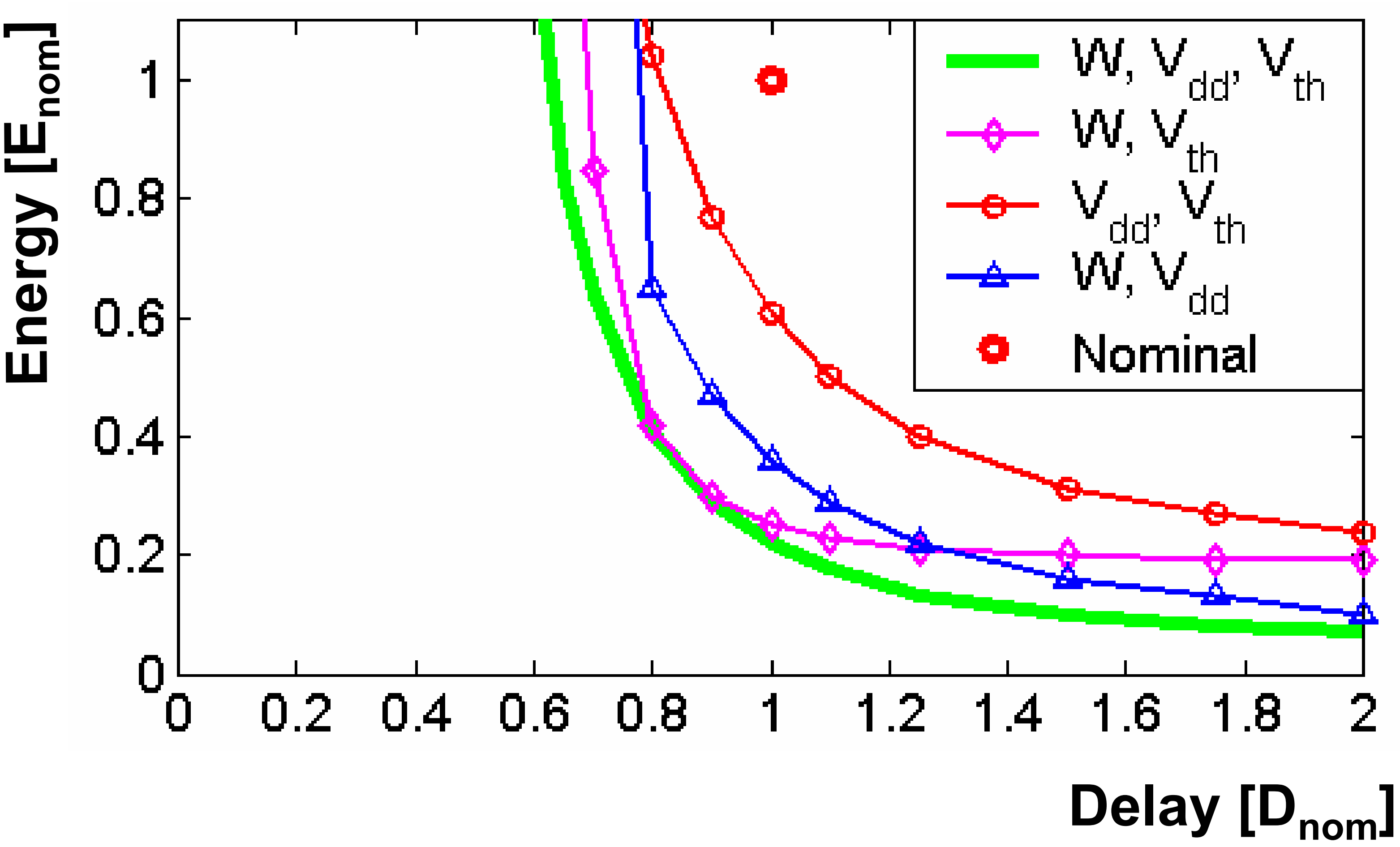
Sens(V_{dd})=16

Sens(V_{dd})=
Sens(*W*)=
Sens(V_{th})=1

Sens(V_{th})=
Sens(*W*)=22

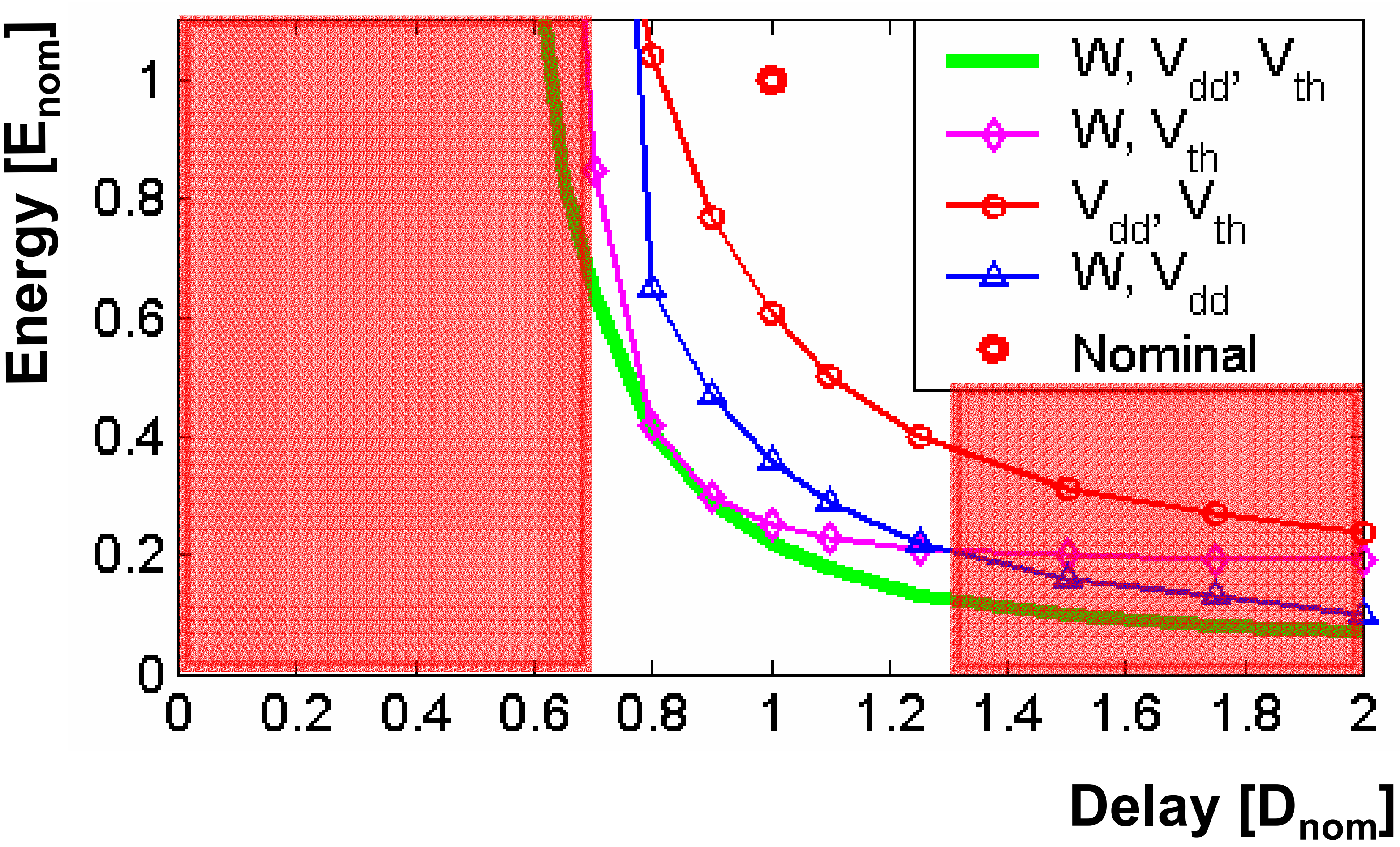
Limited range of tuning variables

Reducing the number of dimensions



Threshold and sizing nearly optimal around the nominal point

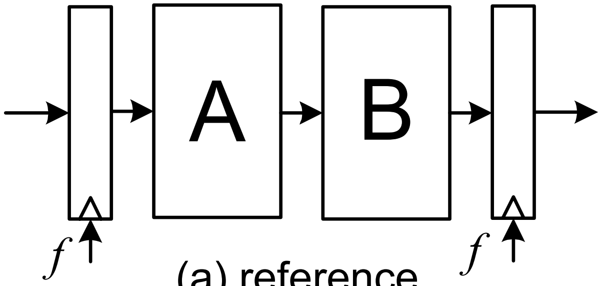
Scope of circuit optimization



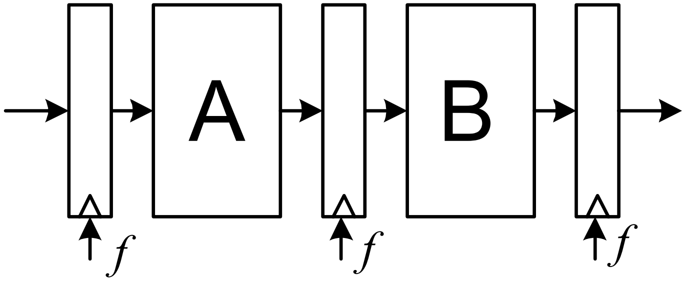
Effective region +/-30% around nominal delay

Circuit & μArchitecture optimization

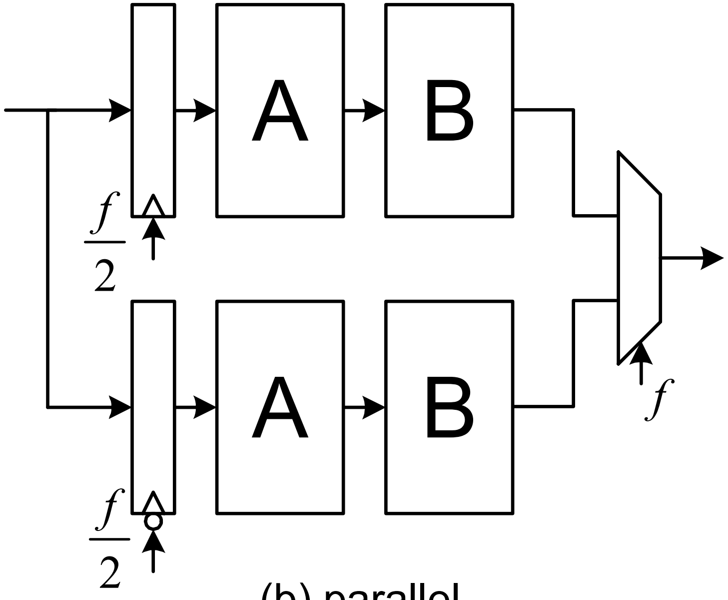
- Revisit the old argument for parallelism



(a) reference



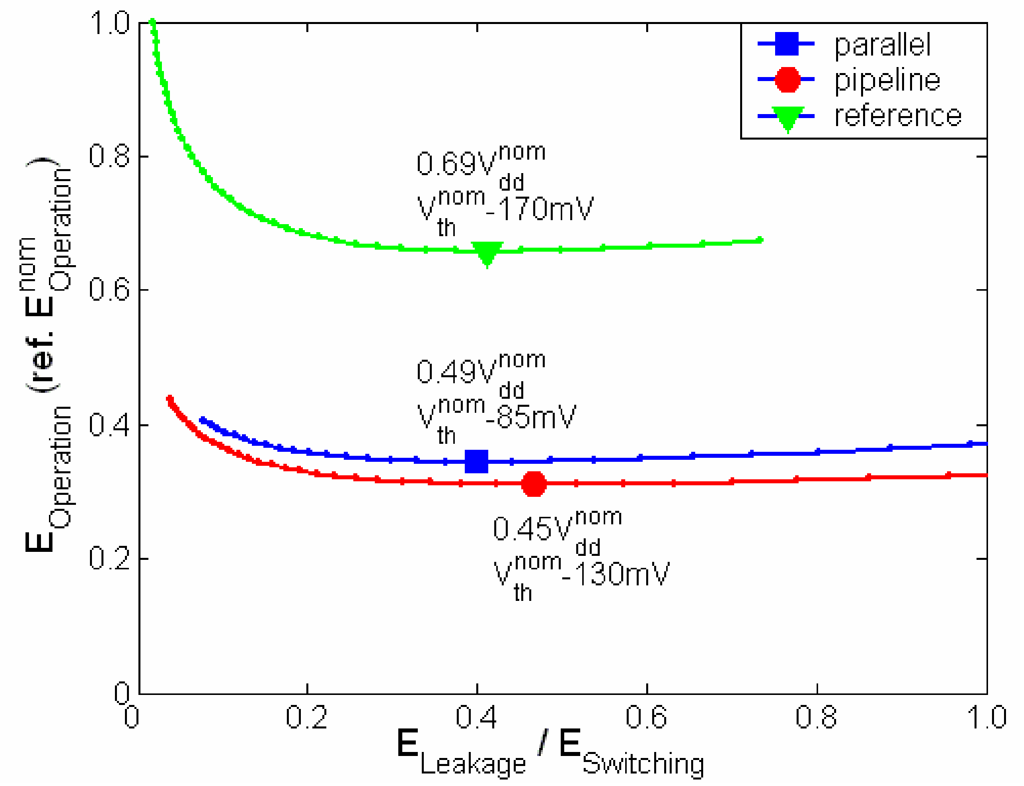
(c) pipeline



(b) parallel

- What happens if we can choose optimal Vdd and Vth for each design?

Balance of leakage and switching energy



$$\left.\frac{E_{Lk}}{E_{Sw}}\right|_{Opt} = \frac{2}{\ln\left(\frac{L_d}{\alpha_{avg}}\right) - K_{tech}}$$

Optimal designs have high leakage current

Conclusions

- All design levels need to be optimized jointly
- Equal marginal costs ⇒ Energy-efficient design
- Peak performance is VERY power inefficient
- Today's designs are not leaky enough to be truly power-optimal
- Pipelining starts to gain advantage over parallelism