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, 5

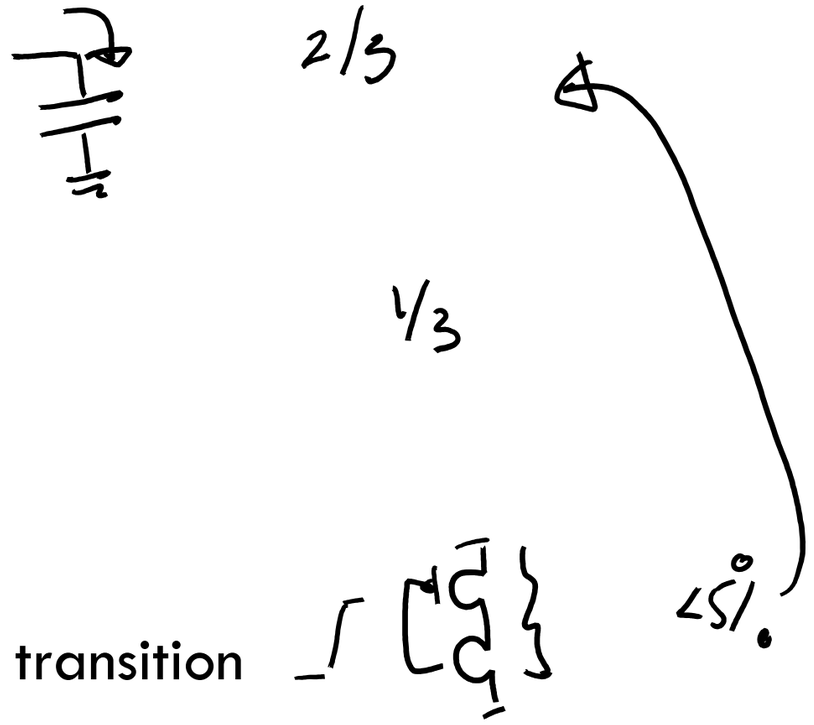
Outline

- 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



(PLLs, SRAM, references)

Summary of Power Dissipation Sources

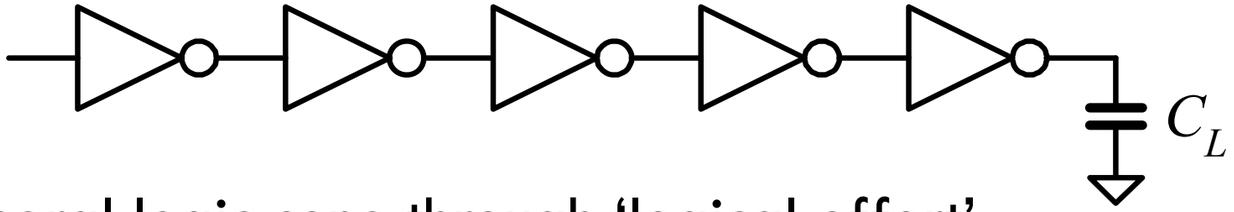
$$P \sim \alpha \cdot (C_L + C_{CS}) \cdot V_{swing} \cdot V_{DD} \cdot f + (I_{DC} + I_{Leak}) \cdot V_{DD}$$

- 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

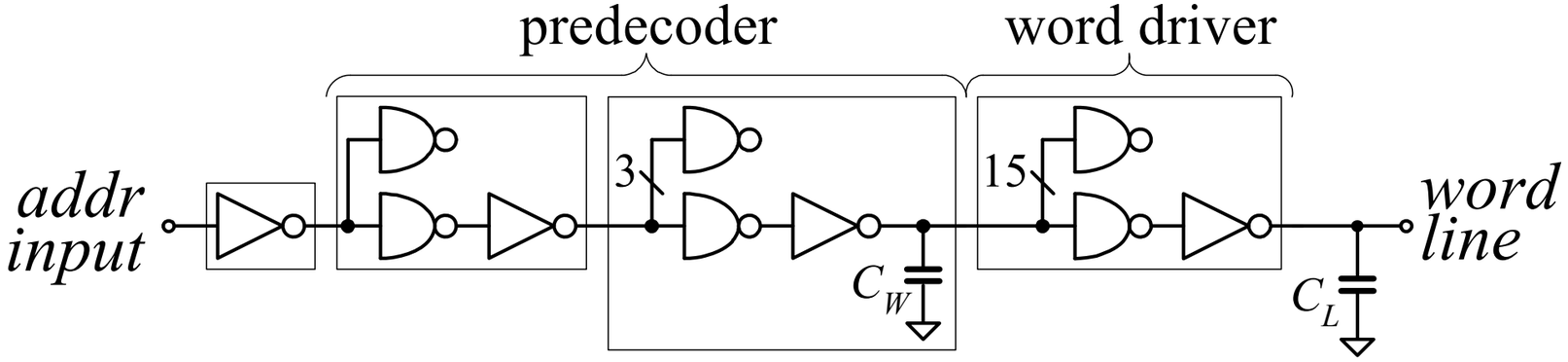
$$P = \frac{energy}{operation} \times rate + static\ power$$

CMOS Performance Optimization

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

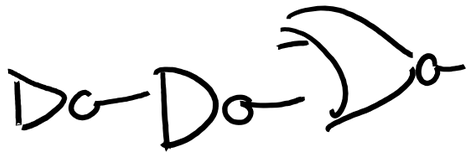


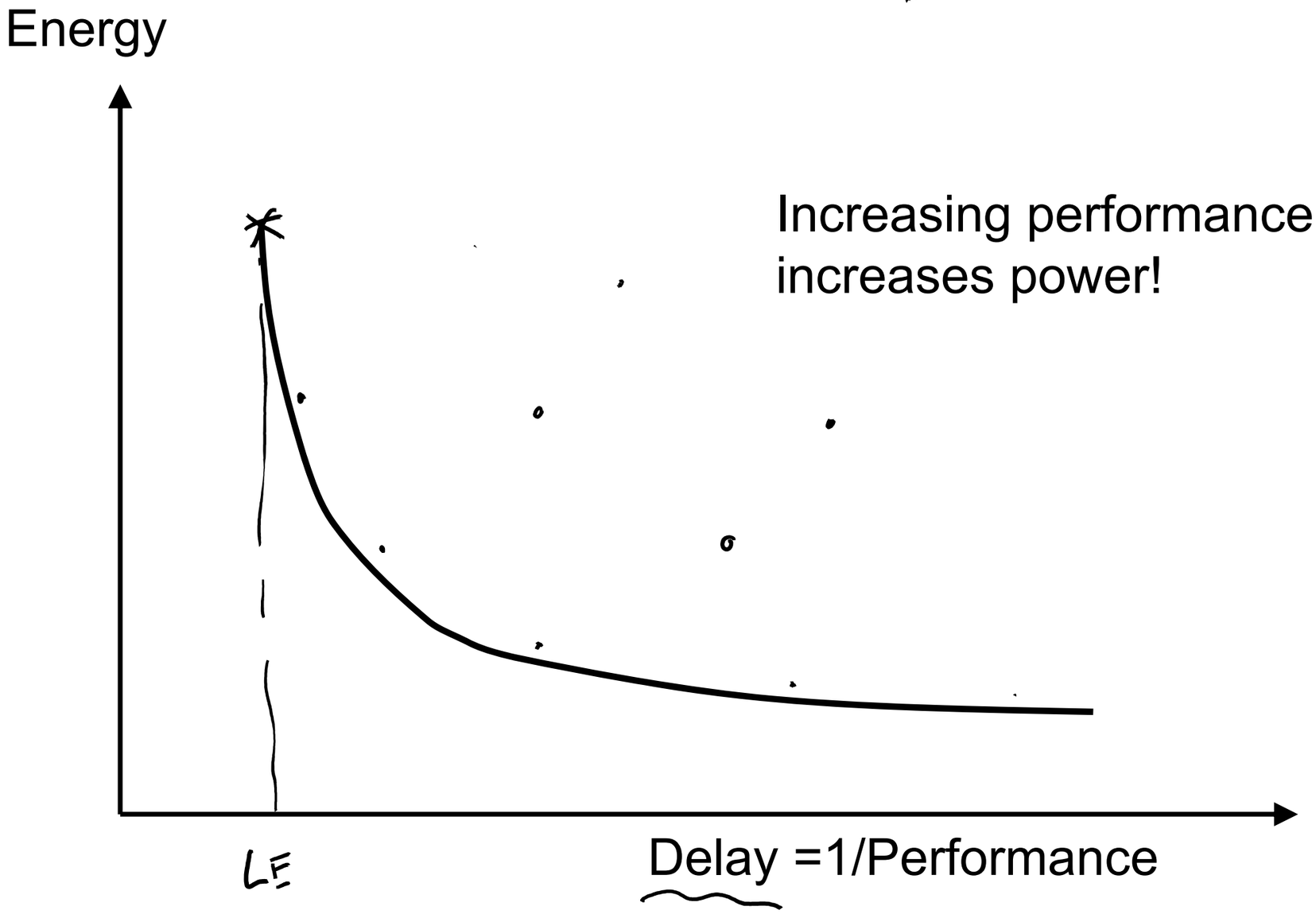
- Extendable to general logic cone through 'logical effort'
- Equal effective fanouts (g_iC_{i+1}/C_i) per stage
- Optimal fanout is around 4



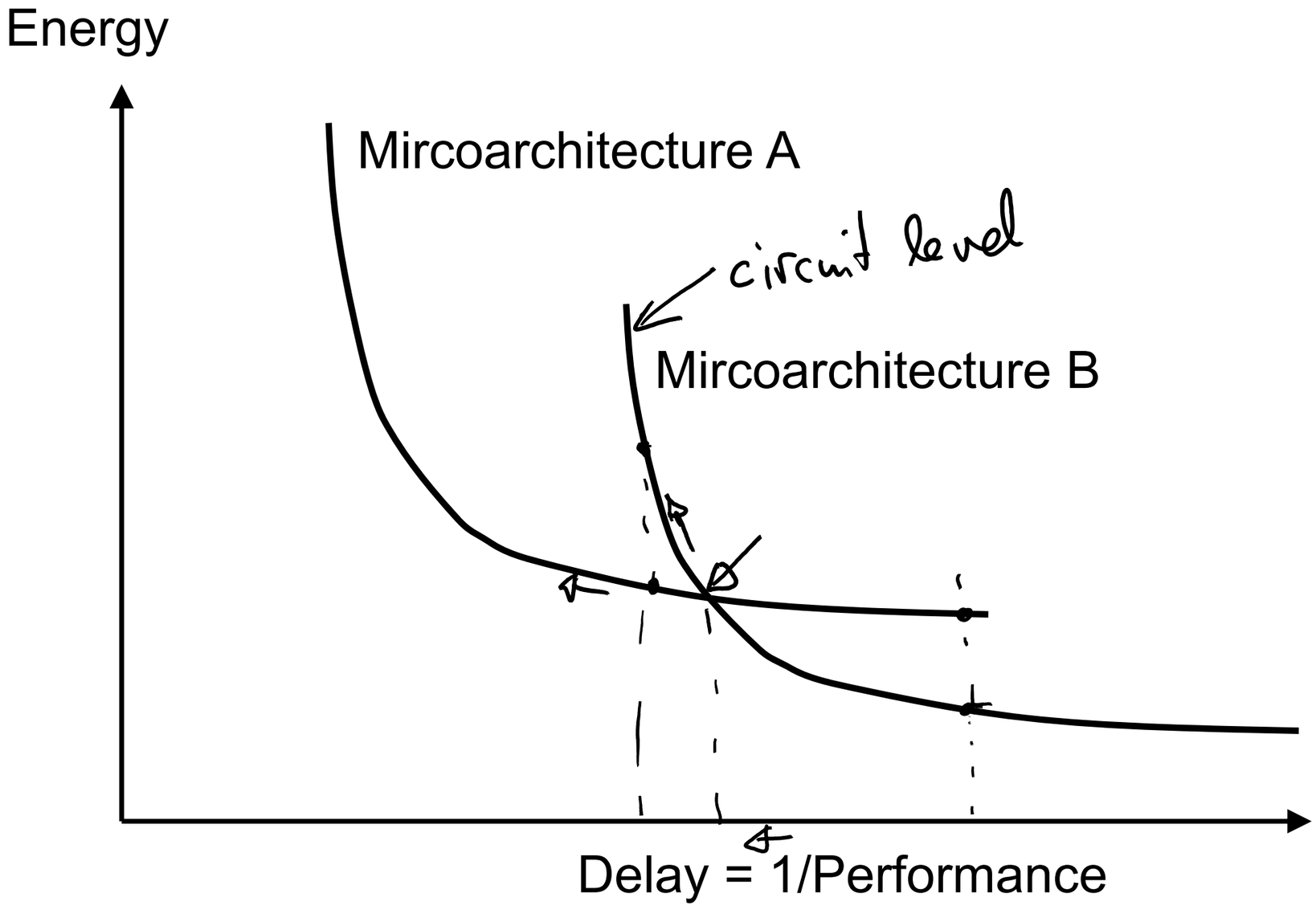
[Ref: I. Sutherland, Morgan-Kaufman'98]

Performance Optimization





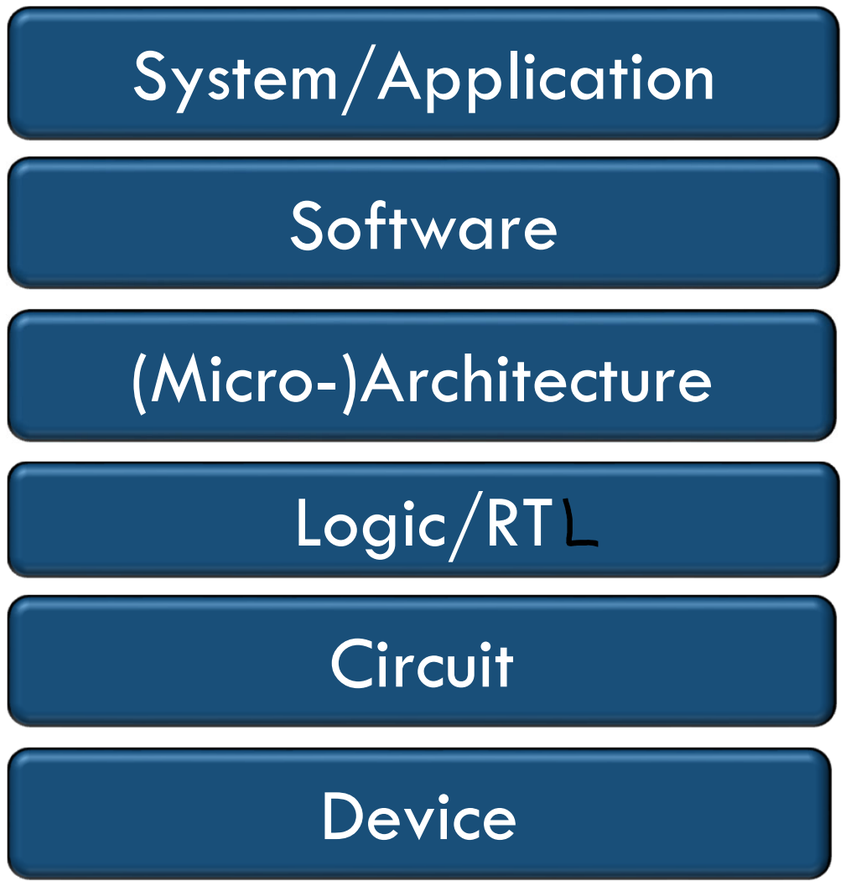
Performance Optimization



The Design Abstraction Stack

A very rich set of design parameters to consider!

It helps to consider options in relation to their abstraction layer



Choice of algorithm

Amount of concurrency

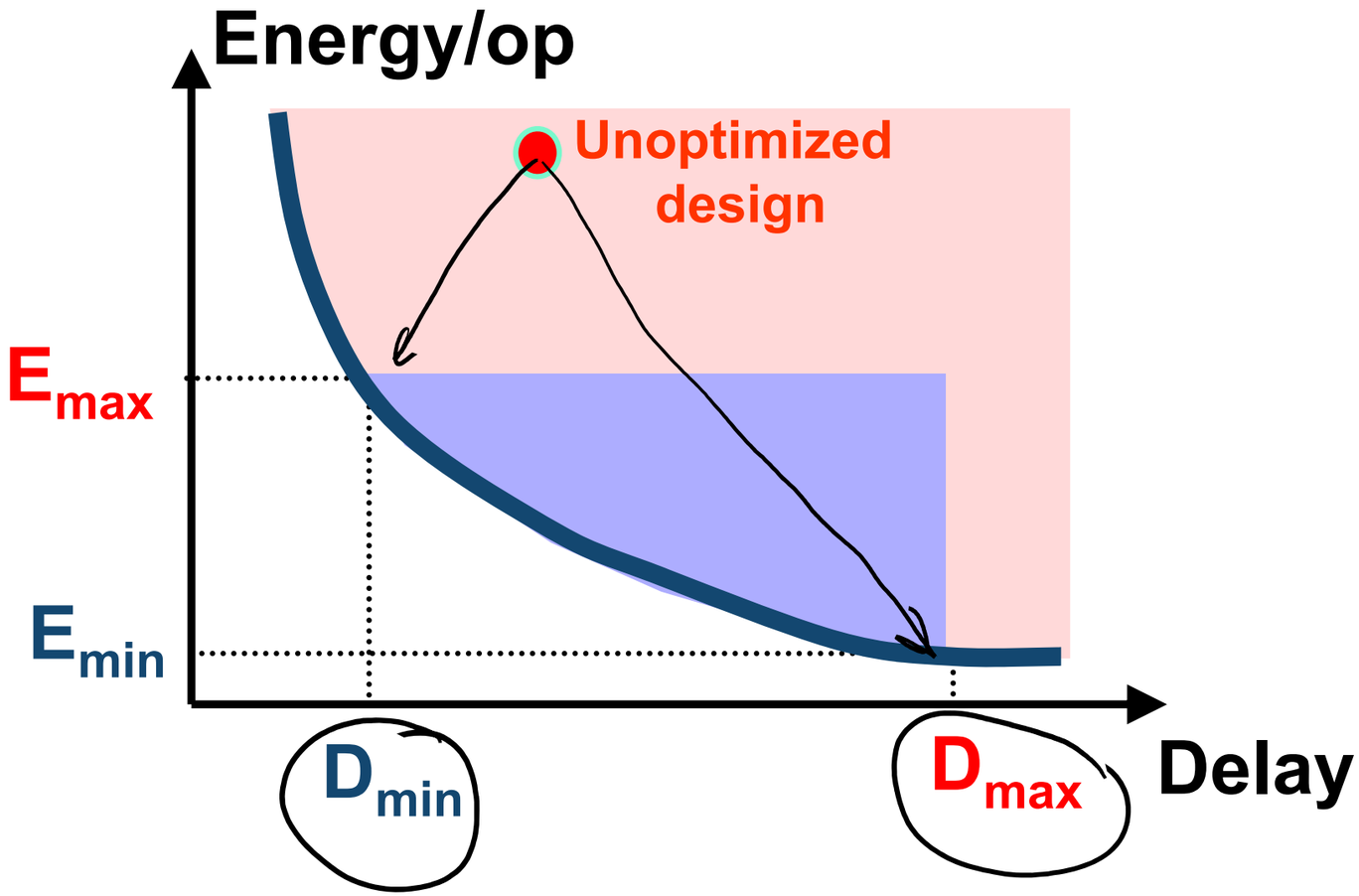
Parallel versus pipelined, general purpose versus application specific

logic family, standard cell versus custom, clocking

sizing, supplies, thresholds

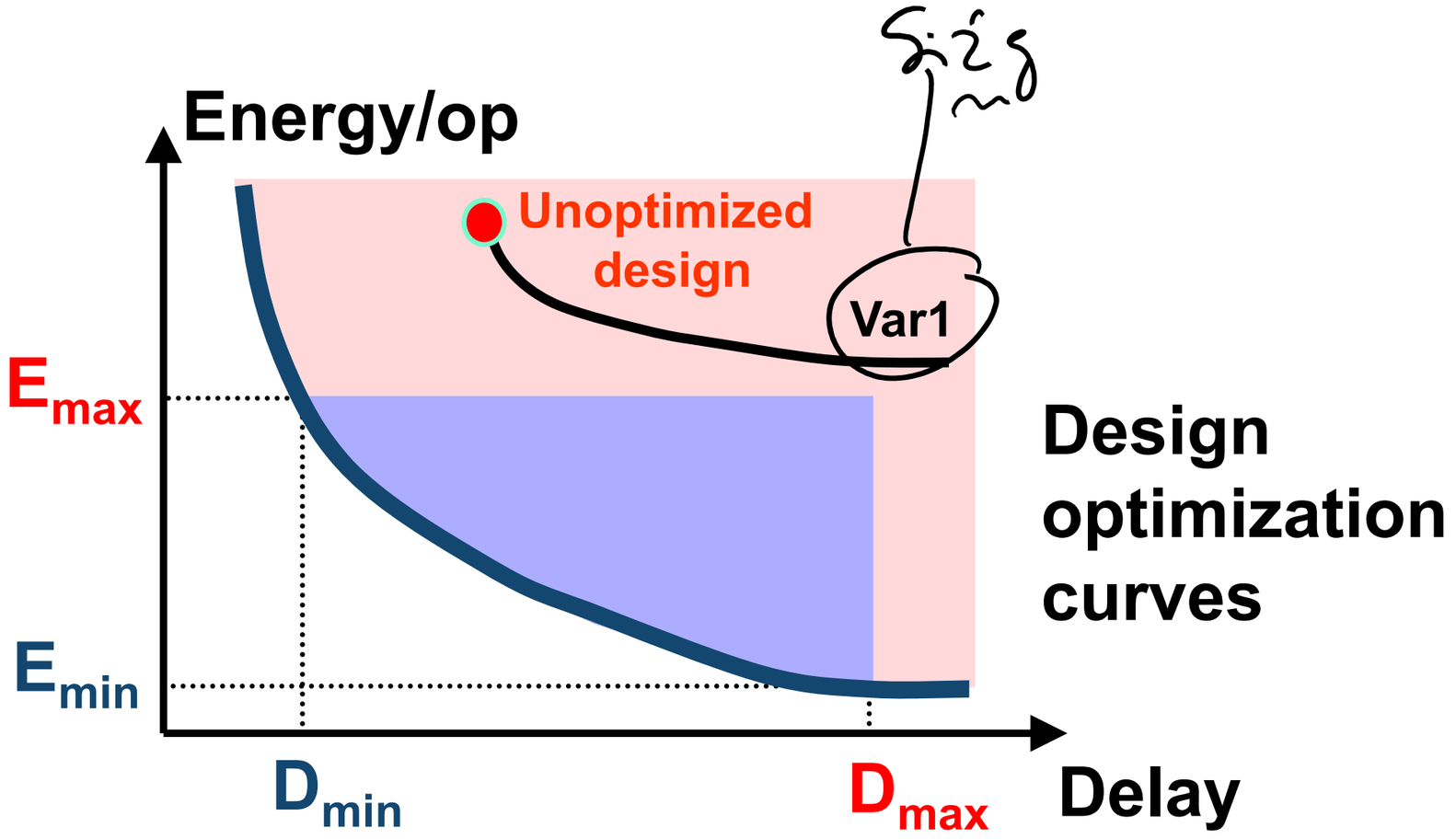
Bulk, PDSOI, FDSOI, finFET

Power-Performance Optimization



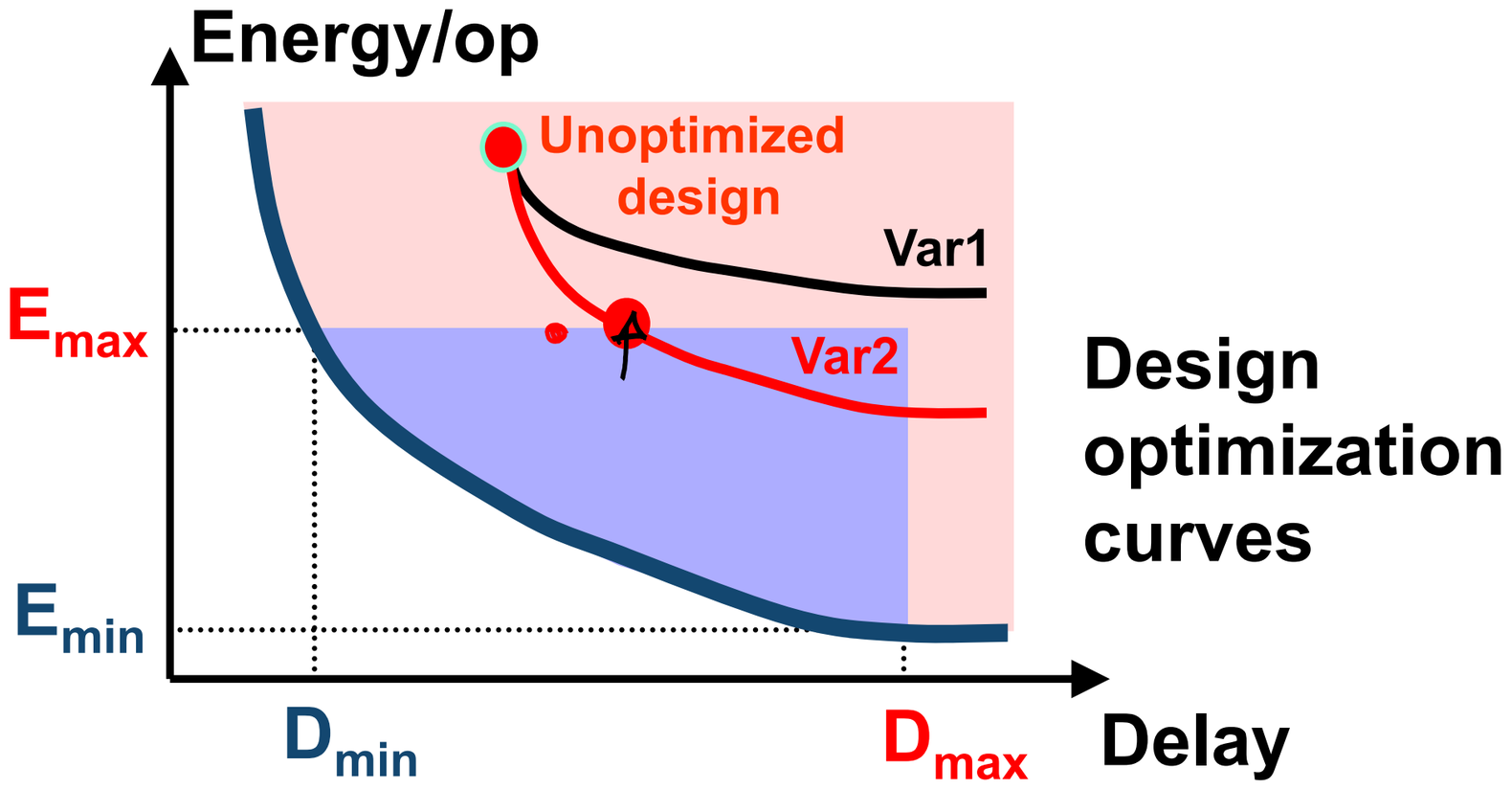
Achieve the highest performance under the power cap

Power-Performance Optimization



Achieve the highest performance under the power cap

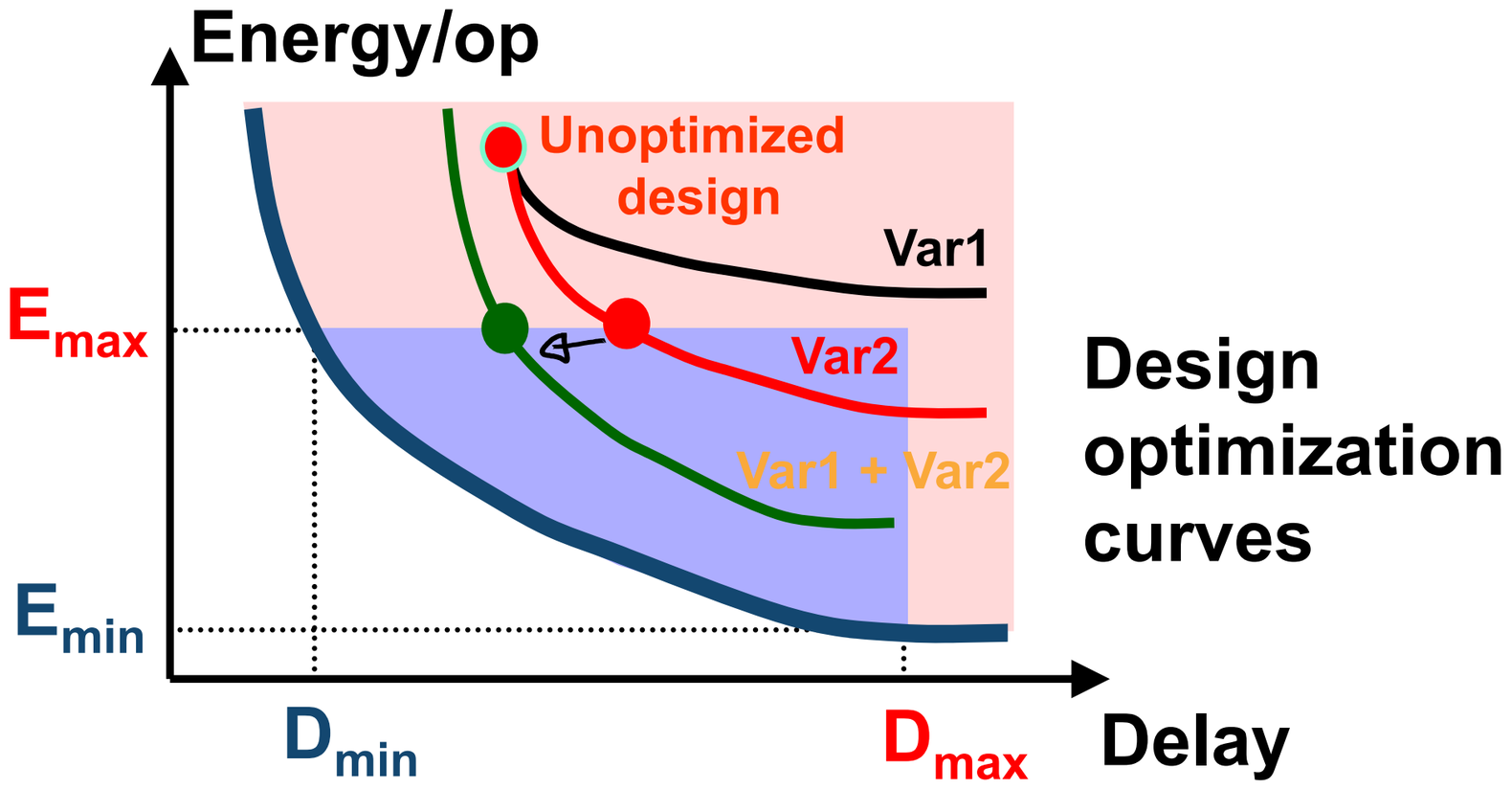
Power-Performance Optimization



Design optimization curves

Achieve the highest performance under the power cap

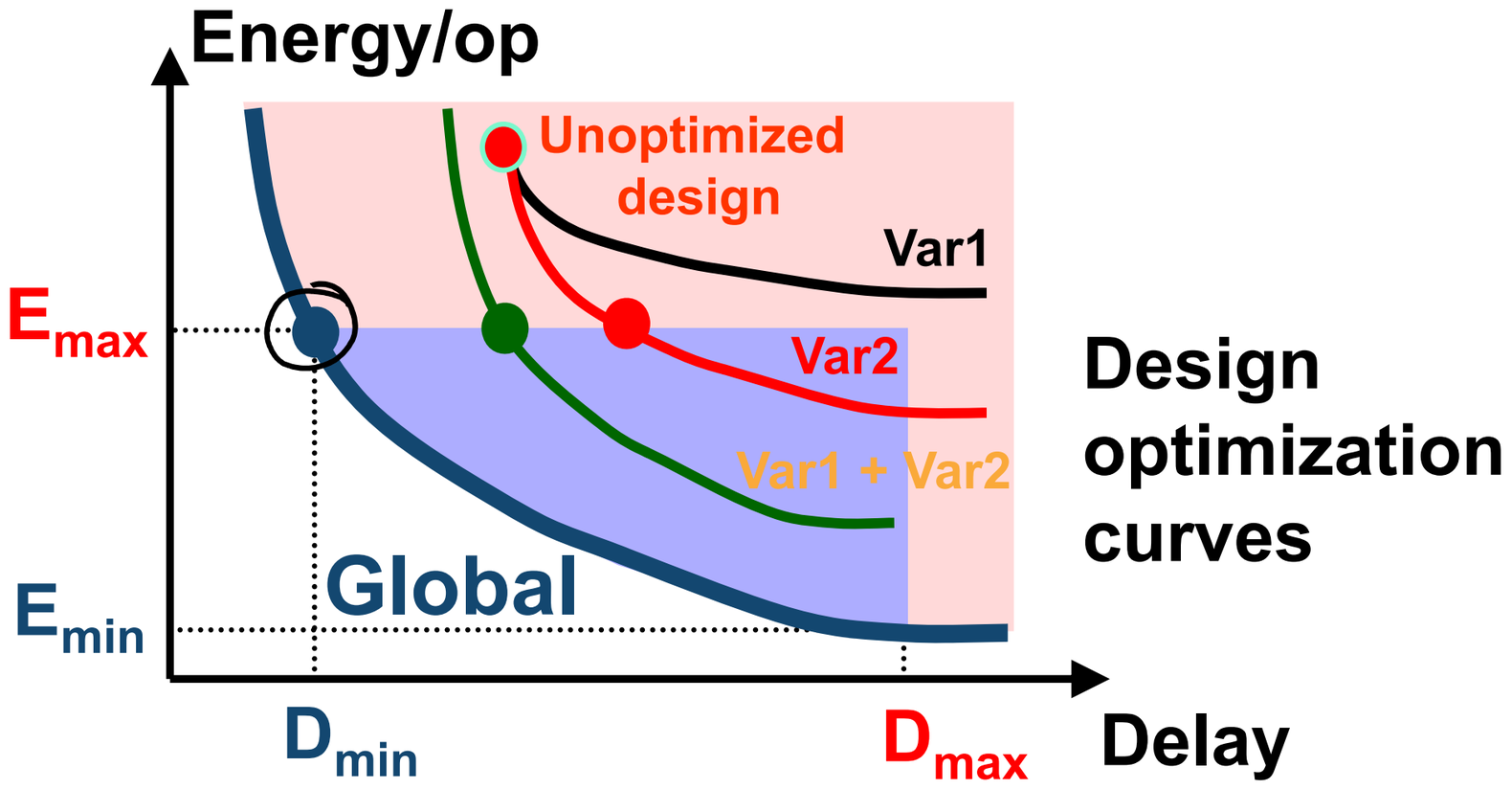
Power-Performance Optimization



Design optimization curves

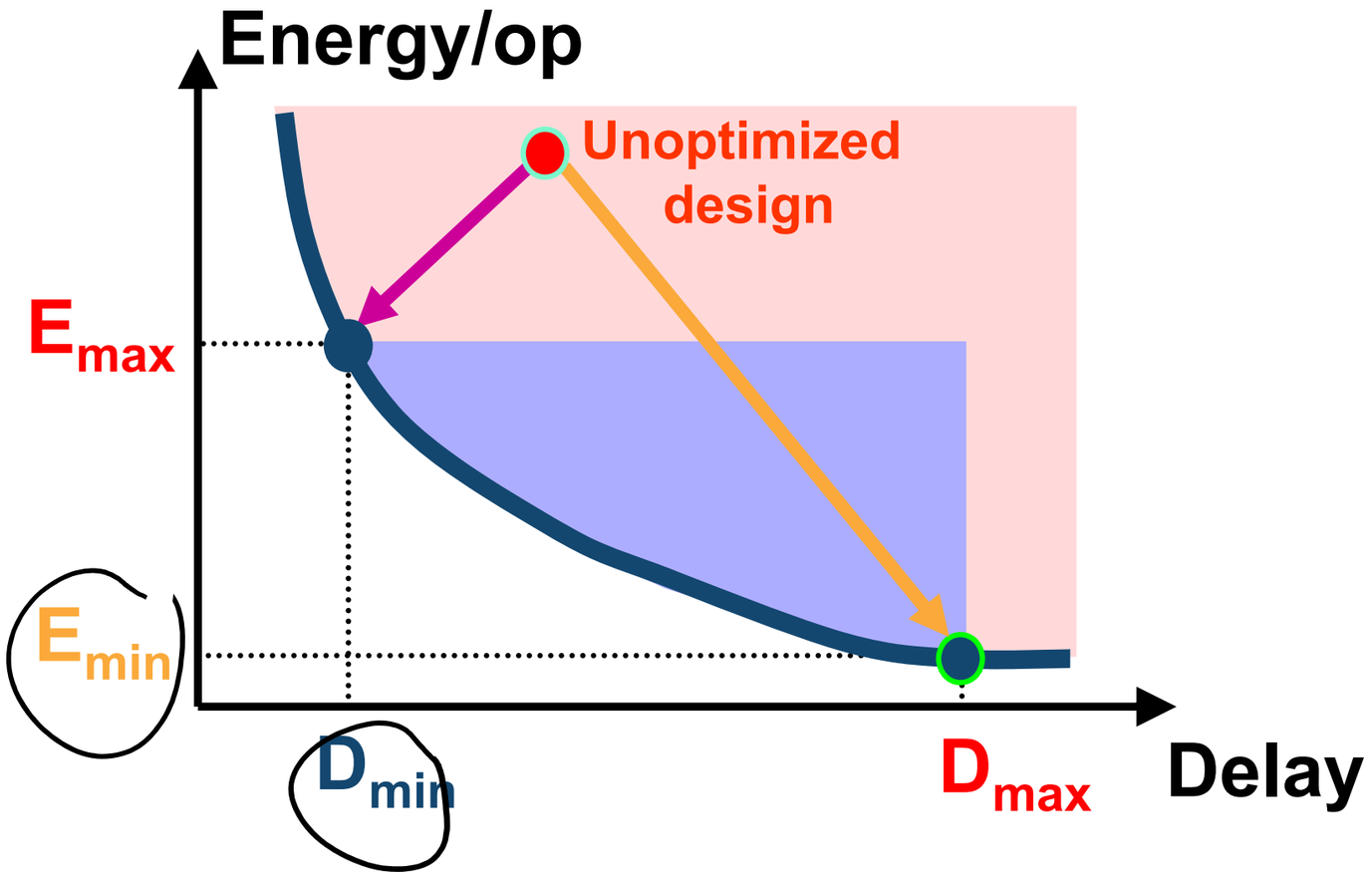
How far away are we from the optimal solution?

Power-Performance Optimization



Global optimum – best performance

Power-Performance Optimization

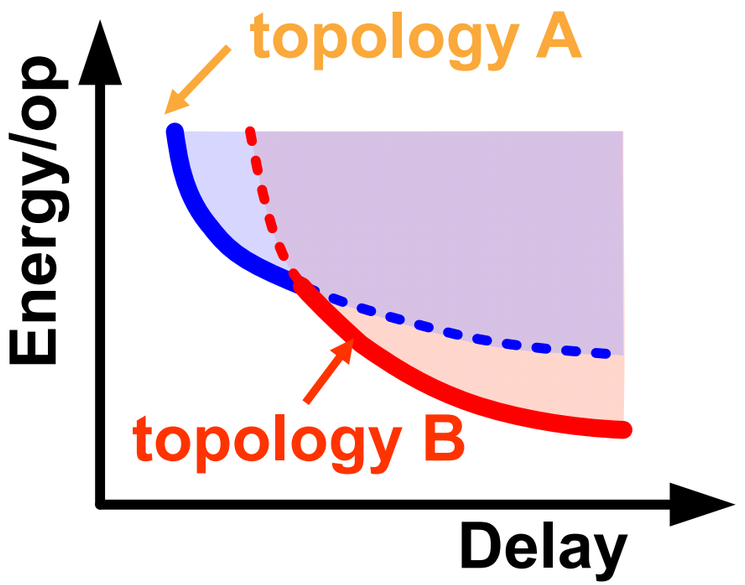


Maximize throughput for given energy or

Minimize energy for given throughput

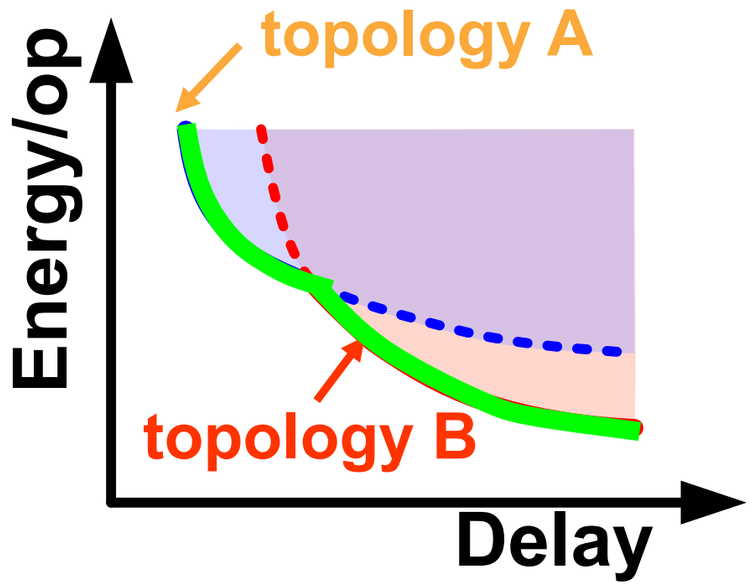
Power-Performance Optimization

- 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)



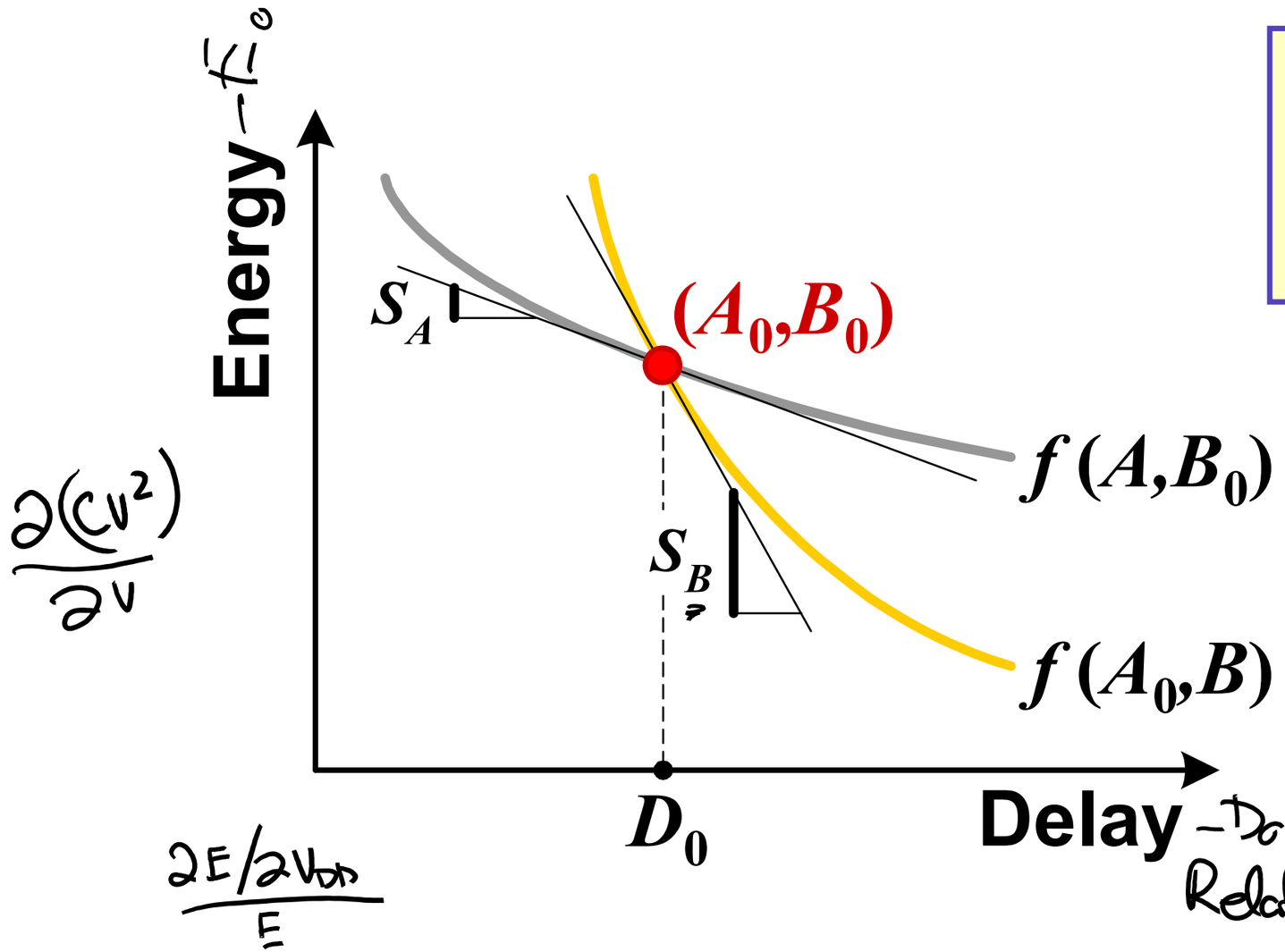
Power-Performance Optimization

- 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



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

$\frac{\partial(CV^2)}{\partial V}$

$\frac{\partial E/\partial V_{DD}}{E}$

$V_A = V_{DD}$

Absolute

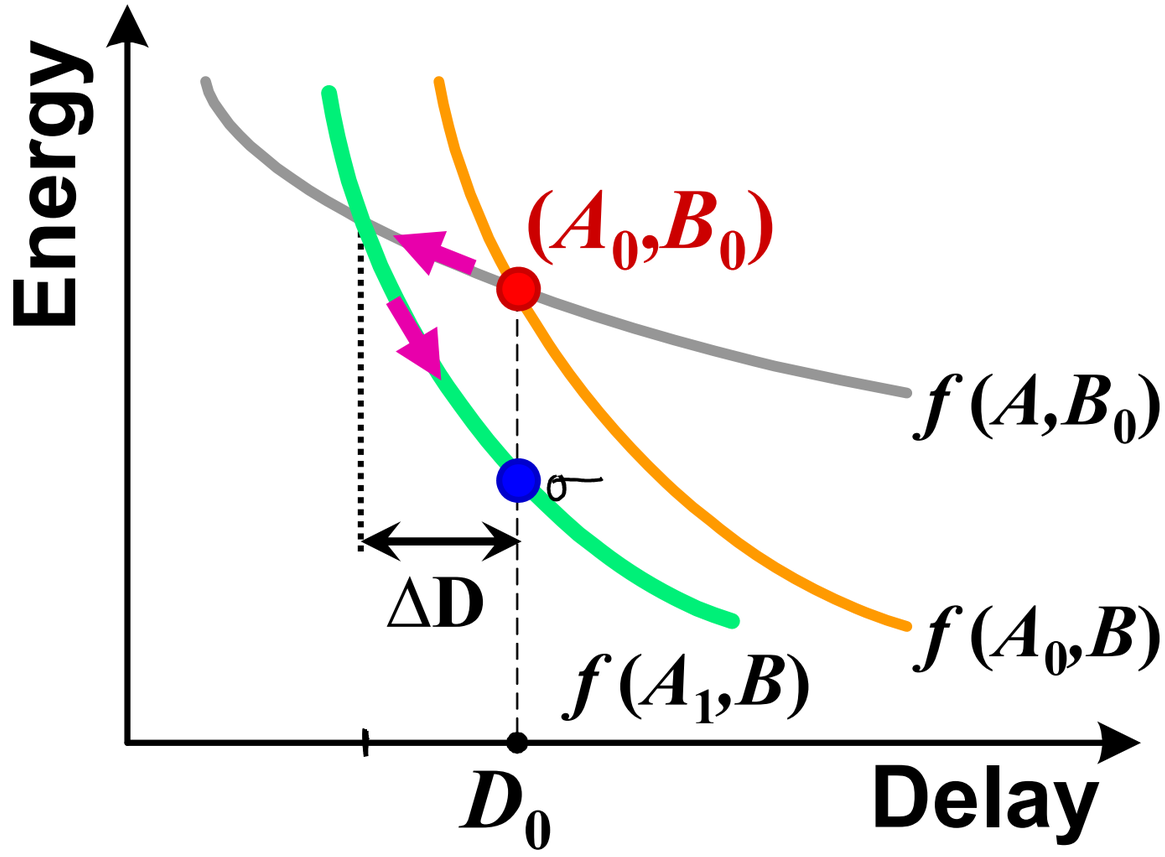
$$\frac{\partial E/\partial V_{DD}}{\partial D/\partial V_{DD}}$$

Relative

$$\frac{\partial E/\partial V_{DD}}{\partial D/\partial V_{DD}} \cdot \frac{D}{E} = \frac{2\%}{1\%}$$

Solution: Equal Sensitivities

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



(or are at that limit)

At the solution point all sensitivities should be equal (slopes)

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

$P \cdot D^n$

- 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 ← improves with V_{DD} ↓
  - n = 3 – minimize energy-(delay)^2 product ← invariant to V_{DD} (@ nom V_{DD})

n = 4 ...

Optimization Problem

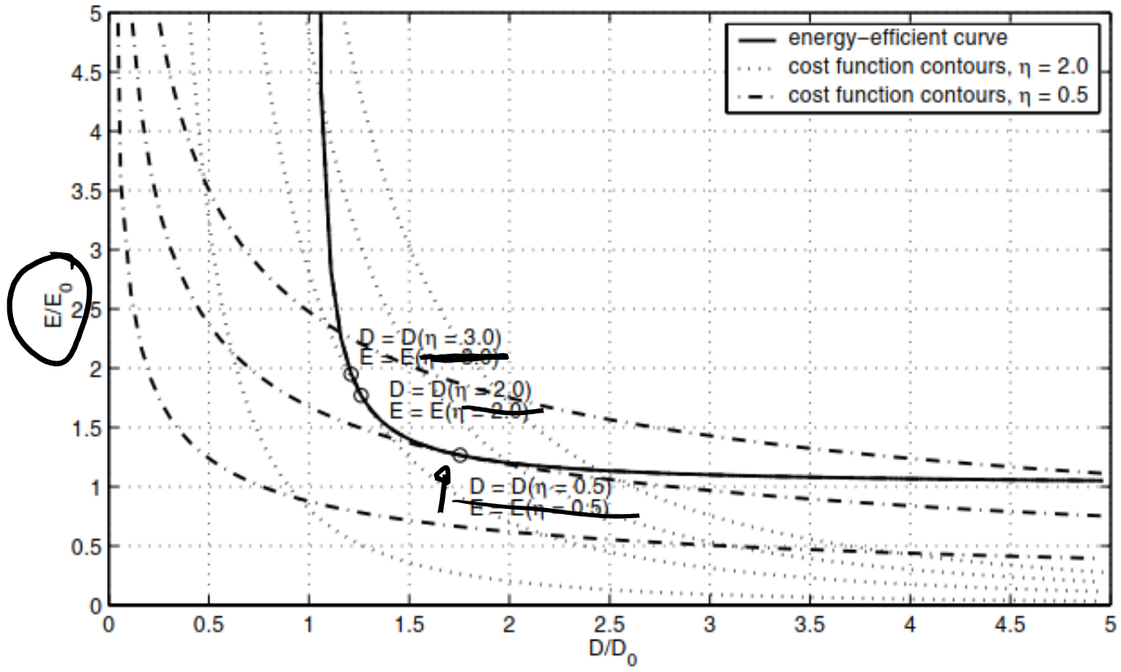
- Set up optimization problem:
  - Maximize performance under energy constraints
  - Minimize energy under performance constraints
- Or minimize a composite function of E^nD^m
  - What are the right n and m?
- $n = 1, m = 1$ is EDP – improves at lower V_{DD}
- $n = 1, m = 2$ is invariant to V_{DD}
 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

Hardware Intesnity

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

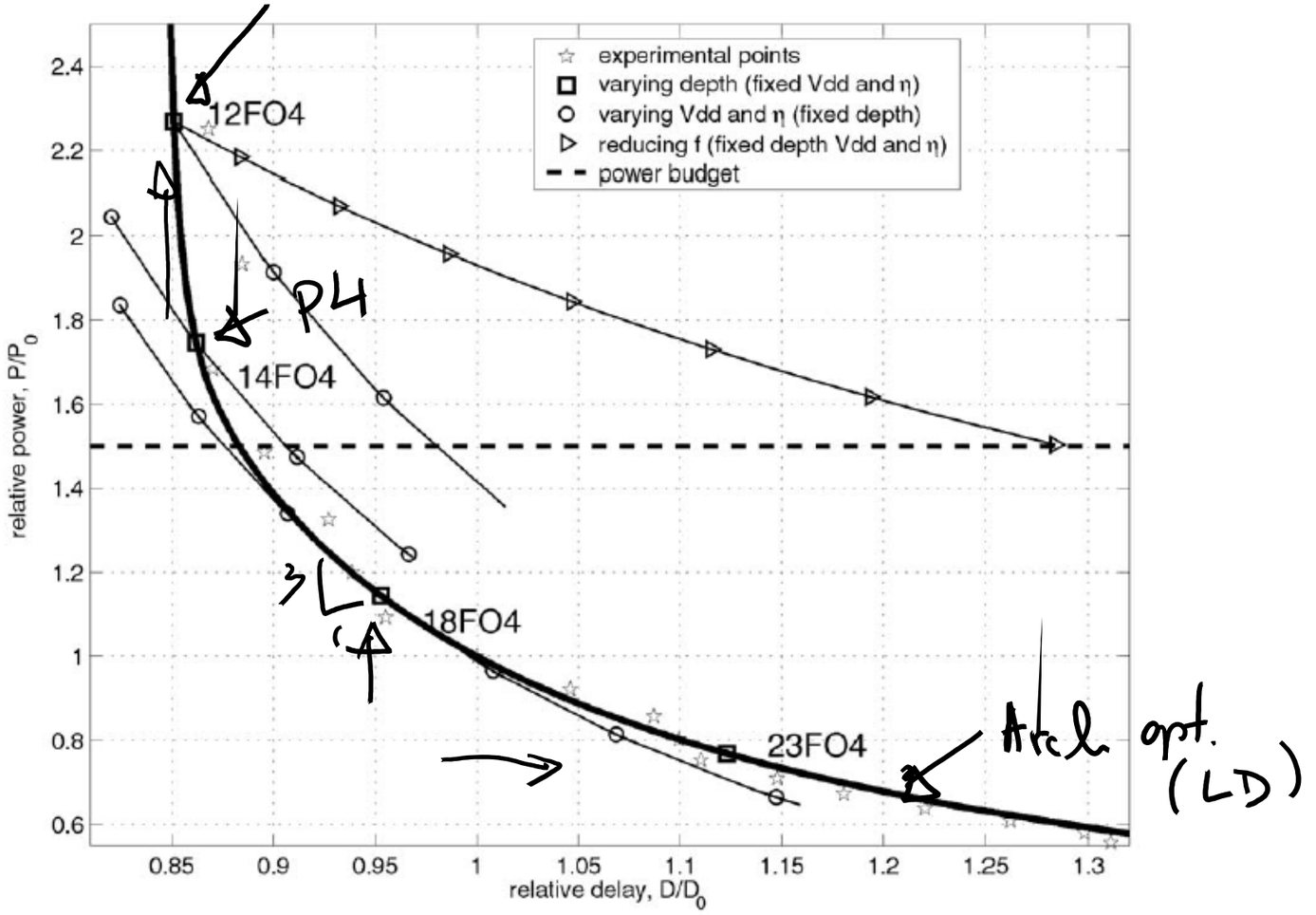
$$F_c = (E/E_0)(D/D_0)^{\eta} \qquad 0 \leq \eta < +\infty,$$

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



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