inst.eecs.berkeley.edu/~ee241b

EE241B : Advanced Digital Circuits

Lecture 18 – Power-Performance Tradeoffs 2

Borivoje Nikolić

MarketWatch, March 28: Opinion: There's no returning to regular schooling as online learning goes mainstream, by Alex Hicks

When in-person education resumes, online learning tools and methods will be entrenched in the system

Announcements

- Project midterm reports due today, March 31
  - Please e-mail me the link to your web page
- Assignment 3 due Thursday, April 2.
  - Quiz next Tuesday
- Reading – req'd
  - Rabaey et al, LPDE, Ch. 4

Outline

- Module 5
  - Power-performance tradeoffs

5.B Power-Performance Tradeoffs

Know Your Enemy

- Where does power go in CMOS?
- Switching (dynamic) power
  - Charging capacitors
- Leakage power
  - Transistors are imperfect switches
- Short-circuit power
  - Both pull-up and pull-down on during transition
- Static currents
  - Biasing currents

Summary of Power Dissipation Sources

$$P \sim \alpha \cdot \left(C_L + C_{CS}\right) \cdot V_{swing} \cdot V_{DD} \cdot f + \left(I_{DC} + I_{Leak}\right) \cdot V_{DD}$$

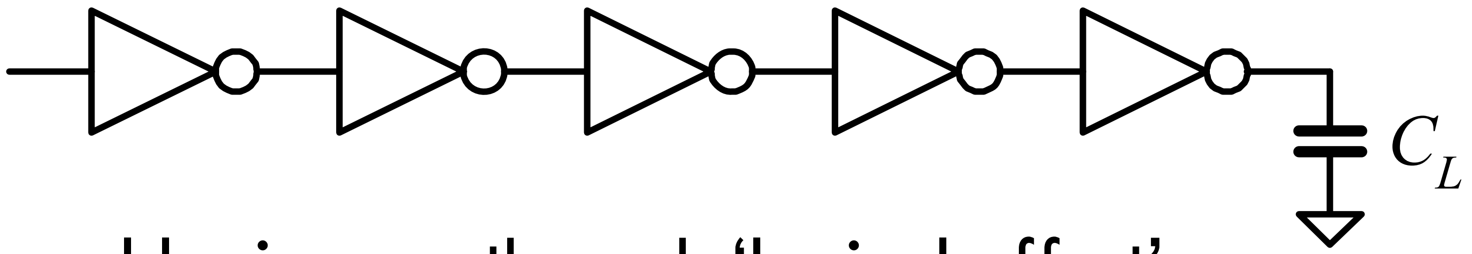
- α – switching activity
- C_L – load capacitance
- C_{CS} – short-circuit “capacitance”
- V_{swing} – voltage swing
- f – frequency

- I_{DC} – static current
- I_{leak} – leakage current

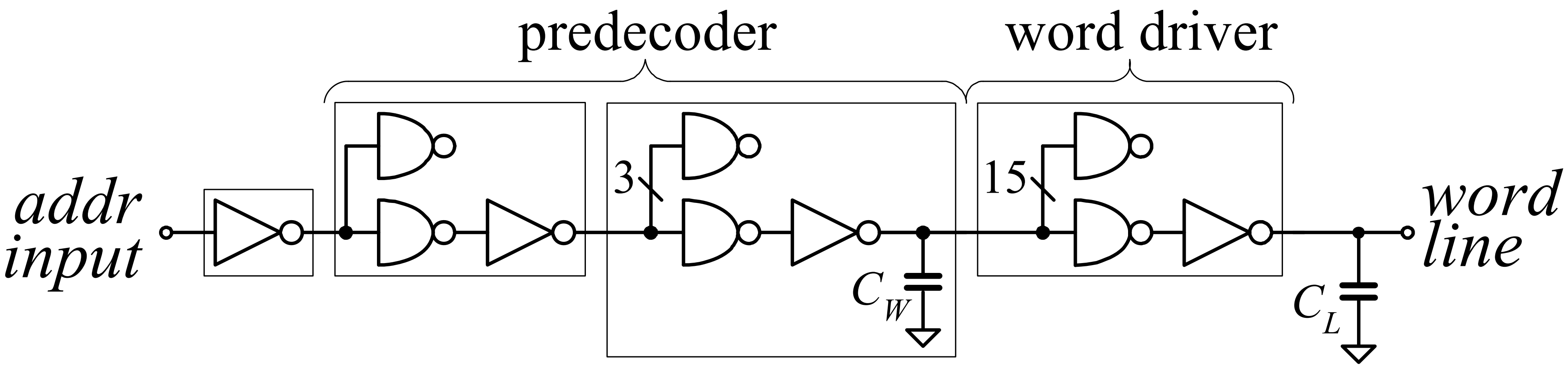
$$P = \frac{energy}{operation} \times rate + static\ power$$

CMOS Performance Optimization

- Reminder - sizing: Optimal performance with equal fanout per stage

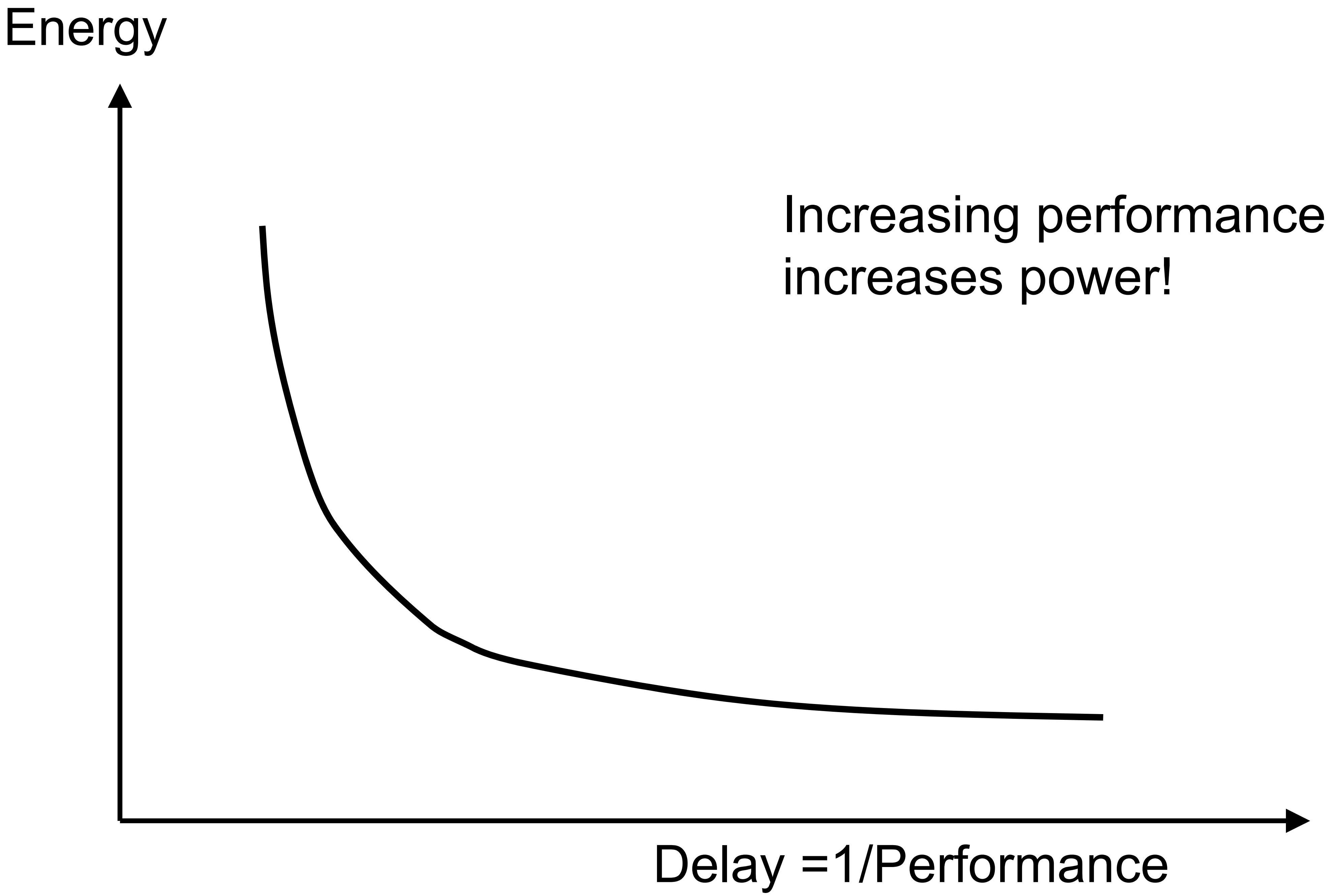


- Extendable to general logic cone through 'logical effort'
- Equal effective fanouts ($g_i C_{i+1}/C_i$) per stage
- Optimal fanout is around 4



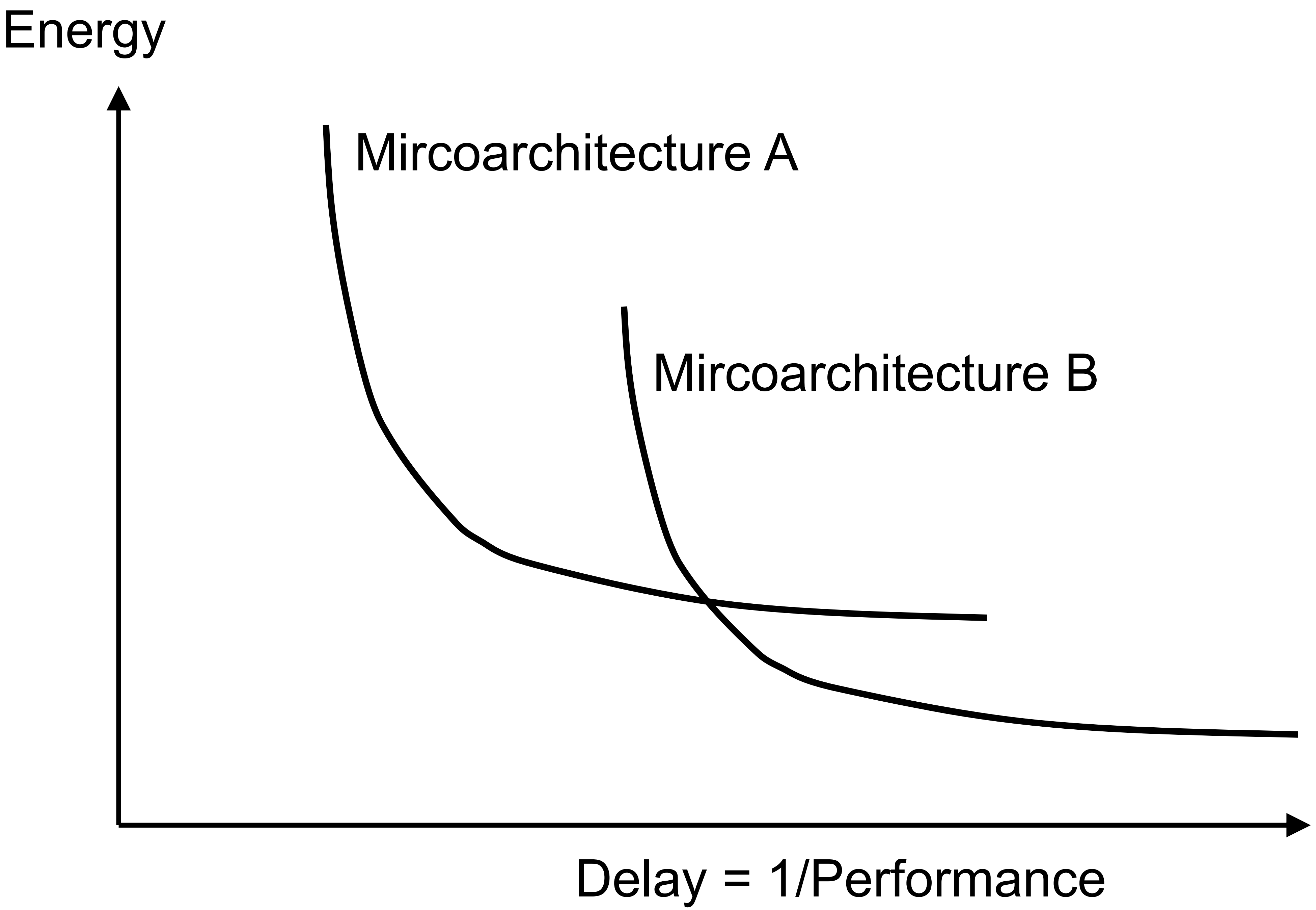
[Ref: I. Sutherland, Morgan-Kaufman'98]

Performance Optimization



Increasing performance increases power!

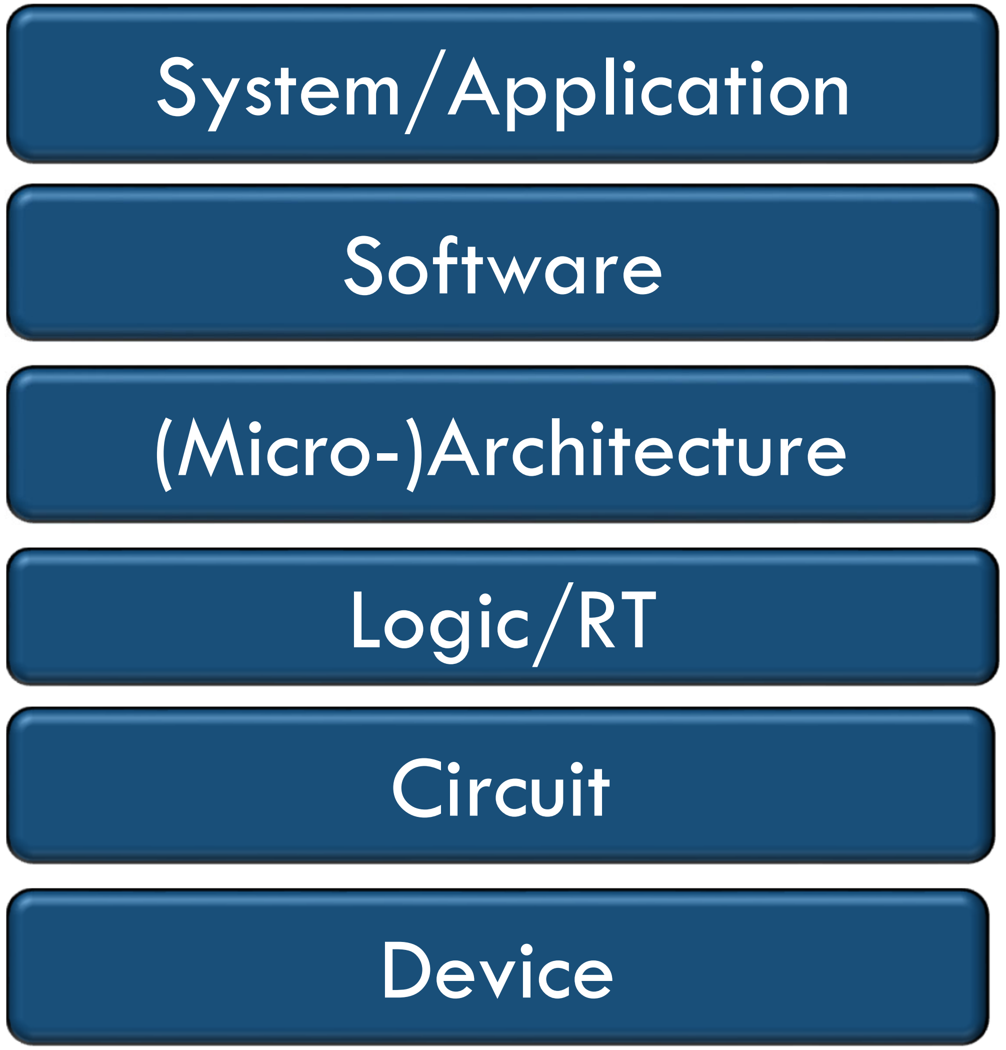
Performance Optimization



The Design Abstraction Stack

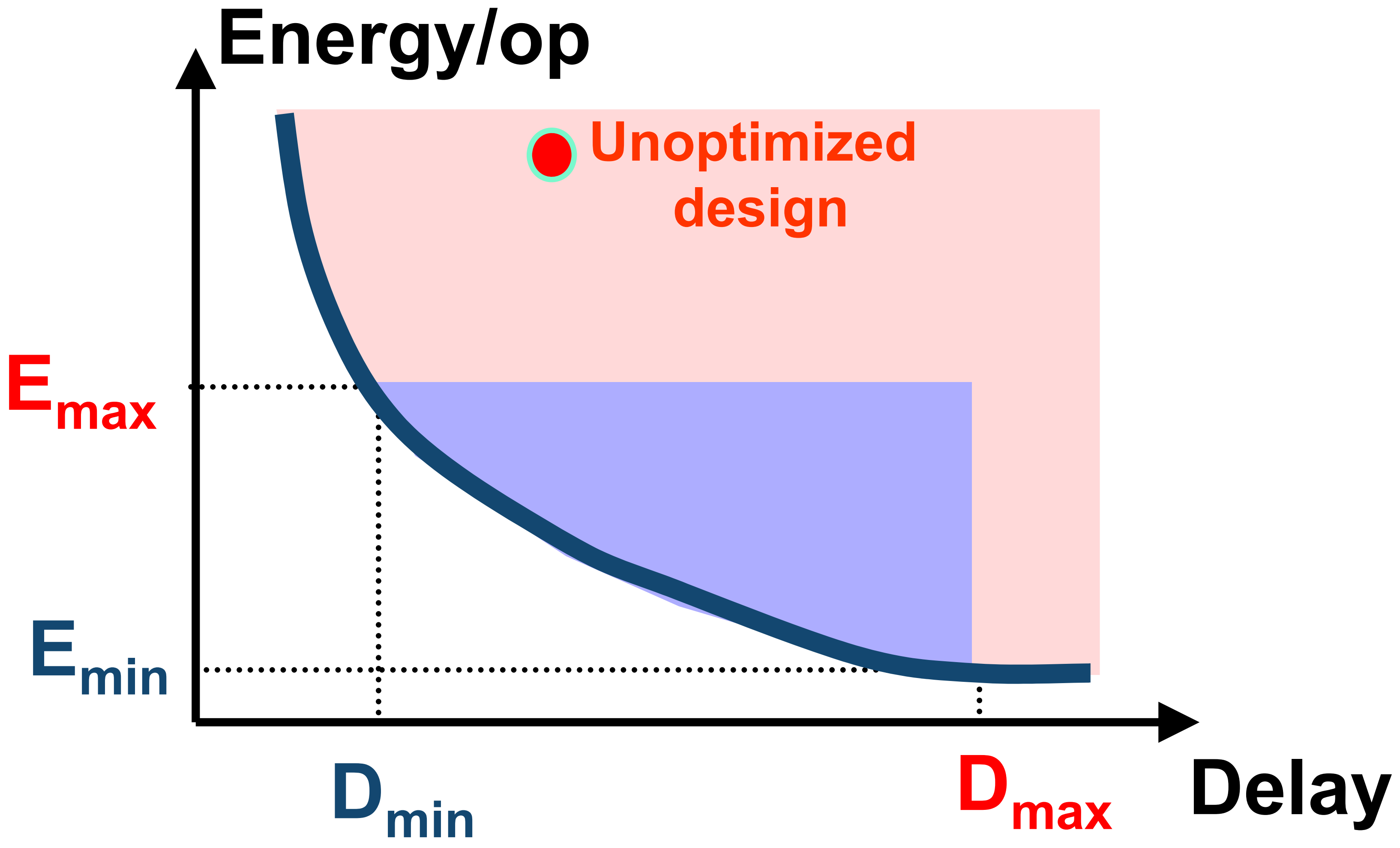
A very rich set of design parameters to consider!

It helps to consider options in relation to their abstraction layer



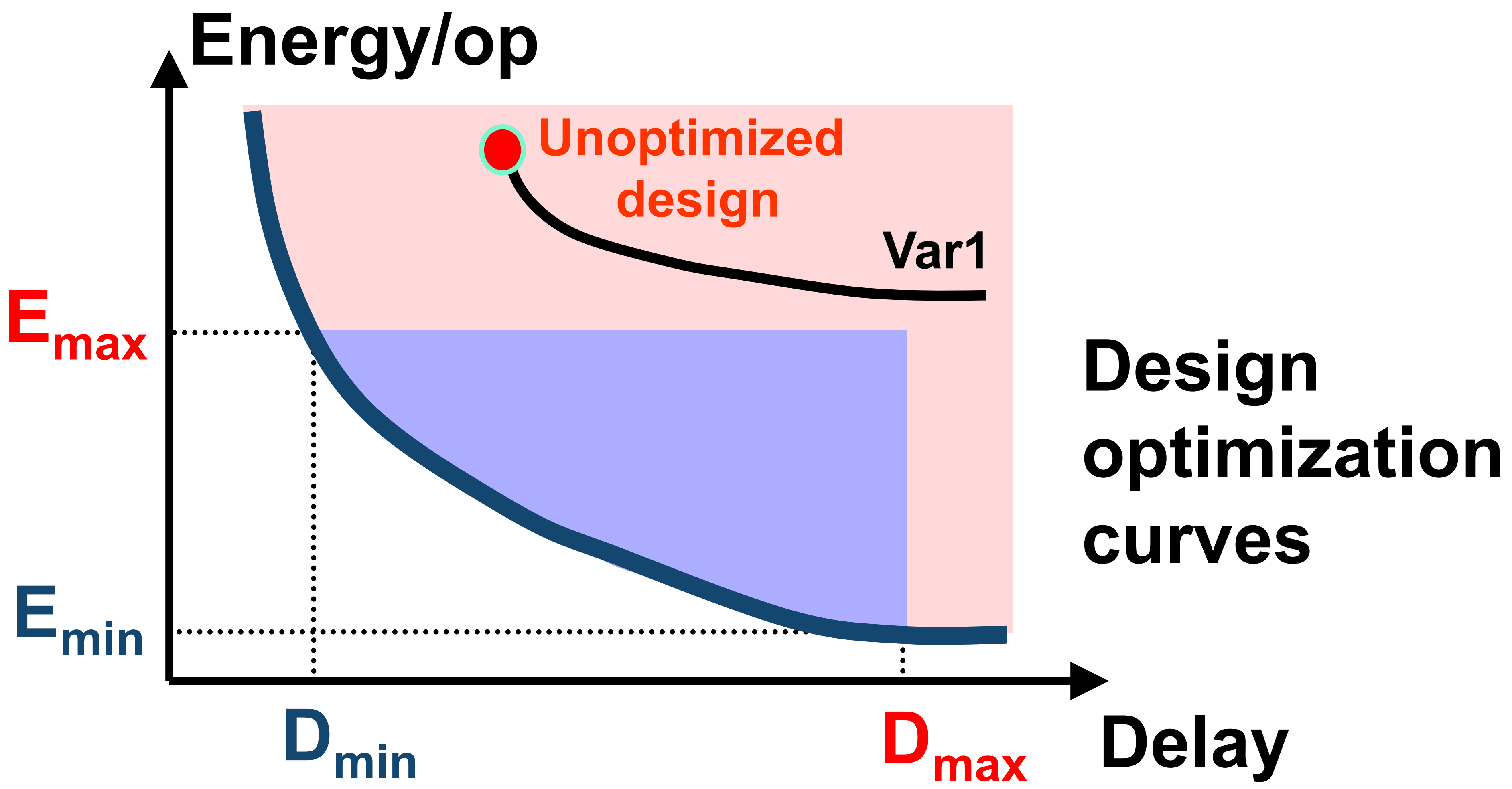
System/Application	Choice of algorithm
Software	Amount of concurrency
(Micro-)Architecture	Parallel versus pipelined, general purpose versus application specific
Logic/RT	logic family, standard cell versus custom, clocking
Circuit	sizing, supplies, thresholds
Device	Bulk, PDSOI, FDSOI, finFET

Power-Performance Optimization



Achieve the highest performance under the power cap

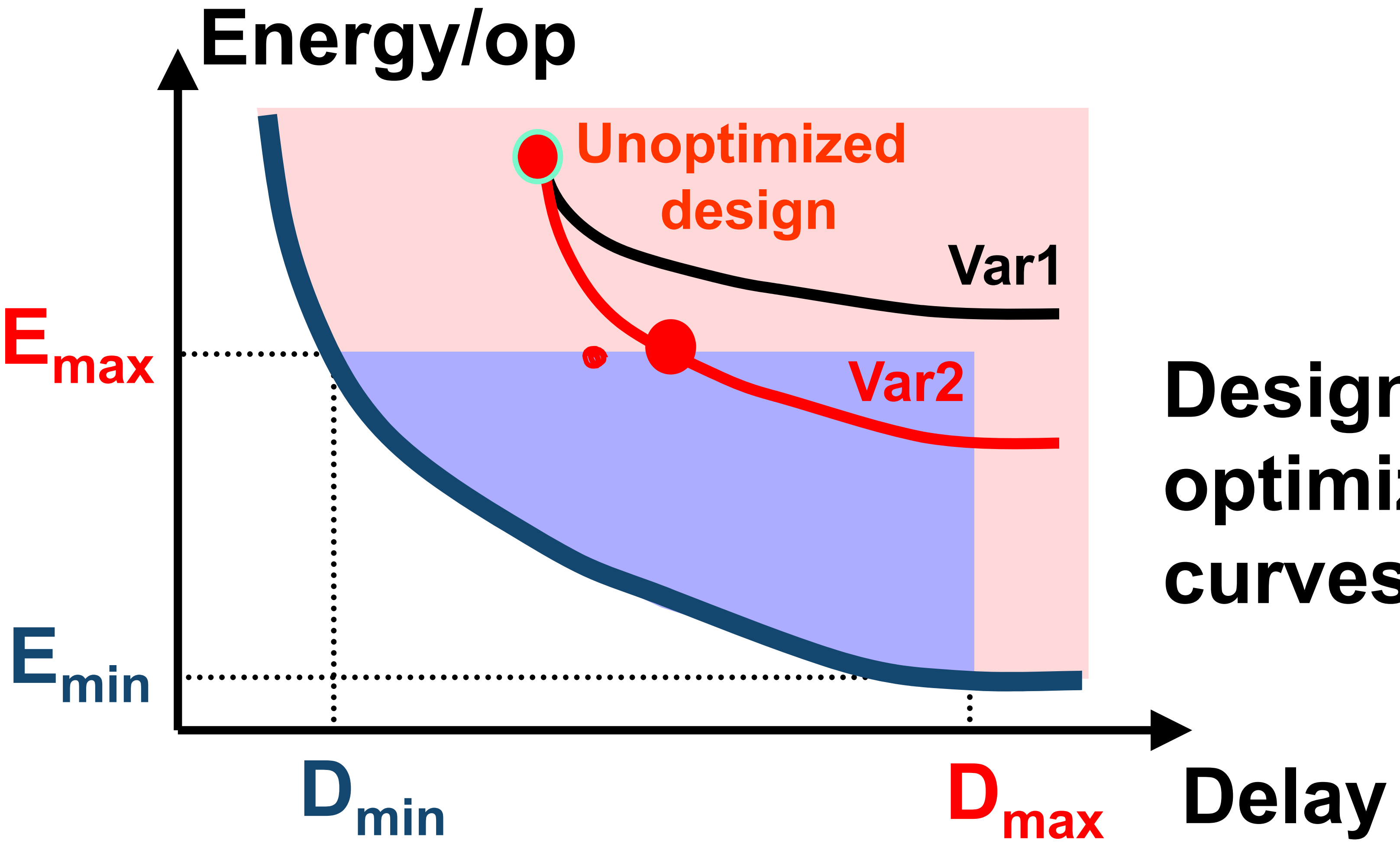
Power-Performance Optimization



Design optimization curves

Achieve the highest performance under the power cap

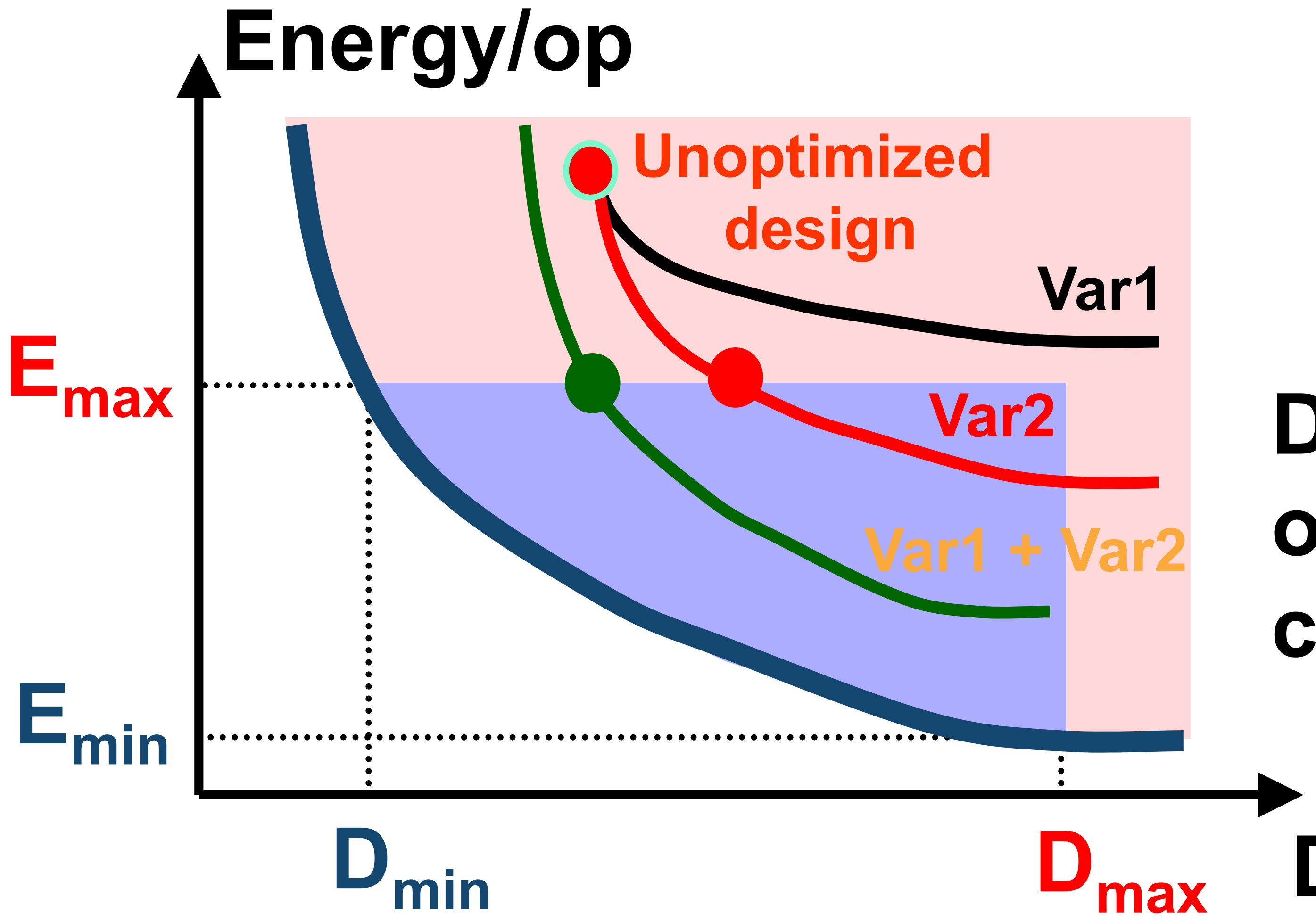
Power-Performance Optimization



Design optimization curves

Achieve the highest performance under the power cap

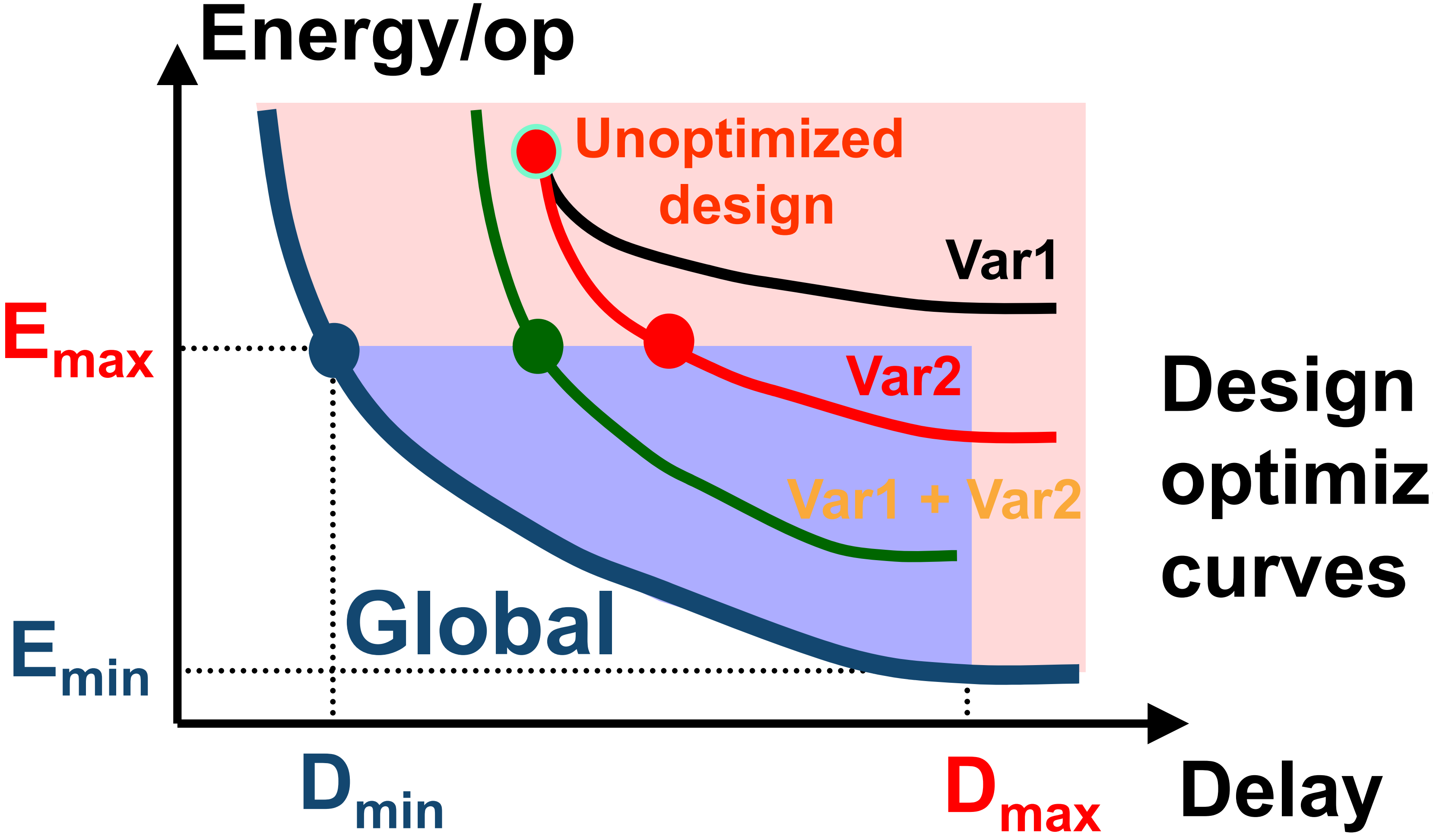
Power-Performance Optimization



Design optimization curves

How far away are we from the optimal solution?

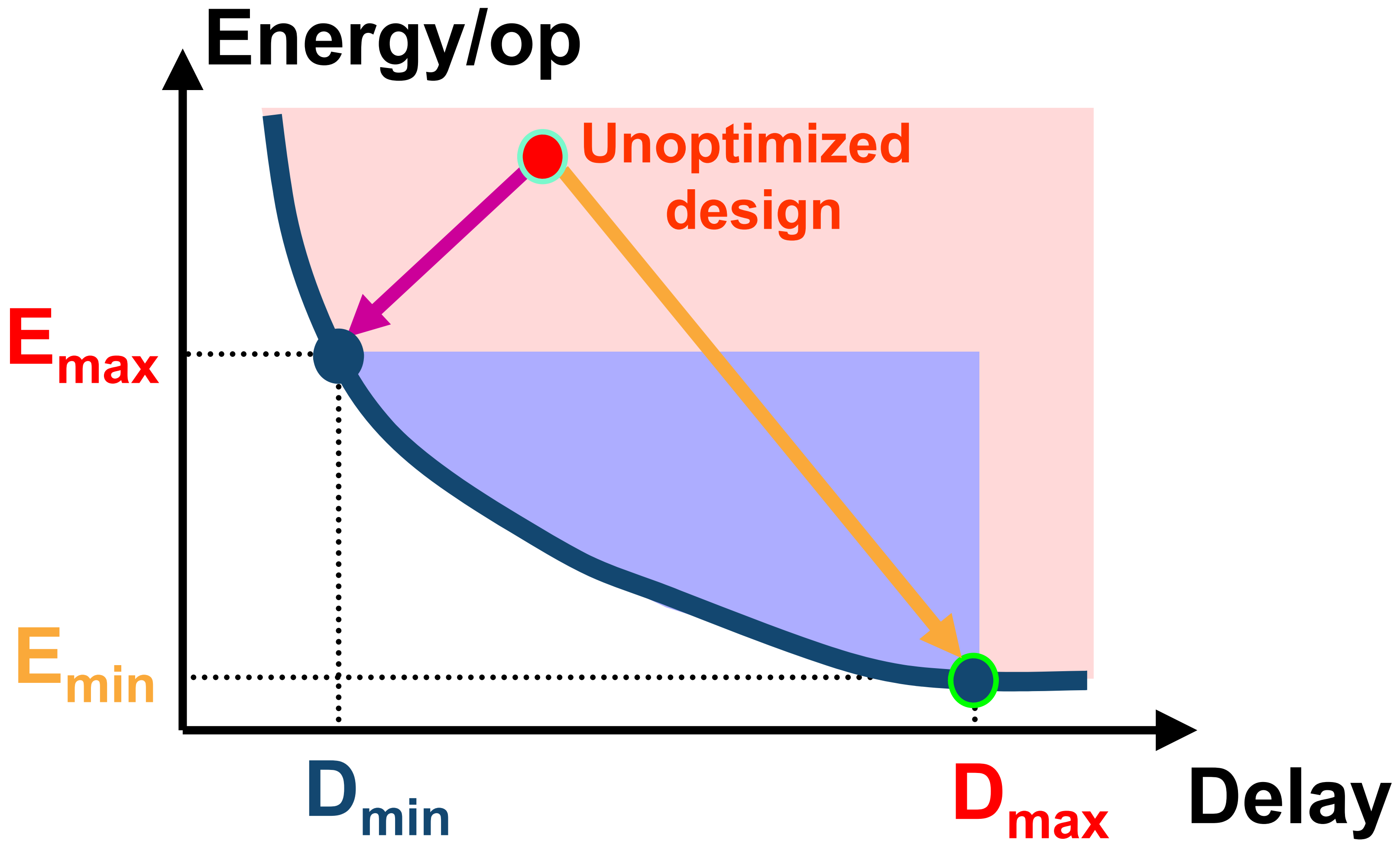
Power-Performance Optimization



Design optimization curves

Global optimum – best performance

Power-Performance Optimization



Maximize throughput for given energy or

Minimize energy for given throughput

Power-Performance Optimization

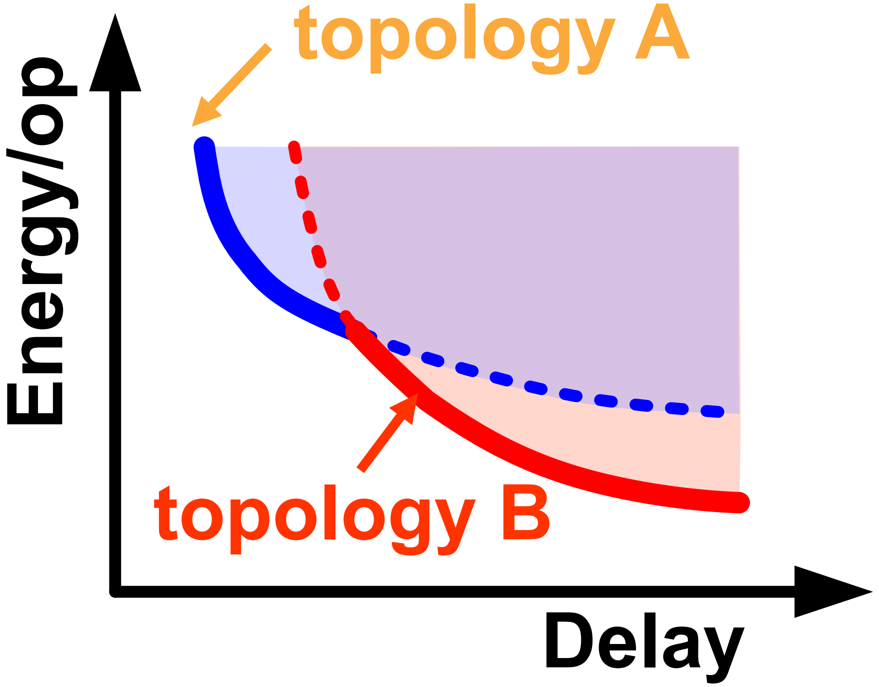
- There are many sets of parameters to adjust
  - Tuning variables
  - Circuit

    (sizing, supply, threshold)
  - Logic style

    (std. cells, custom , …)
  - Block topology

    (adder: CLA, CSA, …)
  - Micro-architecture

    (parallel, pipelined)



Power-Performance Optimization

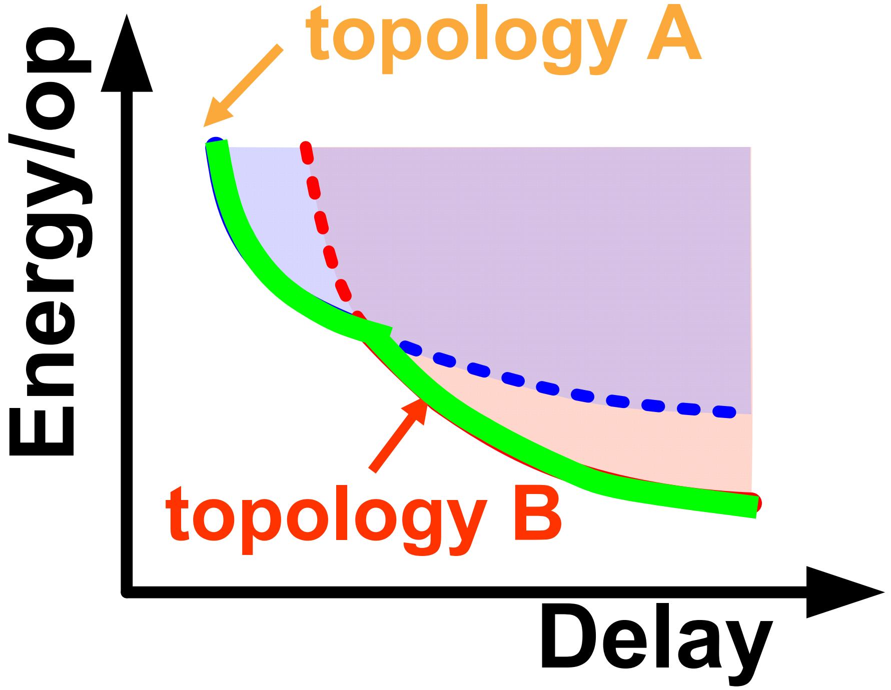
- There are many sets of parameters to adjust
  - Tuning variables
  - Circuit

    (sizing, supply, threshold)
  - Logic style

    (std. cells, custom , …)
  - Block topology

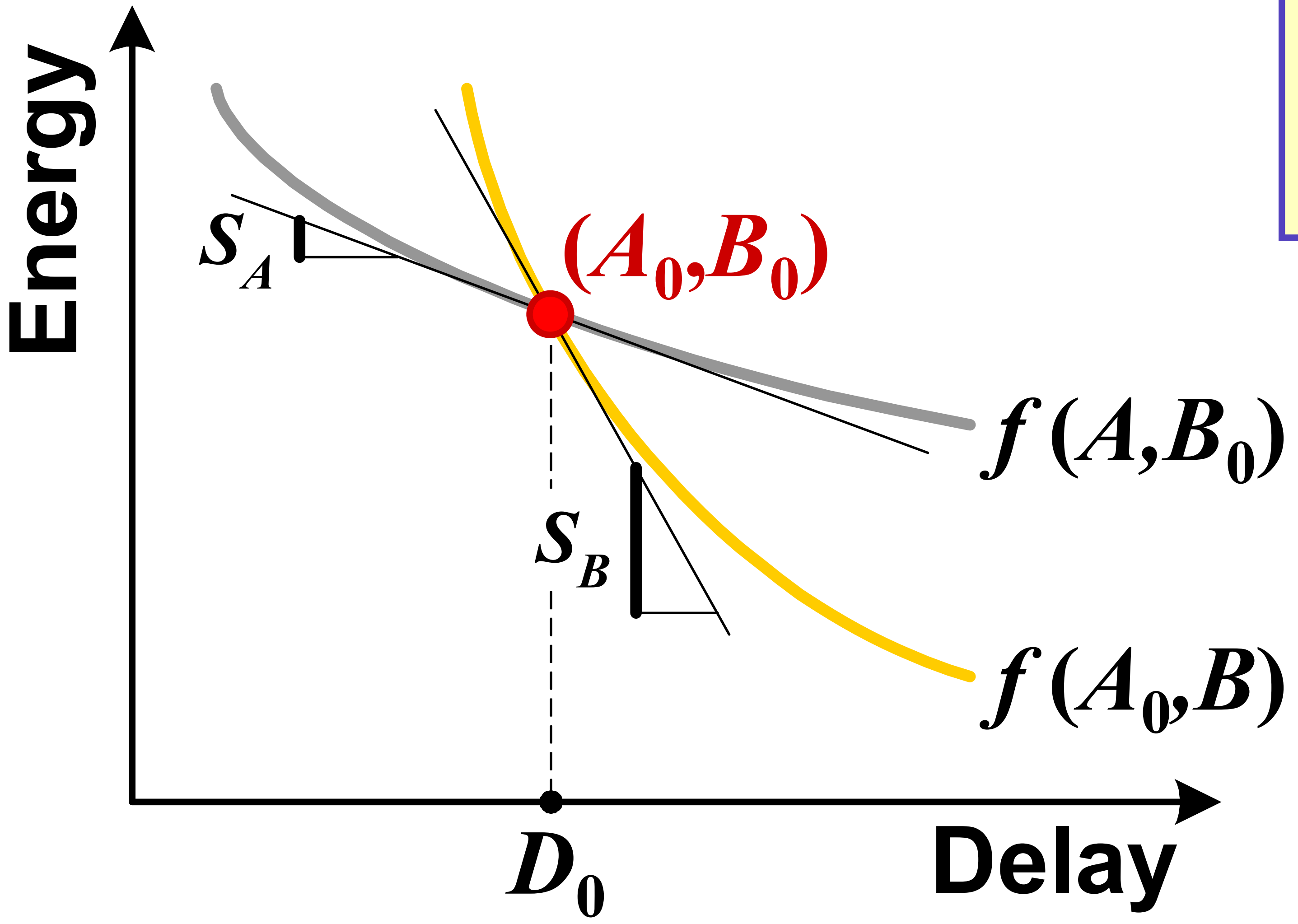
    (adder: CLA, CSA, …)
  - Micro-architecture

    (parallel, pipelined)



Globally optimal power-performance curve for a given function

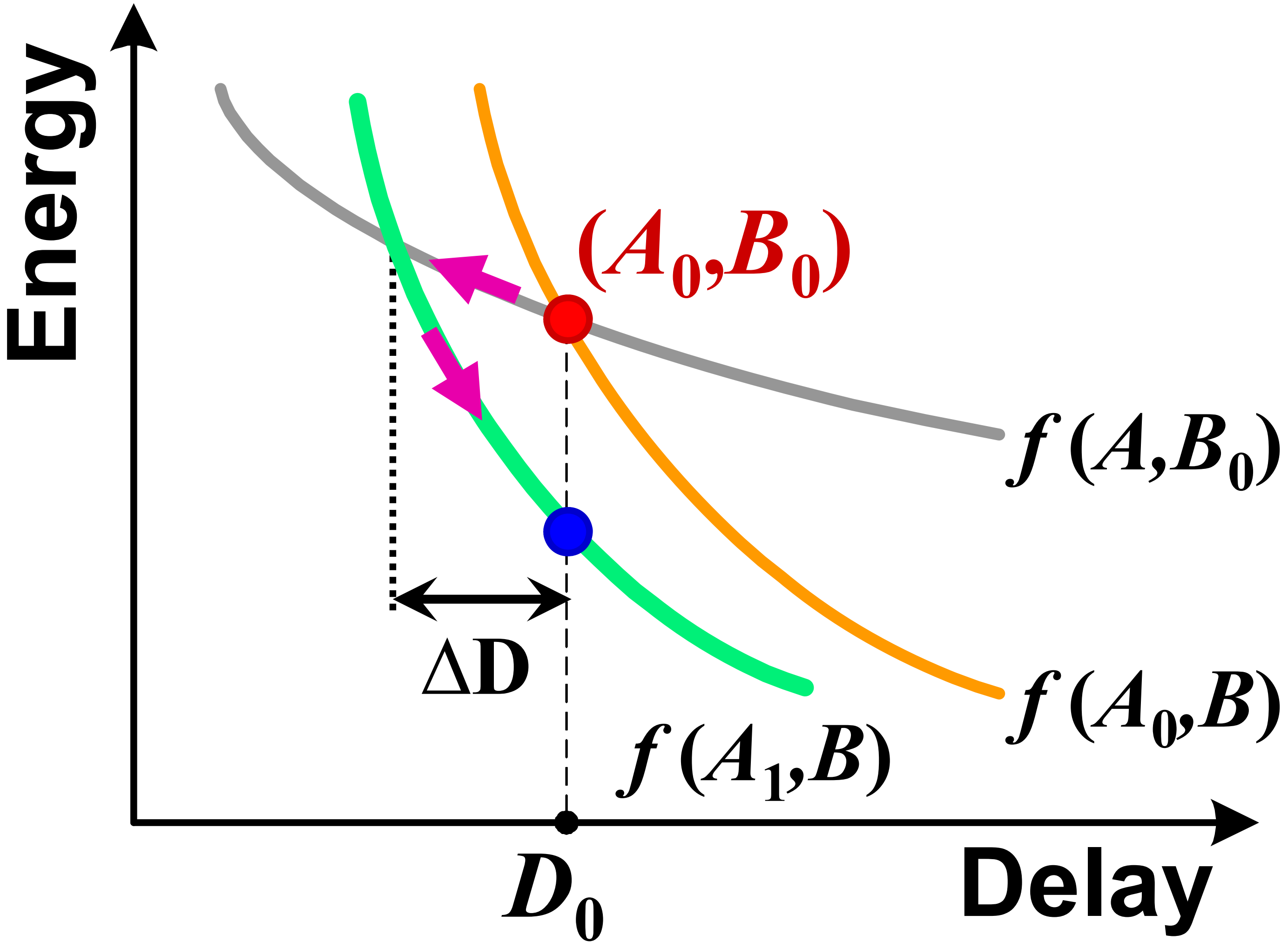
Energy-Delay Sensitivity



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

Solution: Equal Sensitivities

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



At the solution point all sensitivities should be equal

5. C Architectural Optimization

Optimal Processors

- Processors used to be optimized for performance
  - Optimal logic depth was found to be 8-11 FO4 delays in superscalar processors
  - 1.8-3 FO4 in sequentials, rest in combinatorial
    - Kunkel, Smith, ISCA'86
    - Hriskesh, Jouppi, Farkas, Burger, Keckler, Shivakumar, ISCA'02
    - Harstein, Puzak, ISCA'02
    - Sprangle, Carmean, ISCA'02
- But those designs are have very high power dissipation
  - Need to optimize for both performance and power/energy

From System View: What is the Optimum?

- How do sensitivities relate to more traditional metrics:
  - Power per operation (MIPS/W, GOPS/W, TOPS/W)
  - Energy per operation (Joules per op)
  - Energy-delay product
- Can be reformatted as a goal of optimizing power x delay^n
  - n = 0 – minimize power per operation
  - n = 1 – minimize energy per operation
  - n = 2 – minimize energy-delay product
  - n = 3 – minimize energy-(delay)^2 product

Optimization Problem

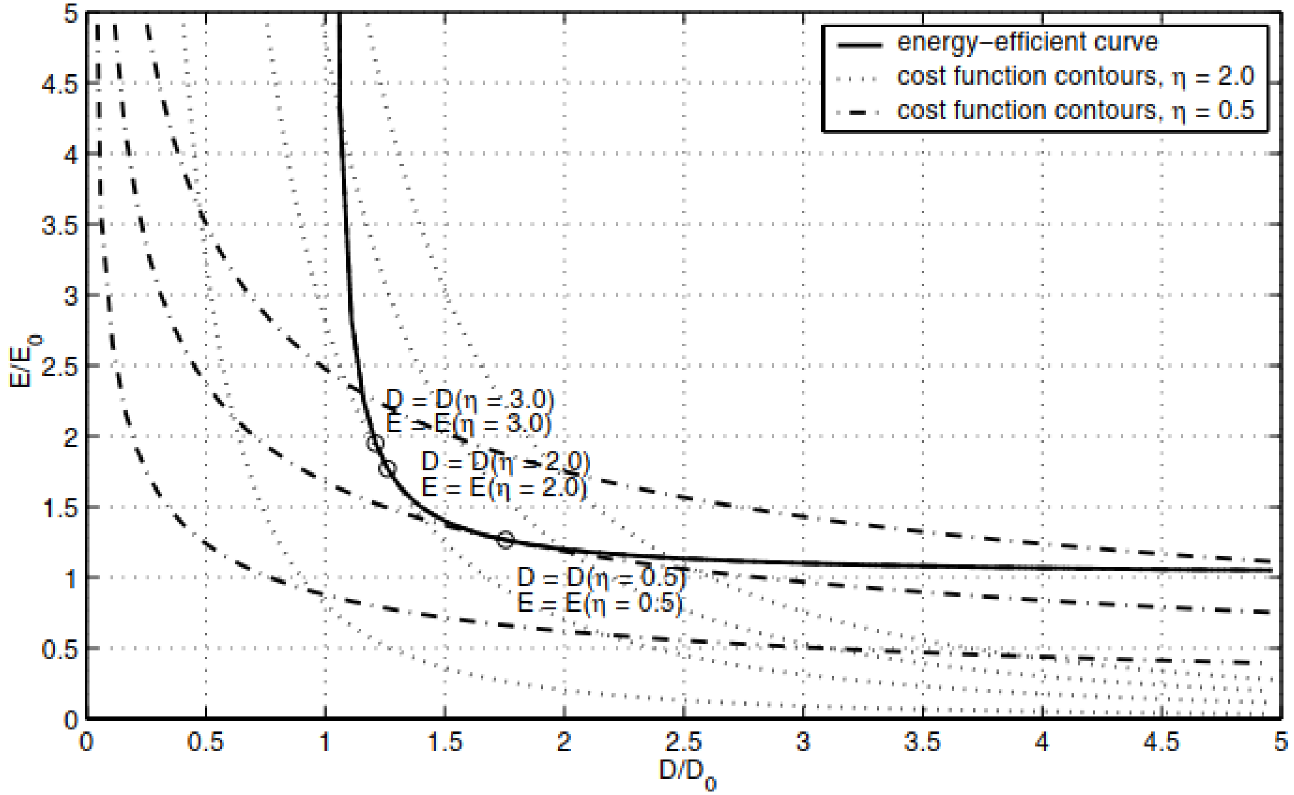
- Set up optimization problem:
  - Maximize performance under energy constraints
  - Minimize energy under performance constraints
- Or minimize a composite function of E^nD^m
  - What are the right n and m?
- n = 1, m = 1 is EDP – improves at lower V_{DD}
- n = 1, m = 2 is invariant to V_{DD}
  - $E \sim CV_{DD}^2$
  - $D \sim 1/V_{DD}$

Hardware Intesnity

- Introduced by Zyuban and Strenski in 2002.
- Measures where is the design on the Energy-Delay curve
- Parameter in cost function optimization

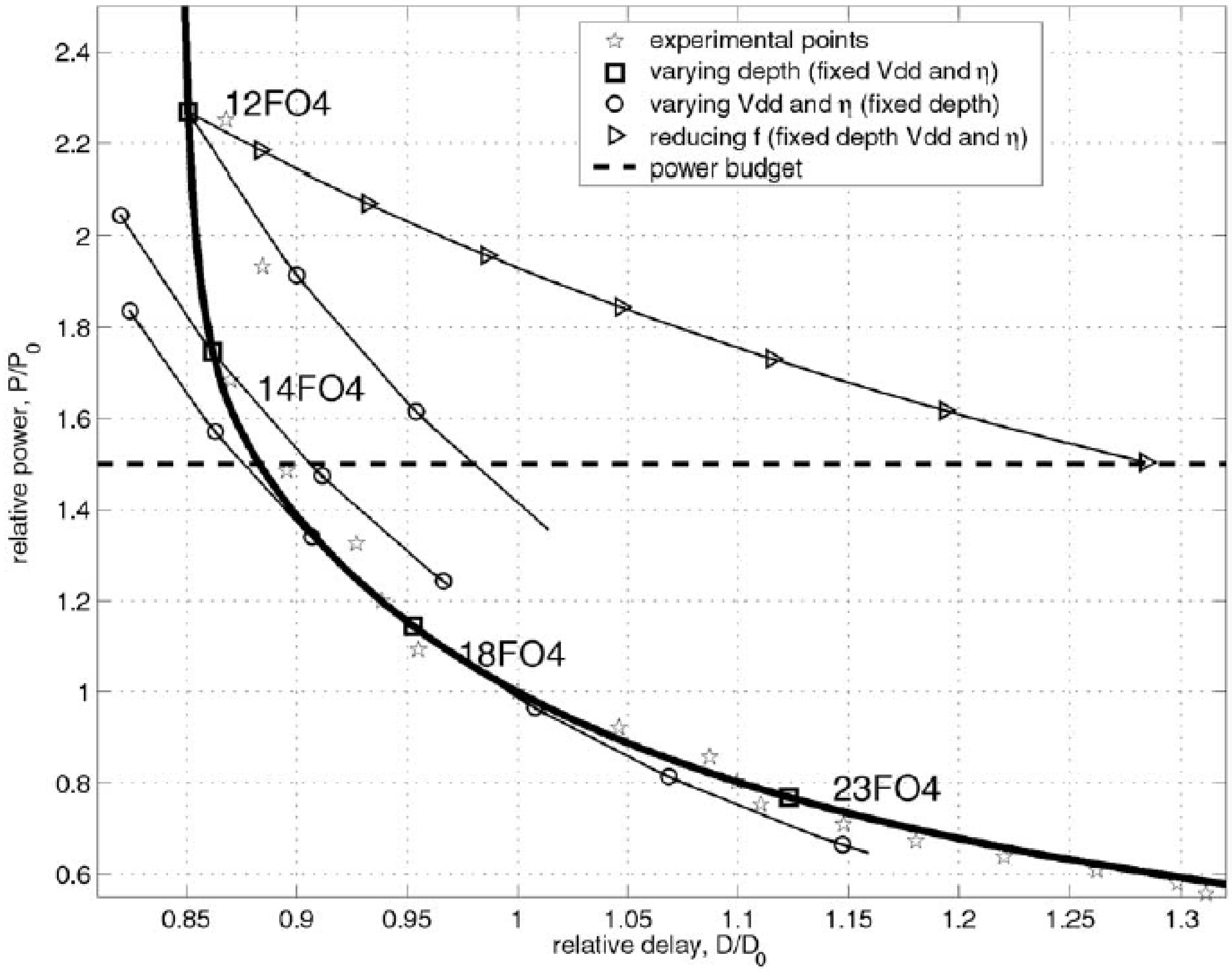
$$F_c = (E/E_0)(D/D_0)^{\eta} \qquad 0 \leq \eta < +\infty,$$

$$\eta = -\left.\frac{D\partial E}{E\partial D}\right|_{v}$$



Slope of the optimal E-D curve at the chosen design point

Optimum Across Hierarchy Layers



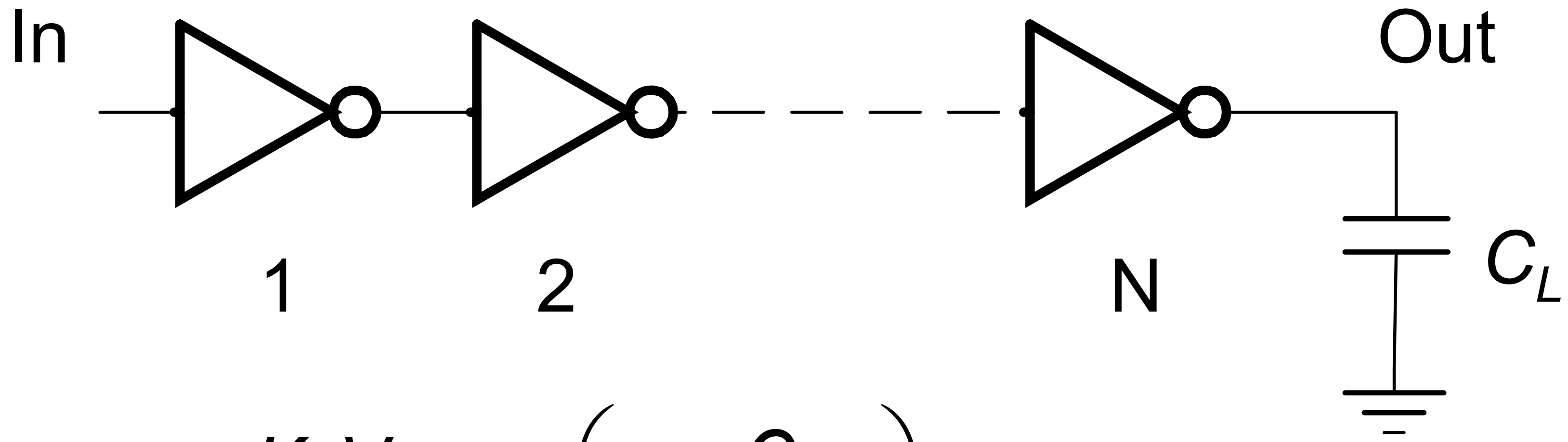
Optimal logic depth in pipelined processors is ~18FO4

Relatively flat in the 16-22FO4 range

Zyuban et al, TComp'04

5.D Circuit-Level Tradeoffs

Alpha-Power Based Delay Model



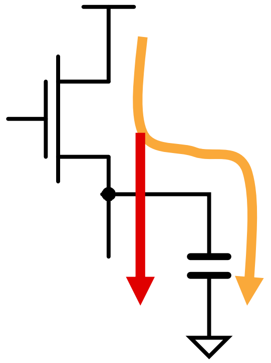
$$t_{pi} = \frac{K_d V_{DD}}{(V_{DD} - V_{Th})^{\alpha}} \left(1 + \frac{C_{L,i}}{C_{in,i}} \right)$$

$$D = \sum t_{pi} = \sum \frac{K_d V_{DD}}{(V_{DD} - V_{Th})^{\alpha}} \left(1 + \frac{W_{L,i}}{W_{in,i}} \right)$$

Energy Models

- **Switching**

$$E_{Sw} = \alpha_{0\to1}\left(C_{L,i} + C_{\text{int},i}\right)V_{DD}^{\ 2}$$



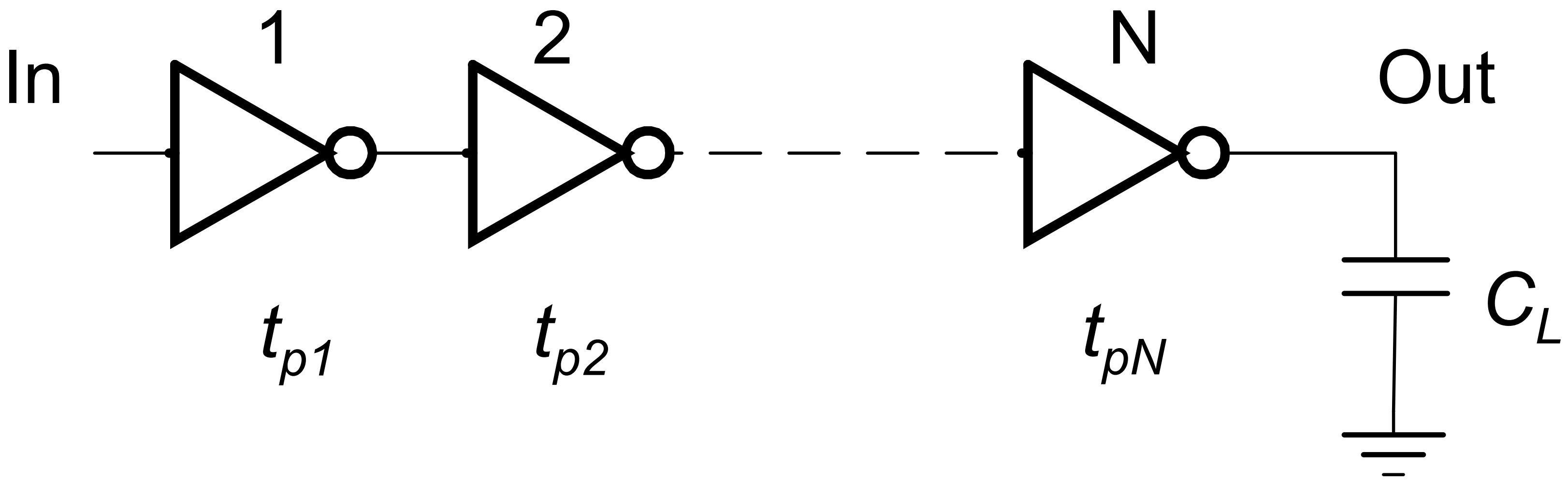
- **Leakage**

$$E_{Lk} = W_{ln} I_0 e^{\frac{-(V_{Th}-\gamma V_{DD})}{nV_t}} V_{DD} D$$

Sizing, Supply, Threshold Optimization

- Transistor sizing can yield large power savings with small delay penalties
  - Gate sizing
  - Beta-ratio adjustments $\beta = Wp/Wn$
  - (Stack resizing)
- Supply voltage affects both active and leakage energy
- Threshold voltage affects primarily the leakage

Apply to Sizing of an Inverter Chain



Unconstrained energy: find min D = Σt_{pi}

$$C_{gin,j} = \sqrt{C_{gin,j-1} C_{gin,j+1}} \qquad W_j = \sqrt{W_{j-1} W_{j+1}}$$

Constrained energy: find min D, under E < E_{max}

Where E = Σe_i

Constrained Optimization

- Find min(D) subject to $E = E_{max}$
  - *Constrained function minimization*

- E.g. Lagrange multipliers

Or dual:

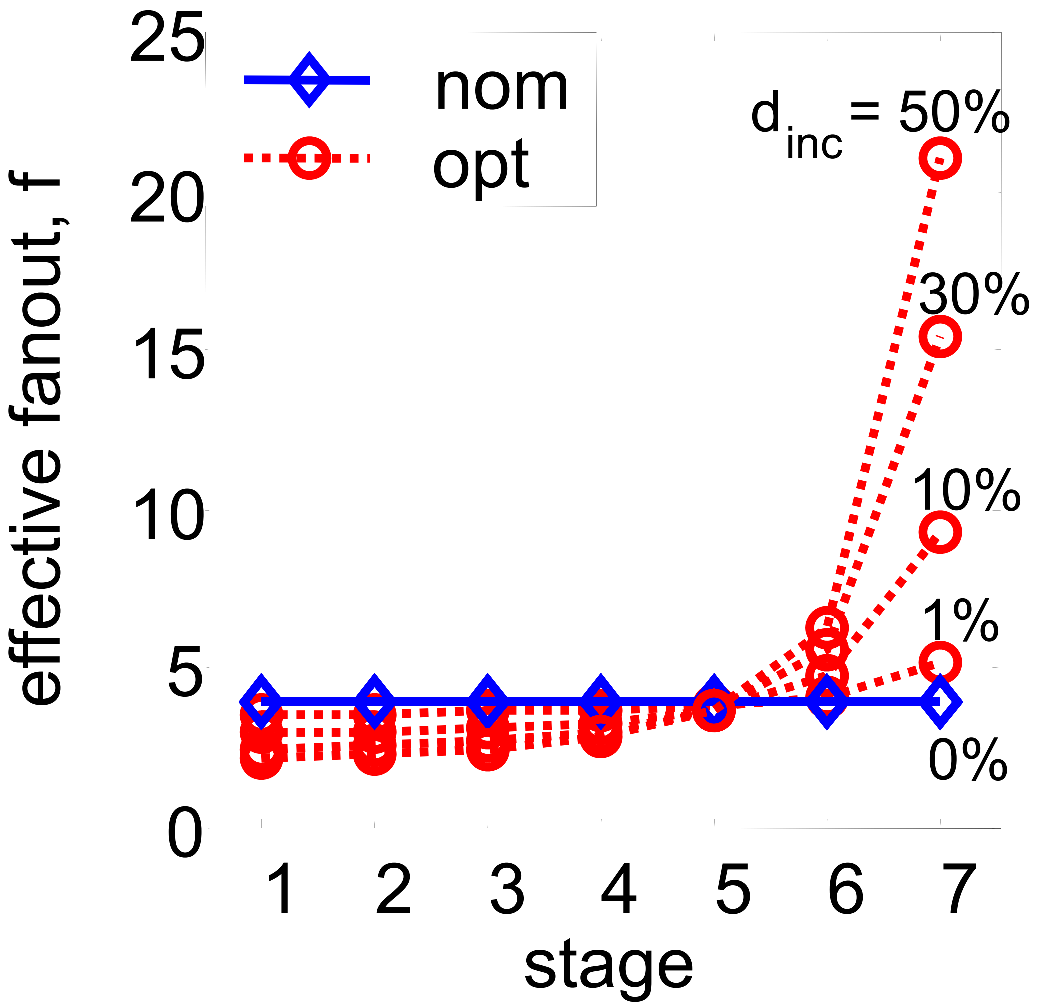
$$\Lambda(x) = D(x) + \lambda\left(E(x) - E_{\max}\right) \qquad \mathrm{K}(x) = E(x) + \lambda\left(D - D_{\max}\right)$$

$$\frac{\partial \Lambda}{\partial x} = 0$$

- Can solve analytically for $x = W_i, V_{DD}, V_{Th}$

Inverter Chain: Sizing Optimization

Inverter Chain: Sizing Optimization



$$W_j = \sqrt{\frac{W_{j-1}W_{j+1}}{1+\lambda W_{j-1}}}$$

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

$$\lambda = -\frac{2KV_{DD}^{\ 2}}{\tau_{nom}S_W}$$

$$S_W \propto \frac{e_j}{f_j - f_{j-1}}$$

e_i – energy per stage
f_i – fanout per stage

Stojanovic, ICCAD'02

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

Sensitivity to Sizing and Supply

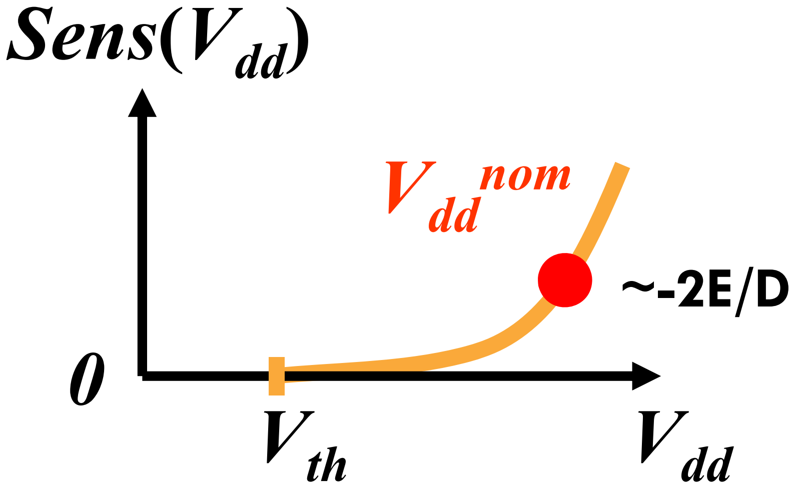
- Gate sizing (W_i)

$$-\frac{\partial E_{sw} / \partial W_j}{\partial D / \partial W_j} = \frac{e_j}{\tau_{nom}\left(f_j - f_{j-1}\right)}$$

∞ for equal f_{eff} (D_{min})

- Supply voltage (V_{dd})

$$-\frac{\partial E_{sw} / \partial V_{DD}}{\partial D / \partial V_{DD}} = \frac{E_{sw}}{D} 2 \frac{1 - x_V}{\alpha - 1 + x_V}$$



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

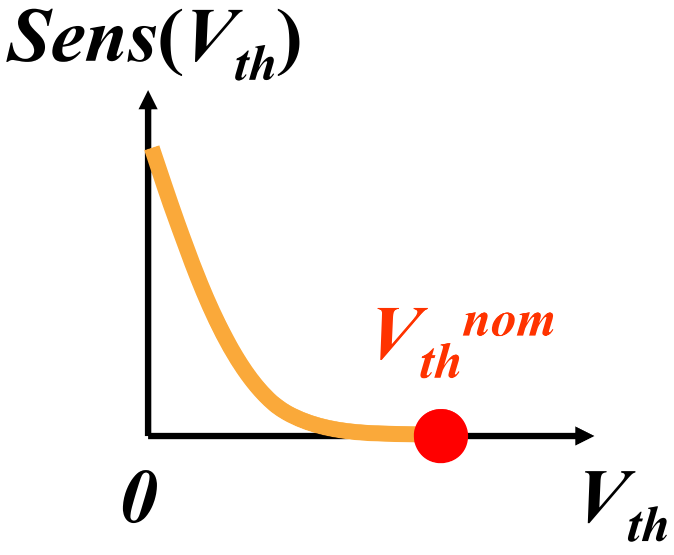
Sensitivity to V_{th}

- Threshold voltage (V_{th})

$$-\frac{\partial E / \partial \Delta V_{Th}}{\partial D / \partial \Delta V_{th}} = P_{Lk}\left(\frac{V_{DD} - V_{Th} - \Delta V_{Th}}{\alpha n V_t} - 1\right)$$

Low initial leakage

⇒ speedup comes for “free”



Next Lecture

- Low-power design
  - Lowering supply voltage