EECS251B : Advanced Digital Circuits and Systems

Lecture 20 – Power-Performance

Borivoje Nikolić

OpenAI and Microsoft reportedly planning $100 billion datacenter project for an AI supercomputer

March 29, 2024. Microsoft and OpenAI are reportedly working on a massive datacenter to house an AI-focused supercomputer featuring millions of GPUs. The Information reports that the project could cost "in excess of $115 billion" and that the supercomputer, currently dubbed "Stargate" inside OpenAI, would be U.S.-based.

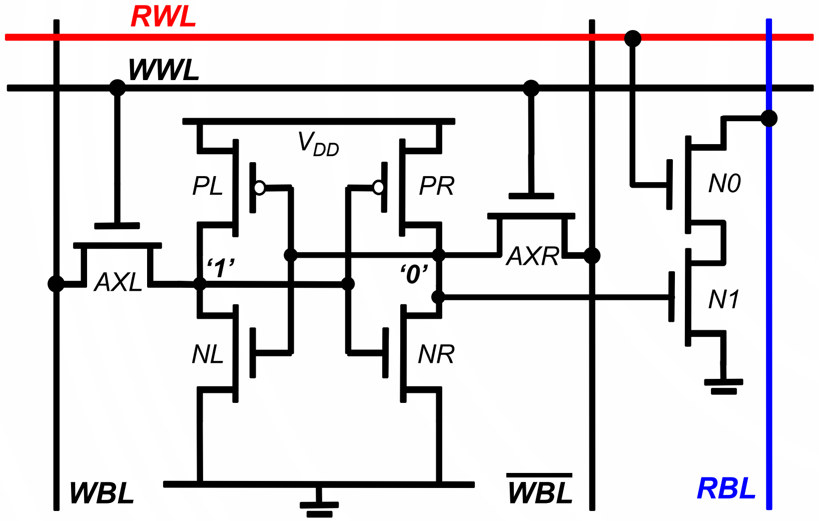
Image credit: Getty Images
https://www.tomshardware.com/tech-industry/artificial-intelligence/openai-and-microsoft-reportedly-planning-dollar100-billion-datacenter-project-for-an-ai-supercomputer

Announcements

- Quiz 3 is on Thursday
- Homework 4 released
- Project
  - Preliminary design review next Tuesday
  - Starting at 9am, so everyone can present
  - 6 minutes per team
- Lab 5 due this week

6T SRAM Alternatives

8T-SRAM



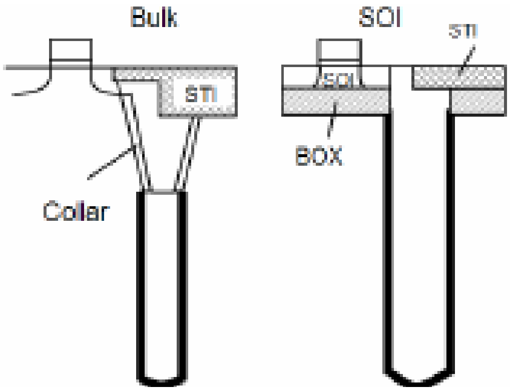
- Read circuit?
- Interleaving?

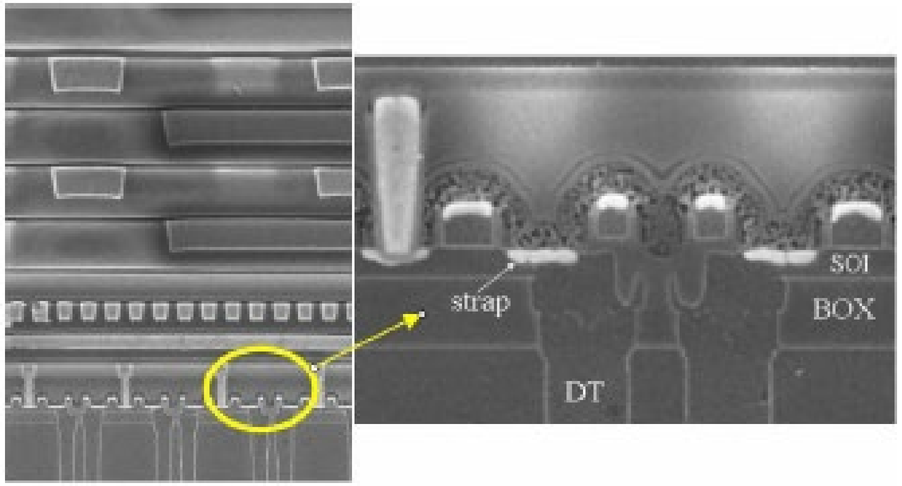
- Dual-port read/write capability (register-file-like cells)
- N0, N1 separates read and write
  - No Read SNM constraint
  - Half-selected cells still undergo read
- Stacked transistors reduce leakage

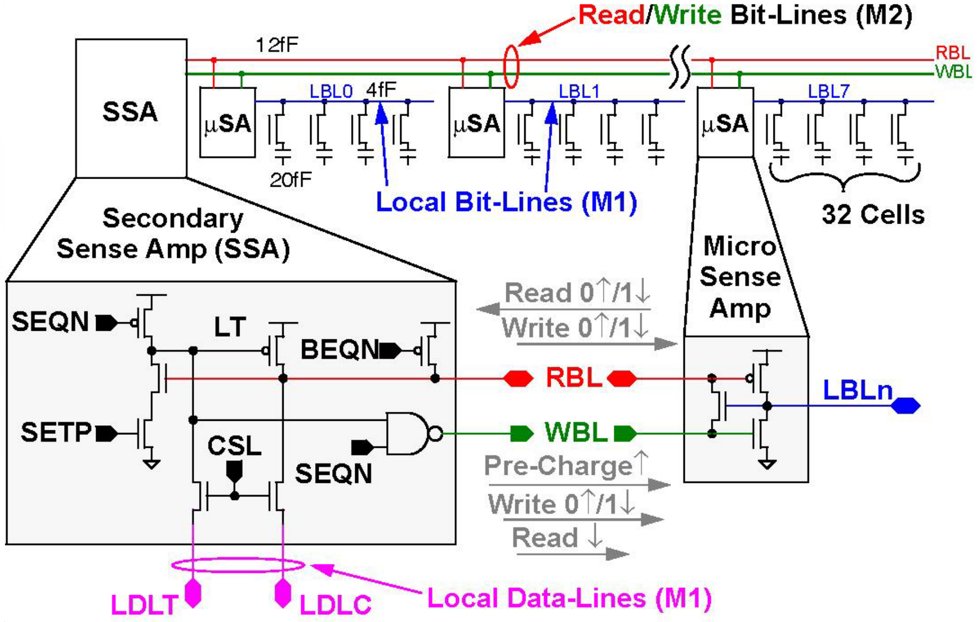
L. Chang, *VLSI Circuits* 2005

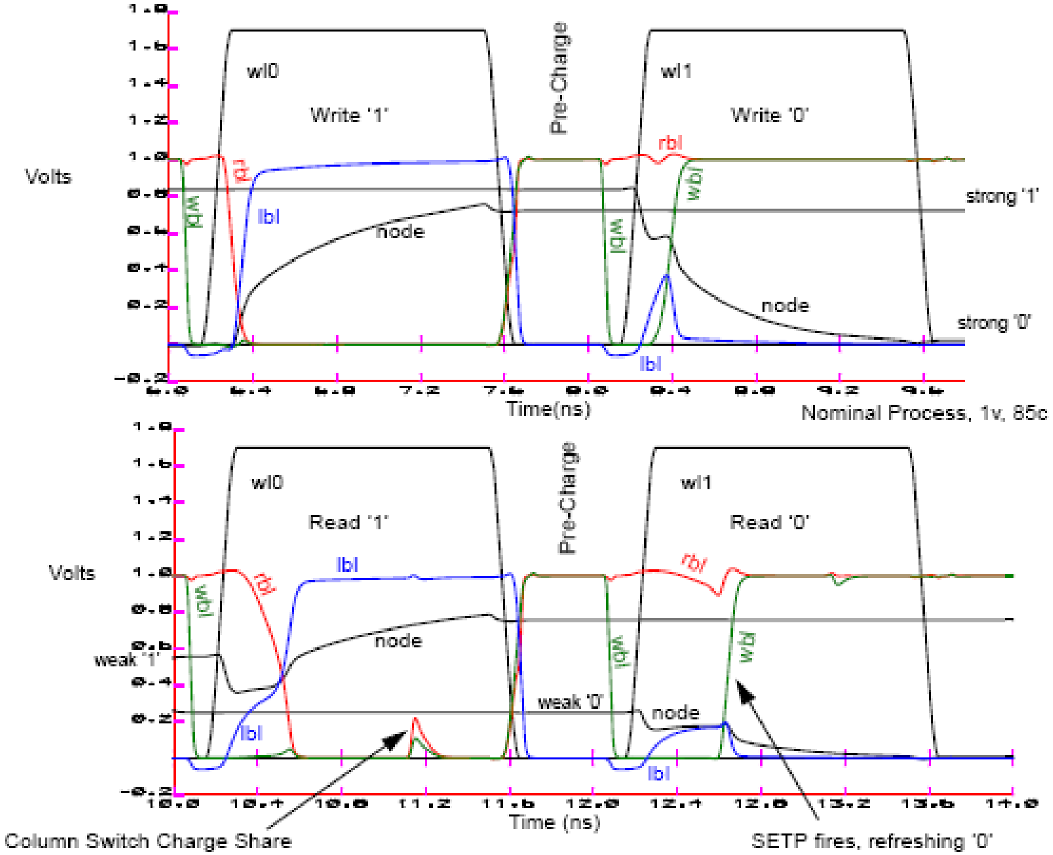
eDRAM

- Process cost: Added trench capacitor









Barth, ISSCC'07, Wang, IEDM'06

Crosspoint Memories

- Barrett, IRE Trans. Comp. 1961.

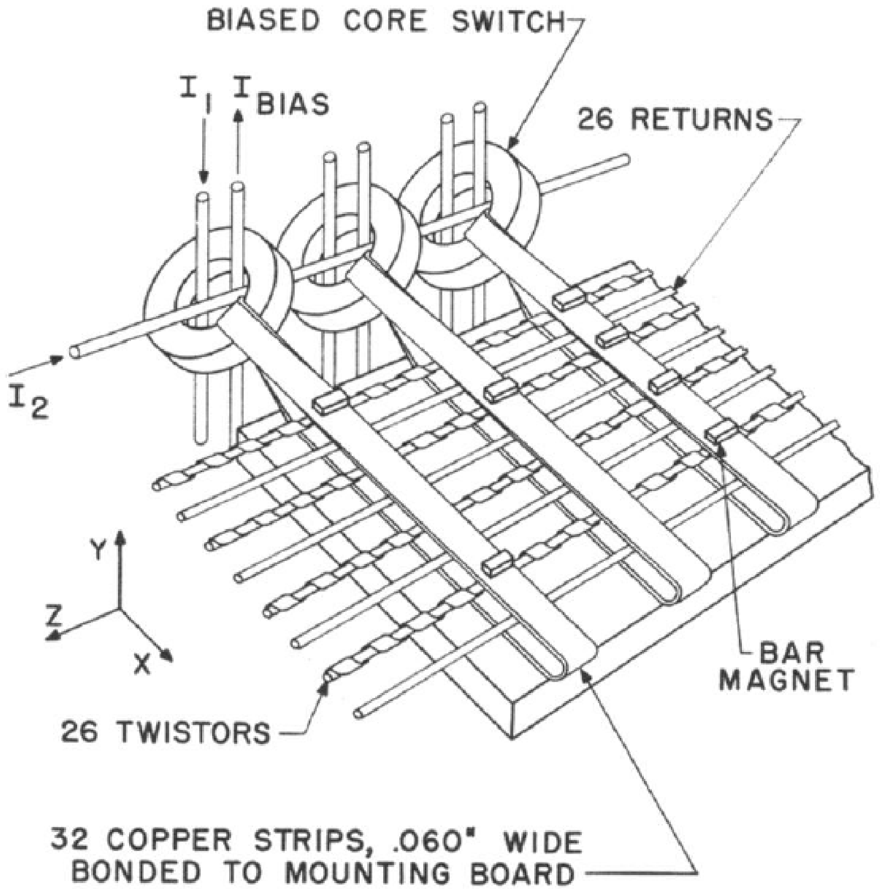
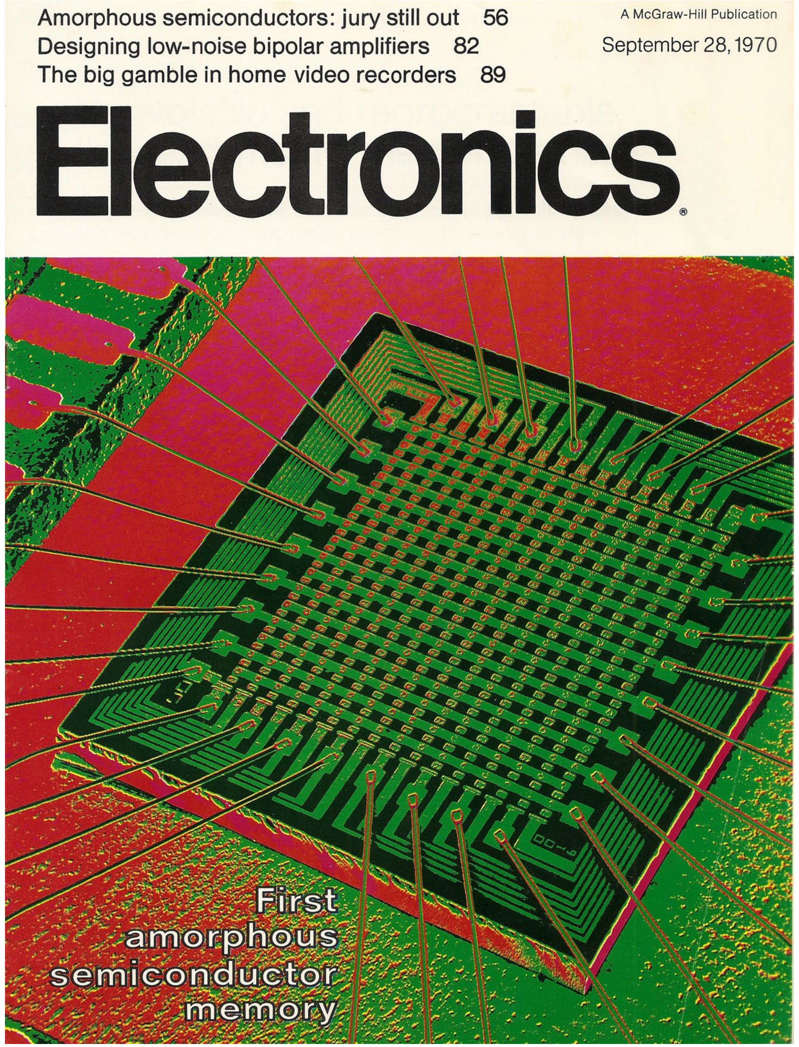


Fig. 2—Memory structure. I_1 and I_2 are access drive currents to core-selection switch. Presence or absence of a magnet over a twistor-strip solenoid crosspoint yields a "zero" or "one." Signals observed between twistor and return wire.

Crosspoint Memories



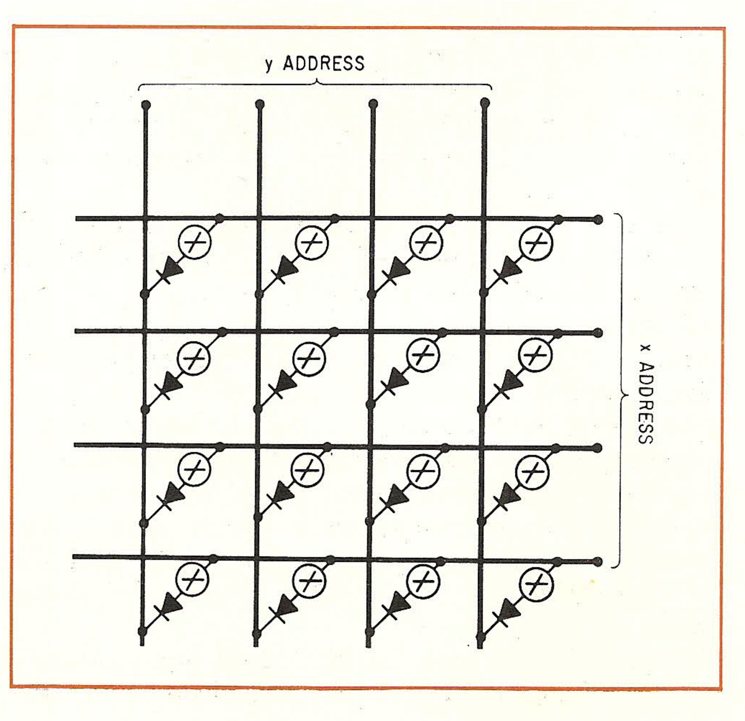
Amorphous semiconductors: jury still out 56
Designing low-noise bipolar amplifiers 82
The big gamble in home video recorders 89

A McGraw-Hill Publication
September 28, 1970

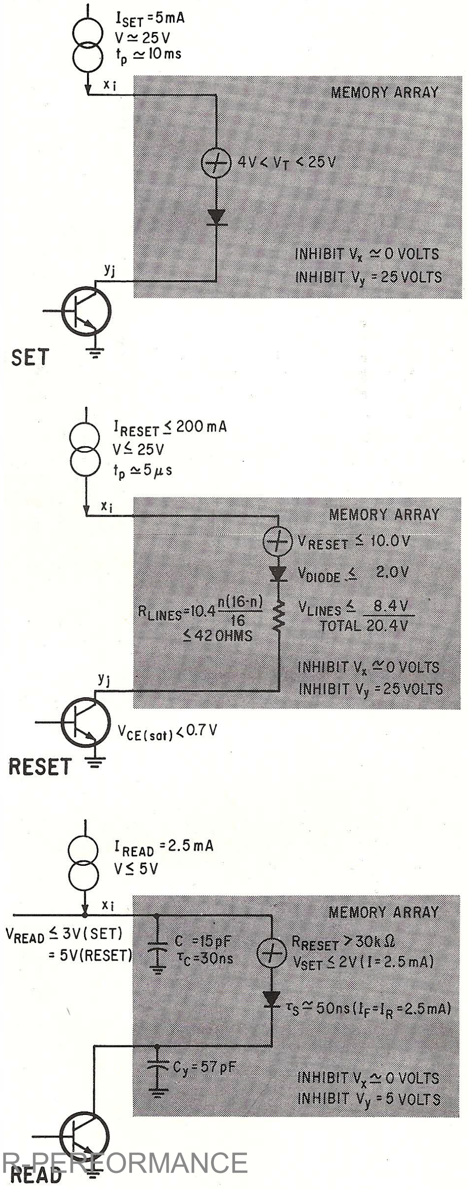
Electronics

First amorphous semiconductor memory

- Neale, Nelson, Moore, Electronics'70
    - 16 x 16 array (256b) of 'read-mostly memory'

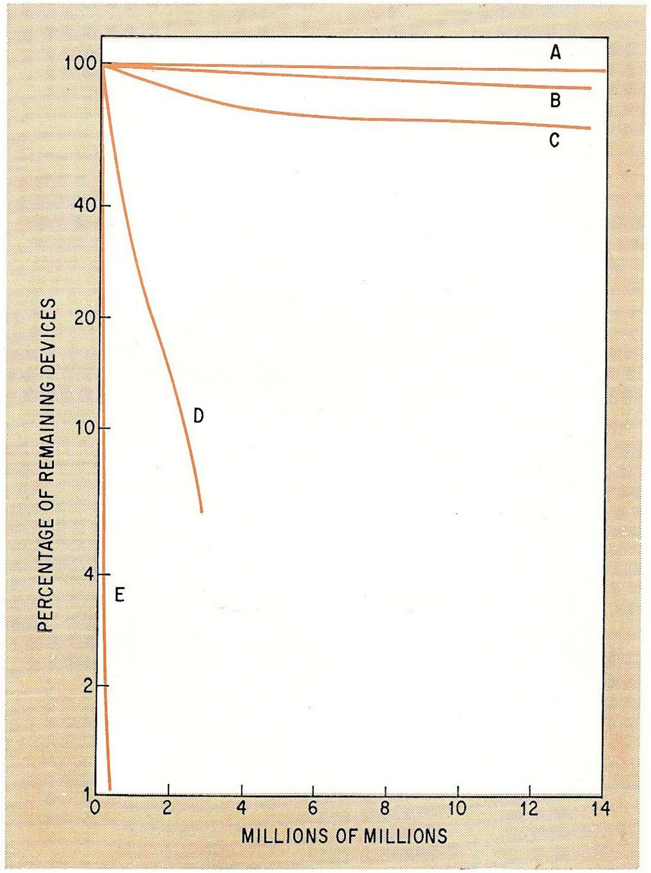


Crosspoint Memory



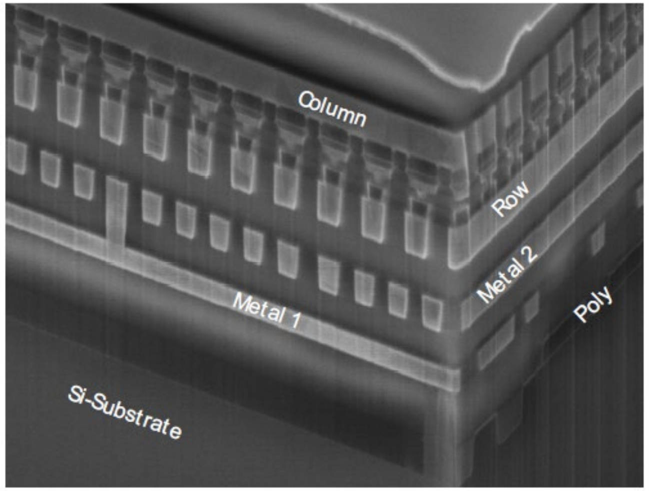
- Four modes
  - Form
  - Set
  - Reset
  - Read

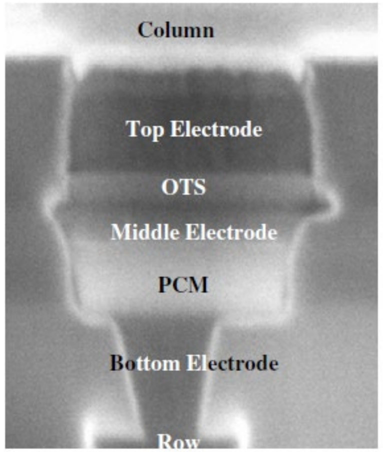
- **Endurance**

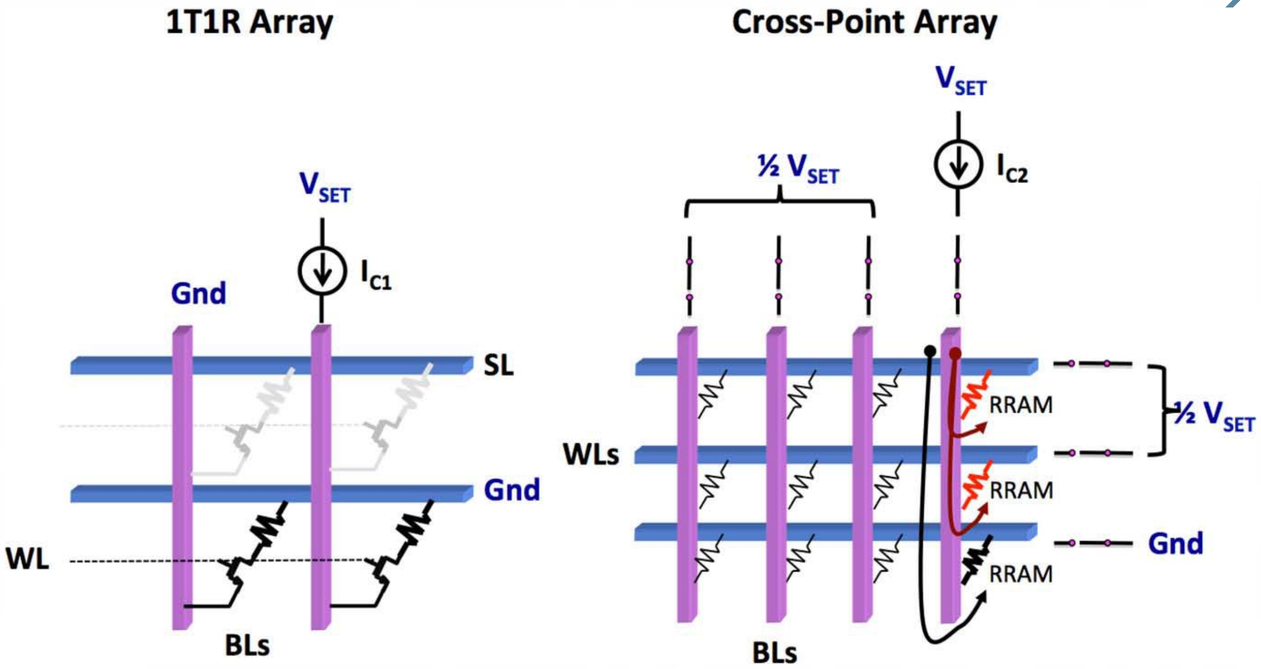


3D Crosspoint Arrays

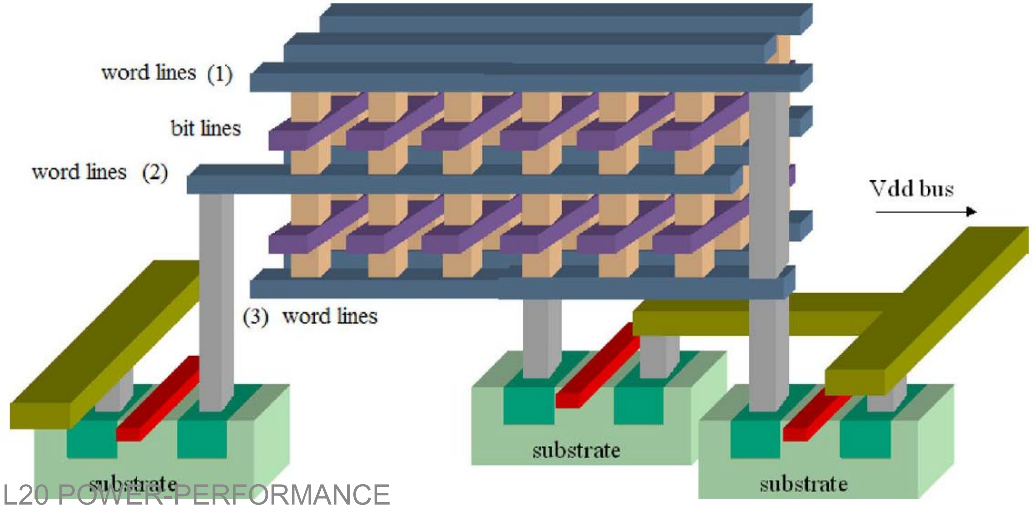
- Kau, IEDM'09







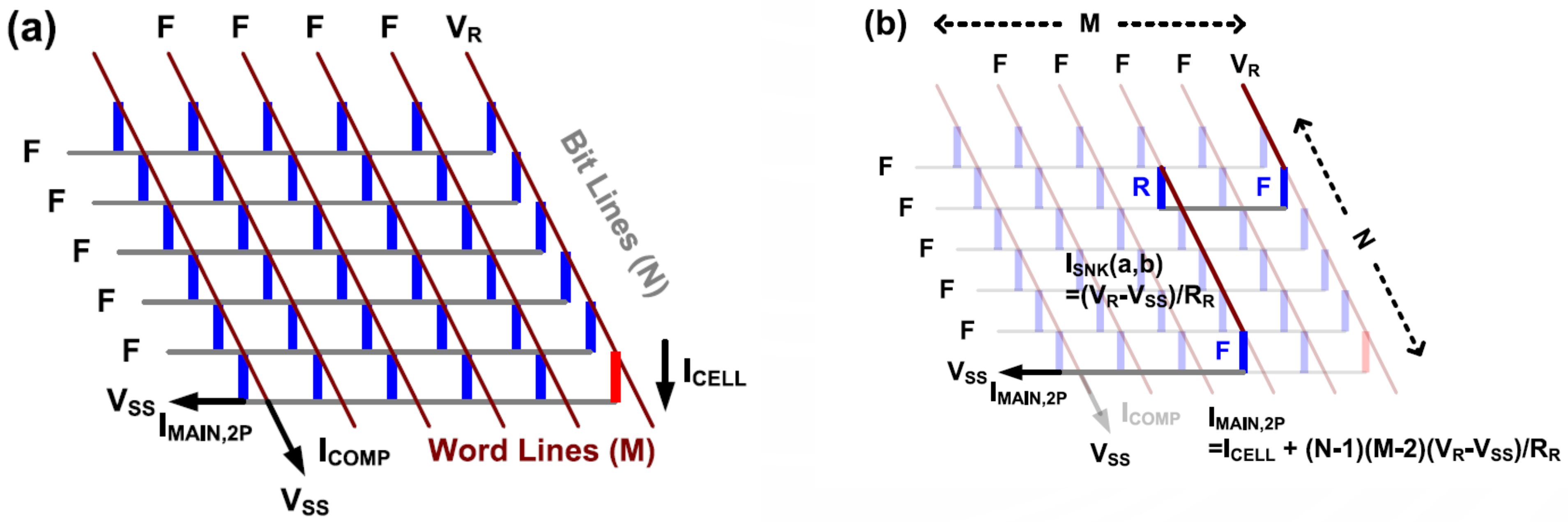
- Yeh, JSSC'15



- Ou, JSSC'11

Crosspoint Arrays

- Read and sneak currents



Bae, TED 4/17

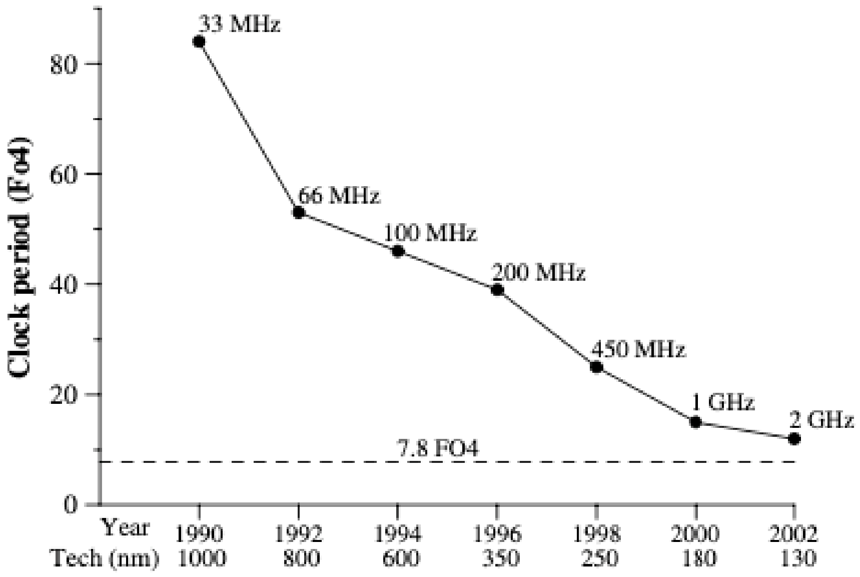
Low-Power Design

Importance of Power Awareness

- Energy: Crucial for Portable Applications
  - Determines battery lifetime
  - Amount of computation that can be performed
  - Performance is what sells products
- Power: Crucial for High-Performance Applications
  - Determines cooling and energy costs
  - Most designs today are power limited
  - Still need maximum performance

The Old Design Philosophy

- Maximum **performance** is primary goal
  - Architecture implements the required function with target throughput, latency
  - At circuit level, supplies, thresholds set to achieve **maximum performance**, subject to reliability constraints
  - Performance achieved through optimum sizing, logic mapping, architectural transformations



Hrishikesh, ISCA'02

Constant Field Scaling Model

Traditional scaling model

If 1/ alpha = 0.7 and chip_size_increase = 1.14^2 = 1.3

Vdd ~ $1/\alpha$ = 0.7

Cckt ~ $1/\alpha$

Cchip ~ α * (chip_size_increase) = 0.91

f ~ α = 1.43

Pckt ~ $\alpha * (1/\alpha) * (1/\alpha)^2 = (1/\alpha)^2$

Ackt ~ $(1/\alpha)^2$

Pckt/Ackt ~ 1

Pchip ~ $\alpha * (1/\alpha)^2 * \alpha$ * (chip_size_increase) = 1.3

Maintaining the frequency scaling model of 1990s

f ~ 2

Pckt/Ackt ~ $2 * (1/\alpha) * (1/\alpha)^2/(1/\alpha)^2 = 2 * 0.7 = 1.4$

Pchip ~ Pckt/Ackt * (chip_size_increase) = 1.8

While slowing down voltage scaling

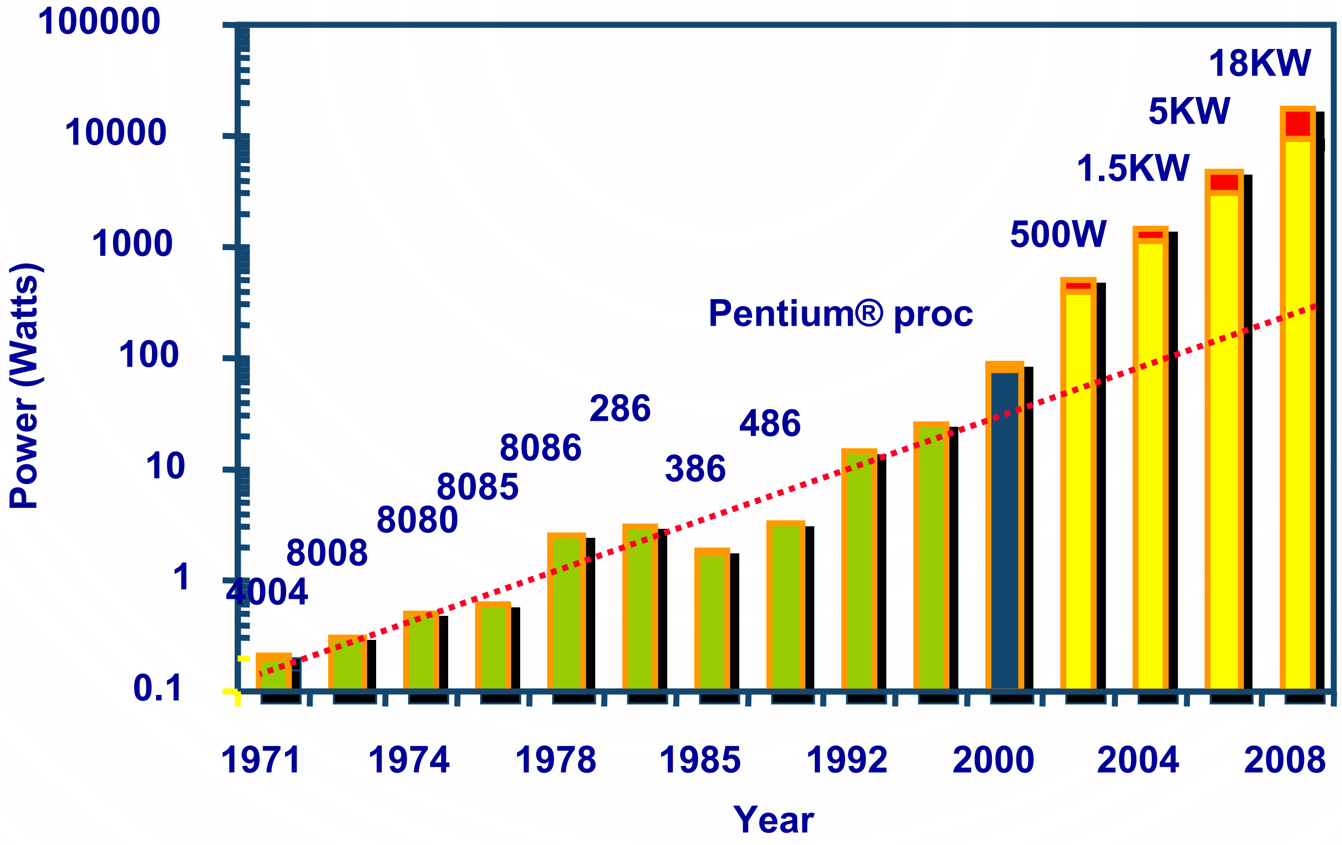
f ~ 2

Vdd ~ 0.9

Pckt/Ackt ~ $2 * (1/\alpha) * (0.9)^2/(1/\alpha)^2 = (0.9)^2 * 1.43 * 2 = 2.3$

Pchip ~ Pckt/Ackt * (chip_size_increase) = 3

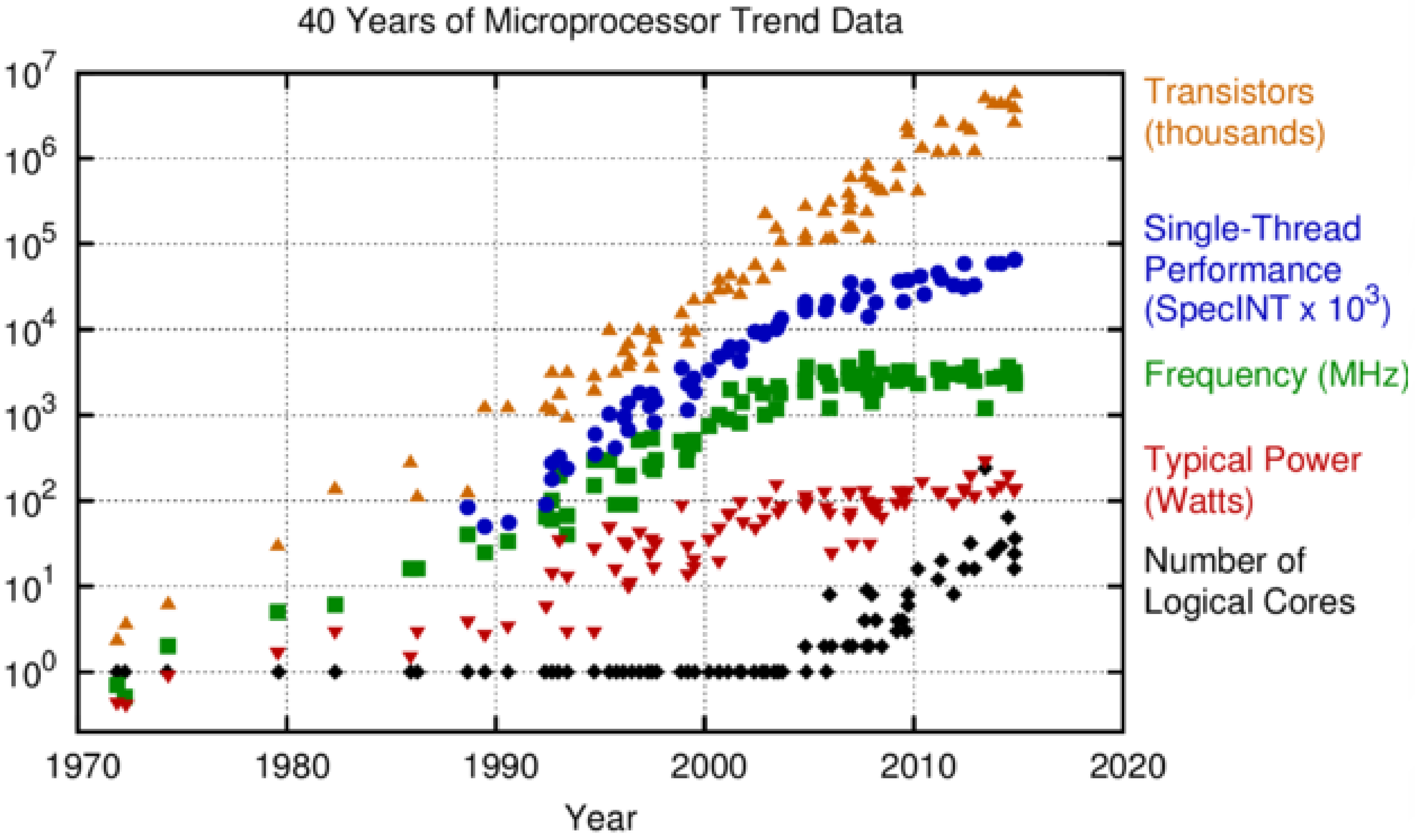
2001 Picture: Power As a Problem



S. Borkar

Power delivery and dissipation will be prohibitive

What actually happened?



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2015 by K. Rupp

The New Design Philosophy

Maximum performance is too power-hungry, and/or not even practically achievable

Extract maximum performance under a power/energy envelope

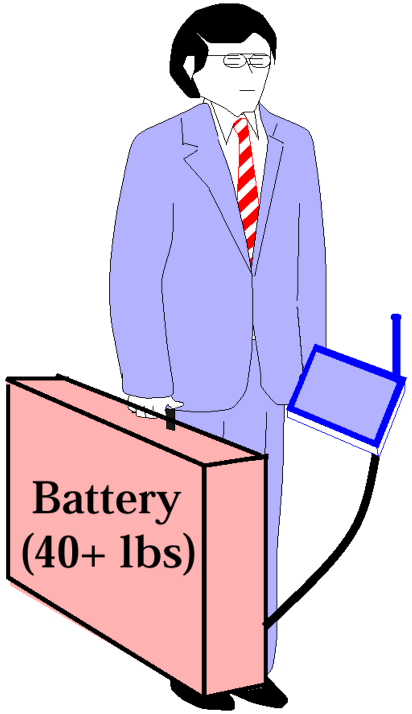
Excess performance (as offered by technology) to be used for energy/power reduction

Trading off speed for power

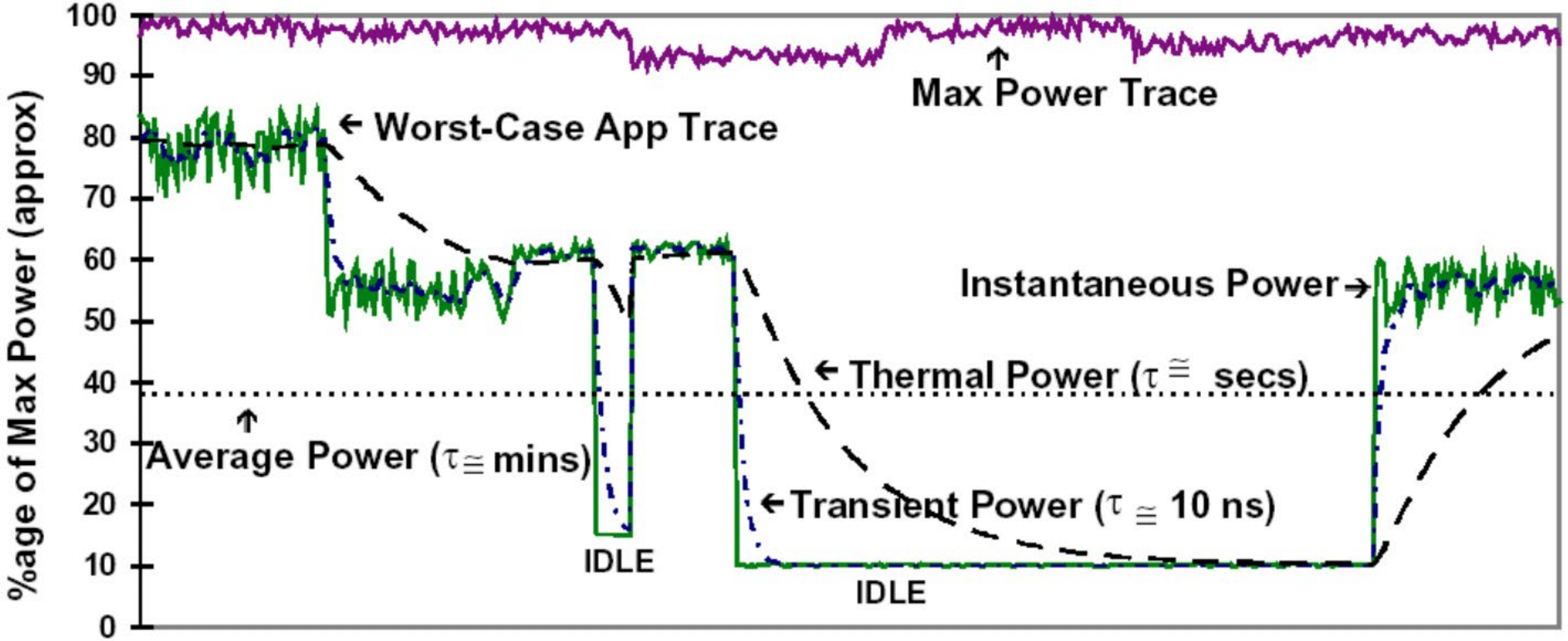
Power and Energy Basics

Energy: Battery Limits

- Little change in basic technology
  - Store energy by using a chemical reaction
- Battery capacity doubles every 10 years
  - Has slowed down
- Energy density/size, safe handling are limiting factor

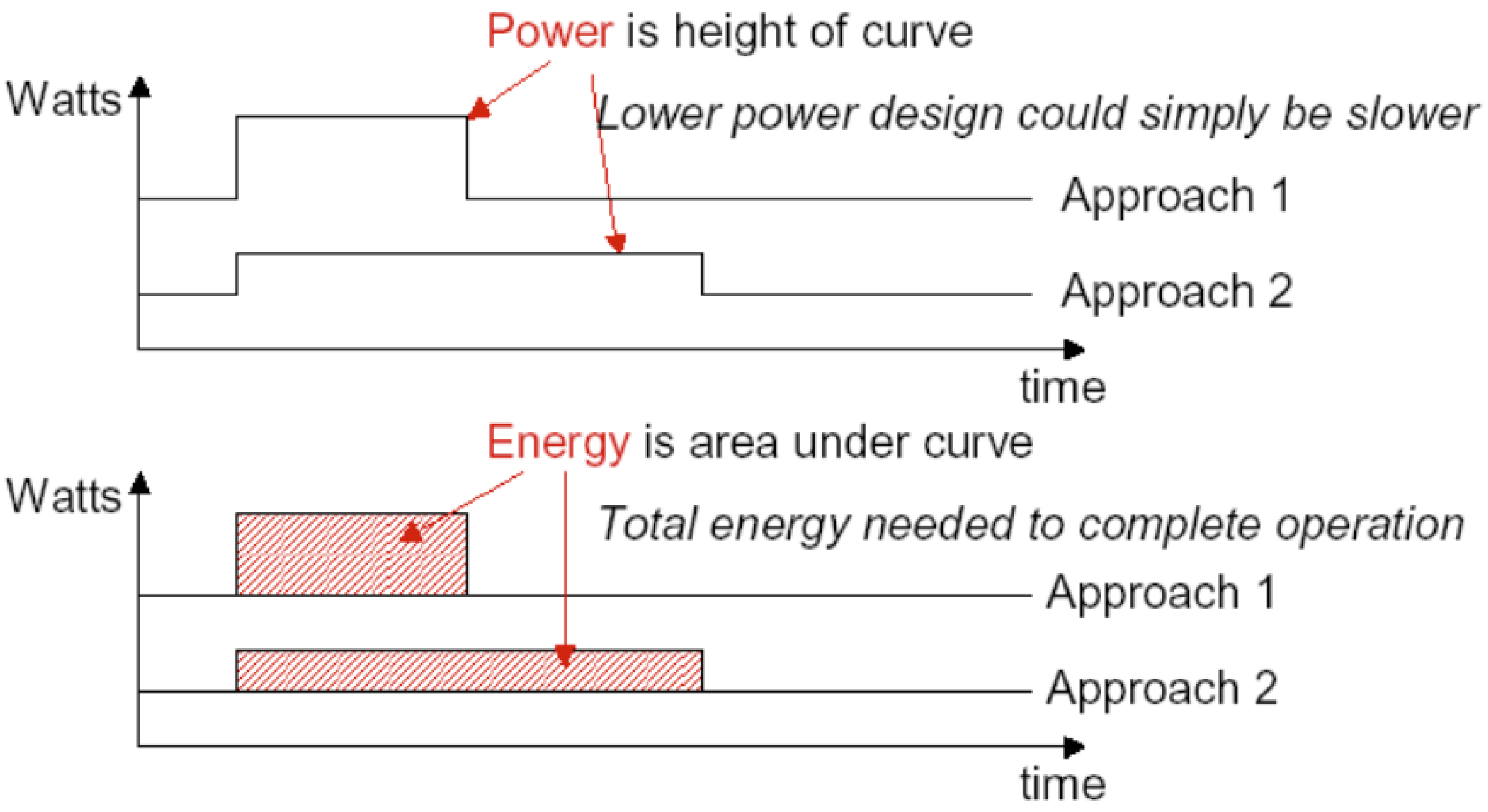


What do we mean by power?



- Max Power: Artificial code generating max activity (power virus).
- Worst-case App Trace: Practical applications worst case power
- Thermal Power: Running average of worst-case app power over a period corresponding to the thermal time constant
- Average Power: Long term average (minutes) of typical apps
- Transient Power: Variability in power consumption for power supply

Power vs Energy


Power is height of curve
Watts
Lower power design could simply be slower
Approach 1
Approach 2
time
Energy is area under curve
Watts
Total energy needed to complete operation
Approach 1
Approach 2
time

- Energy = power * delay (joules = watts * seconds)

Power vs Energy

- **Power-delay Product (PDP/Energy)** = P_avg * T
  - Energy consumed
- **Energy-delay Product (EDP)** = PDP * T
  - Takes into account that one can trade increased delay for lower energy
- **Energy-delay^2 Product (EDDP)** = EDP * T
  - Why do we need so many formulas?!!?
  - We want a voltage-invariant efficiency metric.
    - Power ~ CV^2f, and performance ~ f (and V)

Know Your Enemy

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

Summary of Power Dissipation Sources

$$P \sim \underbrace{\alpha \cdot \left(C_L + C_{CS}\right)\cdot V_{swing} \cdot V_{DD} \cdot f}_{\text{Dynamic power}} + \underbrace{\left(I_{DC} + I_{Leak}\right)\cdot V_{DD}}_{\text{Static power}}$$

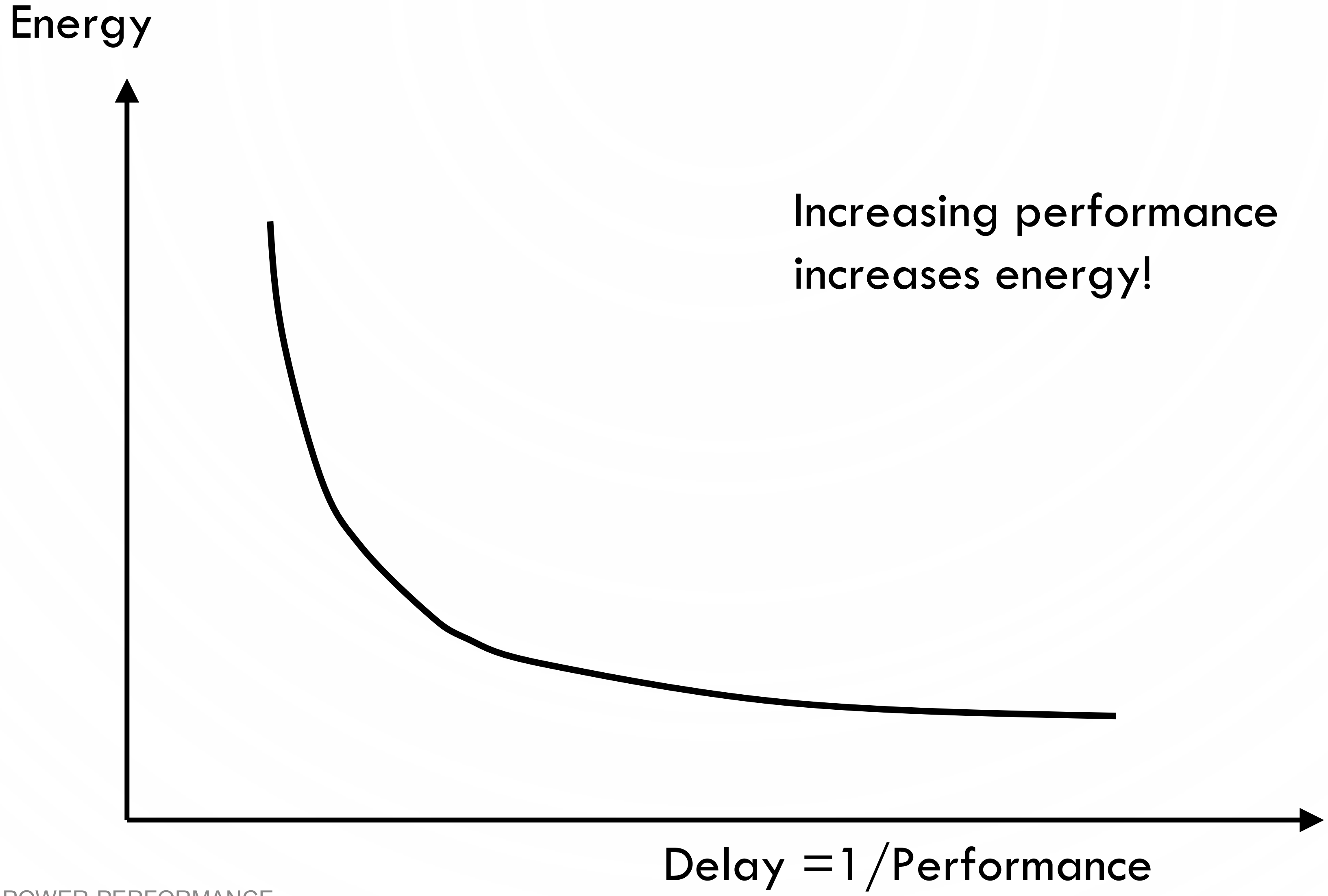
- α – switching activity
- C_L – load capacitance
- C_{CS} – short-circuit "capacitance"
- V_{swing} – voltage swing
- f – frequency

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

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

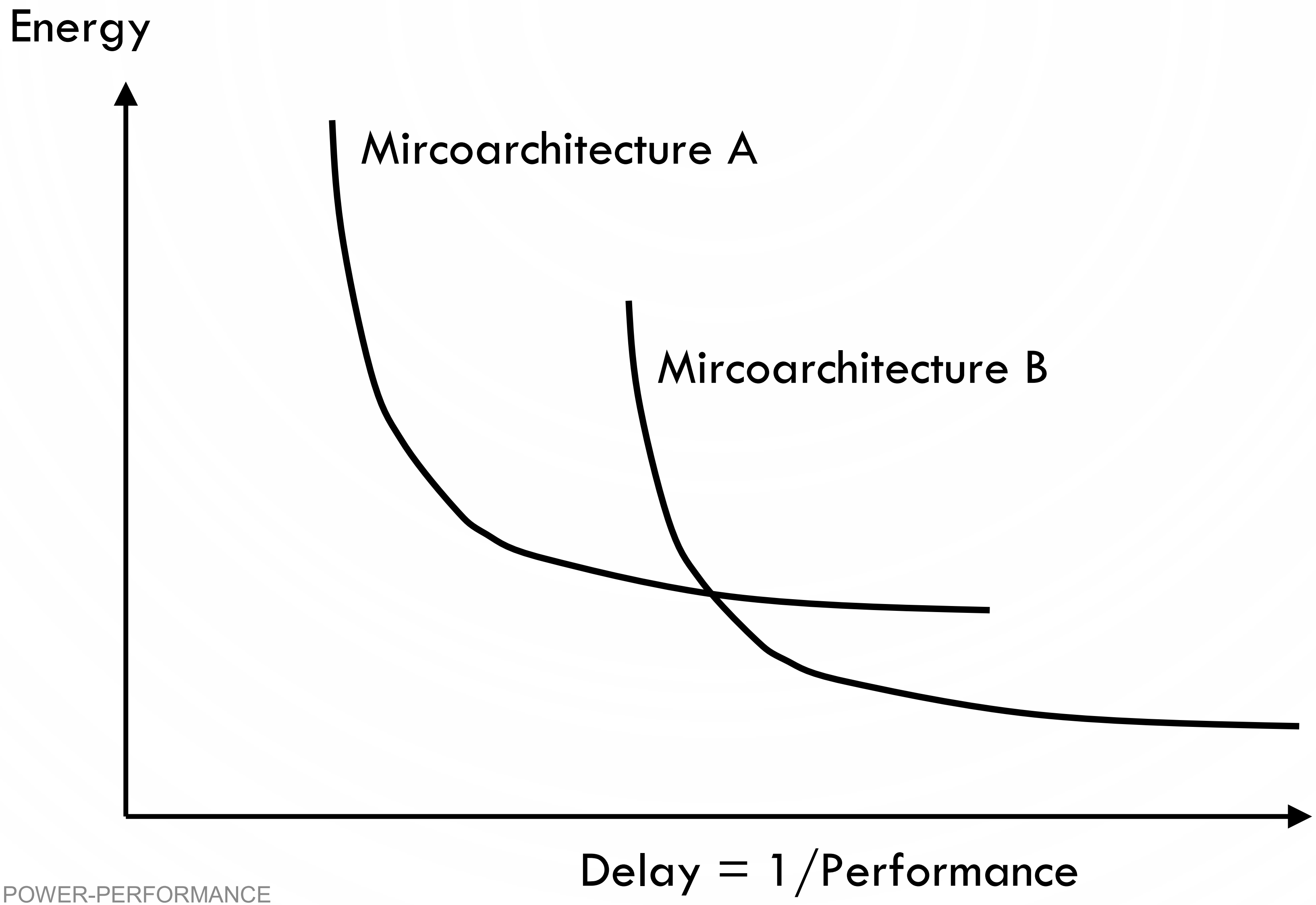
Power-Performance Tradeoffs

Performance Optimization

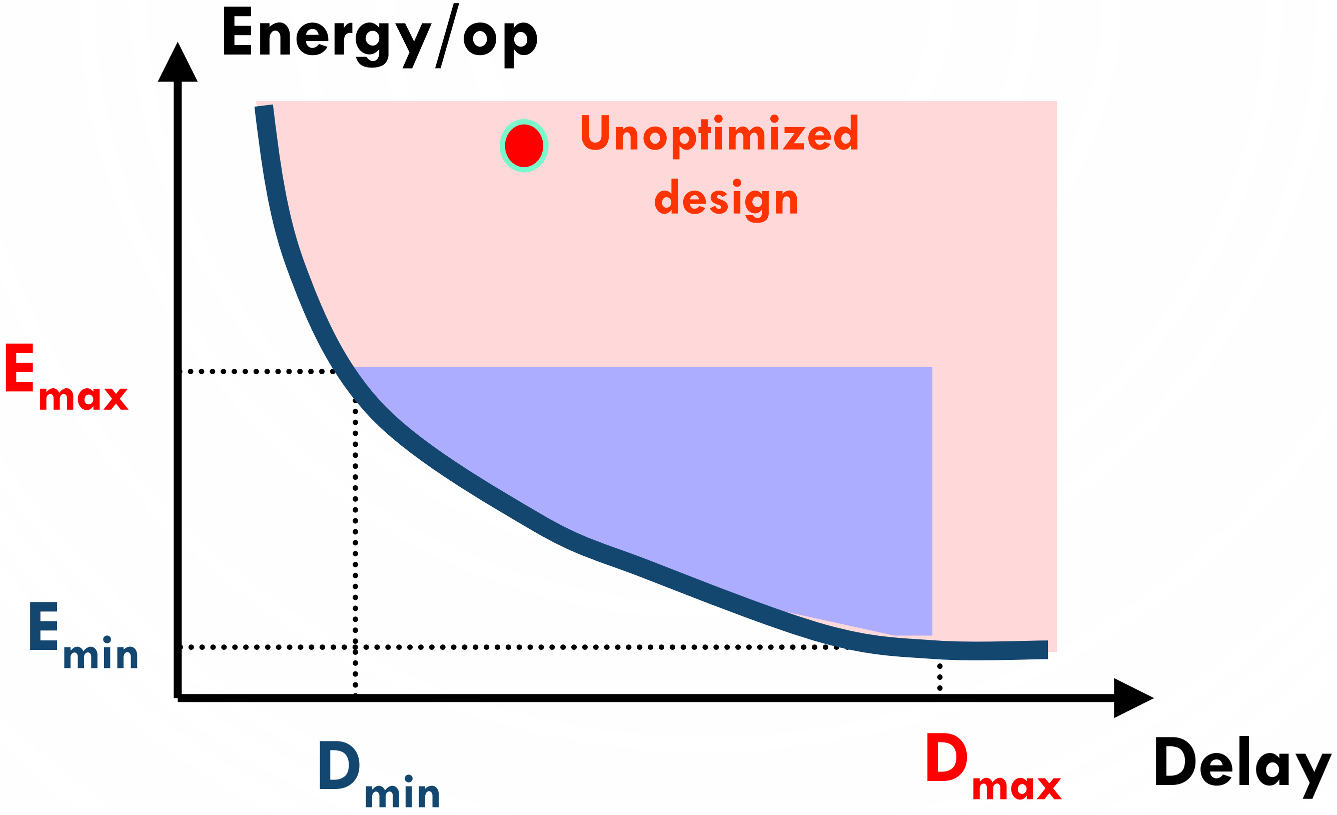


Increasing performance increases energy!

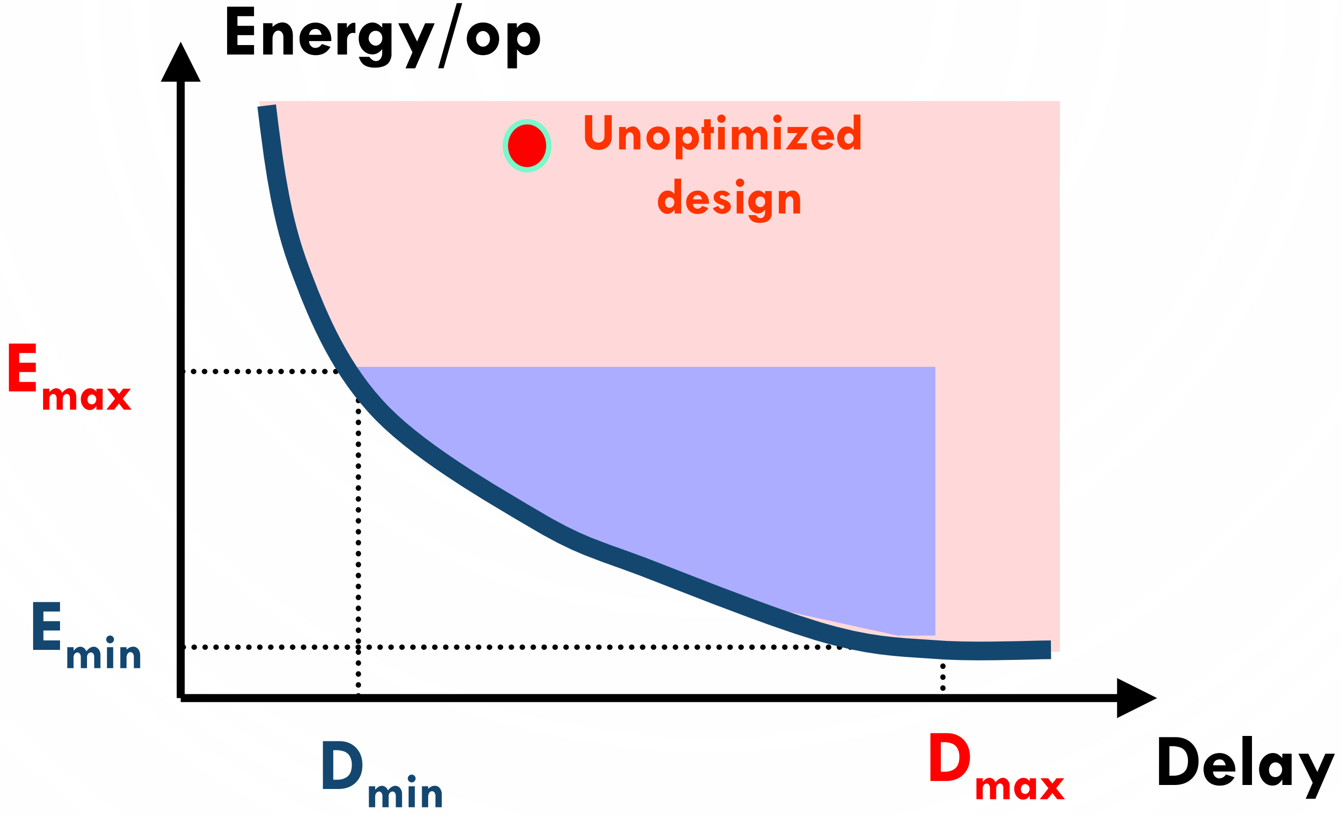
Performance Optimization



Power-Performance Optimization

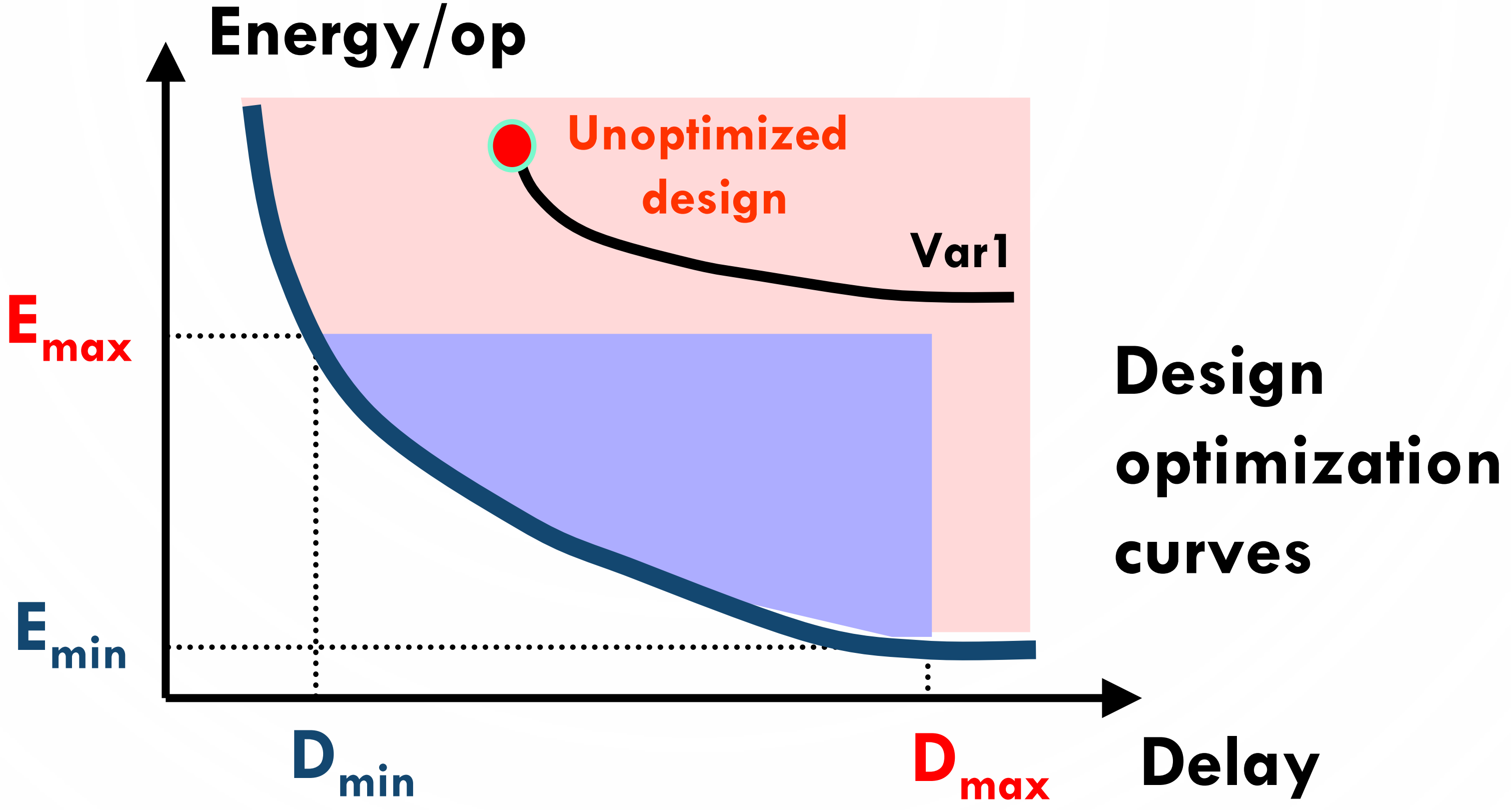


Power-Performance Optimization



Achieve the highest performance under the energy constraint

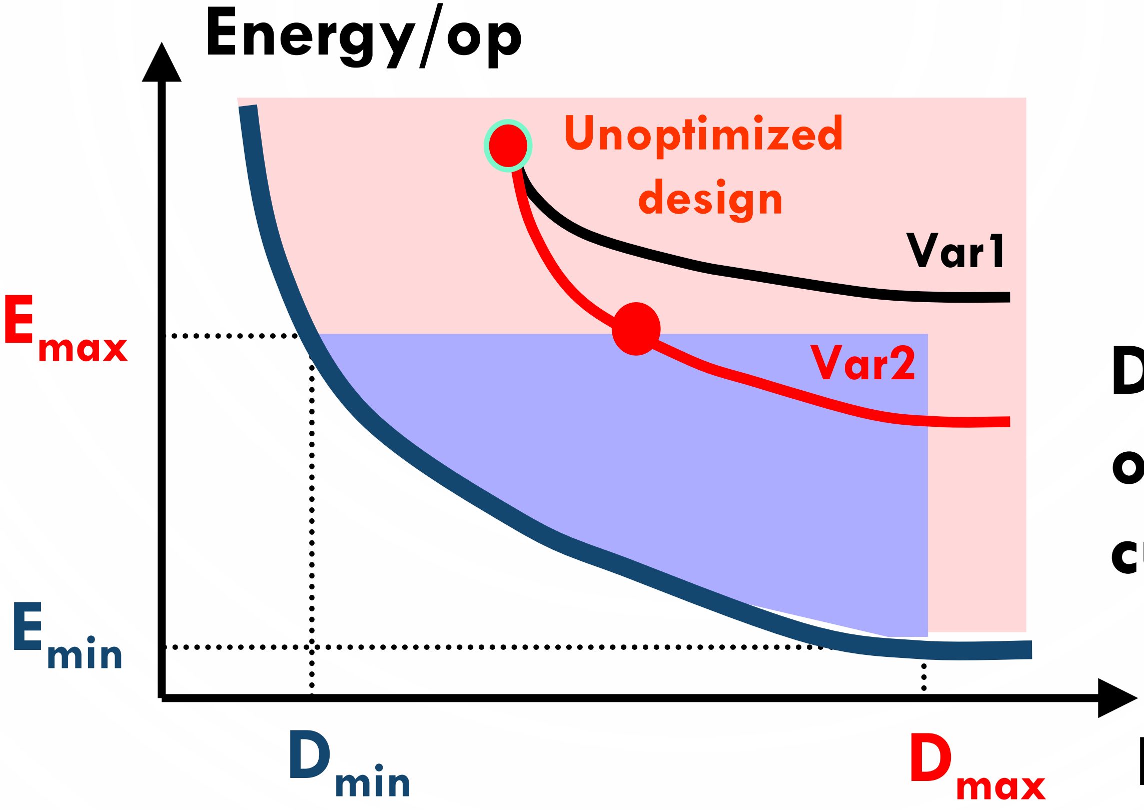
Power-Performance Optimization



Design optimization curves

Achieve the highest performance under the energy constraint

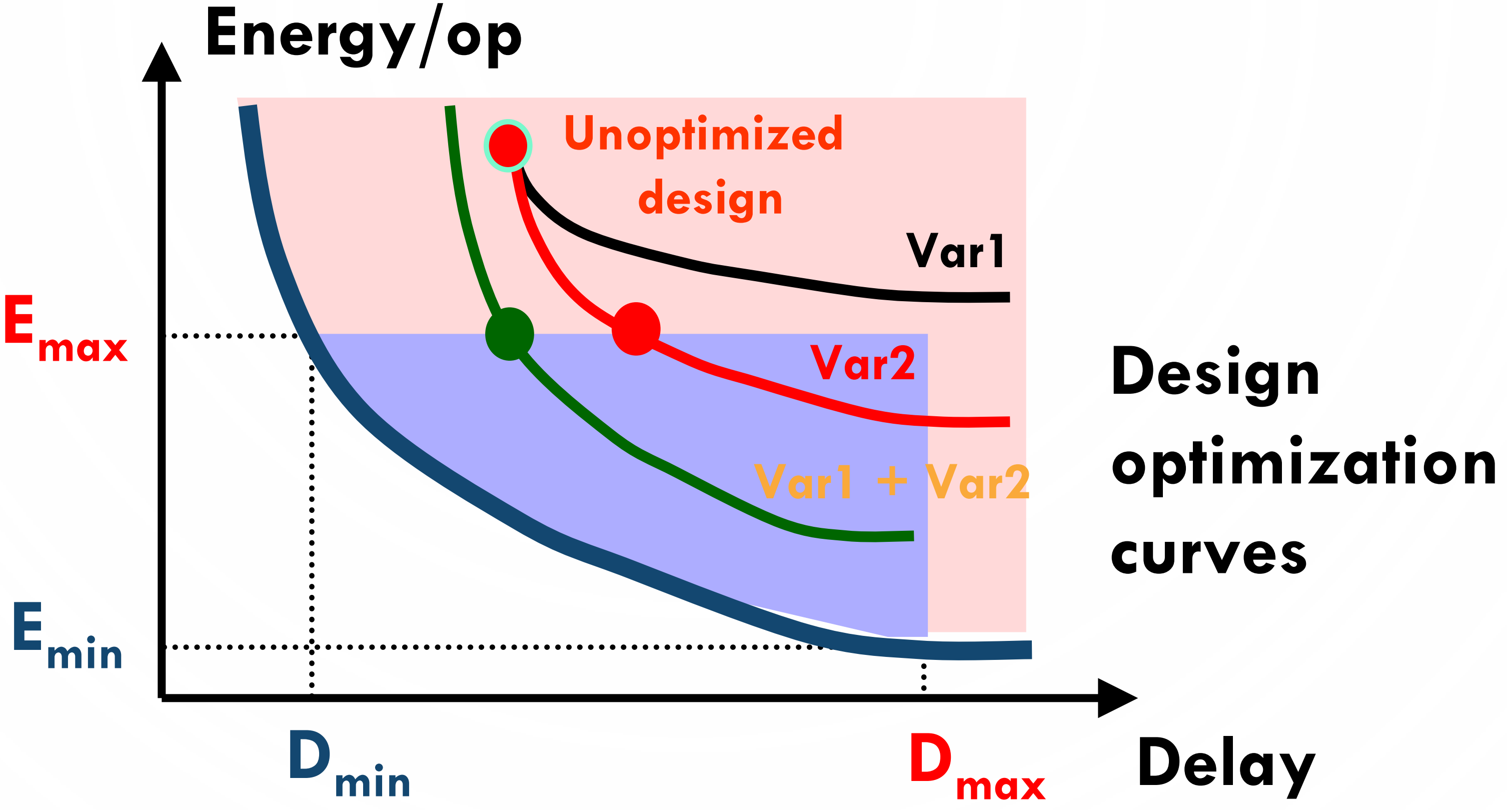
Power-Performance Optimization



Design optimization curves

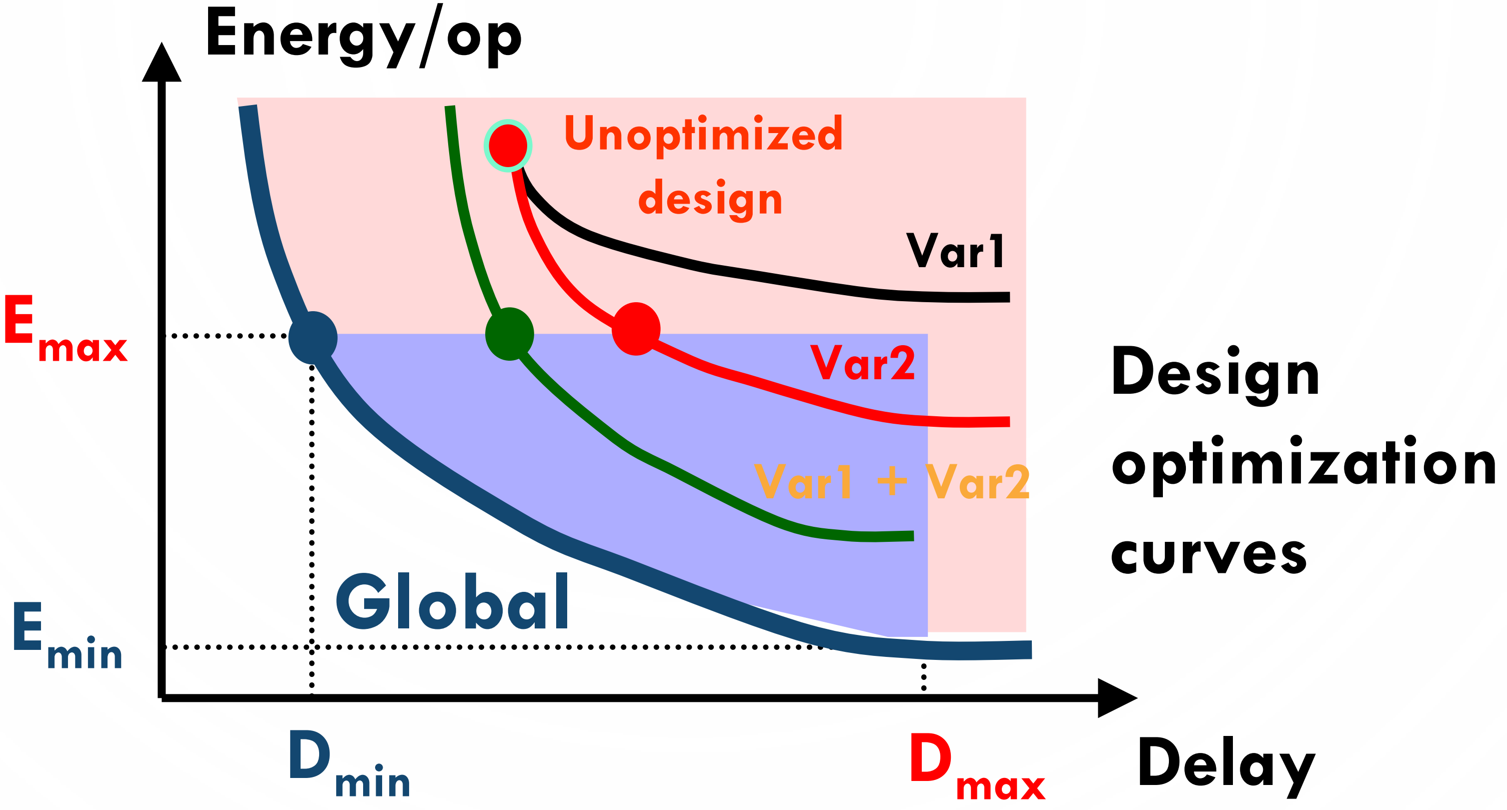
Achieve the highest performance under the energy constraint

Power-Performance Optimization



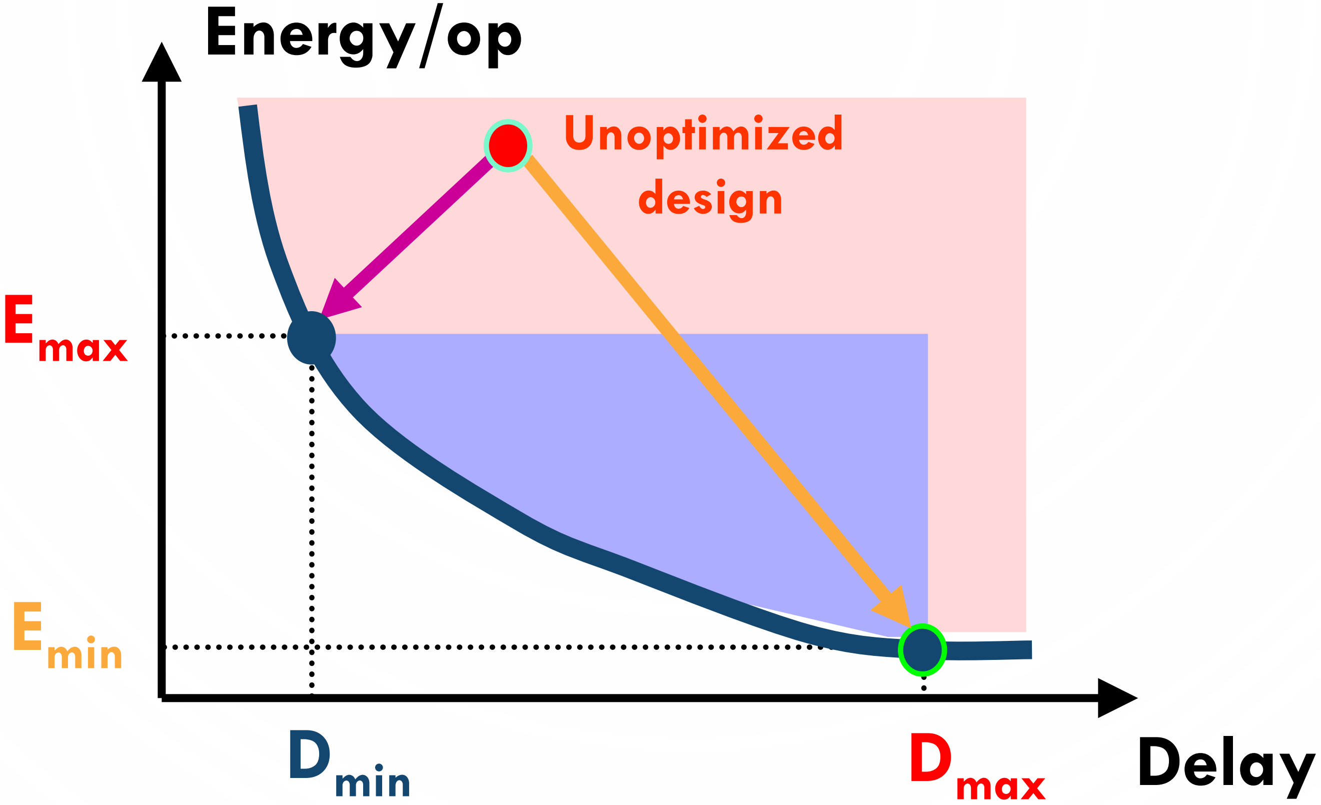
How far away are we from the optimal solution?

Power-Performance Optimization



Global optimum – best performance

Power-Performance Optimization



Maximize throughput for given energy or

Minimize energy for given throughput

Power-Performance Optimization

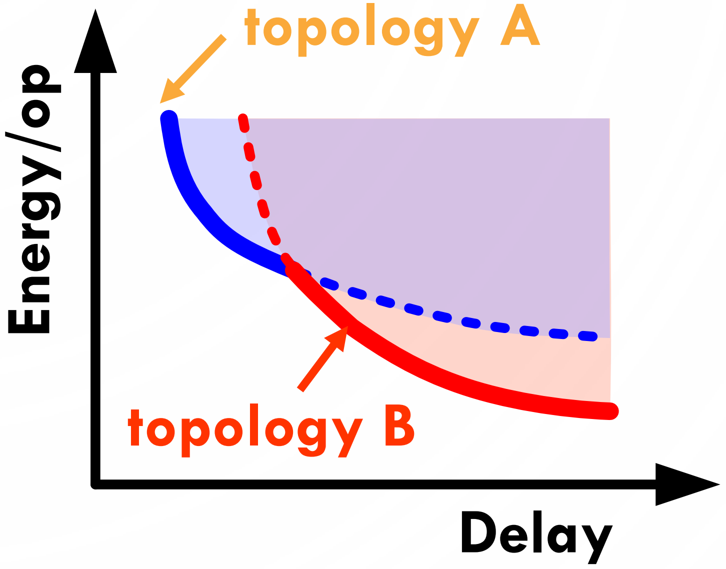
- There are many sets of parameters to adjust
  - Tuning variables
  - Circuit

    (sizing, supply, threshold)
  - Logic style

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

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

    (parallel, pipelined)



Power-Performance Optimization

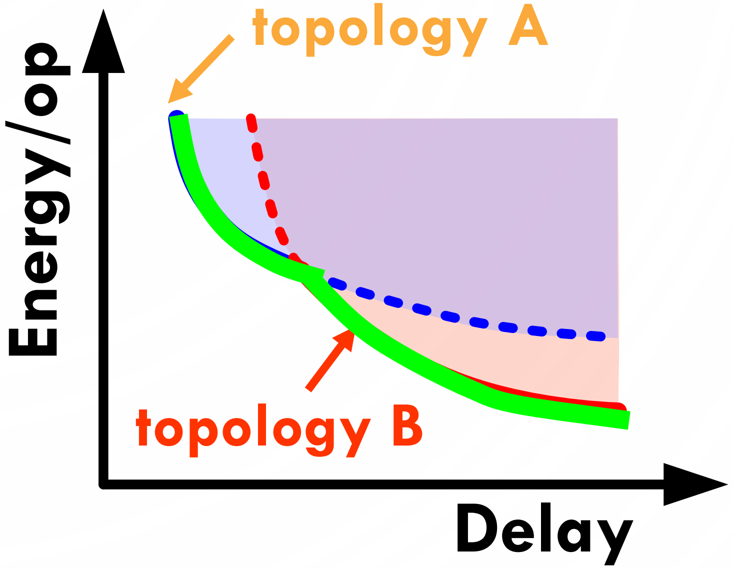
- There are many sets of parameters to adjust
  - Tuning variables
  - Circuit

    (sizing, supply, threshold)
  - Logic style

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

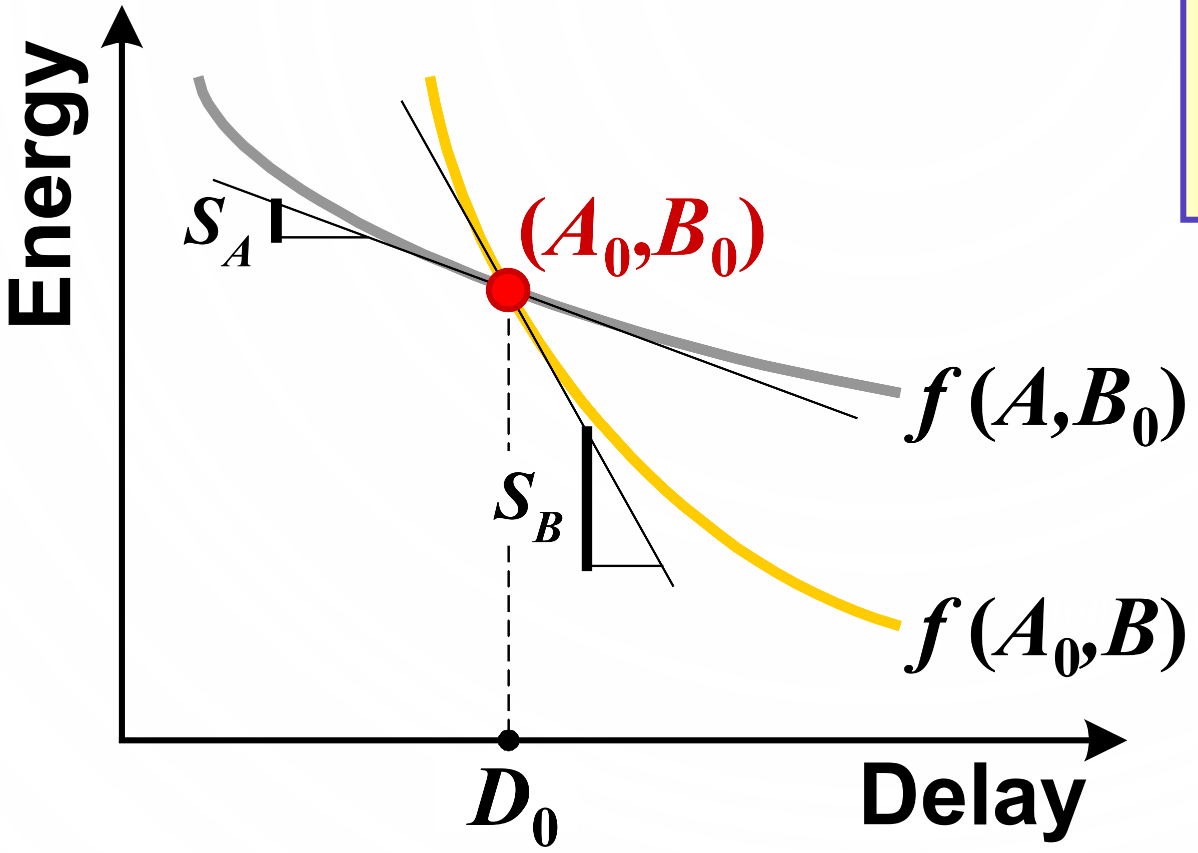
    (adder: CLA, CSA, …)
  - Micro-architecture

    (parallel, pipelined)



Globally optimal power-performance curve for a given function

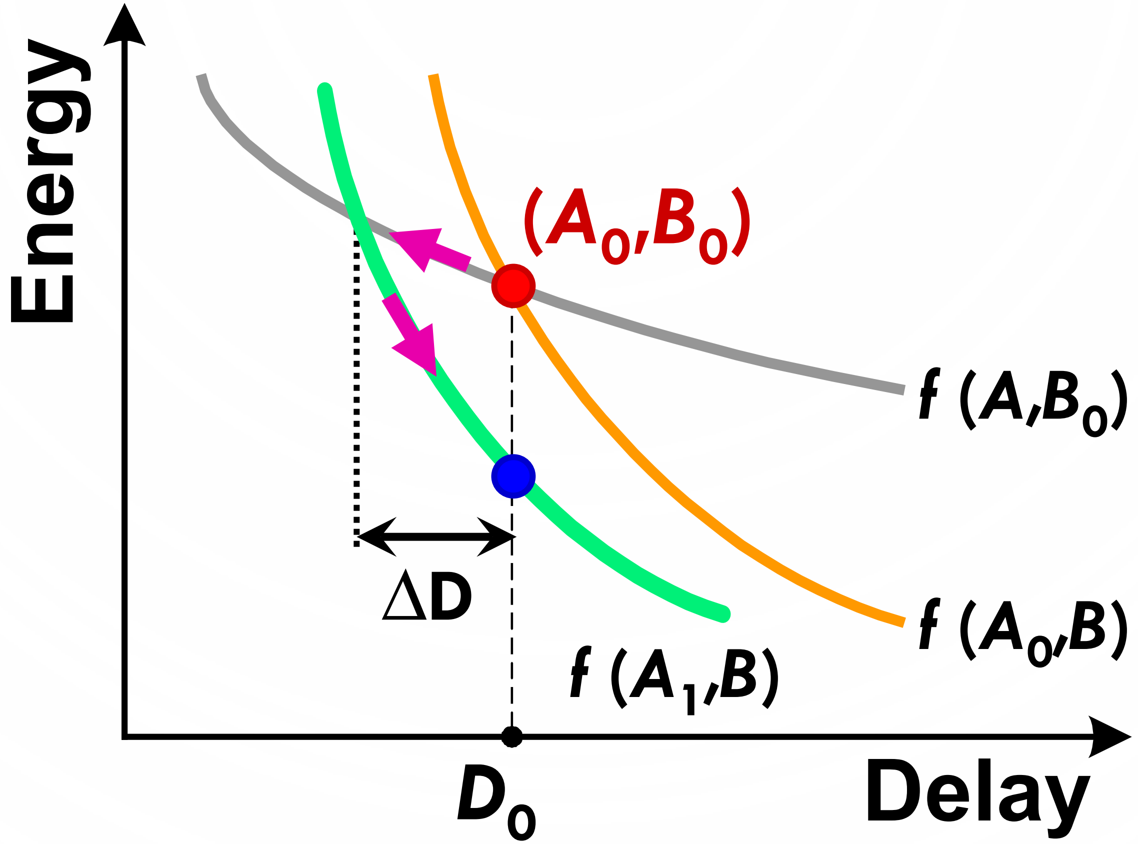
Energy-Delay Sensitivity



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

Solution: Equal Sensitivities

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



At the solution point all sensitivities should be equal

Architectural Optimizations

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 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

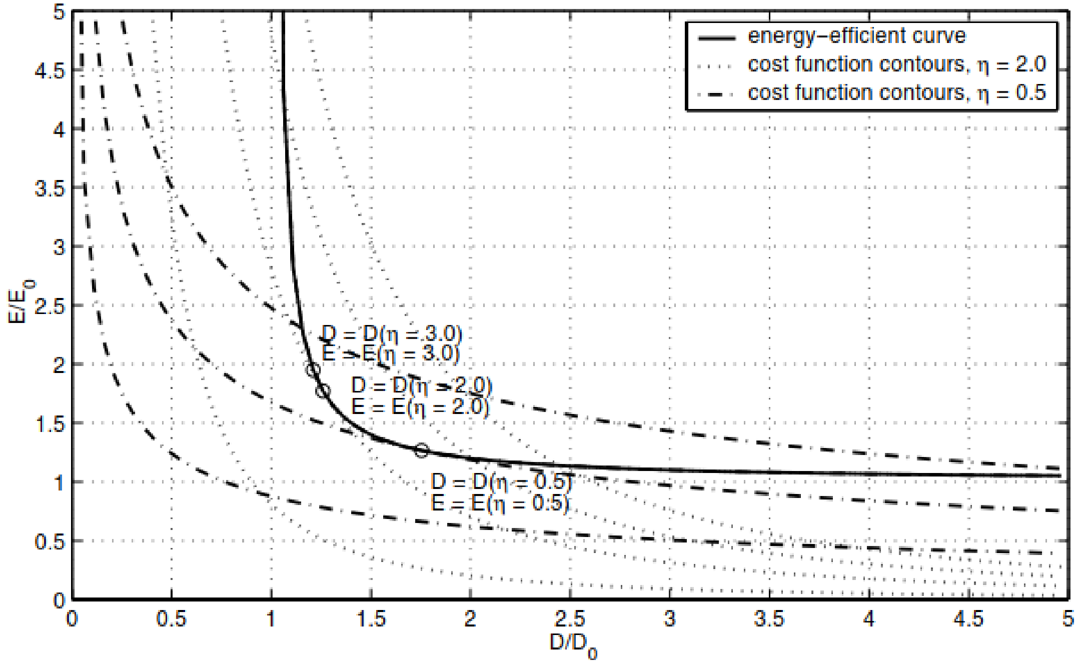
- Set up optimization problem:
  - Maximize performance under energy constraints
  - Minimize energy under performance constraints
- Or minimize a composite function of $E^m D^n$
  - What are the right m and n?
- $m = 1, n = 1$ is EDP – improves at lower V_{DD}
- $m = 1, n = 2$ is invariant to V_{DD}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

Hardware Intensity

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

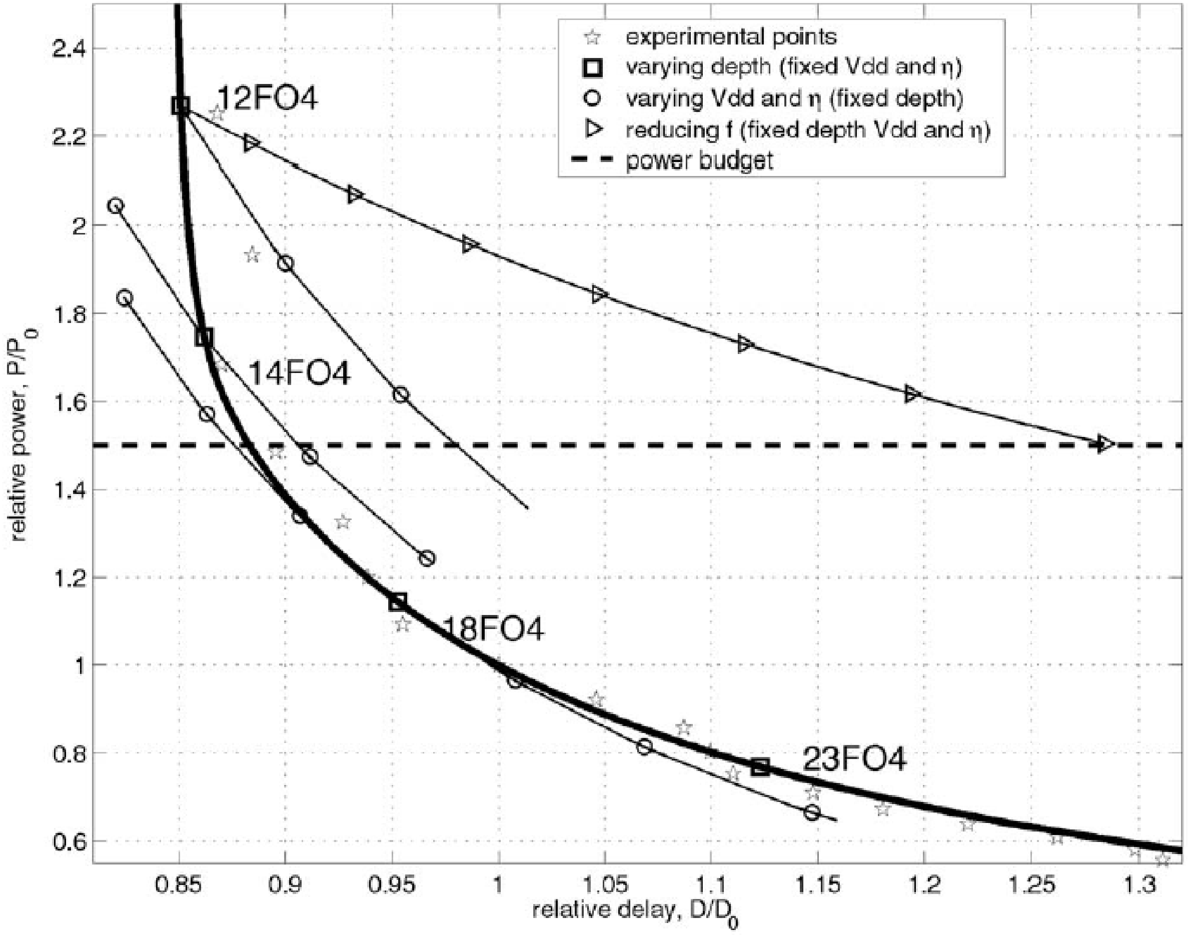
$$F_c = (E/E_0)(D/D_0)^{\eta} \qquad 0 \leq \eta < +\infty,$$

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



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

Optimum Across Hierarchy Layers



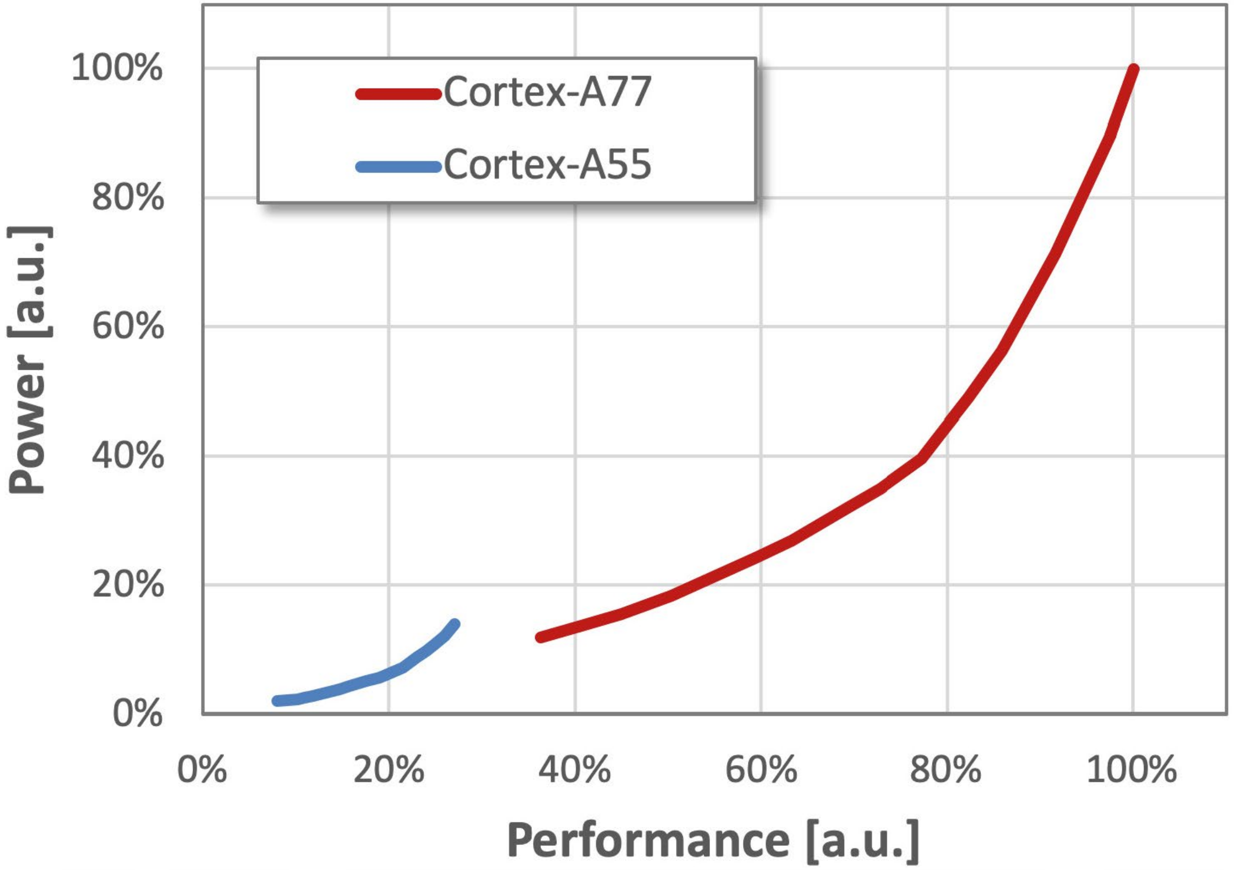
Zyuban et al, TComp’04

Optimal logic depth in pipelined processors is ~18FO4

Relatively flat in the 16-22FO4 range

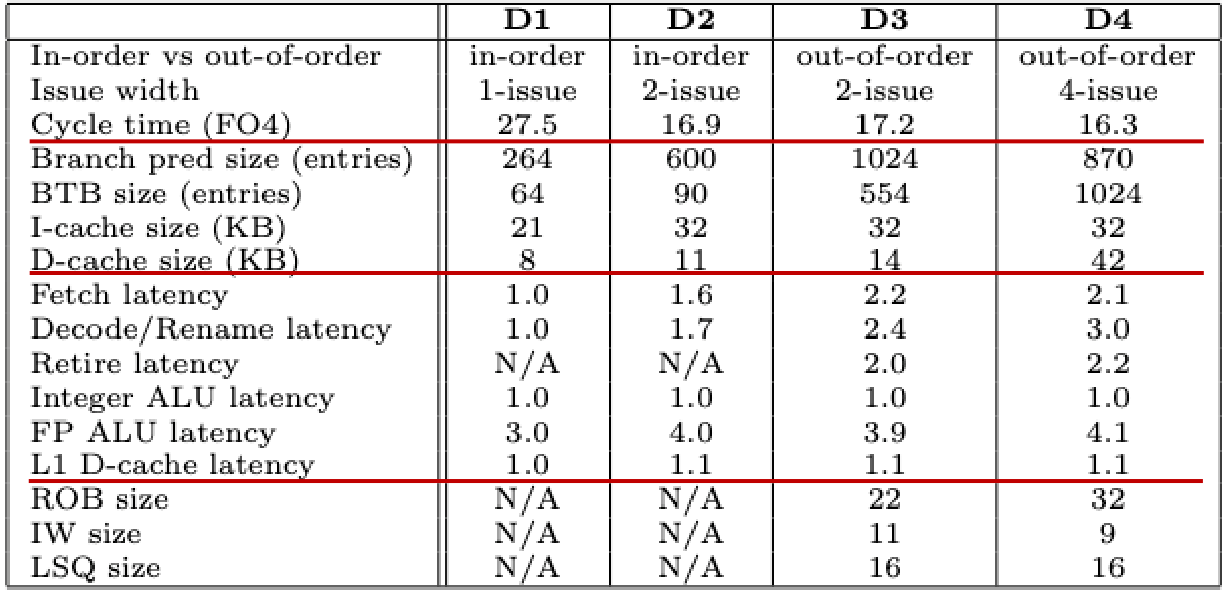
Architectural Tradeoffs

- H, Mair, ISSCC’20

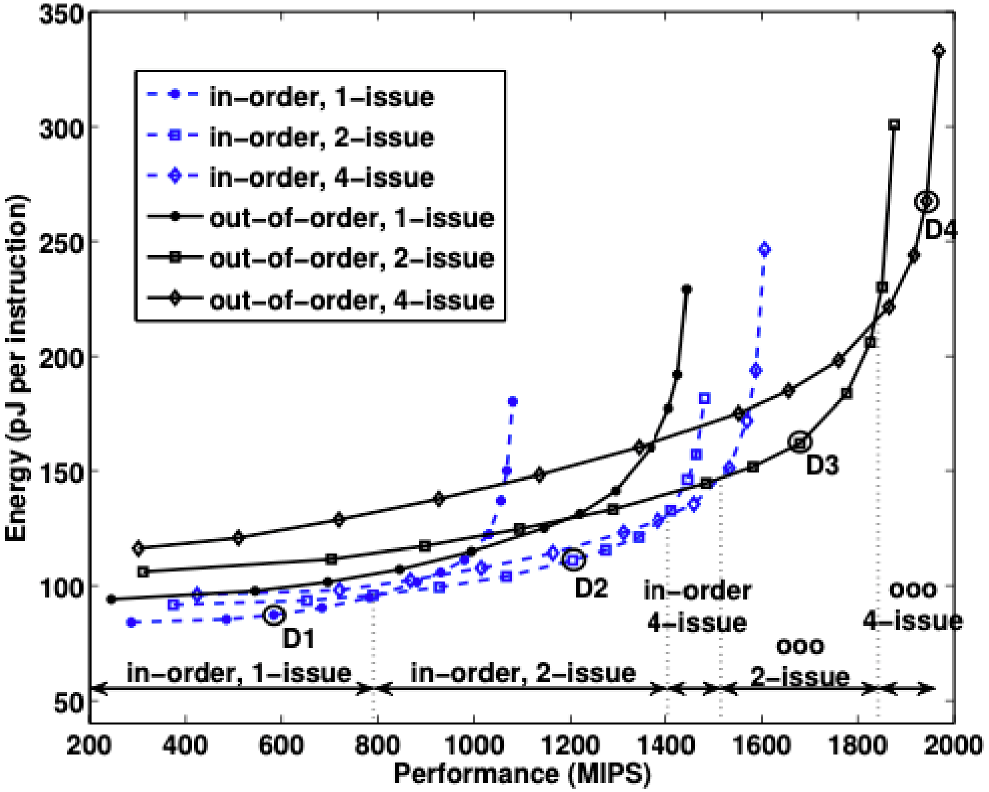


Energy-Delay Tradeoff of Modern Processors

Table 3: *Design Configuration Details For Selected Design Points.*



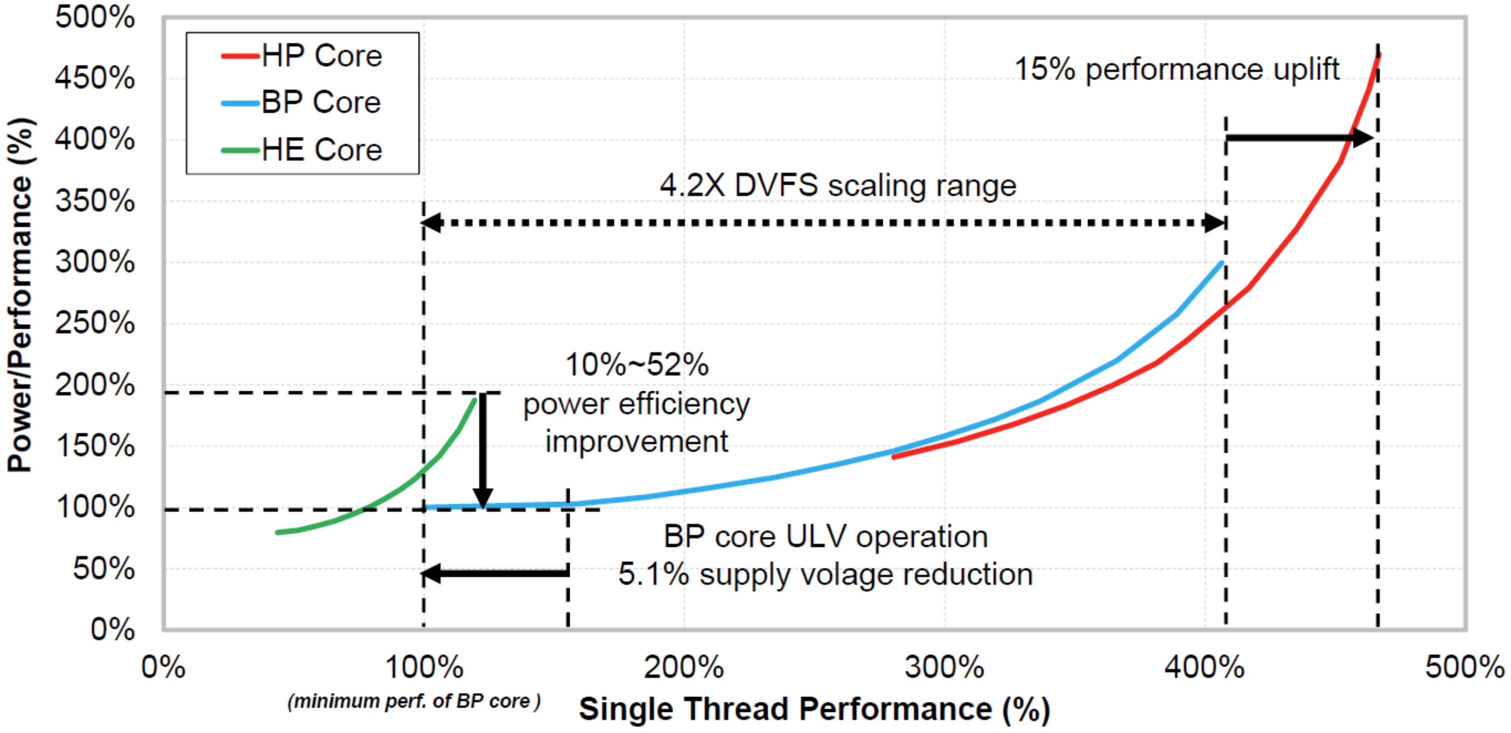
	D1	D2	D3	D4
In-order vs out-of-order	in-order	in-order	out-of-order	out-of-order
Issue width	1-issue	2-issue	2-issue	4-issue
Cycle time (FO4)	27.5	16.9	17.2	16.3
Branch pred size (entries)	264	600	1024	870
BTB size (entries)	64	90	554	1024
I-cache size (KB)	21	32	32	32
D-cache size (KB)	8	11	14	42
Fetch latency	1.0	1.6	2.2	2.1
Decode/Rename latency	1.0	1.7	2.4	3.0
Retire latency	N/A	N/A	2.0	2.2
Integer ALU latency	1.0	1.0	1.0	1.0
FP ALU latency	3.0	4.0	3.9	4.1
L1 D-cache latency	1.0	1.1	1.1	1.1
ROB size	N/A	N/A	22	32
IW size	N/A	N/A	11	9
LSQ size	N/A	N/A	16	16



Azizi, ISCA'10

Architectural Tradeoffs: Tri-Gear

- HP: High performance (ARM Cortex A78, optimized for speed, 3.0GHz)
- BP: Balanced performance (ARM Cortex A78, optimized for power, 2.6GHz))
- HE: High efficiency (ARM A55, 2.0GHz)



Summary

- SRAM alternatives
  - 8-T SRAM
  - eDRAM
  - Crosspoint arrays (e.g. RRAM)
- Understanding energy and power
- Power-performance tradeoffs

Next Lecture

- Low-power design
  - Lowering supplies
  - Parallelism and Pipelining