

Methodology for Energy-Efficient Design of Digital Circuits

Vojin G. Oklobdzija,

Advanced Computer Systems Engineering Laboratory
University of California / University of Texas
TxACE: Center of Excellence for Analog Circuits

<http://www.acsel-lab.com>

IEEE SSCS, Distinguished Lecture
Santa Clara Valley Chapter
April 15, 2010



*with acknowledgment to contributions of my former students:
Dr. Hoan Dao, AMD, Austin, TX and
Dr. Bart Zeydel, Plato Networks

Summary of the Presentation

Energy Efficiency of Digital CMOS Circuits

- o Problems
- o Energy-Delay Relationship
- o Minimizing Energy for a given delay
- o Methodology
- o Determining the best structures for high-performance system
- o Implications on the architecture

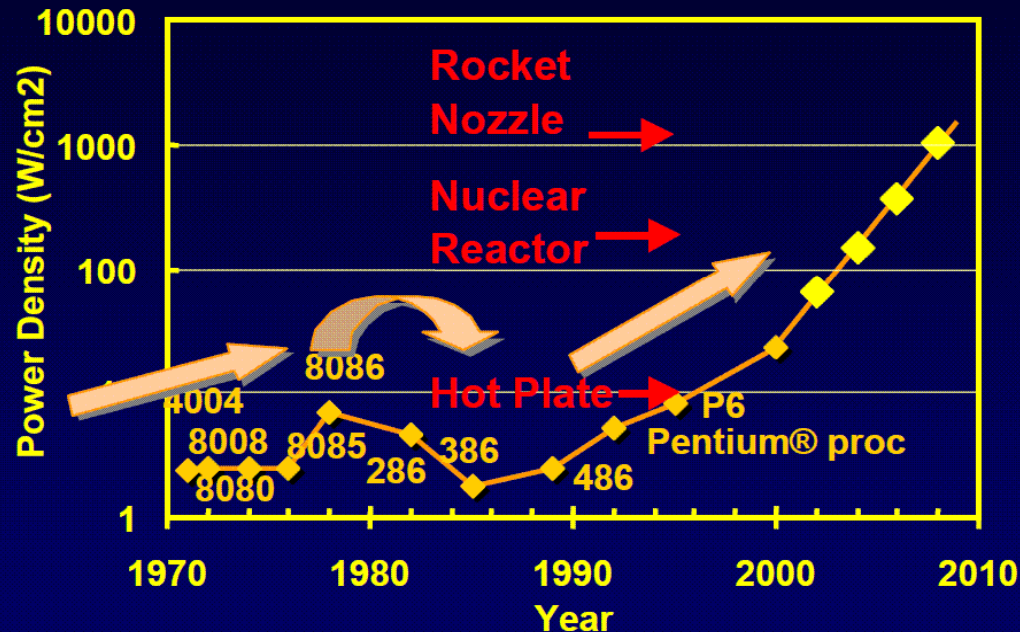
Challenges in High-Performance Design

- ❑ Optimizing for **power** - *not speed!* (or maximizing speed under the power budget)
- ❑ Logical Effort (LE) optimizes for speed, regardless of power i.e. brings us in the worse energy spot.
- ❑ Our method optimizes for:
power @ given speed or ***speed @ given power***
- ❑ We are developing new approaches for power efficiency (overlooked by delay optimization) applicable to:
 - Circuit structures
 - Design techniques
 - Energy-Delay Space
 - Creation of optimal Standard cell (ASIC) libraries

Motivation for Energy Efficient Design

Shekhar Borkar

Power density will increase

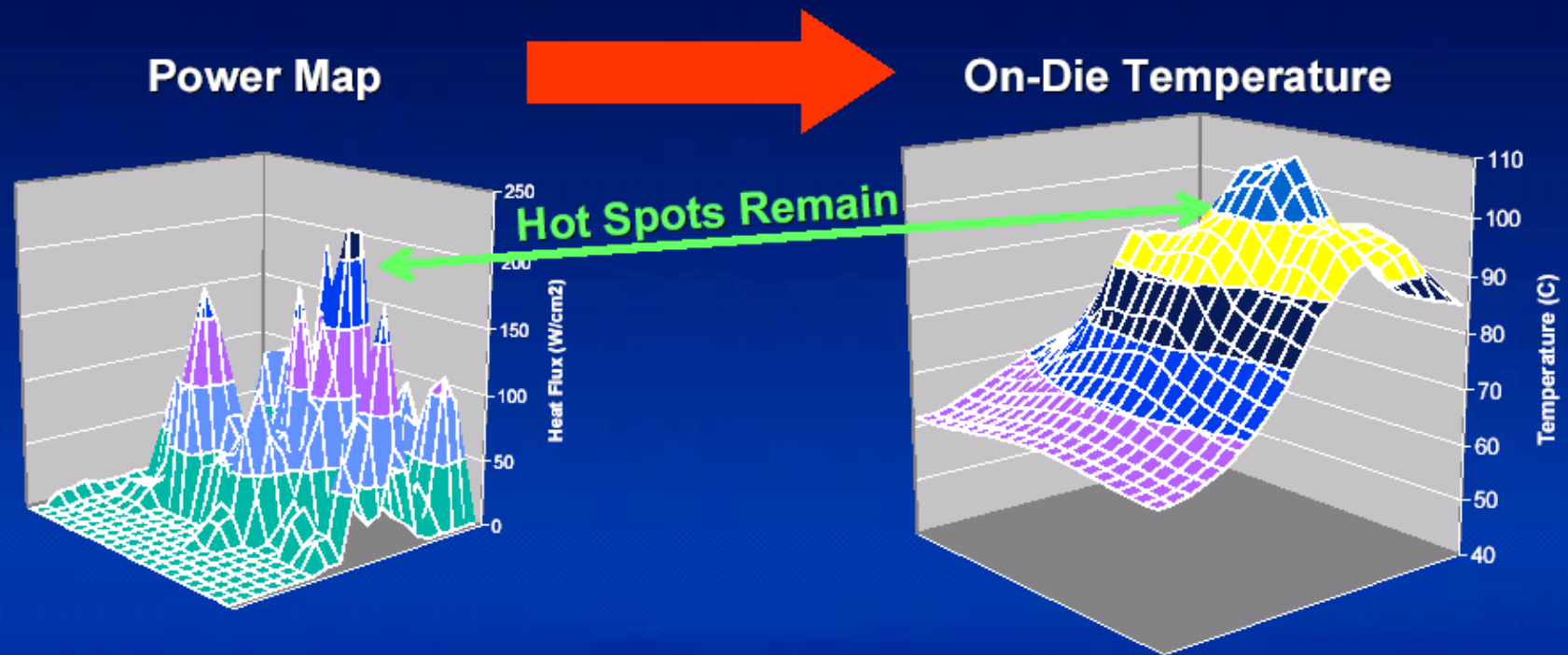


Power density too high to keep junctions at low temp

intel.

- Power density passed the level found in the Nuclear Reactor !
- Power density degrades the reliability and speed.

Power Density: The Future

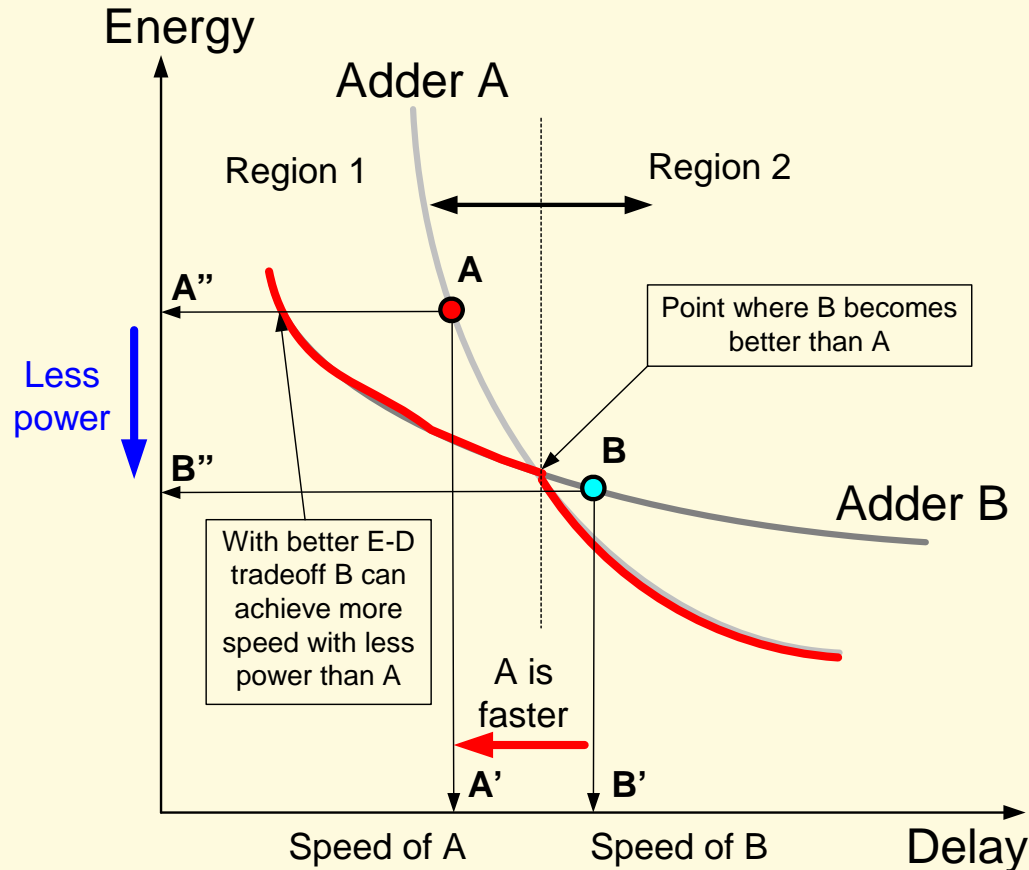


- **With high power density, cannot assume uniformity**
 - As die temperature increases, CMOS logic slows down
 - At high die temp., long-term reliability can be compromised

Energy-Delay Relationship

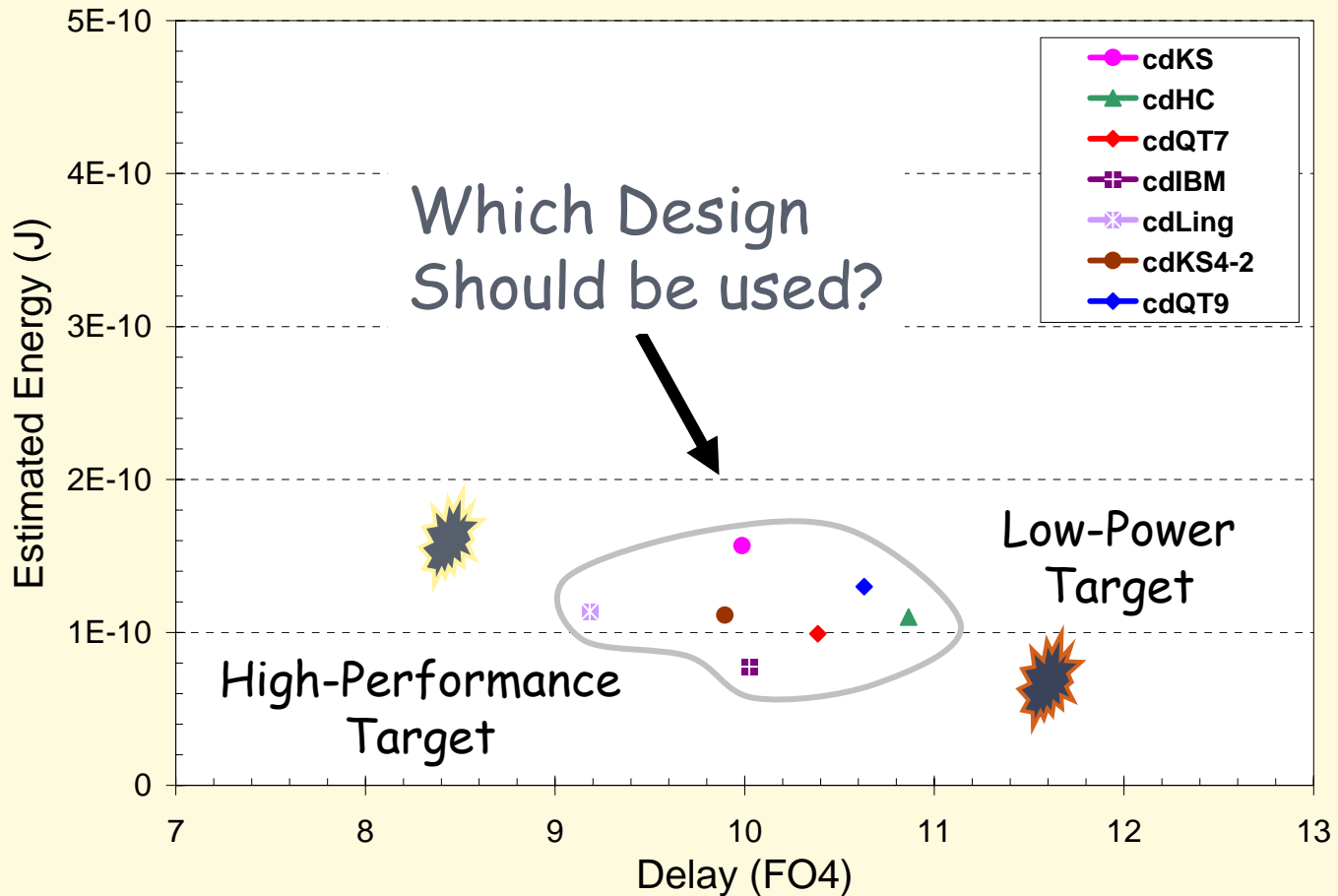


Energy-Delay Space View



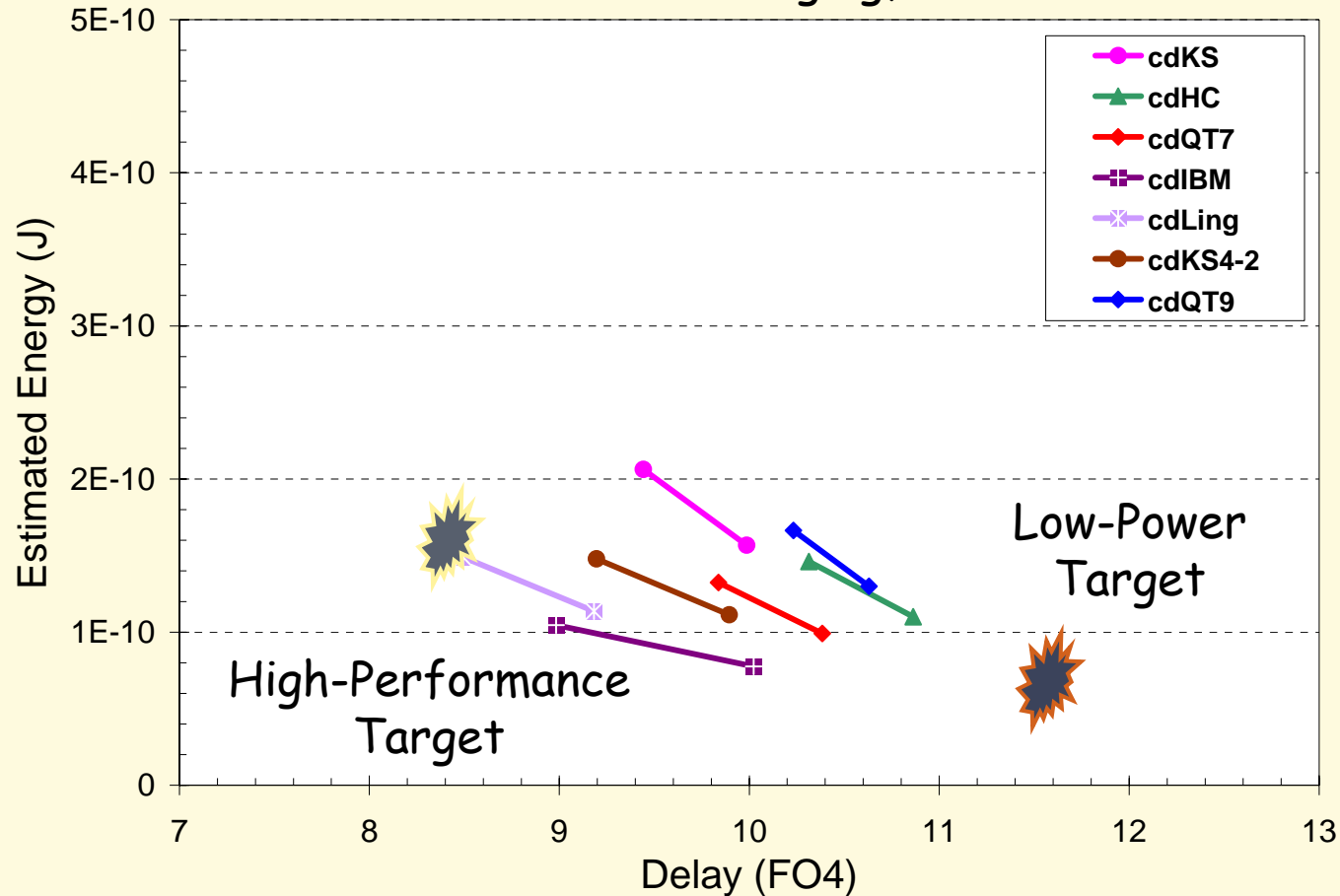
- Must look at Energy-Delay Space of designs

Energy-Delay Space View



Energy-Delay Space View

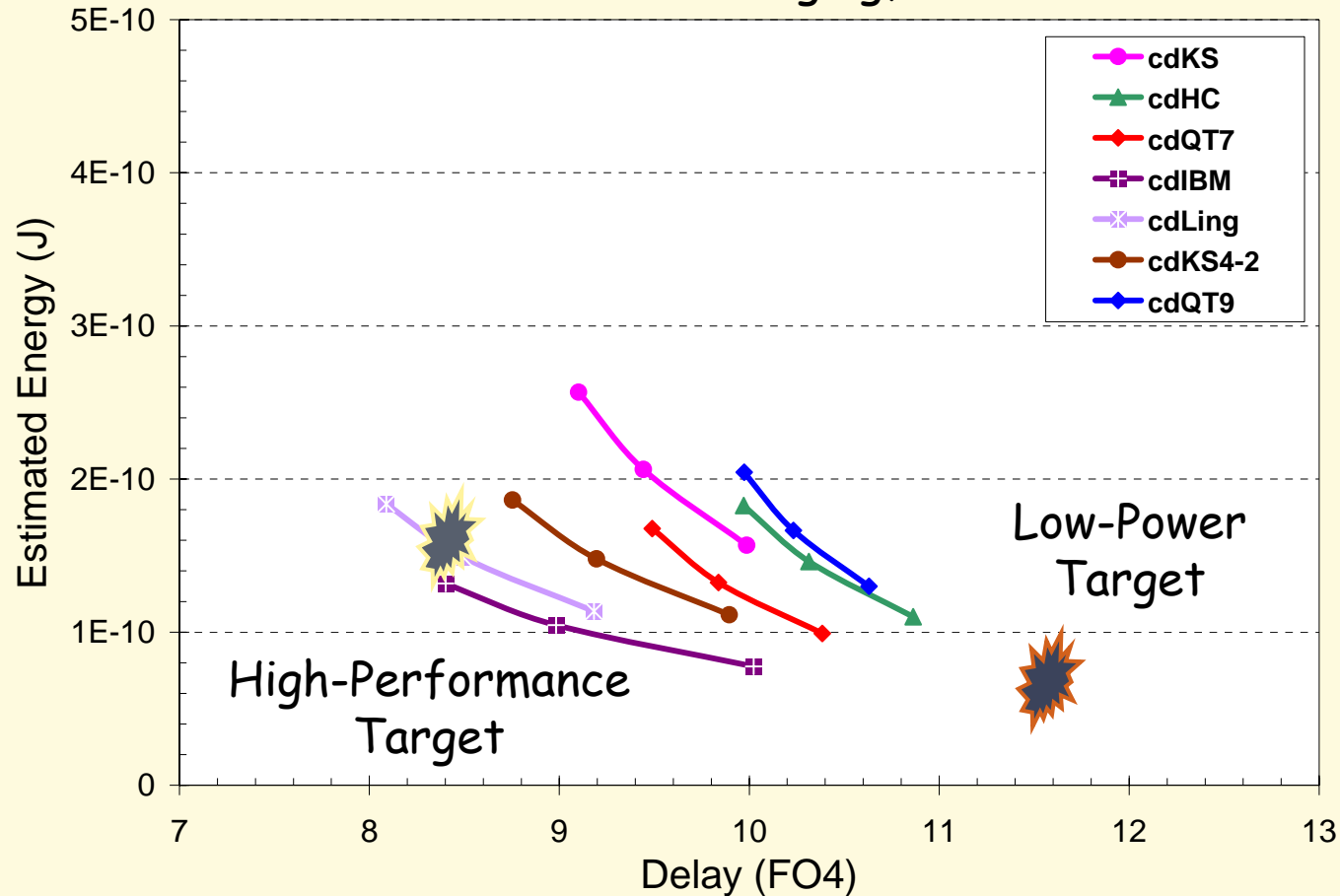
H is changing, w/ C_{out} =constant



- Begin to see characteristics of designs

Energy-Delay Space View

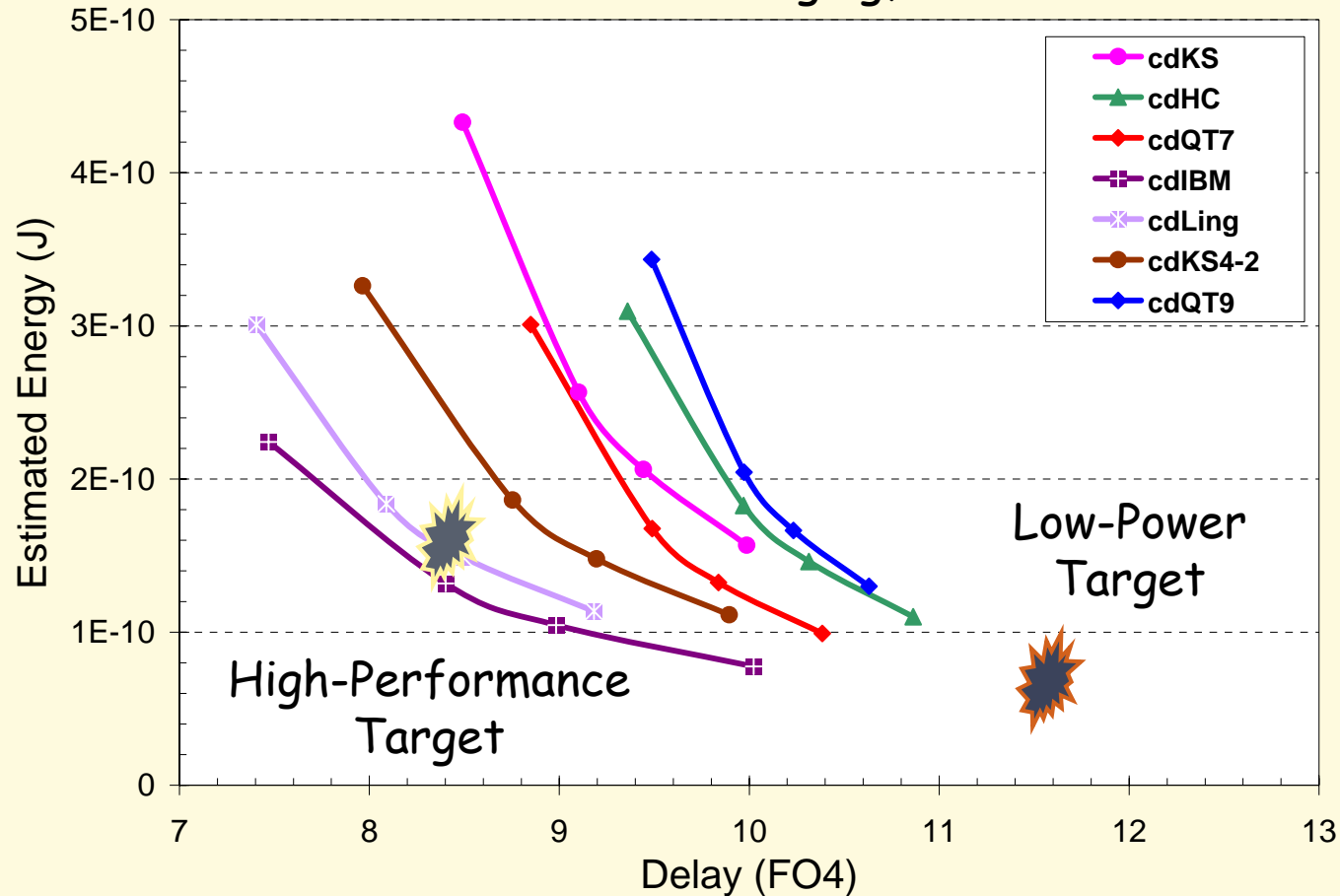
H is changing, w/ C_{out} =constant



- Begin to see characteristics of designs

Energy-Delay Space View

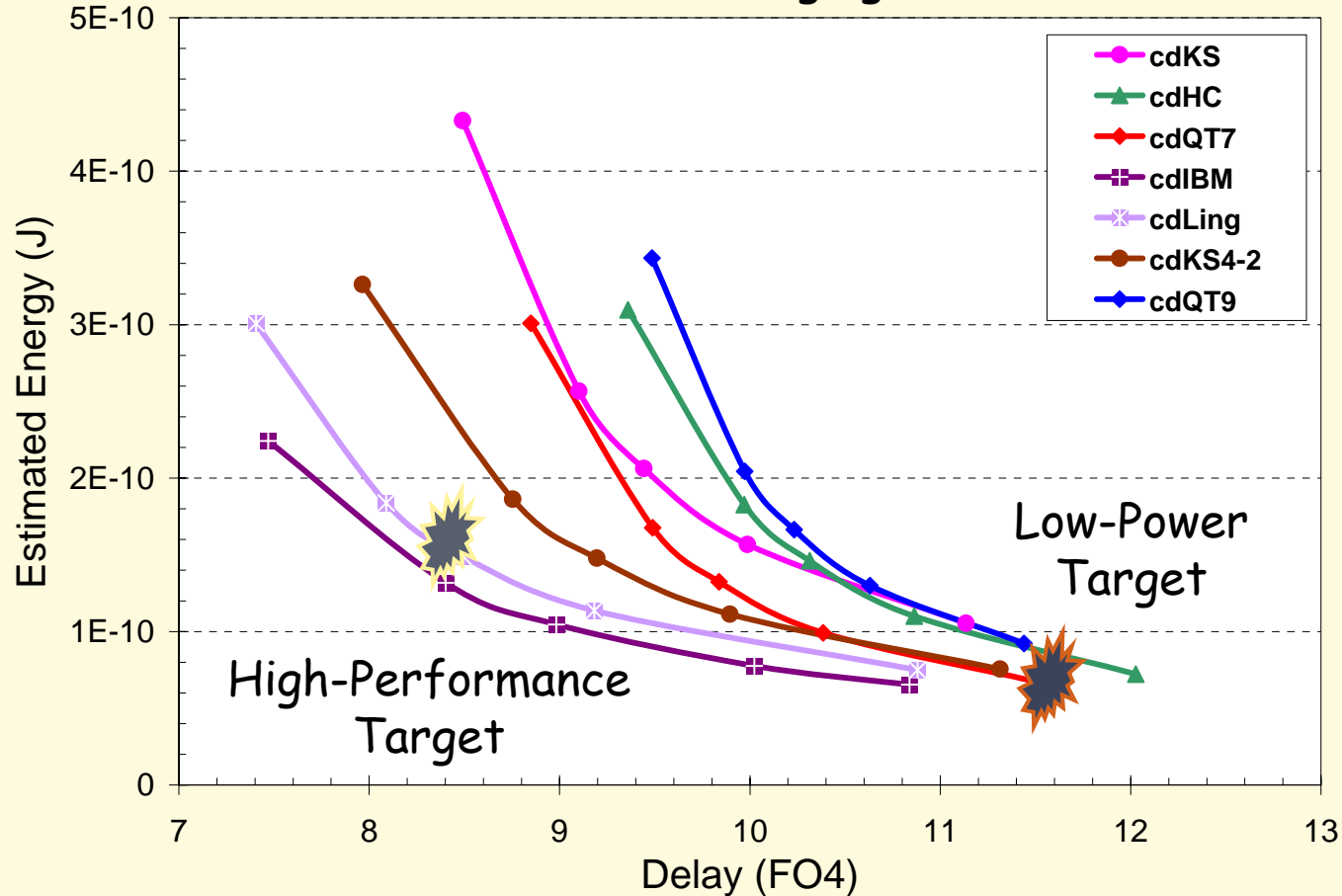
H is changing, w/ C_{out} =constant



- Best High-Performance designs are clearly seen
- Different than what would be chosen from single point

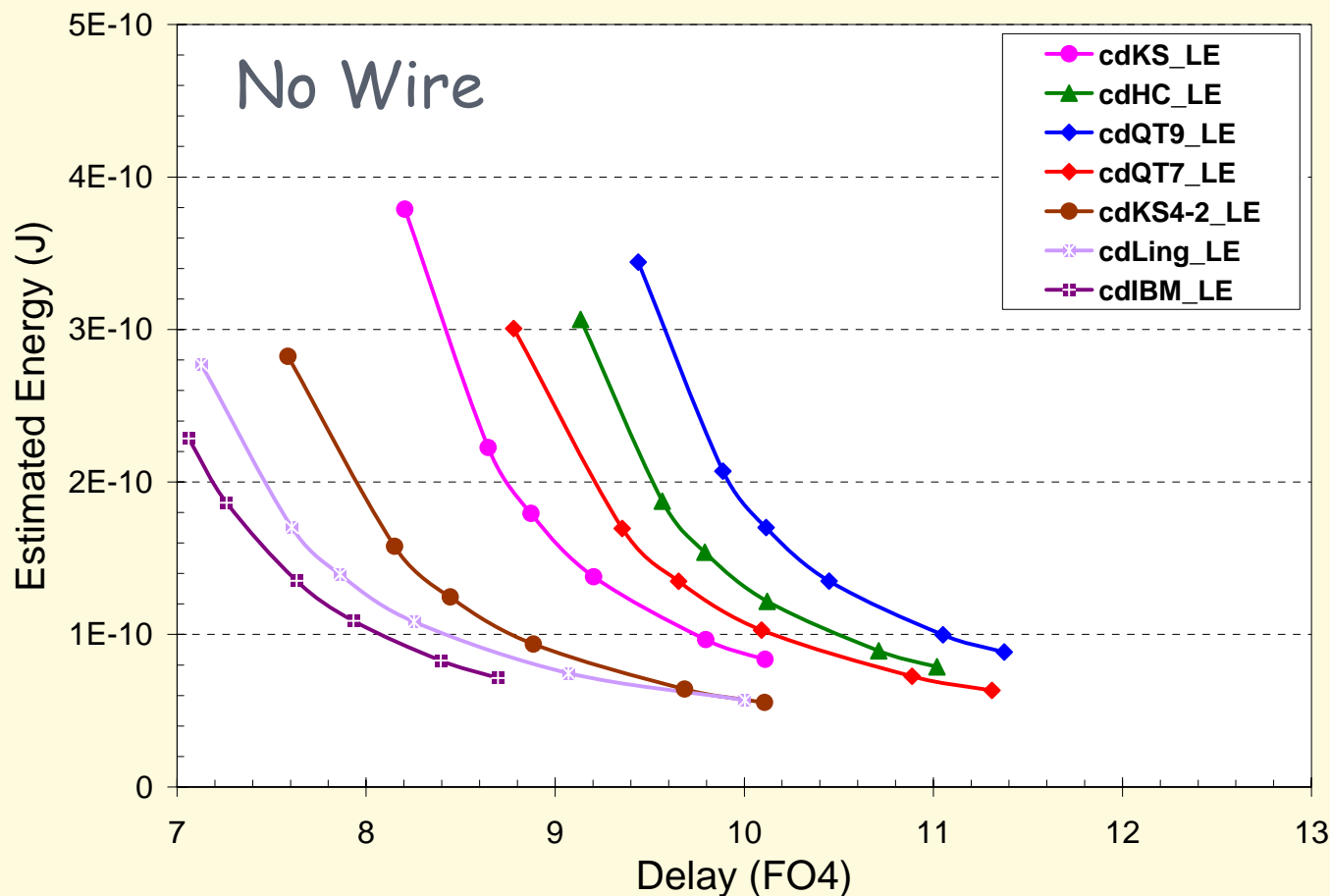
Energy-Delay Space View

H is changing, w/ C_{out} =constant



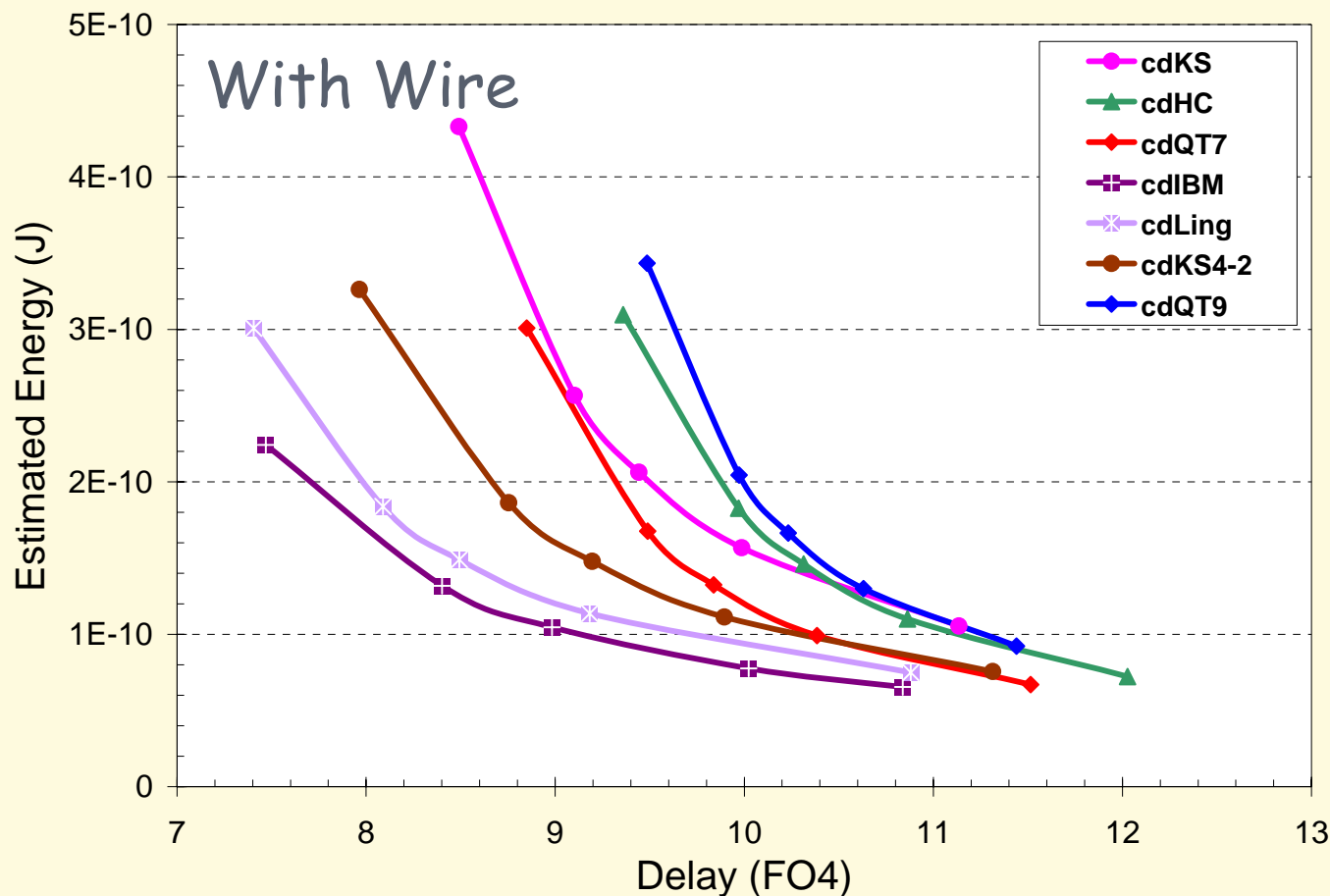
- Also determines best design for Low-Power Target

Contribution of Wire to Delay and Energy should be examined too



- Without wire, differences appear large

Contribution of Wire to Delay and Energy should be examined too

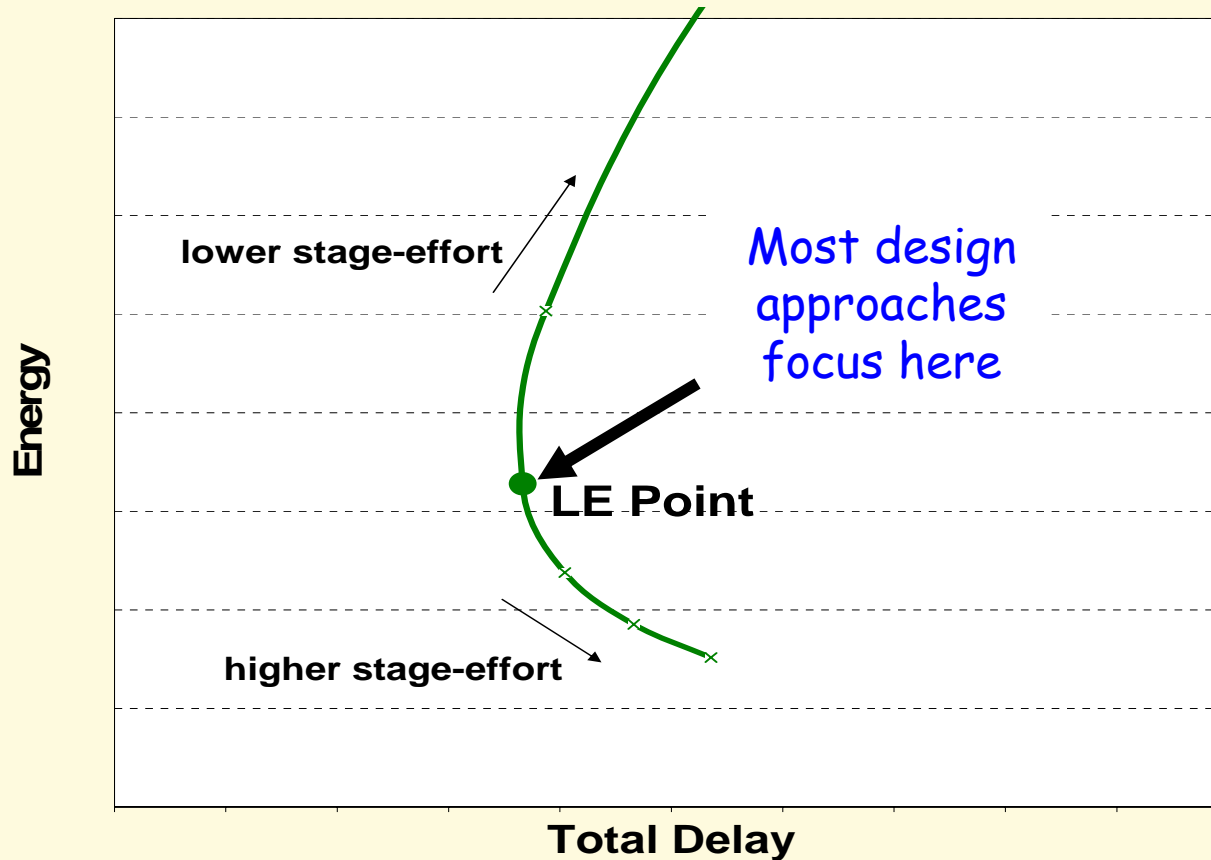


- Wire strongly impacts selection of "best adders"

Energy

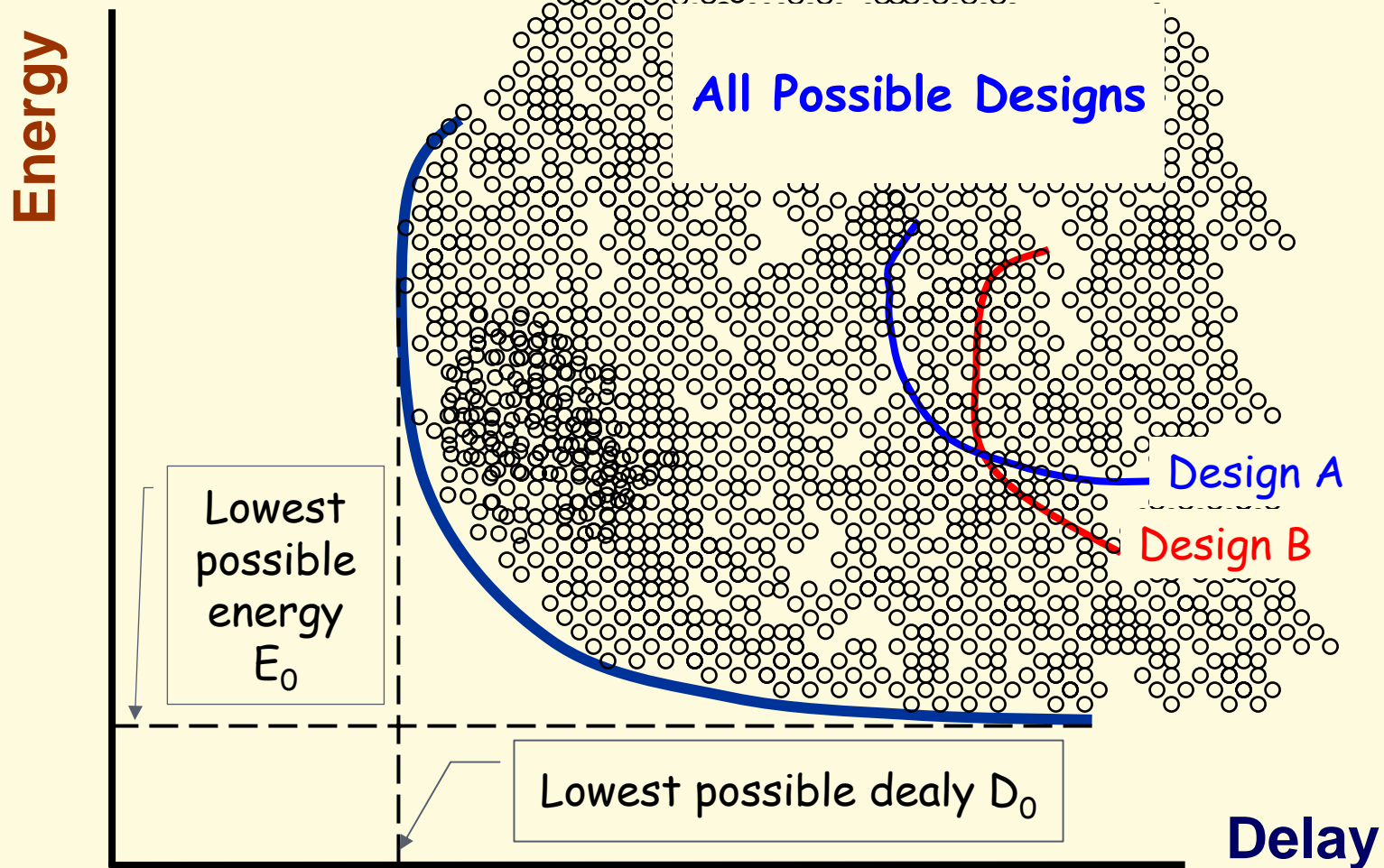


Where does Logical Effort lead us?



- It is possible to lower energy by trading delay? or ...

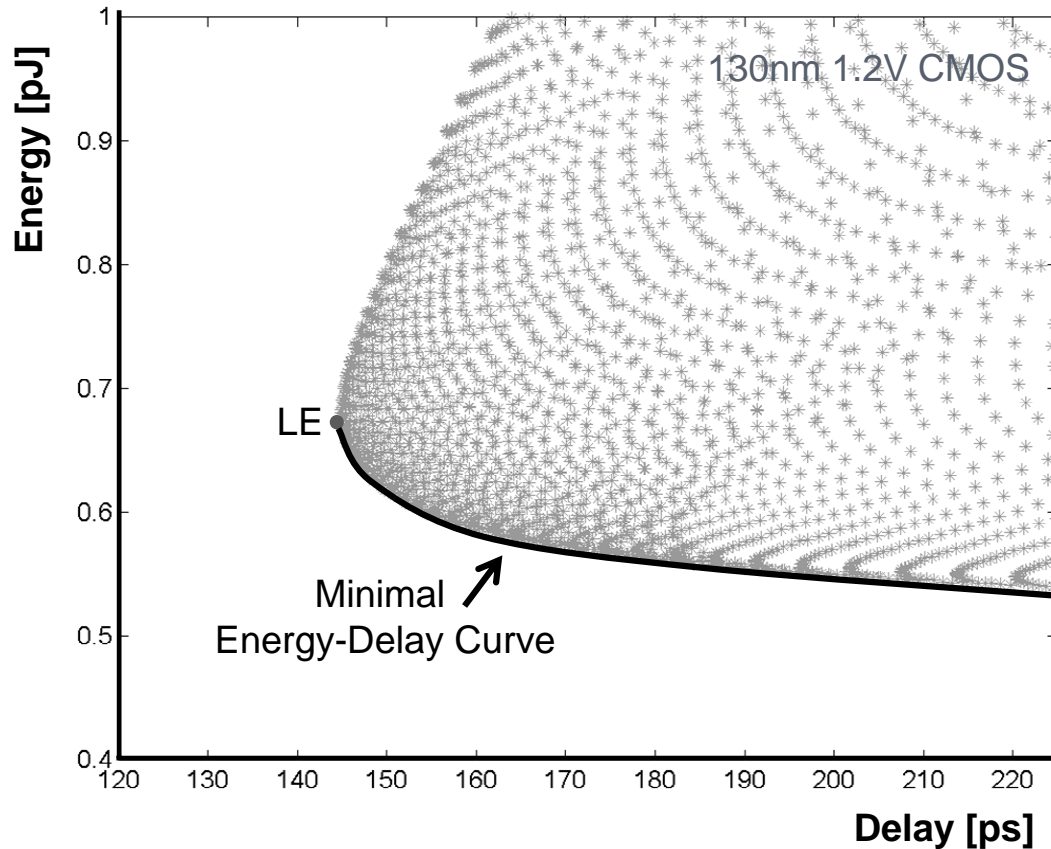
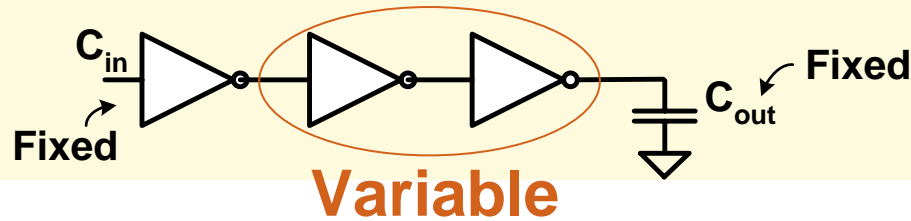
Design in Energy-Delay Space



$$(E-E_0)(D-D_0)=0.2 \times E_0 D_0$$

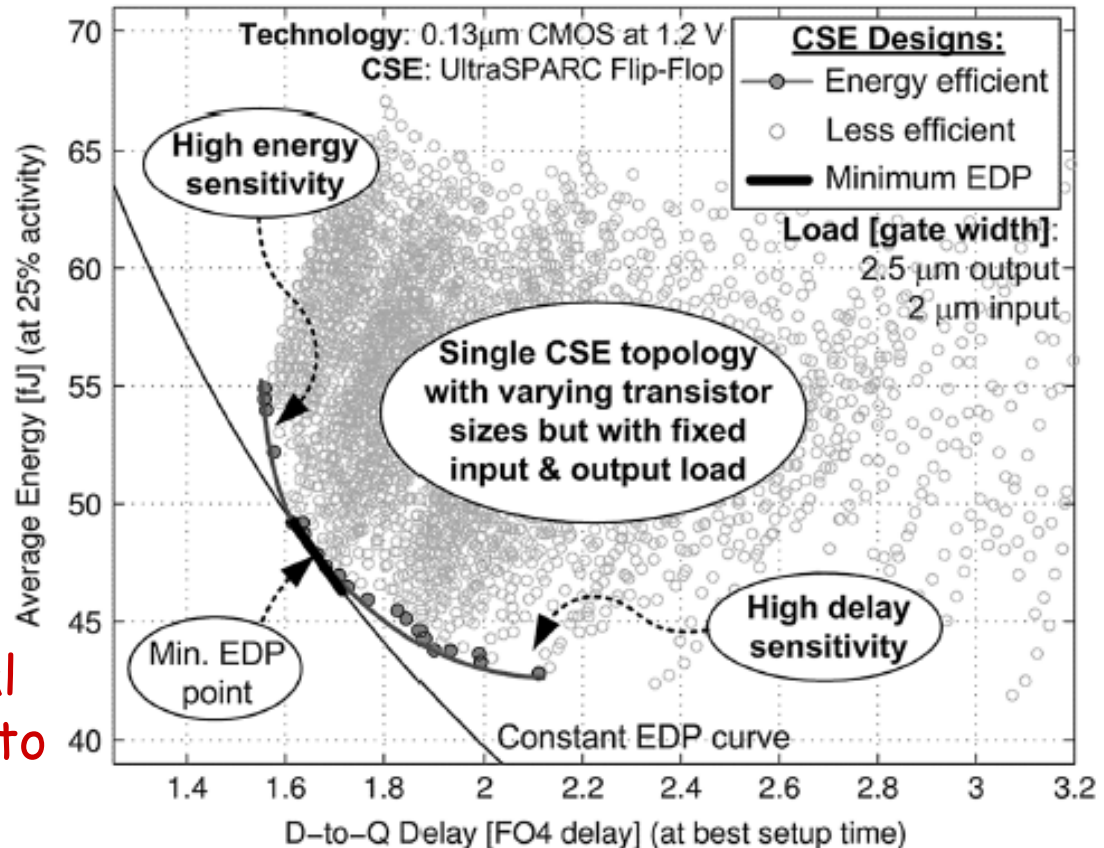
**P. Penzes, Caltech, PhD 2002, V. Zyuban, ISLPED 2002*

Energy-Delay Space of a Circuit (Fixed Input Size and Output Load)



Exhaustive search

A circuit with 10 transistors and 10 possible size for each transistor requires to check 10^{10} possible solutions!



To obtain the minimal E-D curve do I have to check all possible solutions?

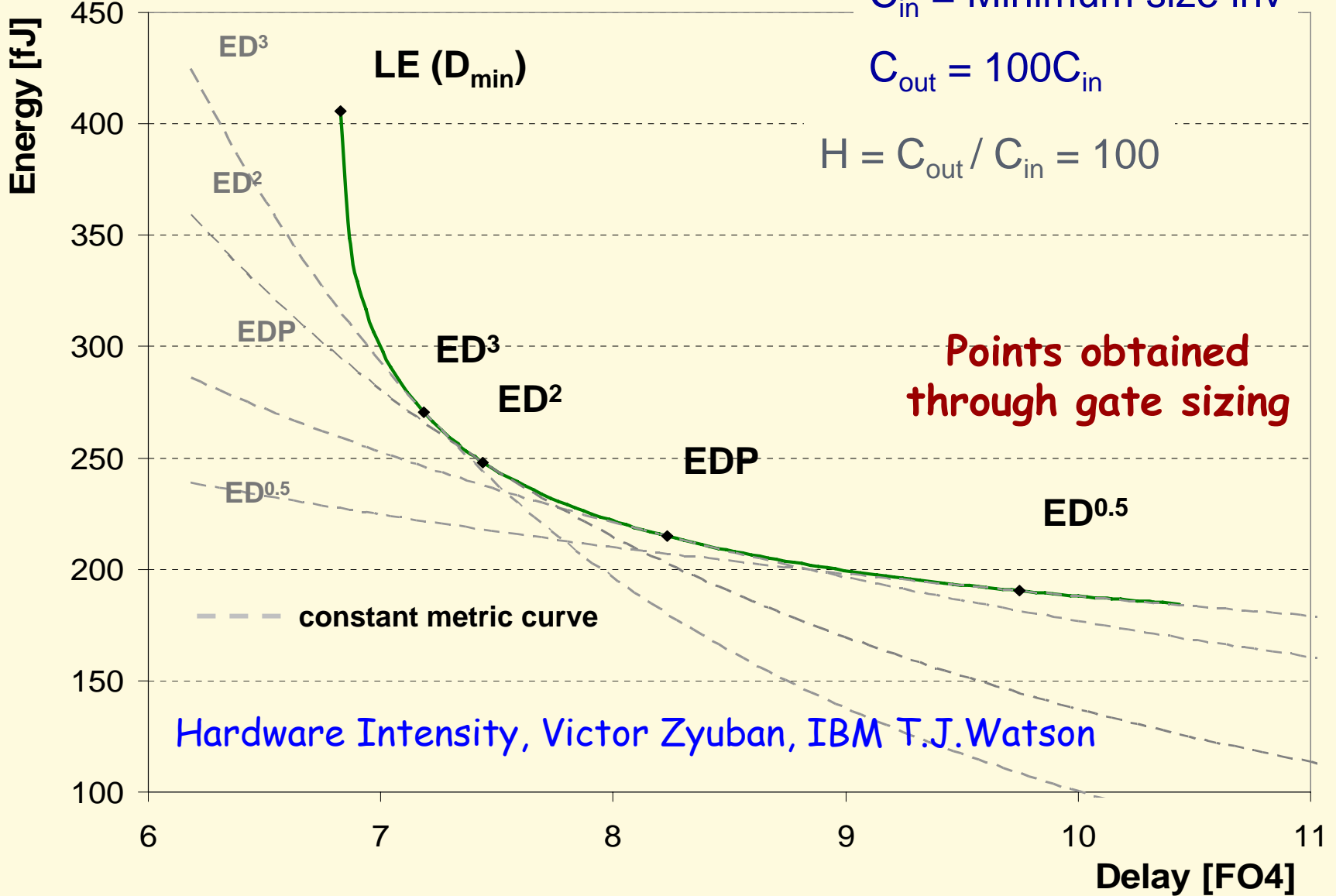
C. Giacomotto, N. Nedovic, and V. G. Oklobdzija, "The Effect of the System Specification on the Optimal Selection of Clocked Storage Elements", IEEE JoSSC, vol. 42, no. 6, June 2007.

Hardware Intensity: V. Zyuban, IBM

C_{in} = Minimum size inv

$C_{out} = 100C_{in}$

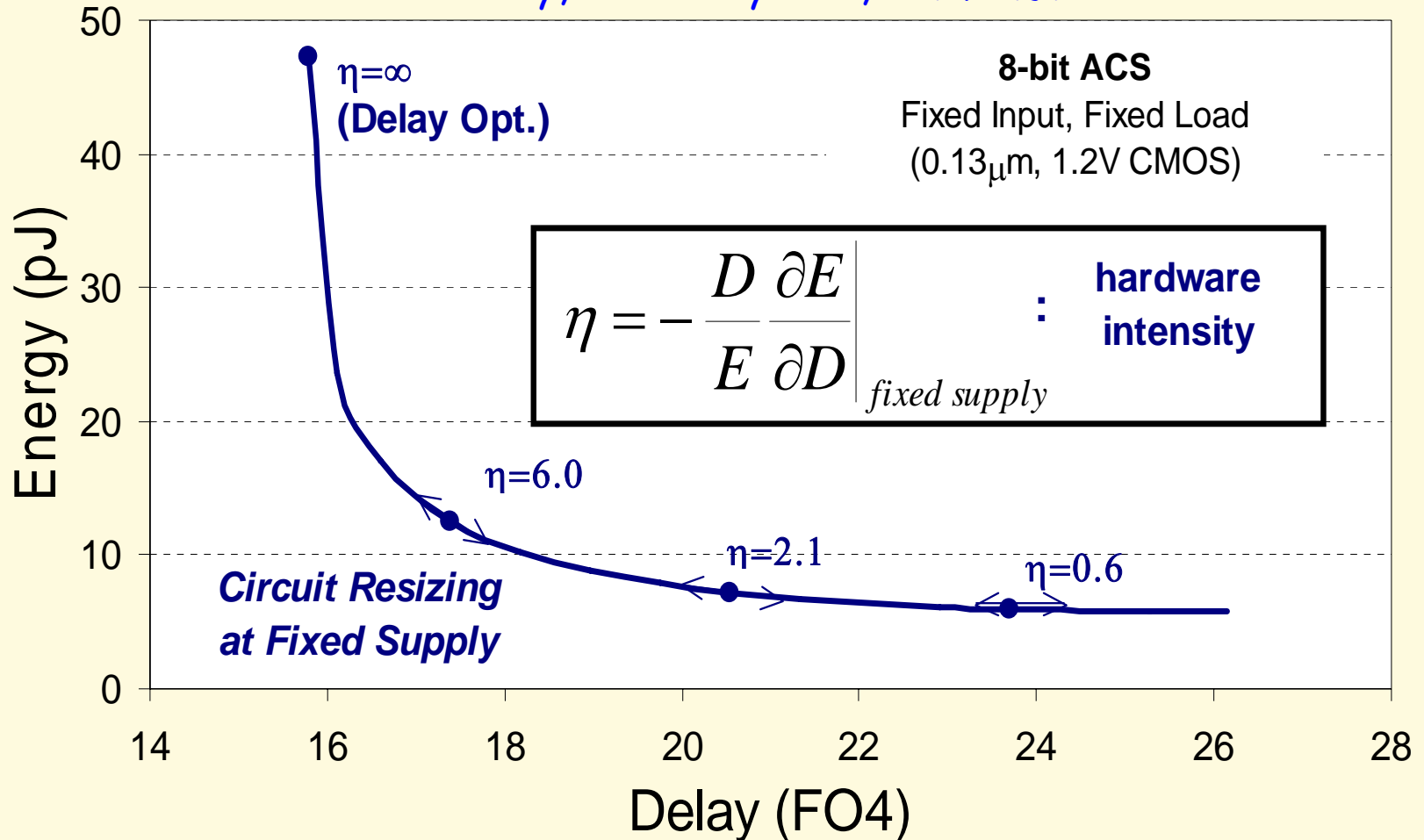
$H = C_{out} / C_{in} = 100$



Hardware Intensity, Victor Zyuban, IBM T.J.Watson

Design Objective at Nominal V_{DD}

Hardware Intensity, Victor Zyuban, IBM T.J.Watson

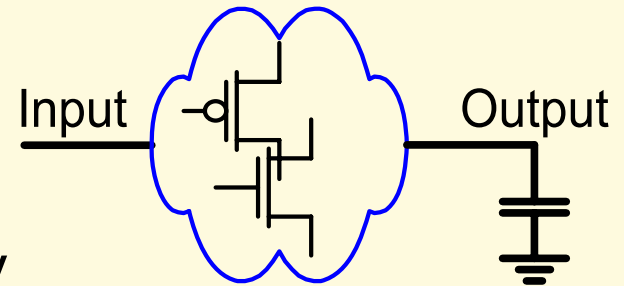


- Design choice depends on (E,D) requirements

Prior Work on Design Optimization

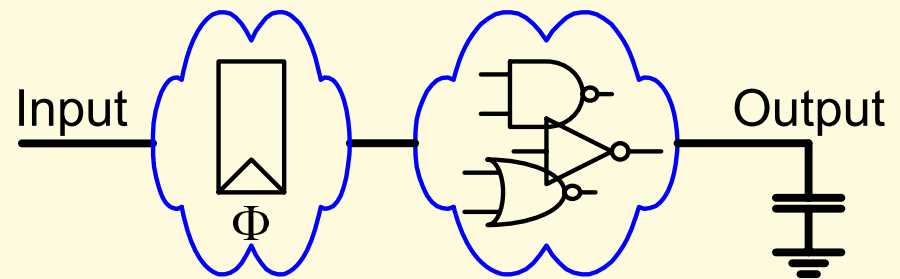
- **Transistor-based** [TILOS]

- Sizing individual transistors
- Growing complexity
- Applicable to small blocks only



- **Block-based** [Zuyban & Strenski]

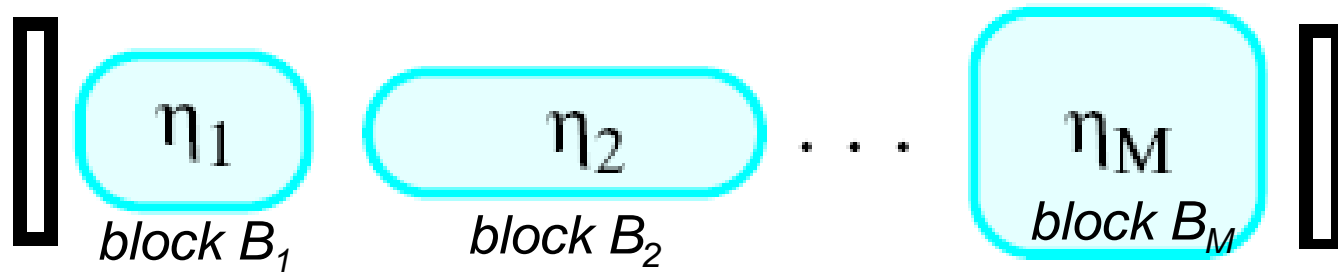
- Blocks: latch & logic
- Trading {energy, delay} of blocks
- CAD tools
- Fixed interface



Transistor-Based Approach

- Optimization problem: ($i=1..M$)
 - Minimize: $\text{Area}(W_1, \dots, W_M) \approx \sum W_i$
 - Constraint: $D_{\text{worst}}(W_1, \dots, W_M) = T$
- Delay modeling
 - Linear (RC-like): TILOS
 - Look-up table: AMPS (Synopsys)
- A convex problem \Rightarrow minimal solution exists
 - Different polynomial algorithms developed
 - May have issues with convergence
 - Long run time with increasing design complexity

Block-Based: Zyuban (IBM)



[Zyuban, IBM T.J. Watson Research]

- Optimization problem for a pipelined stage:

- Minimize energy: $E(B_1, \dots, B_M) = \sum E_{B_i}$

- Delay constraint: $D(B_1, \dots, B_M) = \sum D_{B_i} = T$

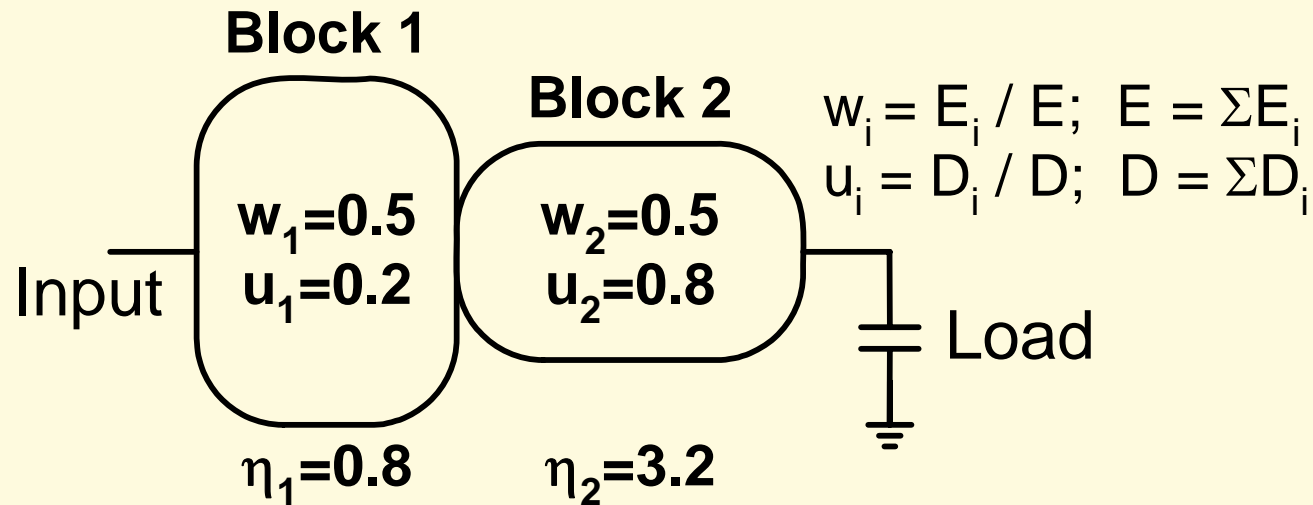
- Optimal criteria:

- $(w_i/u_i) \cdot \eta_i = (w_k/u_k) \cdot \eta_k$ for any (i,k)

with $w_x = E_{B_x}/E$, $u_x = D_{B_x}/T$, $\eta_x = (\partial E_{B_x}/E_{B_x})/(\partial D_{B_x}/D_{B_x})$

- Similar criteria for a simple pipelined system

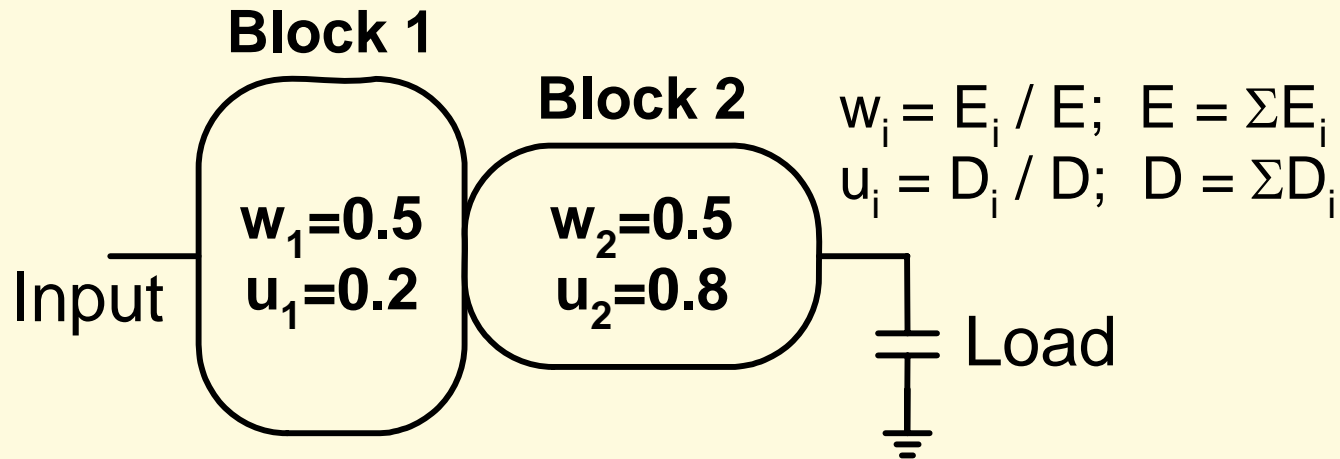
Application: Solution Verification



[Zyuban, IBM T.J. Watson Research]

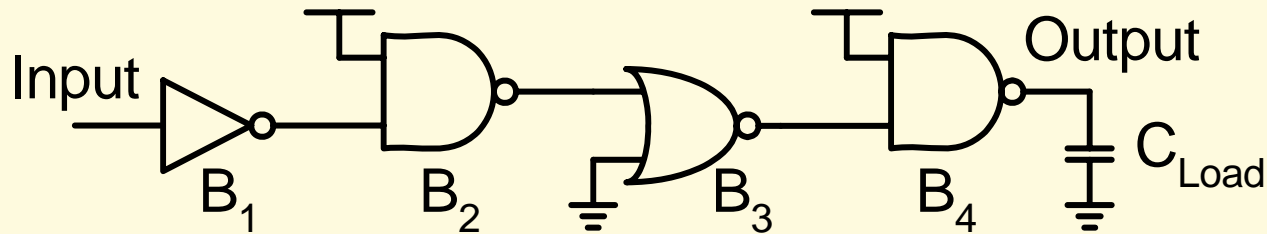
- Verify optimality of solution:
 - Block 1: $(w_1/u_1) \cdot \eta_1 = 2.0$
 - Block 2: $(w_2/u_2) \cdot \eta_2 = 2.0$ } *Equal \Rightarrow optimal!*

Application: Solution Verification



- If $\eta_1 = 3.2$ and $\eta_2 = 0.8$
 - Block 1: $(w_1/u_1) \cdot \eta_1 = 8.0$
 - Block 2: $(w_2/u_2) \cdot \eta_2 = 0.5$ } *Unequal \Rightarrow not optimal*
- Better solution? Relax η_1 and increase η_2

Major Limitation

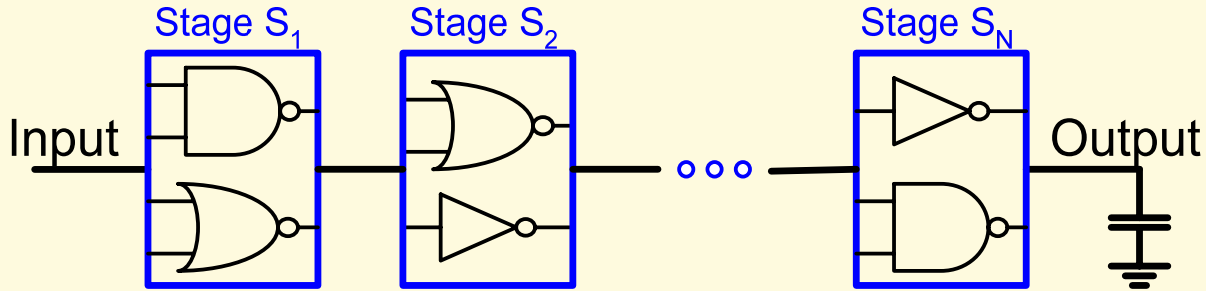


- Zyuban's assumption:
 - Delay & energy independence of each block B_i
- Single path: block \equiv gate

$$\left\{ \begin{array}{l} \text{Delay } T_d \propto \{ C_{out}, C_{in} \} \\ \text{Energy } E \propto \{ C_{out}, C_{in} \} \\ C_{in} : \text{current gate cap} \\ C_{out} : \text{next gate cap} \end{array} \right. : \text{energy, delay dependency of adjacent gates}$$

- Similar dependency between blocks and pipelines
- No analytical solution if accounting dependency

Proposed Stage-Based Approach



- Stage \approx logic depth
- Gates \rightarrow stage
 - Based on maximal distances to input and output
 - Stage delay: $d_{stage} = \max\{d_{gate}\}$
 - Stage energy: $E_{stage} = \sum E_{gate}$
 - Estimated from gate energy & delay models

Delay and Energy Modeling of Gates

- Logical Effort delay model:

reference delay τ
(technology dependent)

$$T_{d,gate} = \left[\frac{\kappa R_{gate} C_{gate}}{\kappa R_{inv} C_{inv}} \circ \frac{C_{out}}{C_{gate}} + \frac{\kappa R_{gate} C_{par}}{\kappa R_{inv} C_{inv}} \right] (\kappa R_{inv} C_{inv}) = (gh + p)\tau$$

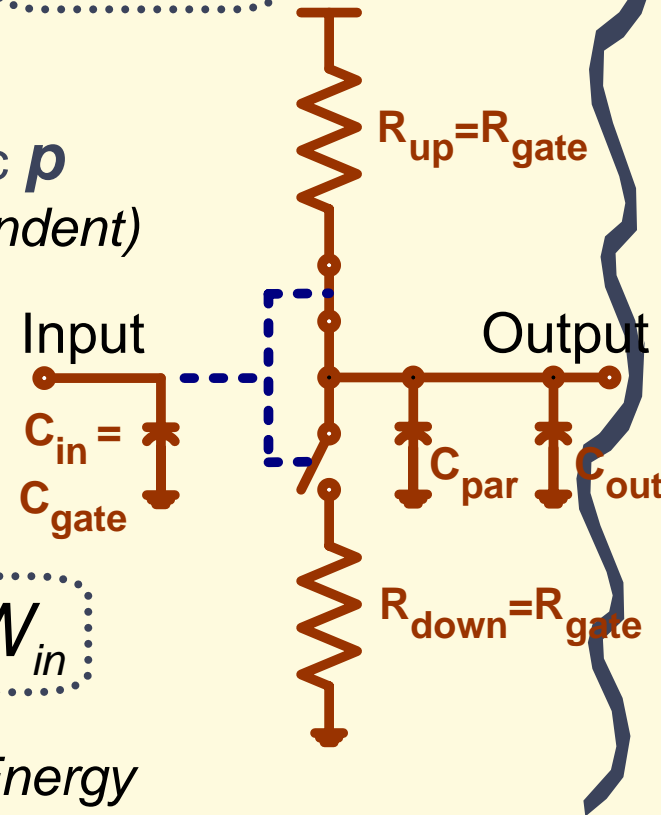
logical effort **g** (gate dependent) electrical effort **h** (load dependent) parasitic **p** (gate dependent)

stage effort **f = g•h**

- Energy model:

$$E_{gate} (W) = E_g W_{out} + E_p W_{in} + P_l T_{CK} W_{in}$$

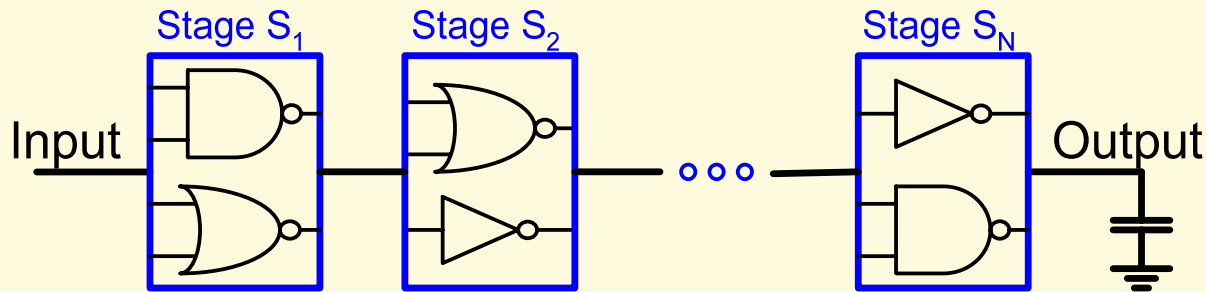
Active Energy Leakage Energy



Pipelined Stage Optimization

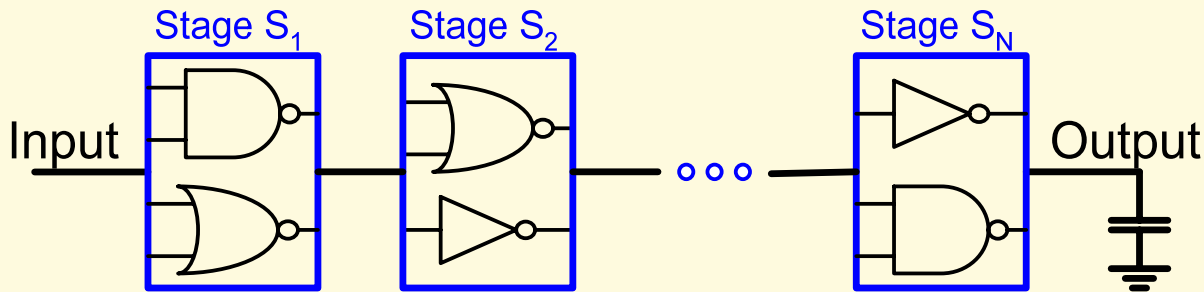


Stage-Based Optimization



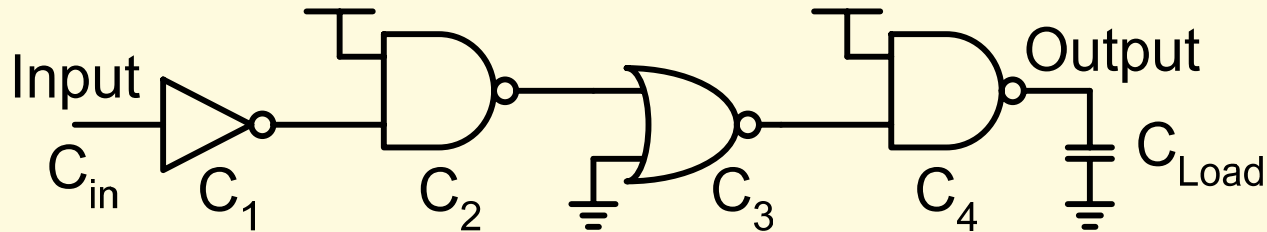
- Optimization functions
 - Delay: $D = \sum D_{\text{Stage}(i)}$
 - Energy: $E = \sum E_{\text{Stage}(i)}$
- Possible design constraints
 - Delay target, D
 - Input size, W_{input}
 - Output load, C_{load}
- Posynomial Problems
 - Solvable with polynomial algorithms

Problem A: Delay Optimization



- Optimization problem
 - Minimize: $D = \sum D_{S_i}$
 - Constraint: $\{\text{Input, Load}\} = \text{const.}$
- Objectives
 - Obtain minimally achievable delay, D_{\min}
 - Wanted in performance-critical designs
 - Disregard energy consumption (*actually*, $\partial E_i / \partial D_i = \infty$)

Single Path: Logical Effort



$$D = \sum_{i=1}^N \left(g_i \frac{C_{i+1}}{C_i} + p_i \right)$$

$C_i = \text{input cap of gate } i$

$C_{in} = C_1 : \text{fixed}$

$C_{N+1} = C_{Load} : \text{fixed}$

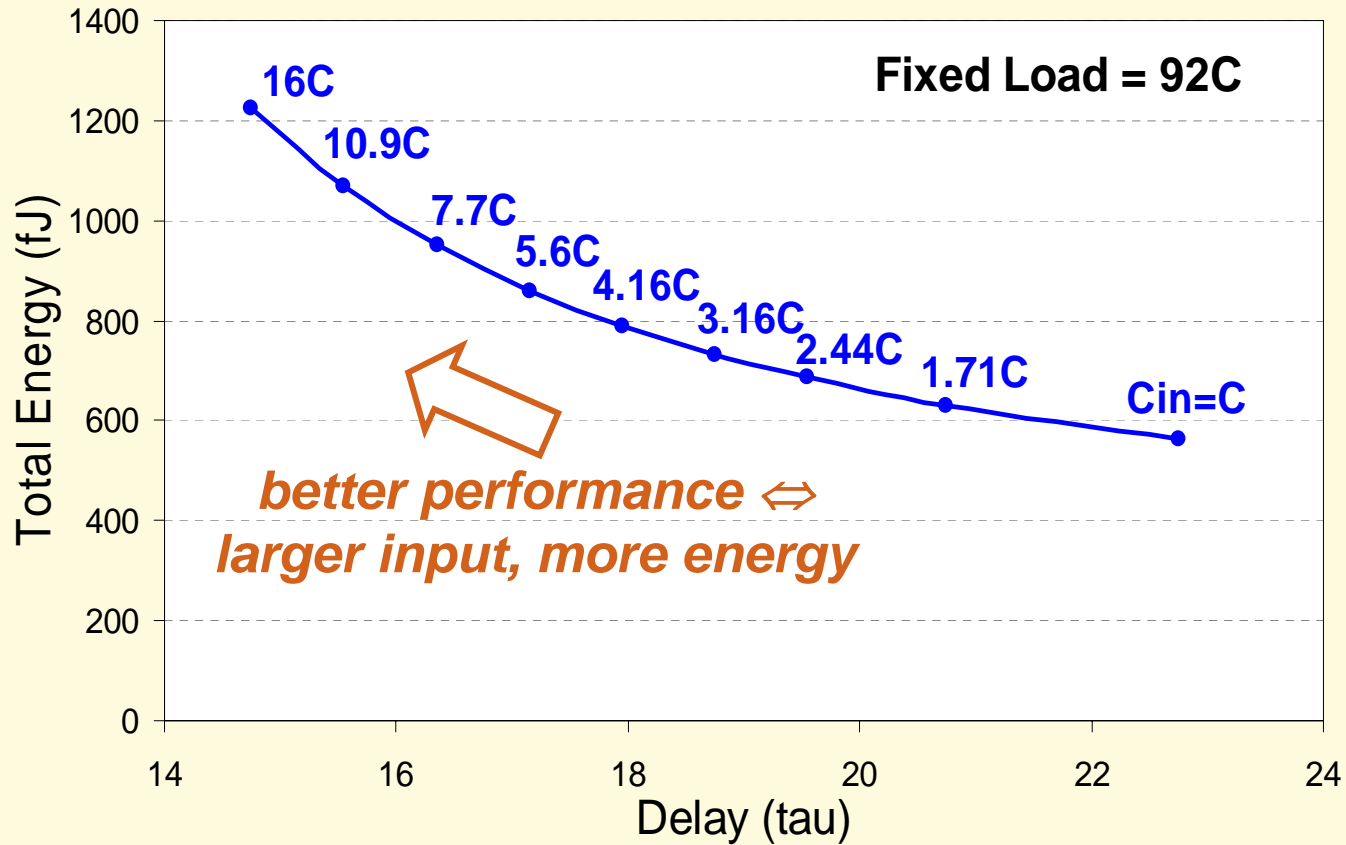
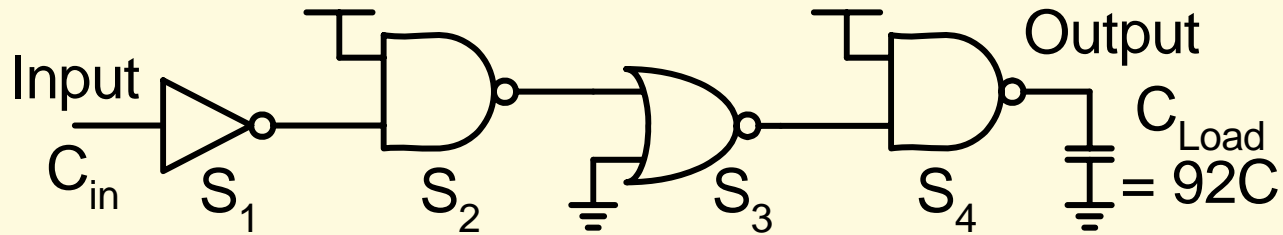
$$dD/dC_i = 0 \ (\forall i = 1..N): \quad f_i = f_{opt} = \left[\left(\prod_{i=1}^N g_i \right) \left(\frac{C_{Load}}{C_{in}} \right) \right]^{1/N}$$

$$D_{min} = N f_{opt} + \sum_{i=1}^N p_i$$

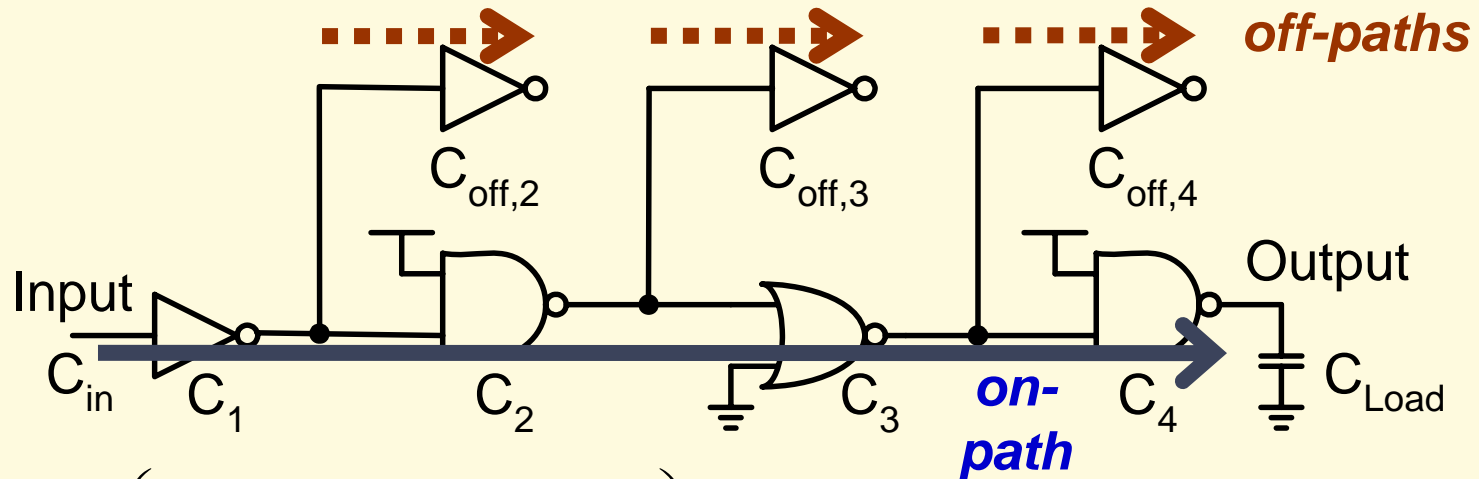
↑
Path Gain, H

- Solution = equal stage effort f (i.e. fan-out)

Energy Cost vs. Total Delay



Multi-Path Circuits



$$D = \sum_{i=1}^N \left(g_i \frac{C_{i+1} + C_{off,i+1}}{C_i} + p_i \right)$$

$$= \sum_{i=1}^N \left(g_i \frac{C_{i+1} + C_{off,i+1}}{C_{i+1}} \frac{C_{i+1}}{C_i} + p_i \right)$$

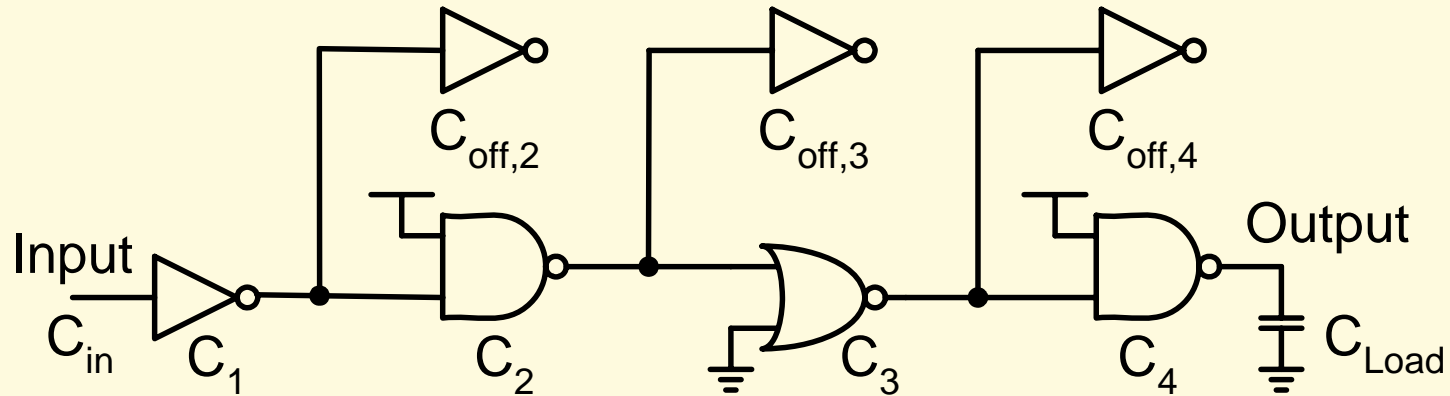
$b_{i+1} = \text{branching factor at } (i+1)^{\text{th}}\text{-gate input}$

$C_i = \text{input cap of gate } i$

$C_{off,i} = \text{input cap of } i^{\text{th}}\text{ off-path gate}$

- Optimal delay depends on off-path load

Linear Branching



- Linear branching: $C_{off,i} / C_i = \mathbf{const.}$

$$dD/dC_i = 0 (\forall i = 1..N): f_i = f_{opt} = \left[\left(\prod_{i=1}^N g_i \right) \left(\prod_{i=1}^N b_i \right) \left(\frac{C_{Load}}{C_{in}} \right) \right]^{1/N}$$

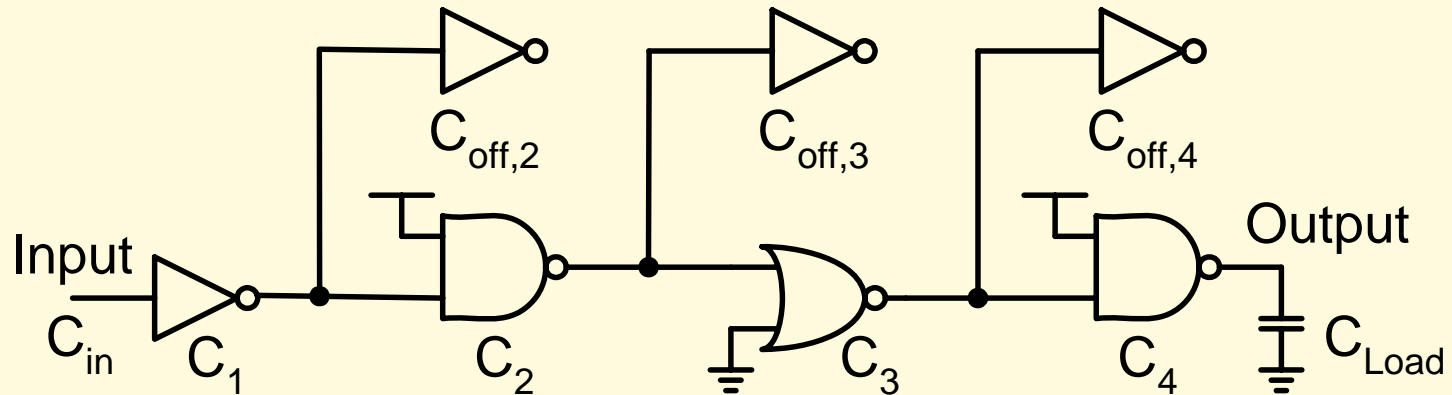
$$D_{min} = N f_{opt} + \sum_{i=1}^N p_i$$



Branching
Factor, B

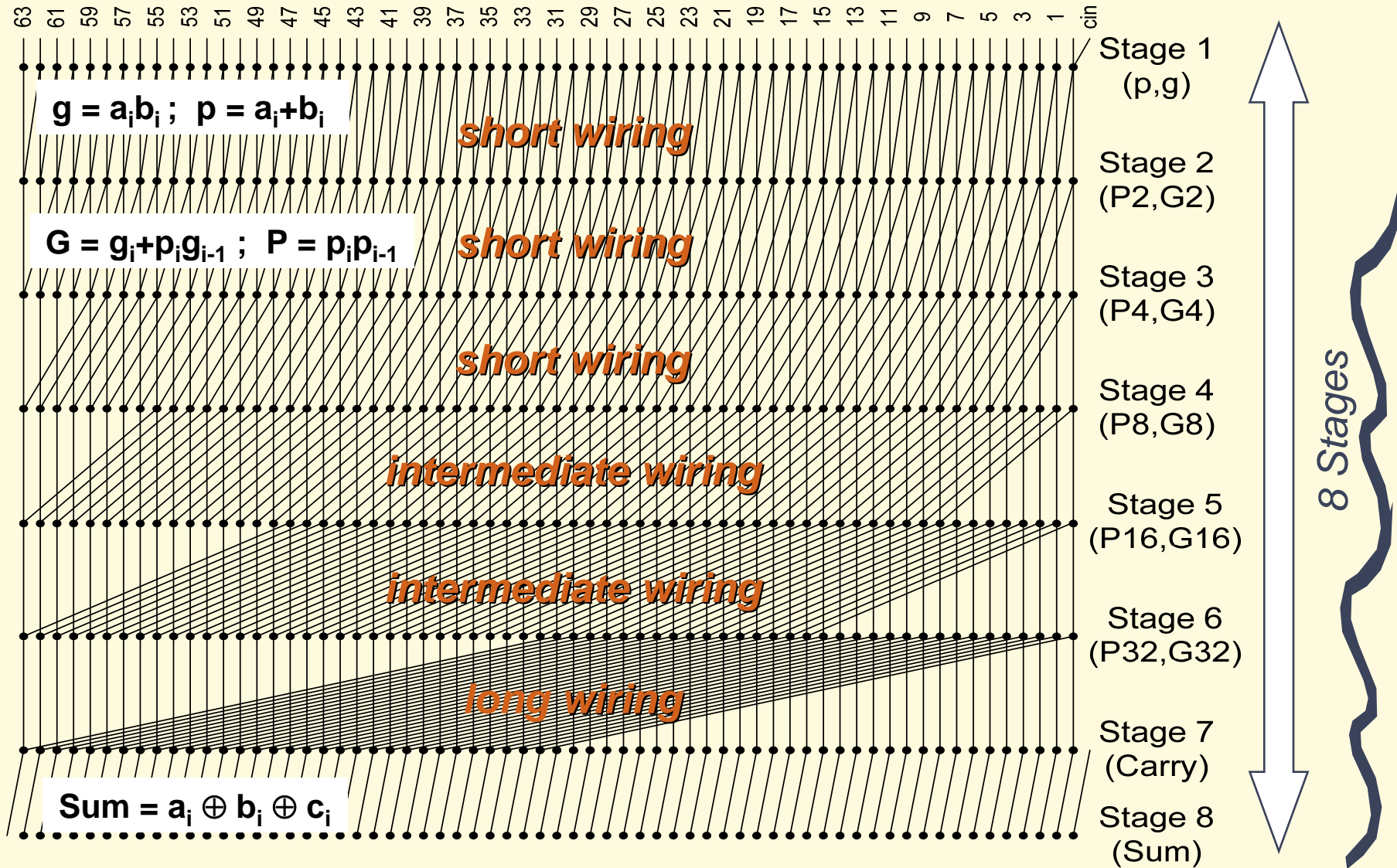
- Similar analytical form for solution

Non-Linear Branching

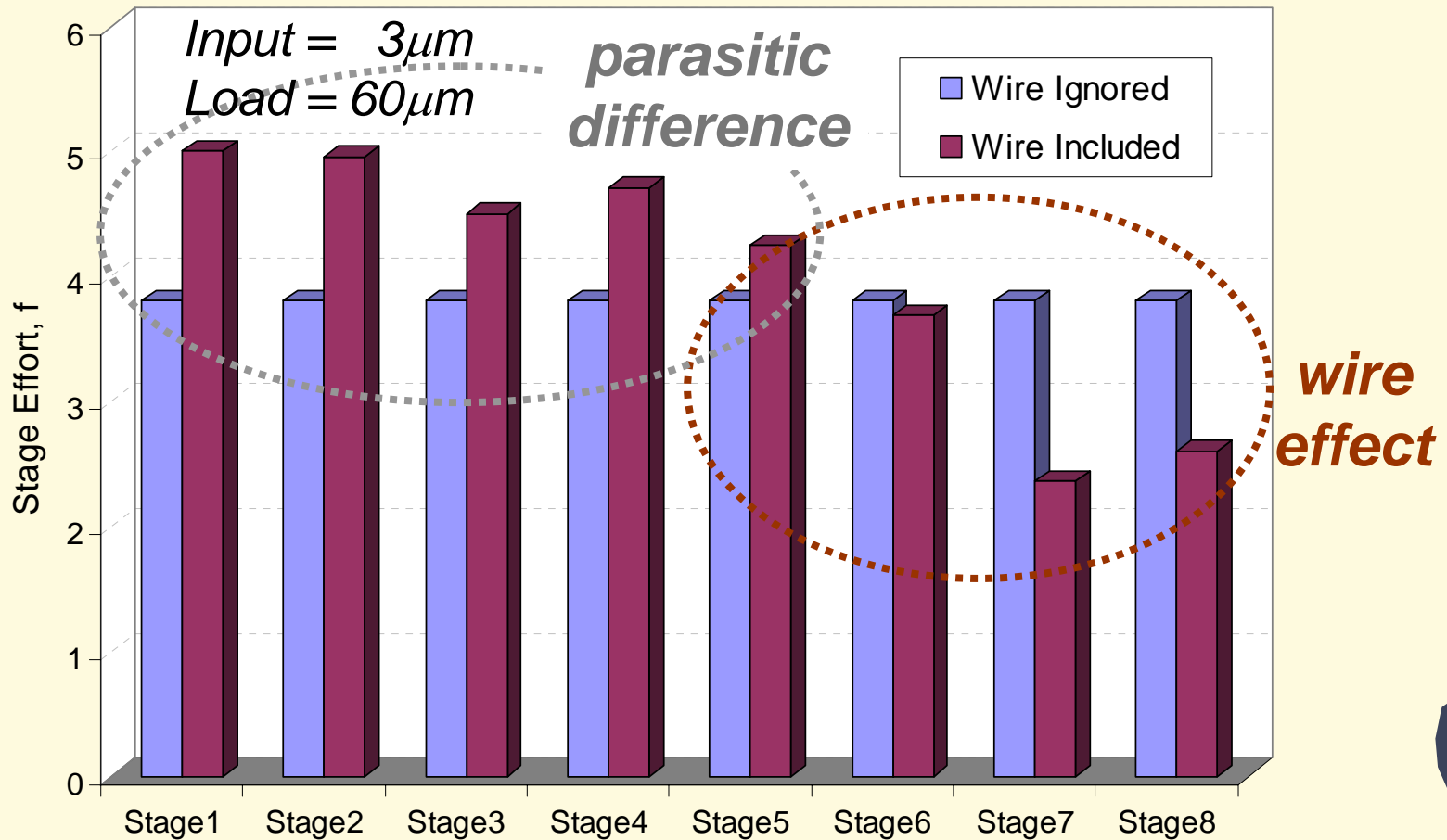


- Nonlinearity due to:
 - Constant off-path load (wire cap, min-size gates)
 - Unequal path lengths
 - Parasitic delay difference of gates
- ***No analytical form***
 - Recursive solving
 - Solution: unequal stage efforts

64-bit Static Kogge-Stone Adder

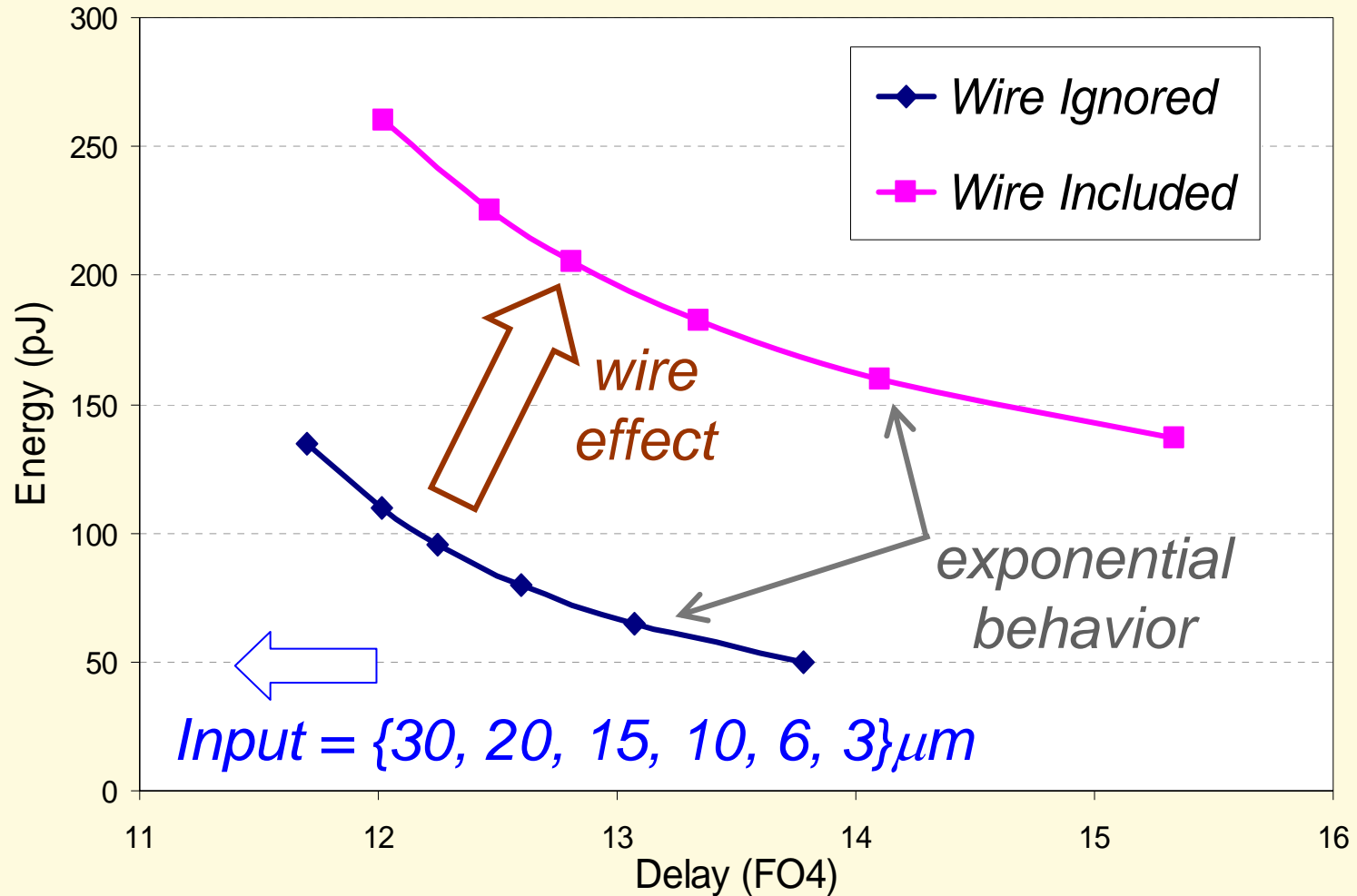


64-b KS: Stage Effort Distribution



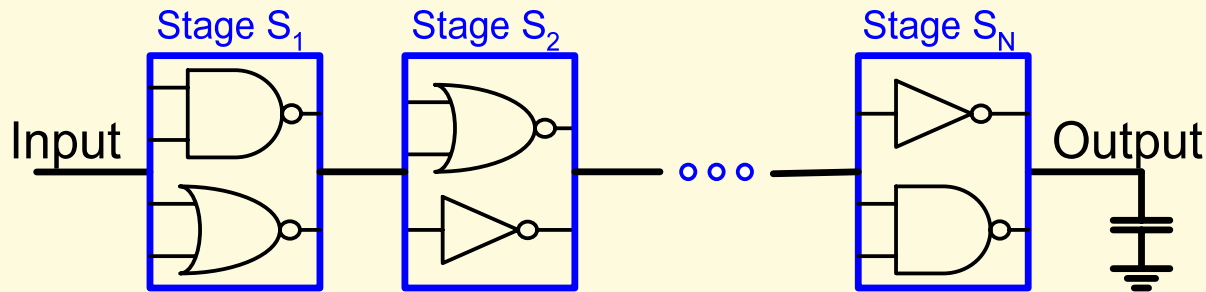
- Nonlinearity causes unequal stage efforts
 - Nonlinear factors: wire load, parasitic delay diff.

64-b KS: Energy versus Delay



- Significant wire effect on delay & energy

Problem B: Energy-Delay Tradeoff



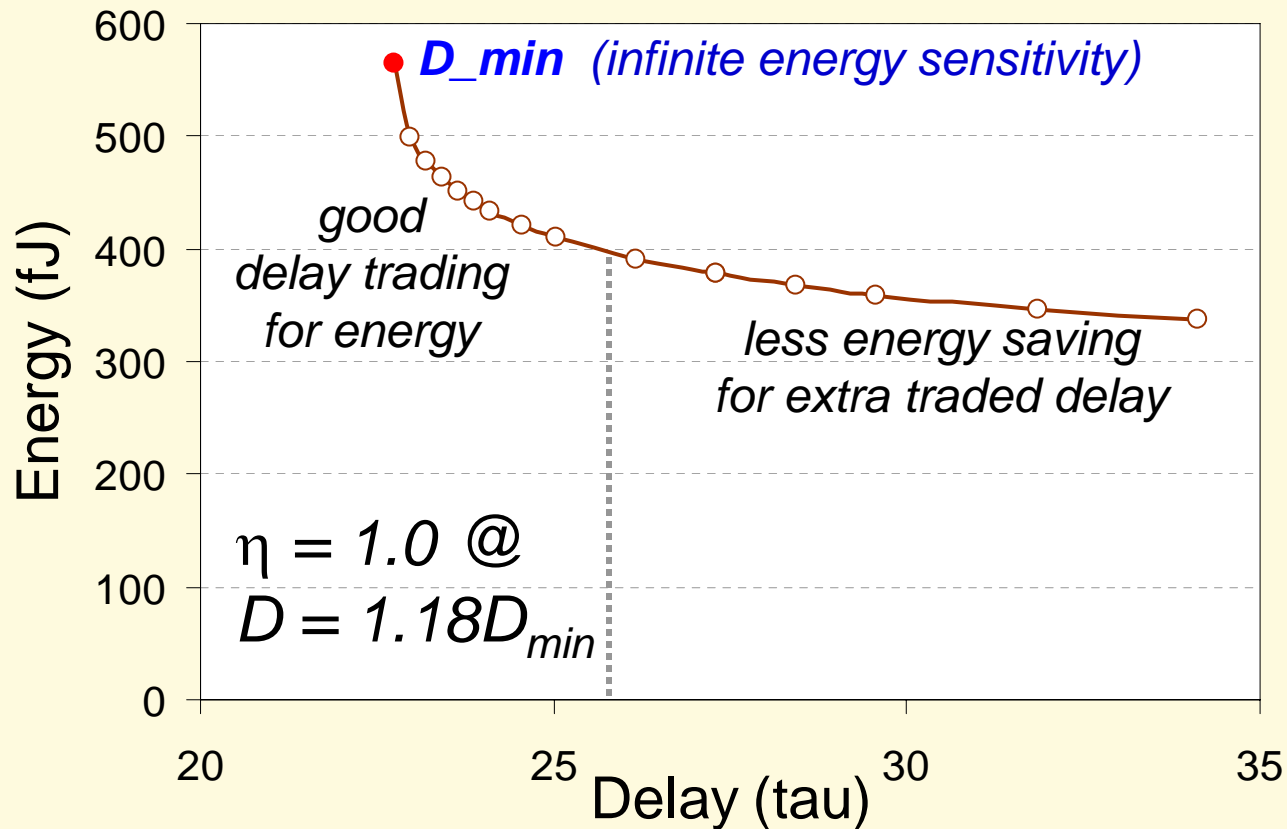
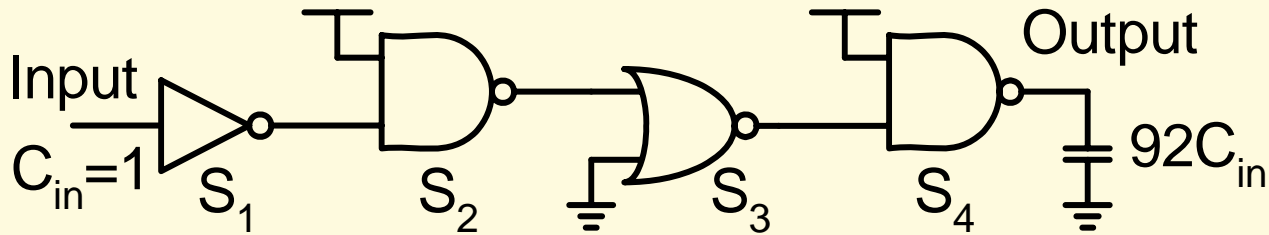
- Optimization Problem

- Minimize: $E = \sum E_{\text{Stage}(i)}$
- Constraint: $\{\text{Input, Load}\} = \text{const.}$
 $\sum D_{\text{Stage}(i)} = D_{\min} + \Delta D$

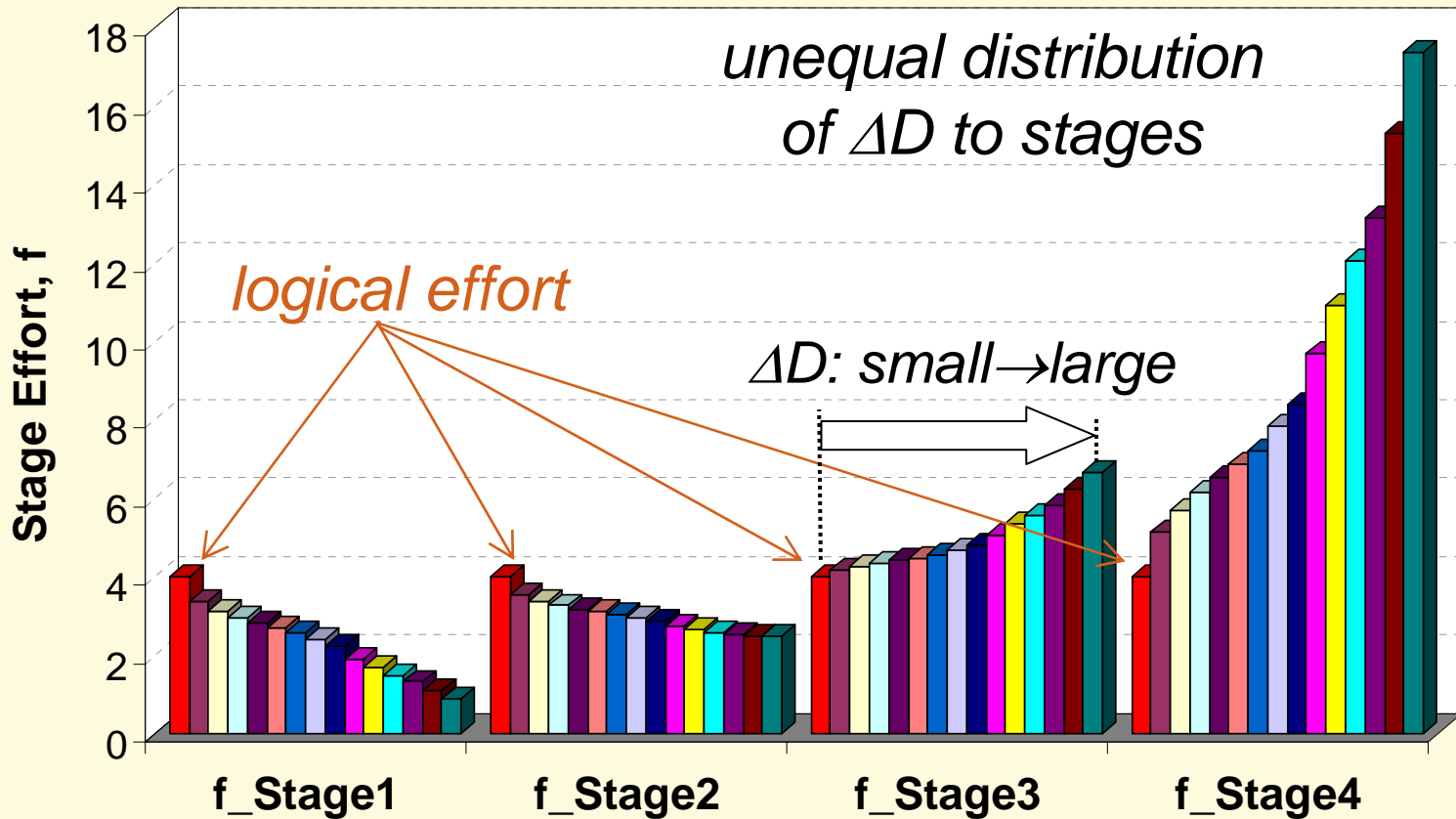
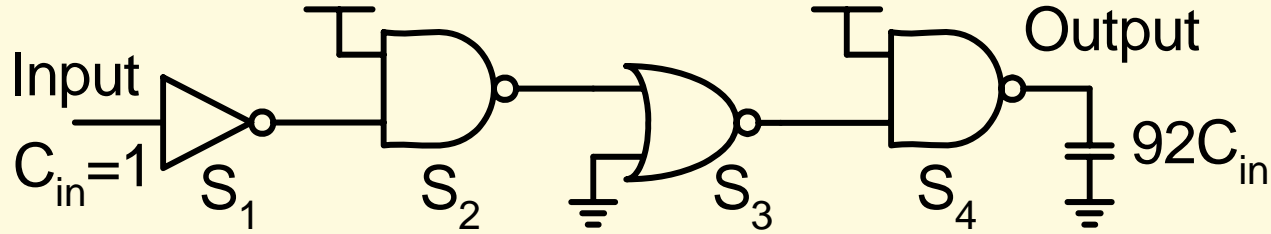
- Objectives

- Avoid infinite energy sensitivity at D_{\min}
- Equalize energy-delay sens.: $(\partial E / \partial D)_{S_i} = (\partial E / \partial D)_{S_k}$
- Trade delay for energy (*traditional approach*)

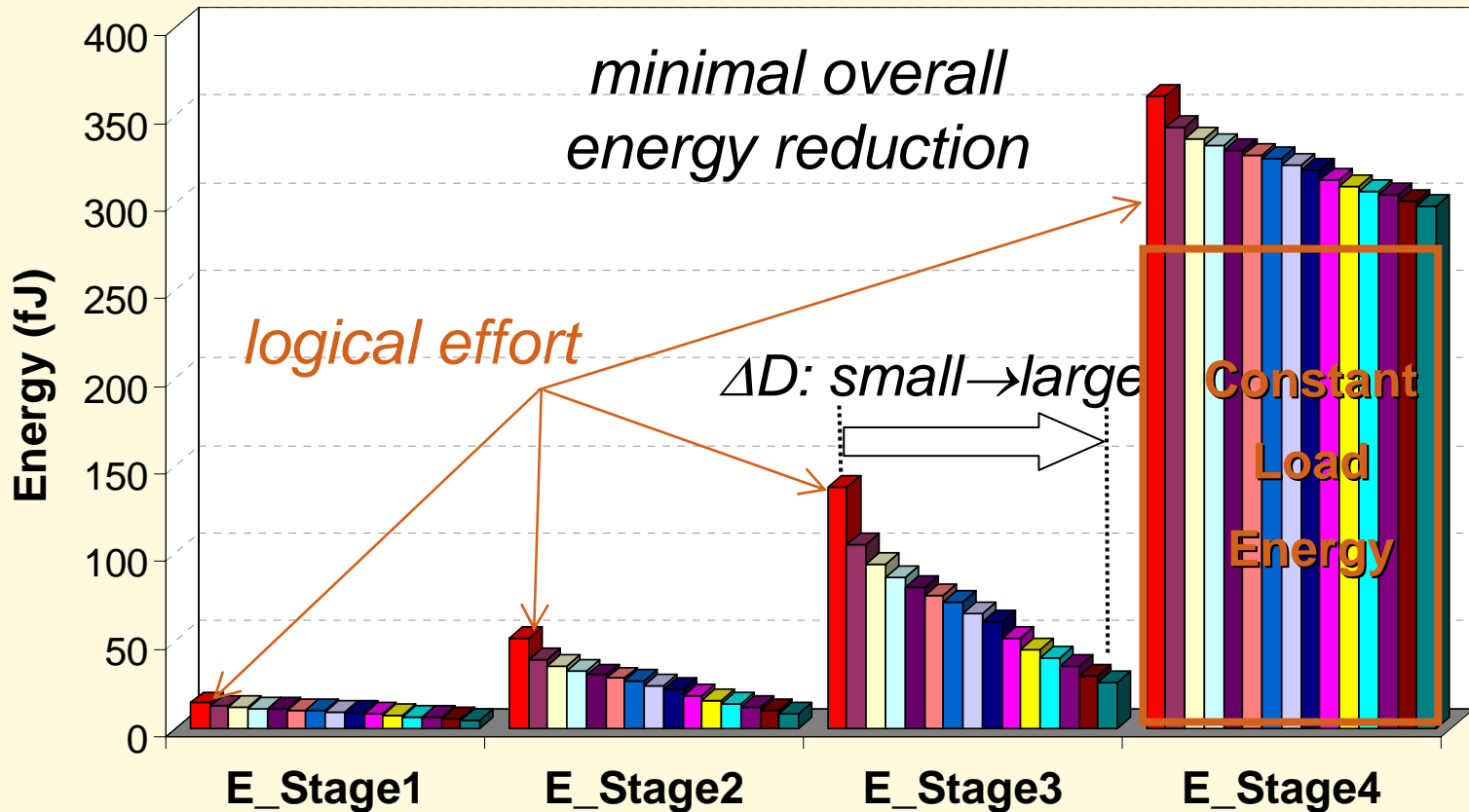
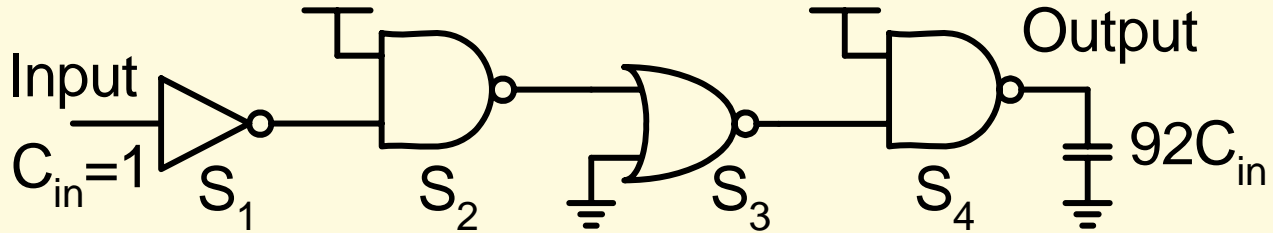
E-D Trade-off: Single Path



E-D Trade-off: Stage Effort Distrib.

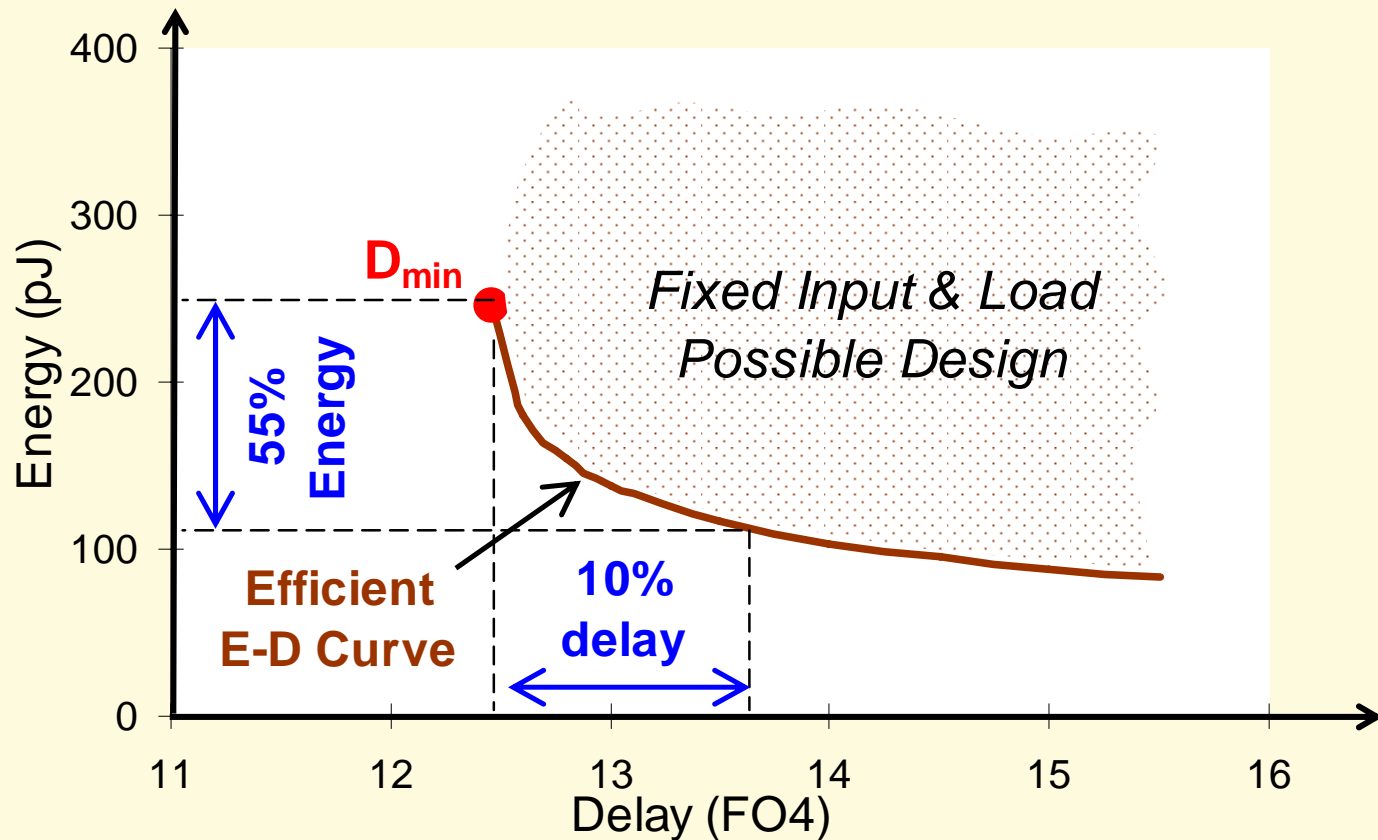


E-D Trade-off: Energy Distrib.



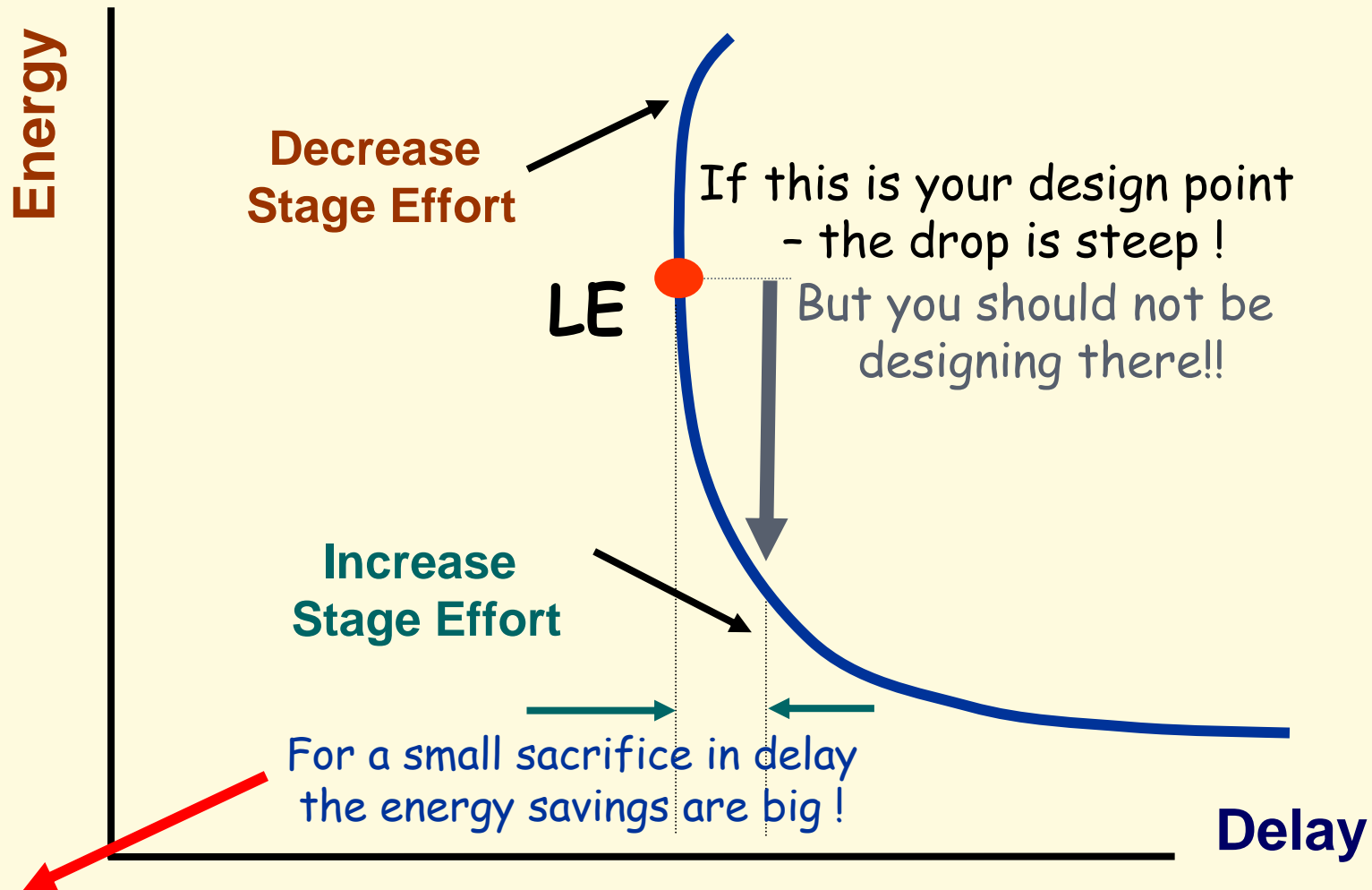
(*) Including constant output load

64-b KS: Energy Delay Tradeoff



- D_{min} solution is very inefficient in energy
- 55% energy saving with 10% delay traded
- Solution = equal stage energy-delay sensitivity

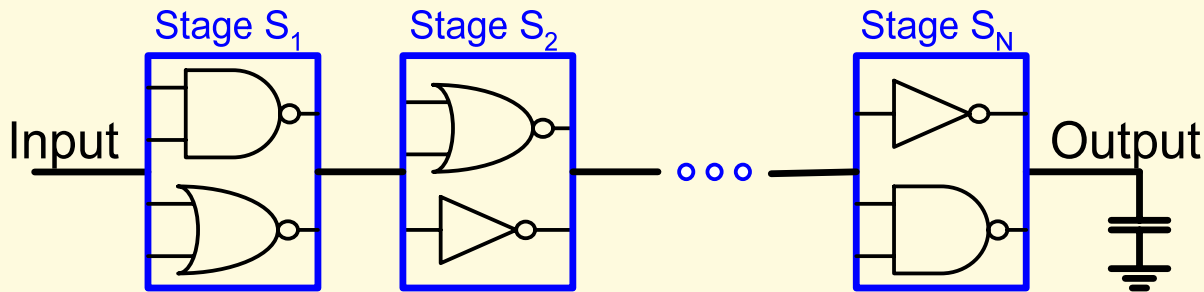
Design in Energy-Delay Space



From:

R.W. Brodersen, M.A. Horowitz, D. Markovic, B. Nikolic, V. Stojanovic, "Methods for True Power Minimization," International Conference on Computer-Aided Design, ICCAD-2002, Digest of Technical Papers, San Jose, CA, November 10-14, 2002, pp. 35-42.

Problem C: Energy Minimization



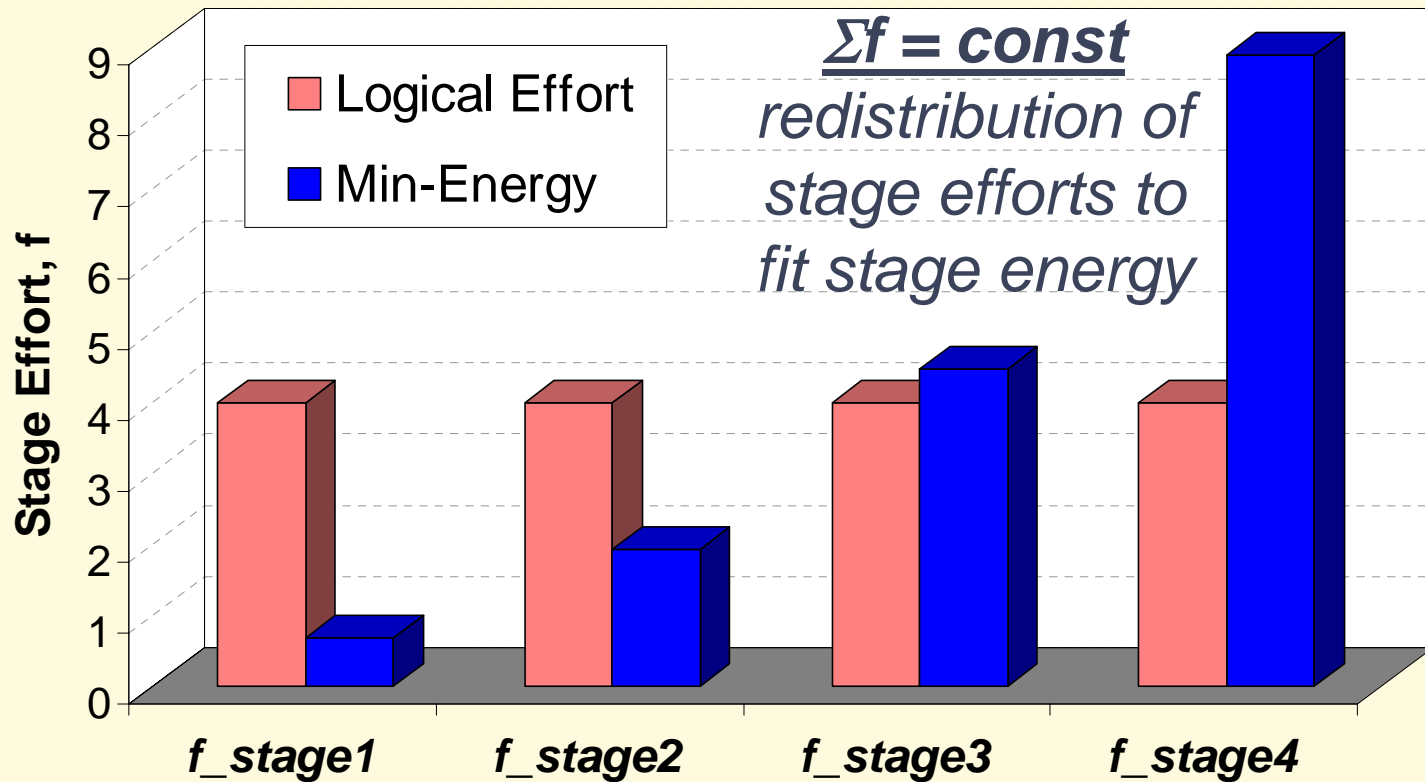
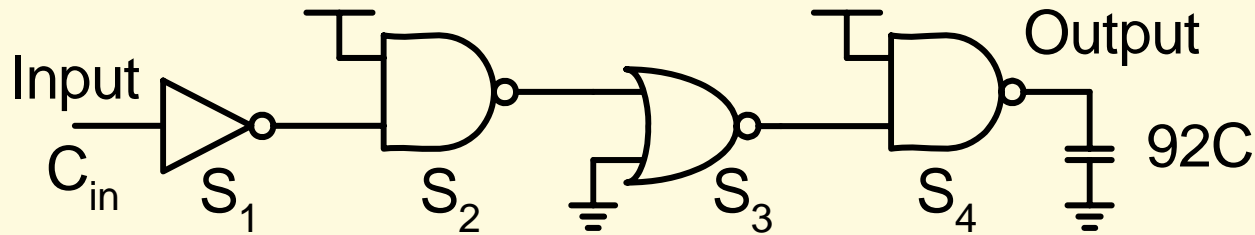
- Problem:

- Minimize: $E = \sum E_{S_i}$
- Constraint: $D = \sum D_{S_j} = \text{const.}$
 $\text{Load} = \text{const.}$

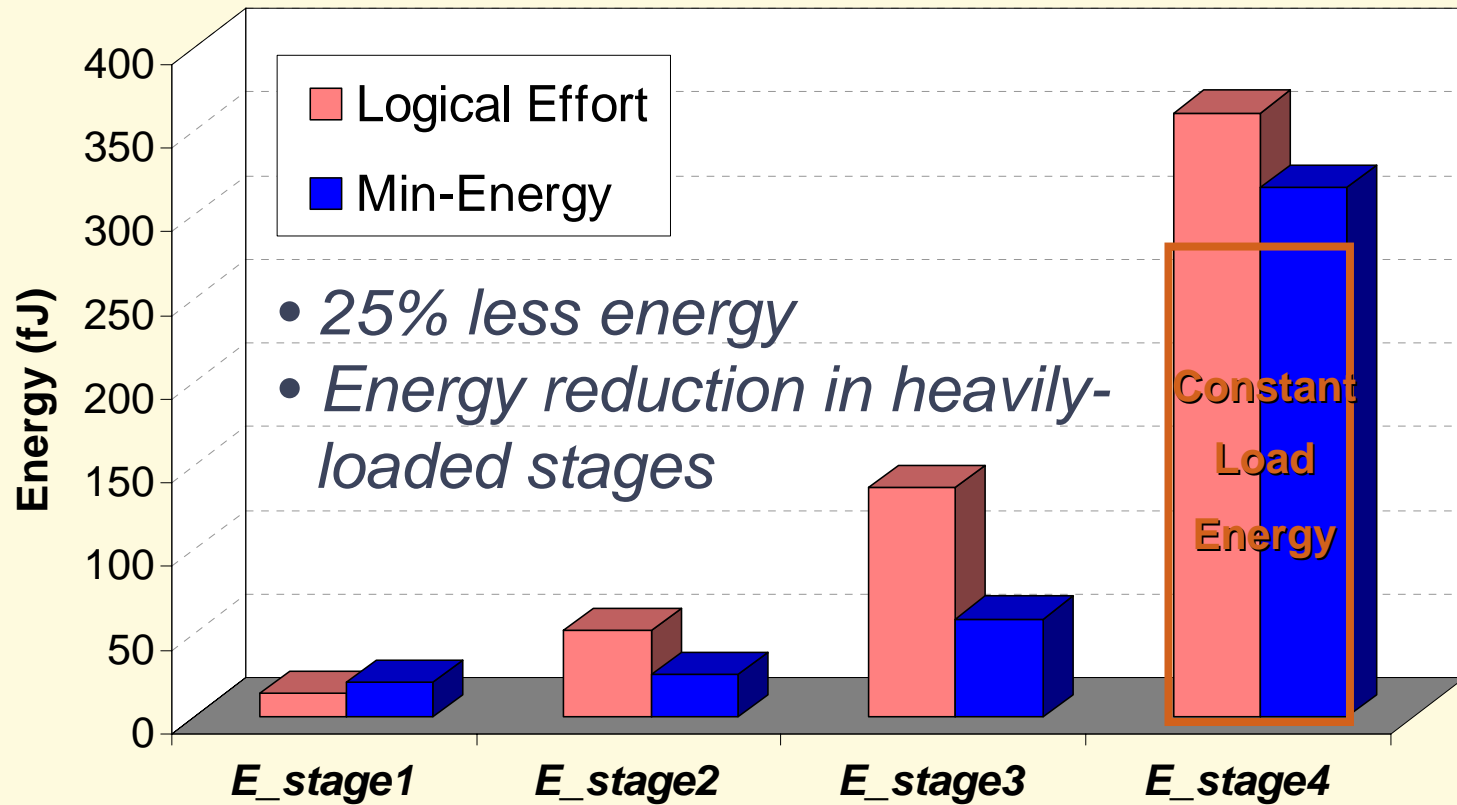
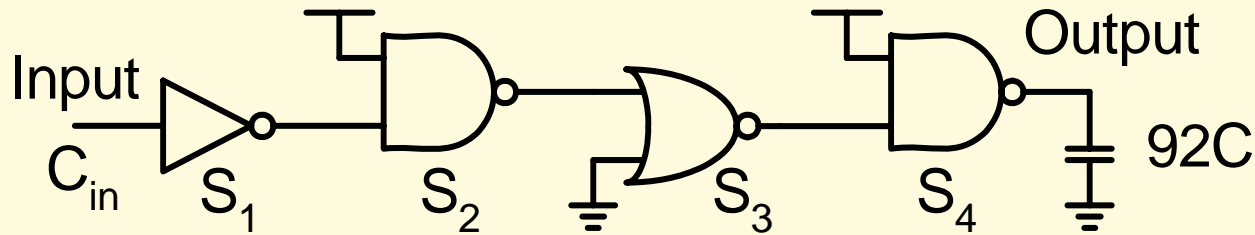
- Objective:

- Obtain absolute minimal energy @ given delay
- Equalize energy-delay sens.: $(\partial E / \partial D)_{S_i} = (\partial E / \partial D)_{S_k}$
- Trade input size such that $(\partial E / \partial \text{Input})_D = 0$

Single Path: Stage Effort Redistrib.

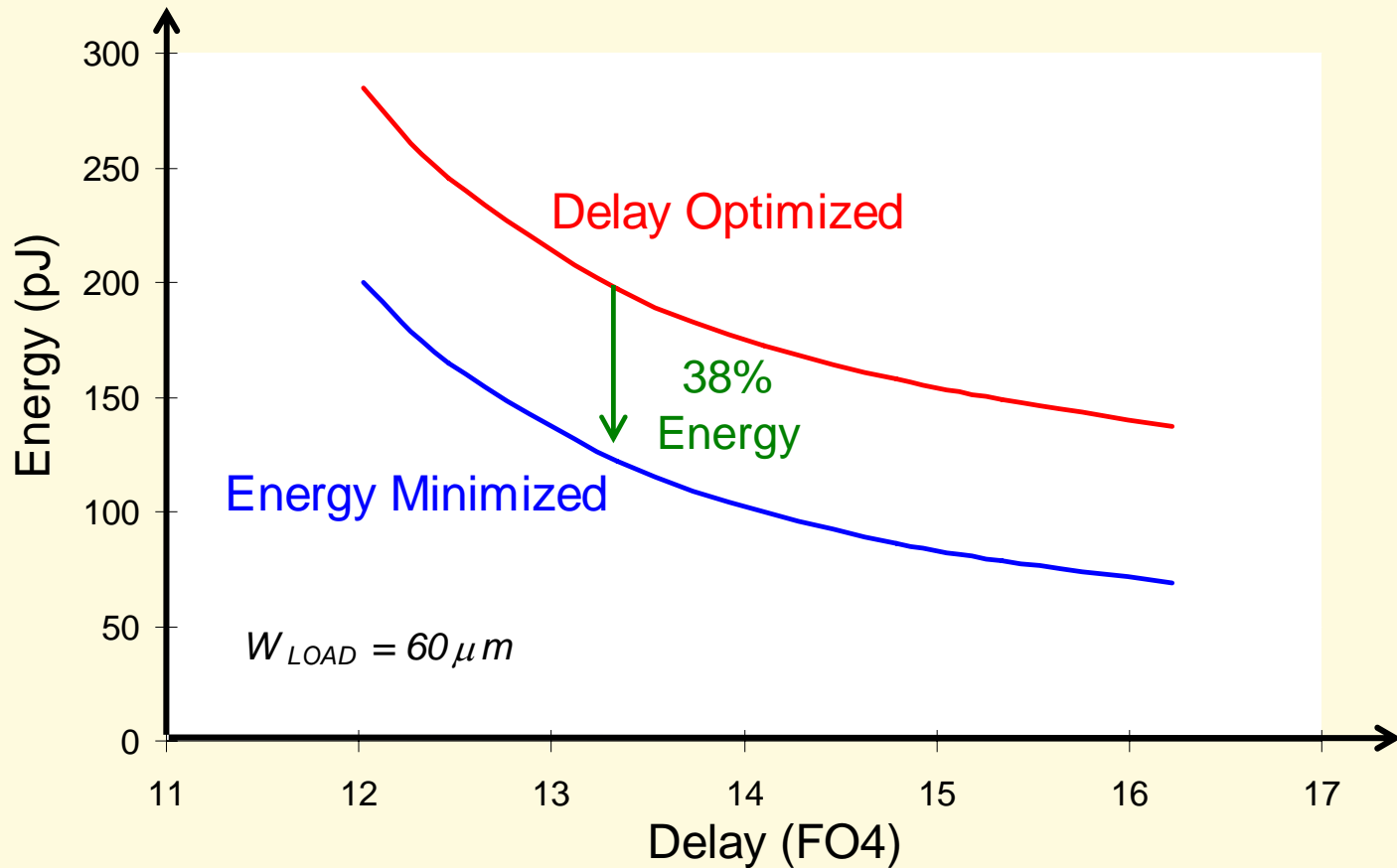


Single Path: Energy Distribution



(*) Including energy consumed on output load

64-b KS: Minimal Energy vs. Delay



- 30 - 50% energy saving @ same performance
- 1.6 - 3.6X input size

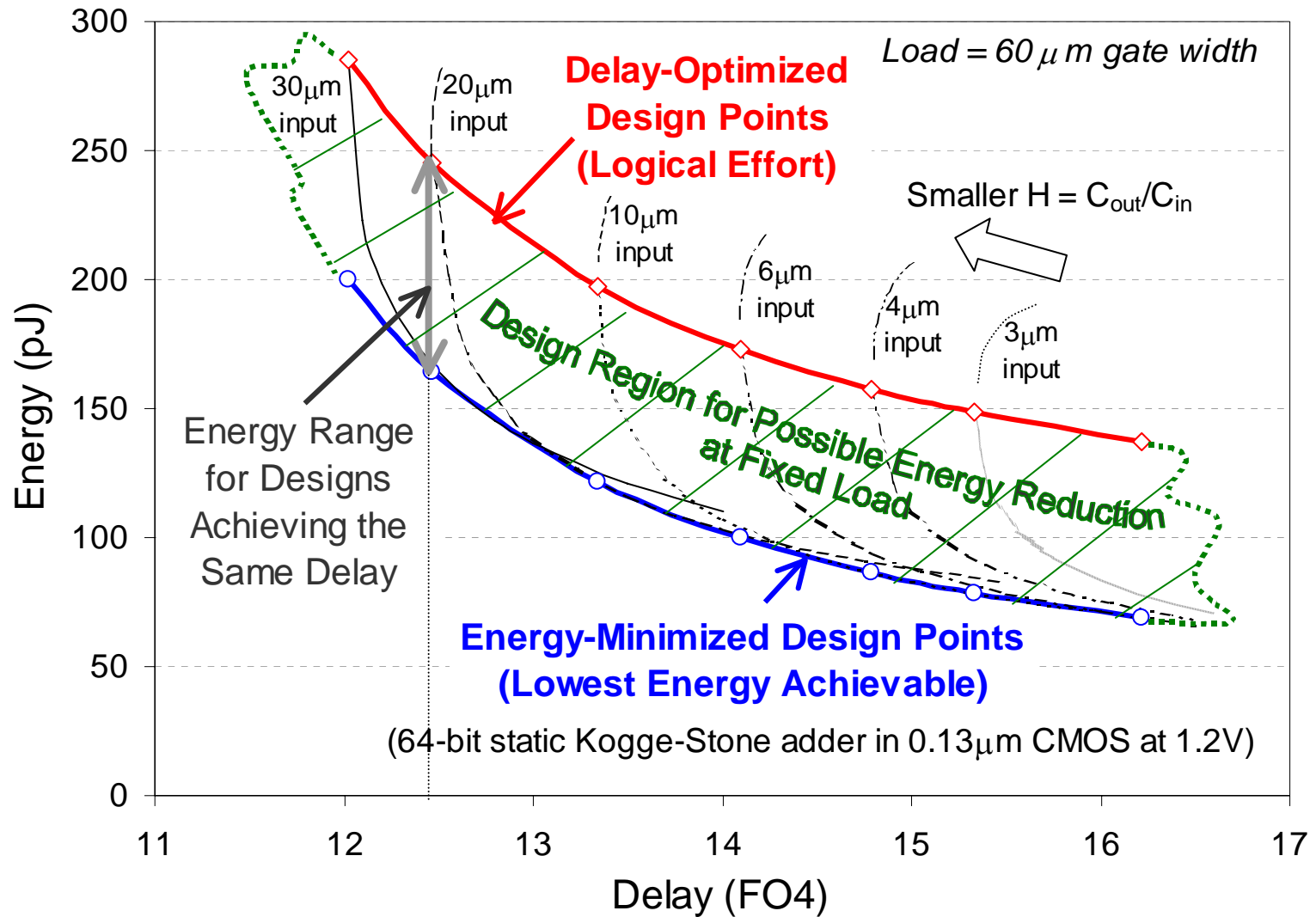
Pipelined System Optimization



Pipelined System Optimization

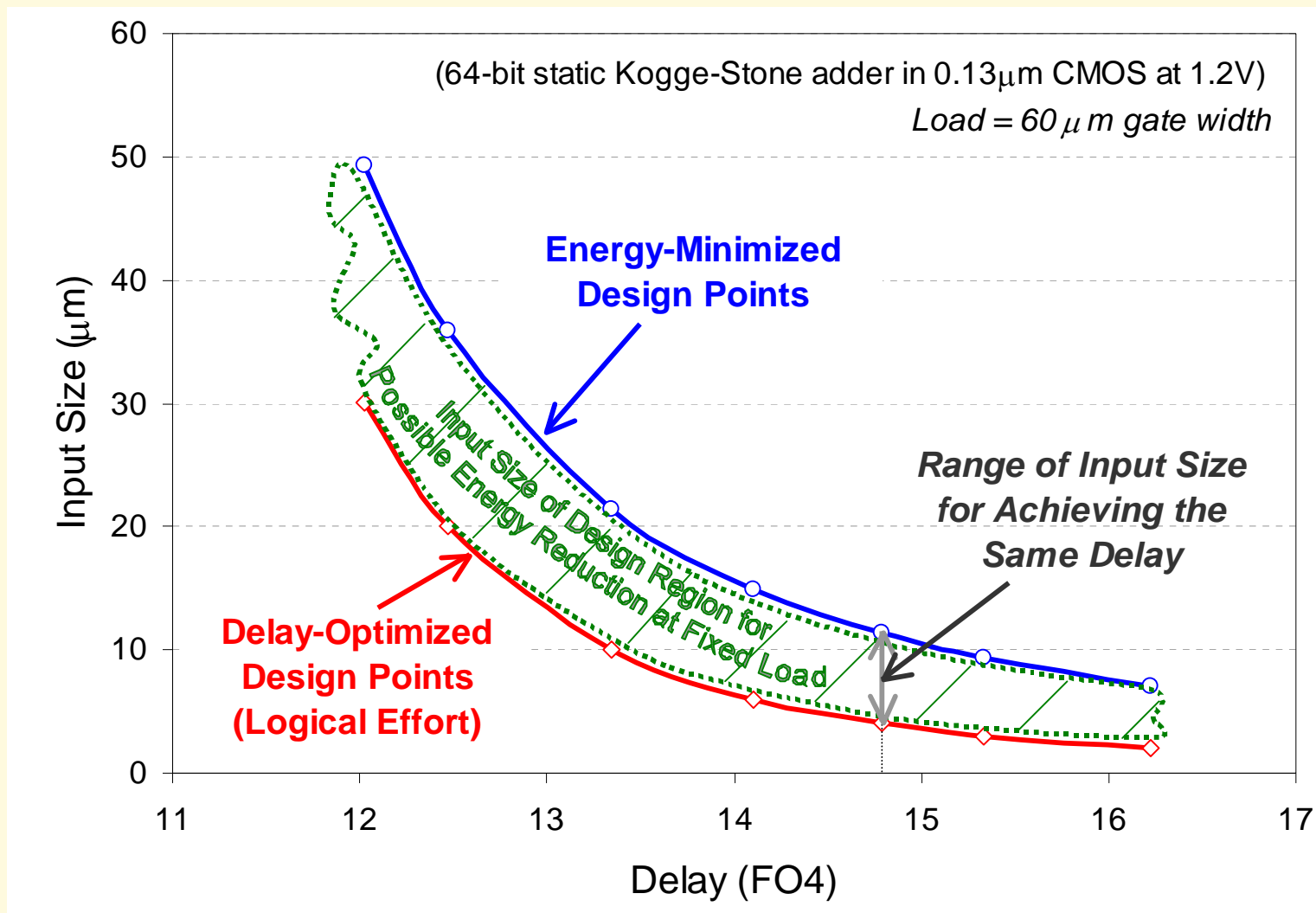
- Design constraints
 - Delay target
 - External I/O constraint
- How to obtain minimal-energy solution?
 - Pipelined stages
 - Minimized for energy
 - Sensitive to input and load variations
 - System level
 - Balancing energy sensitivities at pipelined boundaries
 - Recursive process

Pipelined Stage: Efficient E-D Area



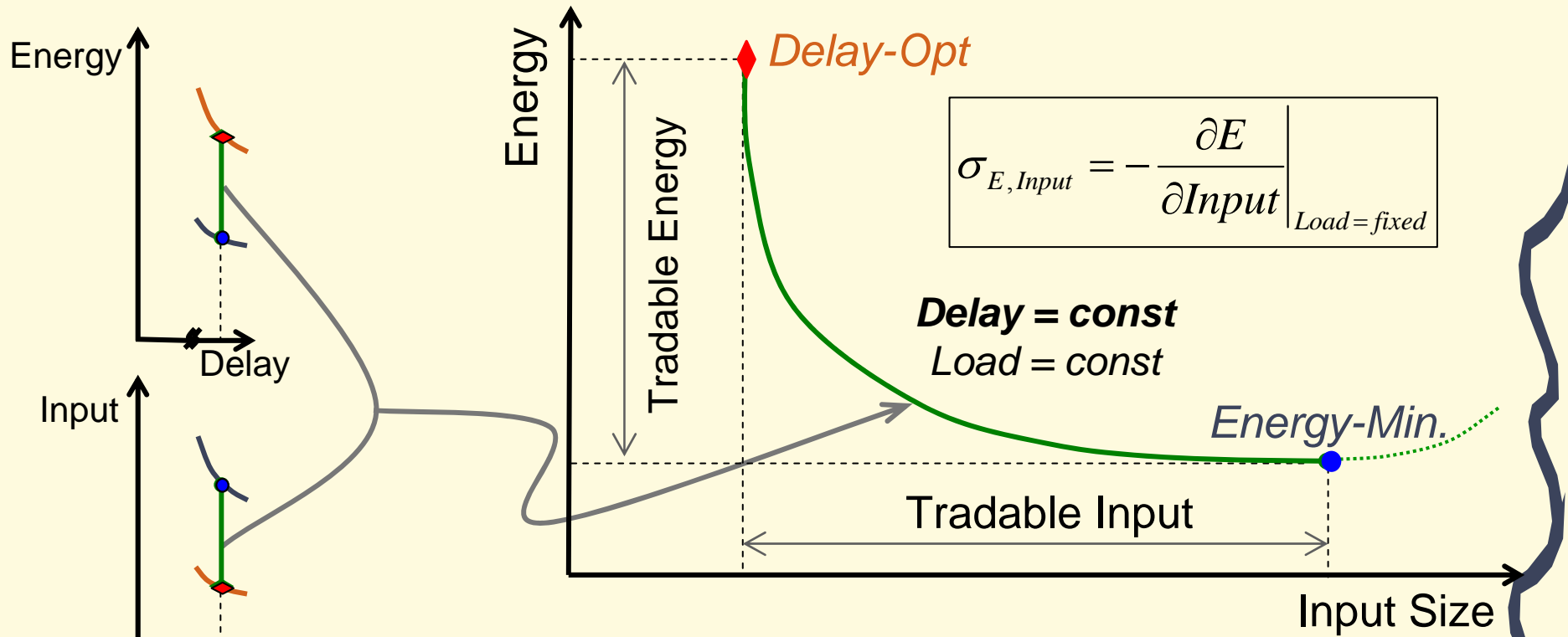
- Efficient E-D area is upper- and lower-bounded

Efficient Input-Delay Area



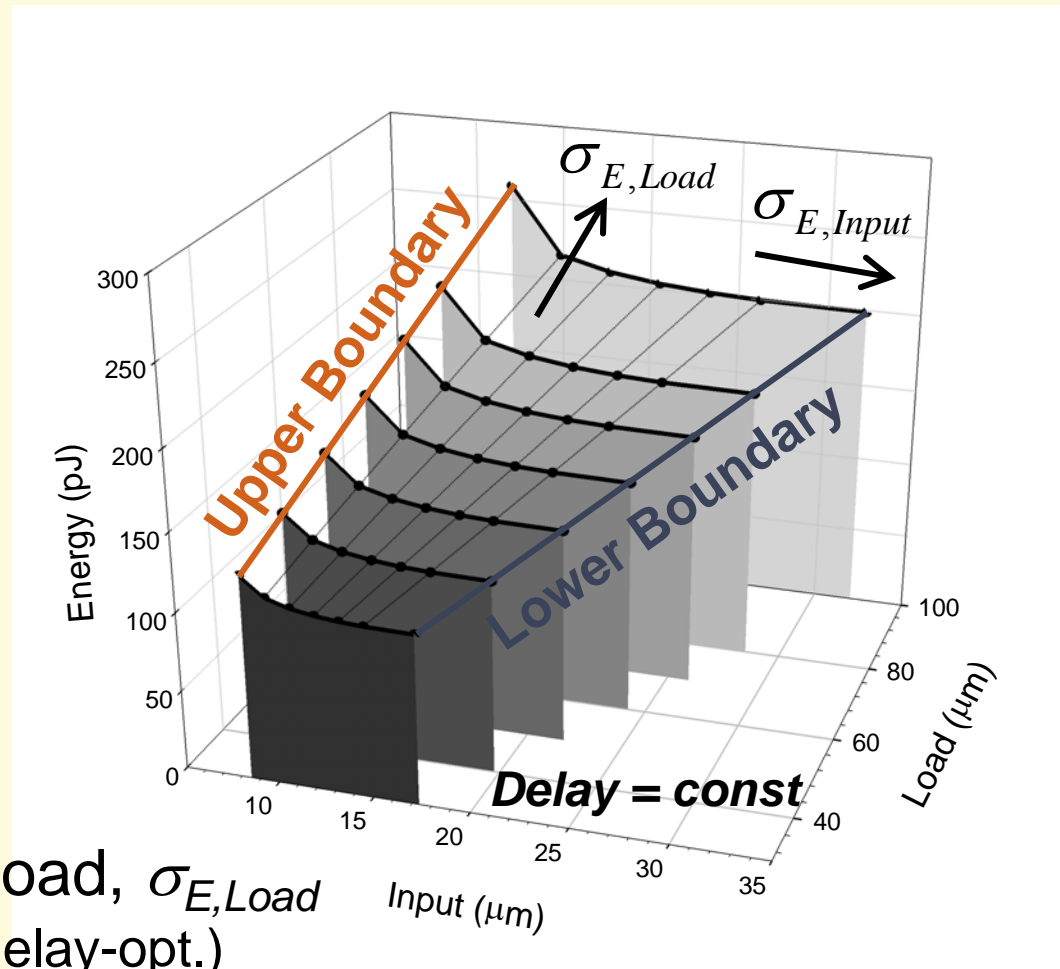
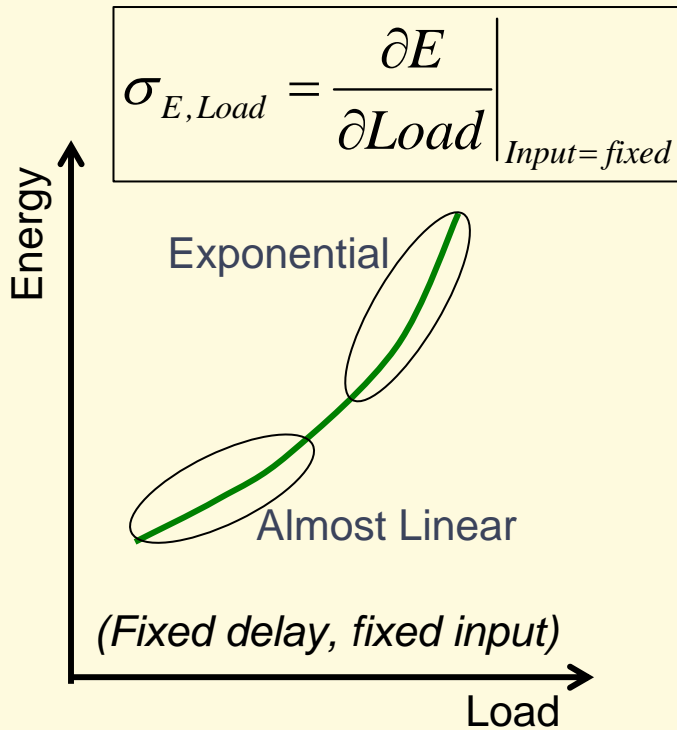
- Larger/smaller input \Leftrightarrow less/more energy

Efficient Energy-Input @ $D = \text{const.}$



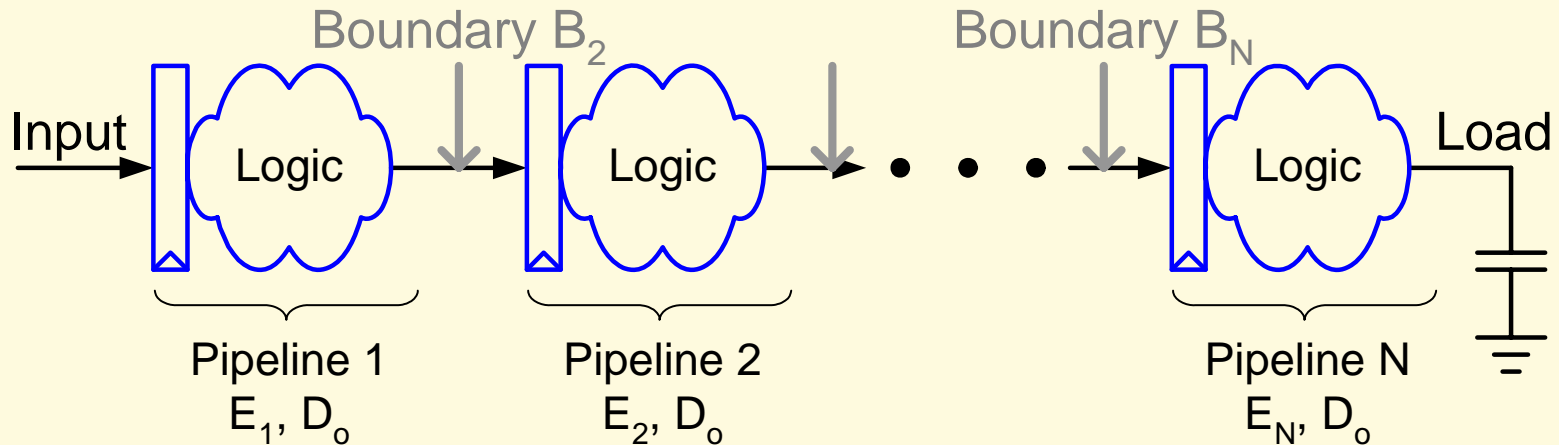
- Energy vs. input size: one-to-one relation
 - Energy sensitivity to input, $\sigma_{E,Input}$
- Flat energy in upper half of input range

Sensitivity to Load @ $D = \text{const.}$



- Energy sens. to output load, $\sigma_{E,Load}$
 - More @ upper bound (delay-opt.)
 - Less @ lower bound (energy-min.)
- Energy components = energy + its sensitivities

System Energy Optimization



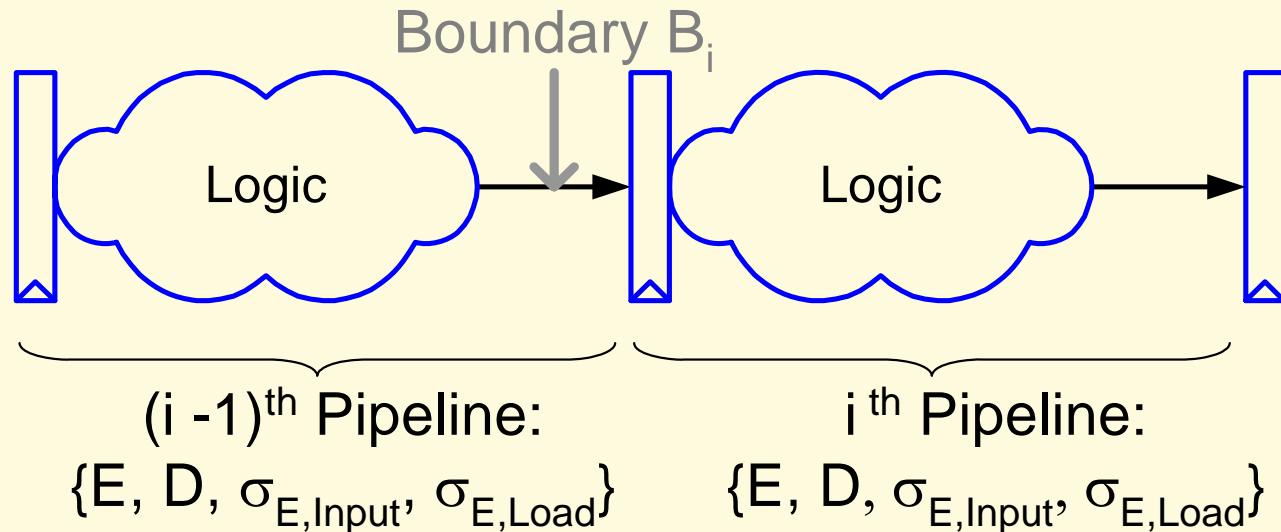
- Energy minimization

- Pipeline: minimize energy @ given input & load
- System: balance energy sensitivities @ boundaries

- Trading elements: input size, output load

- Optimal criteria:
$$\begin{cases} E_{Stage A_i} = \textit{minimal} \\ \sigma_{E, Input A_i} = \sigma_{E, Load A_{i-1}} \end{cases}$$

How to Achieve Less Total Energy



Case A:

- $\sigma_{E,\text{Input}}(i) > \sigma_{E,\text{Load}}(i-1)$
- Increase i^{th} input
 \Rightarrow less E_{Total}

Case B:

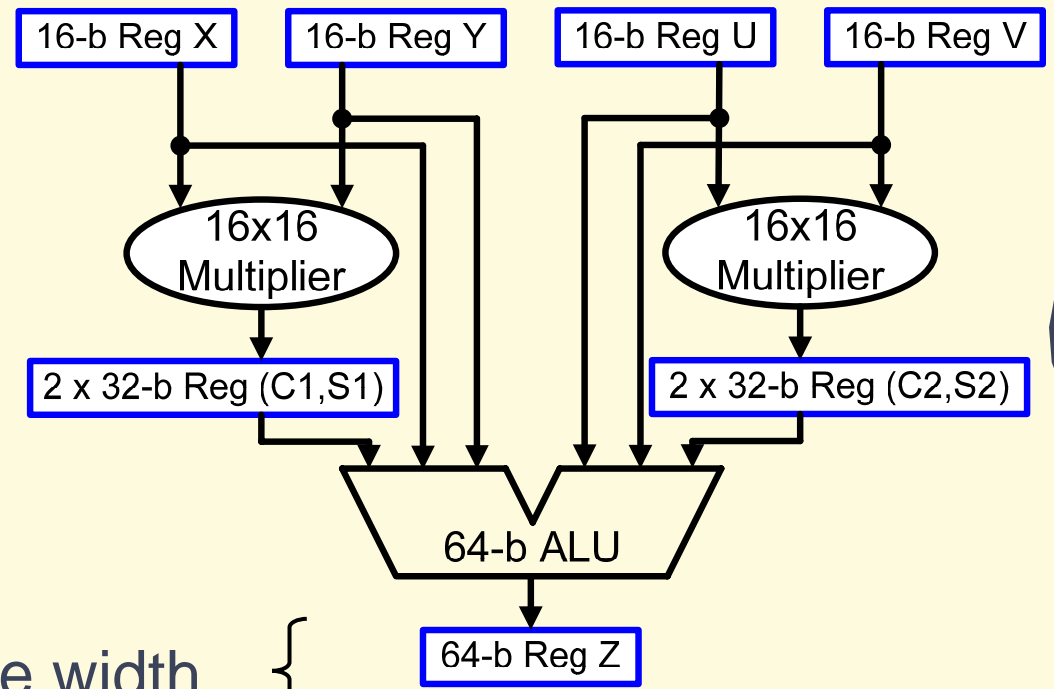
- $\sigma_{E,\text{Input}}(i) < \sigma_{E,\text{Load}}(i-1)$
- Reduce i^{th} input
 \Rightarrow less E_{Total}

$$E_{i-1} + E_i = \min \Leftrightarrow \sigma_{E,\text{Input}}(i) = \sigma_{E,\text{Load}}(i-1)$$

Case Study: Media Datapath

Maximal Input = $30\mu\text{m}$ {

Target Delay = $17FO_4$



Output Load = $60\mu\text{m}$ gate width {

- What is the minimal energy solution?

Case Study: Optimal Criteria

$$\text{MAX}\{X, Y, U, V\} = 30\mu\text{m}$$

$$\{Z\} = 60\mu\text{m}$$

$$\text{Delay} = 17FO_4$$

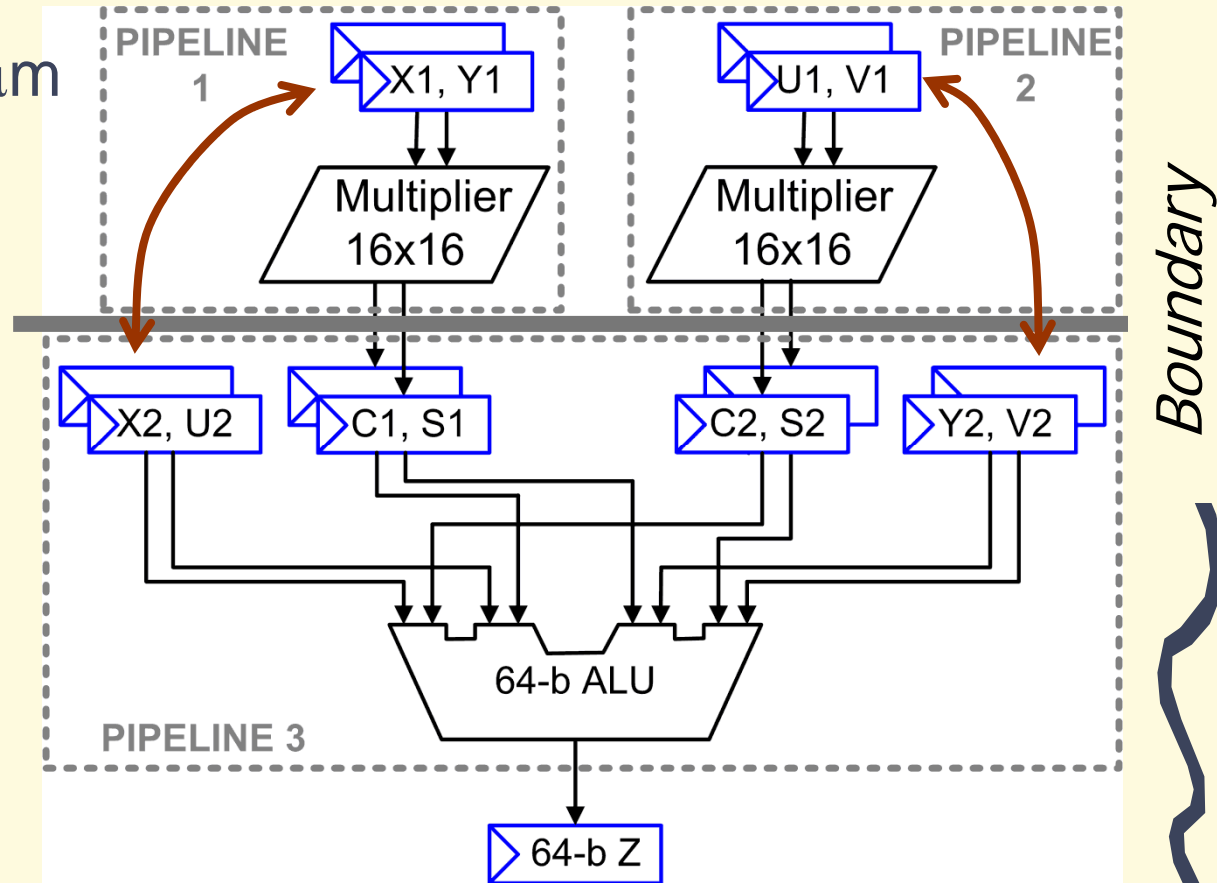
$$E\{P1\} = \text{min.}$$

$$E\{P2\} = \text{min.}$$

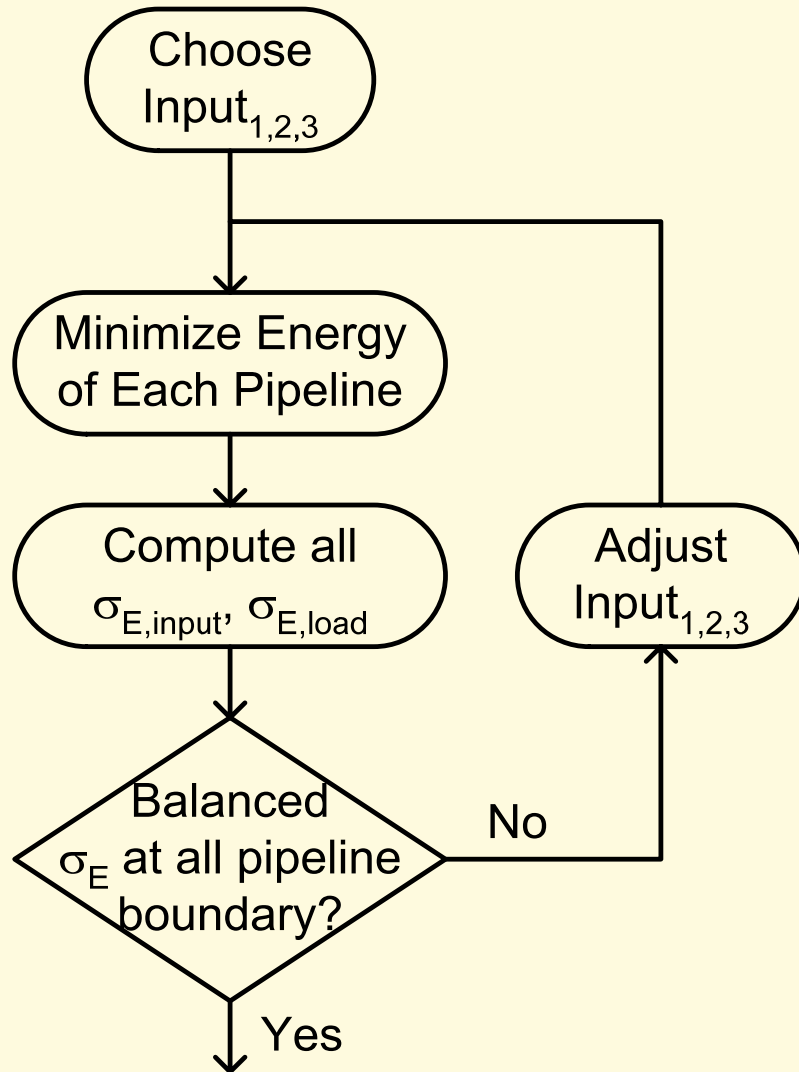
$$E\{P3\} = \text{min.}$$

$$\sigma_{E, \text{Load P1}} + \sigma_{E, \text{Load P2}} = \sigma_{E, \text{Input P3}}$$

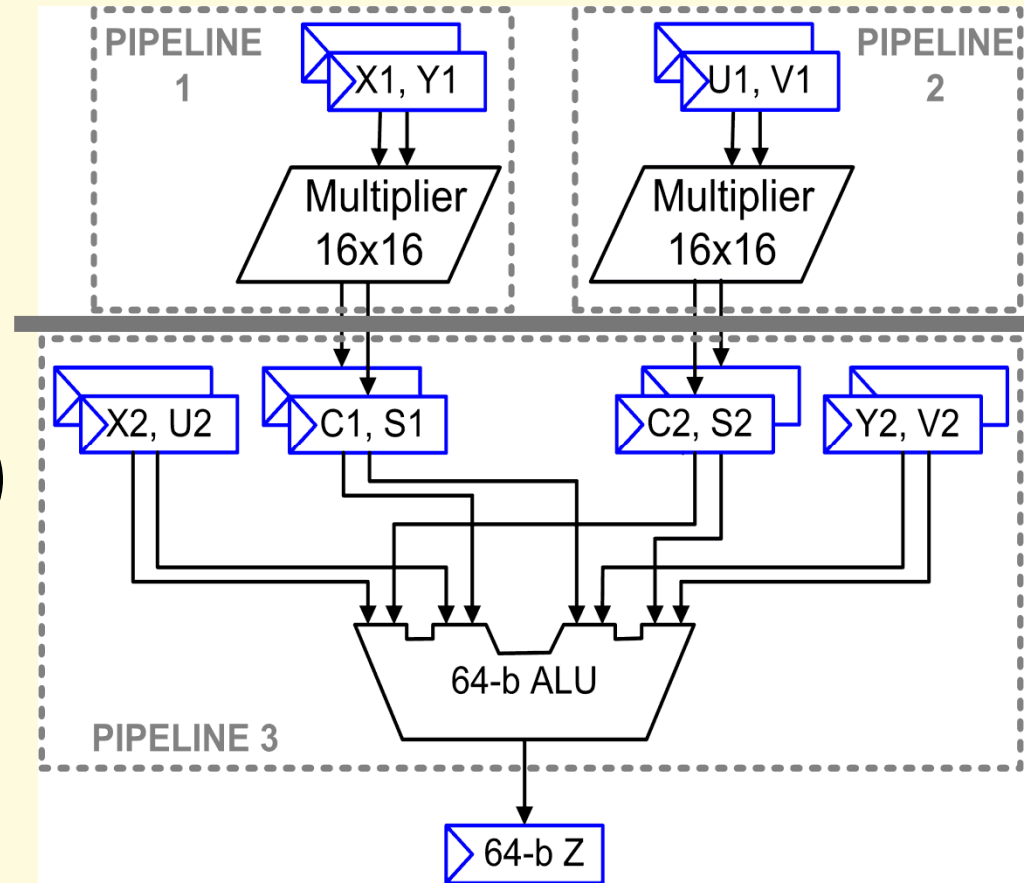
Simplified Datapath



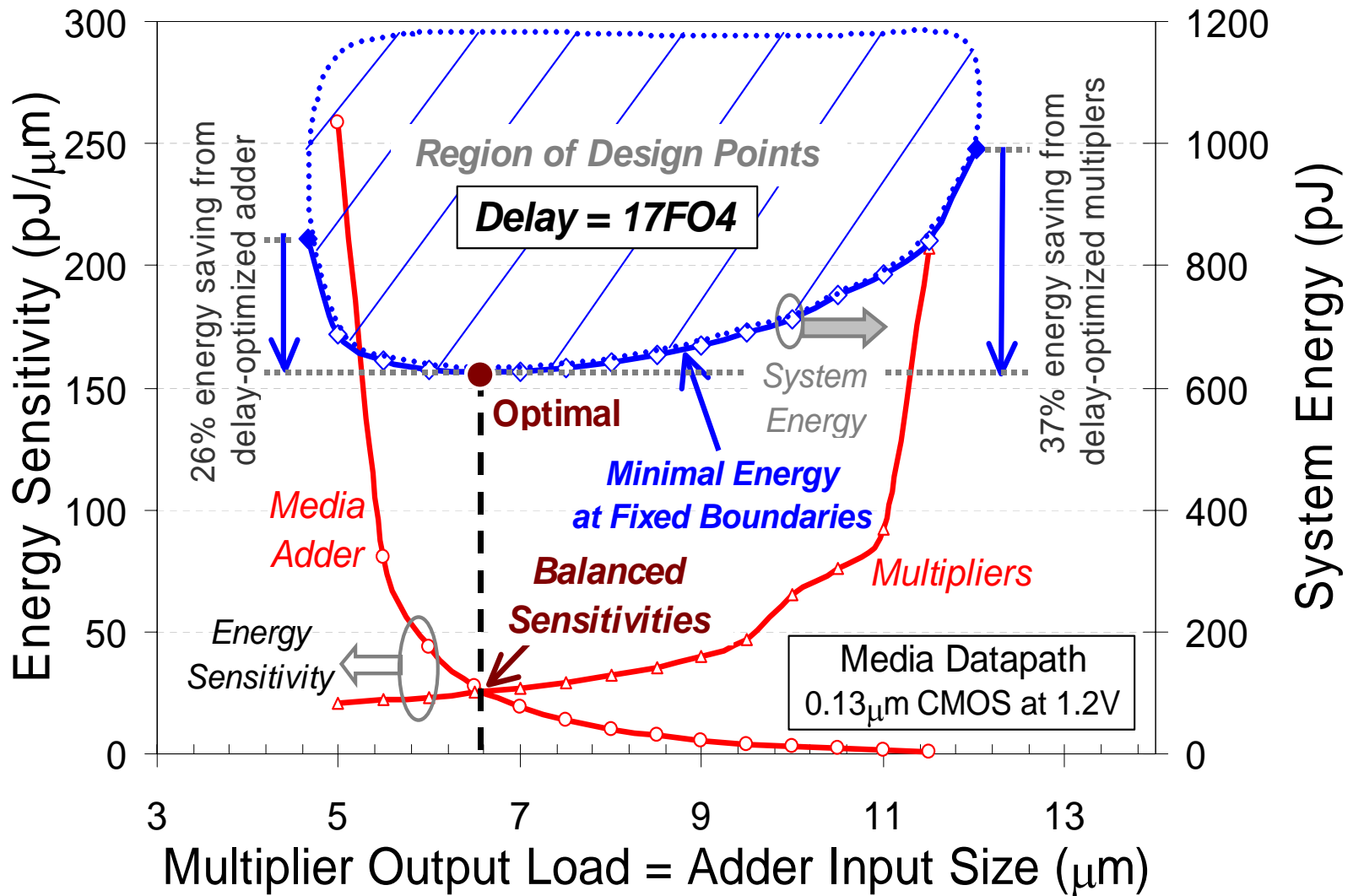
Case Study: Optimal Algorithm



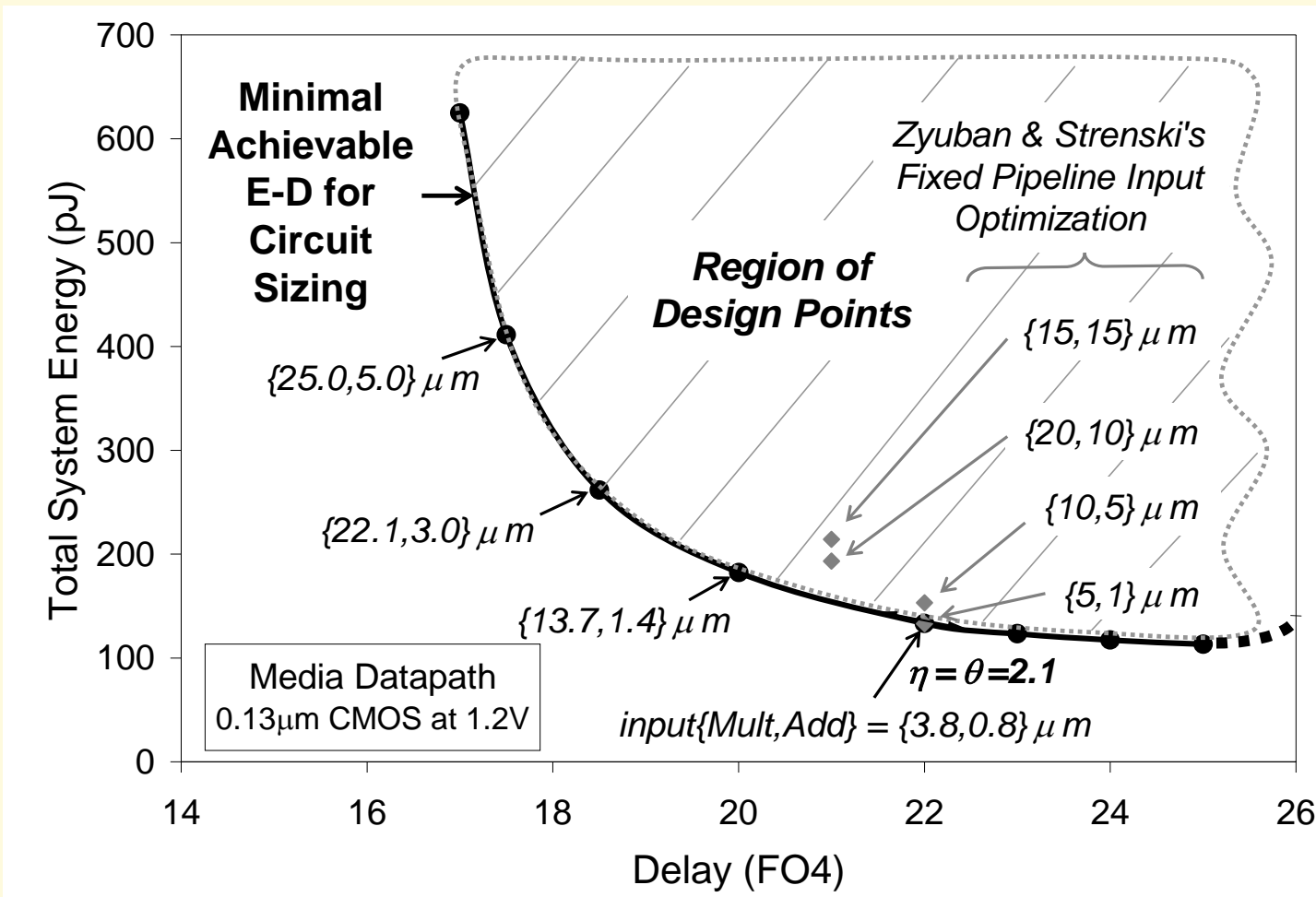
Simplified Datapath



Case Study: Media Datapath Solution

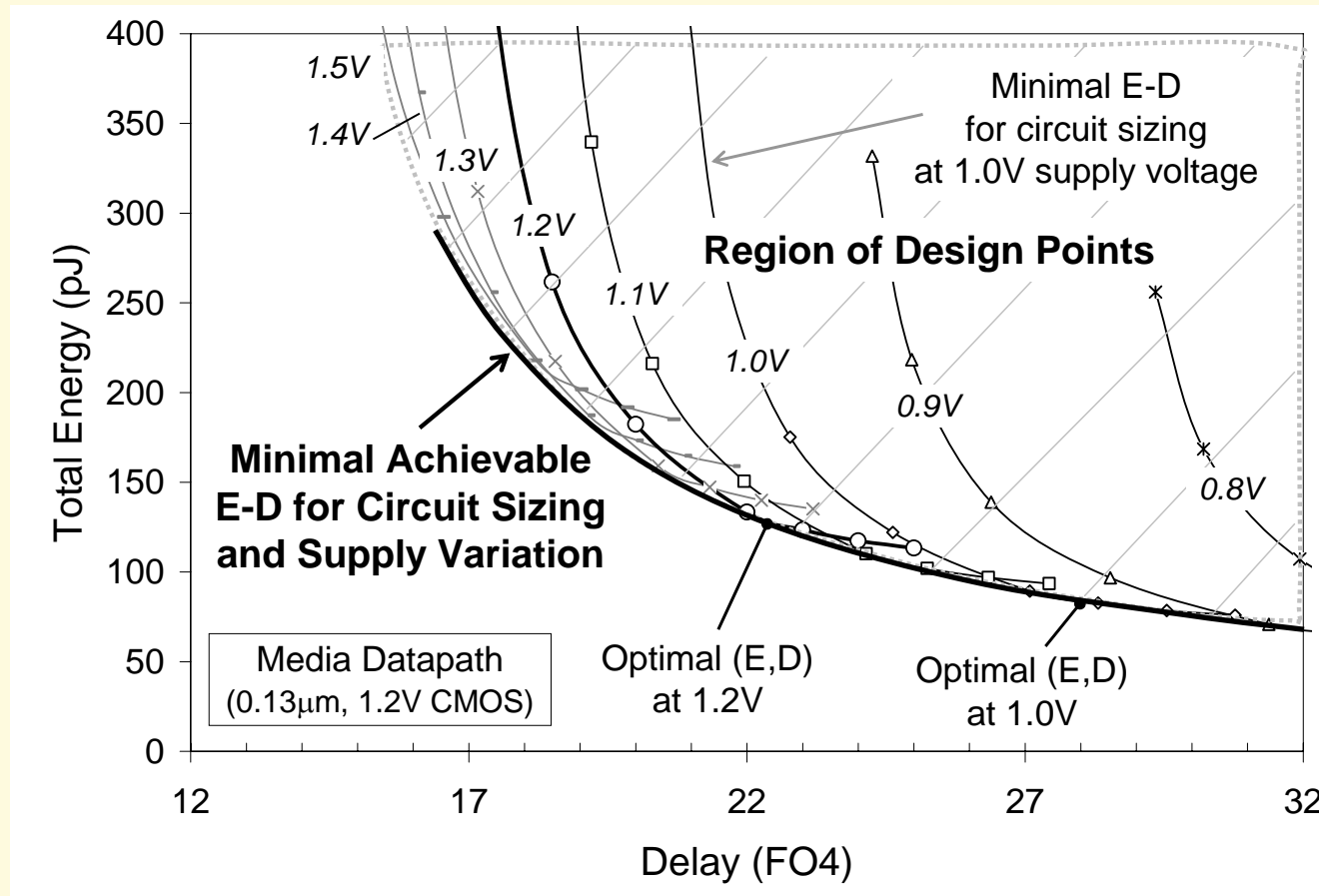


System Energy-Delay @ $V_{DD} = \text{const.}$



- Similar E-D characteristics as single stages
- Possibly less input size @ lower delay

Effect of Supply Scaling

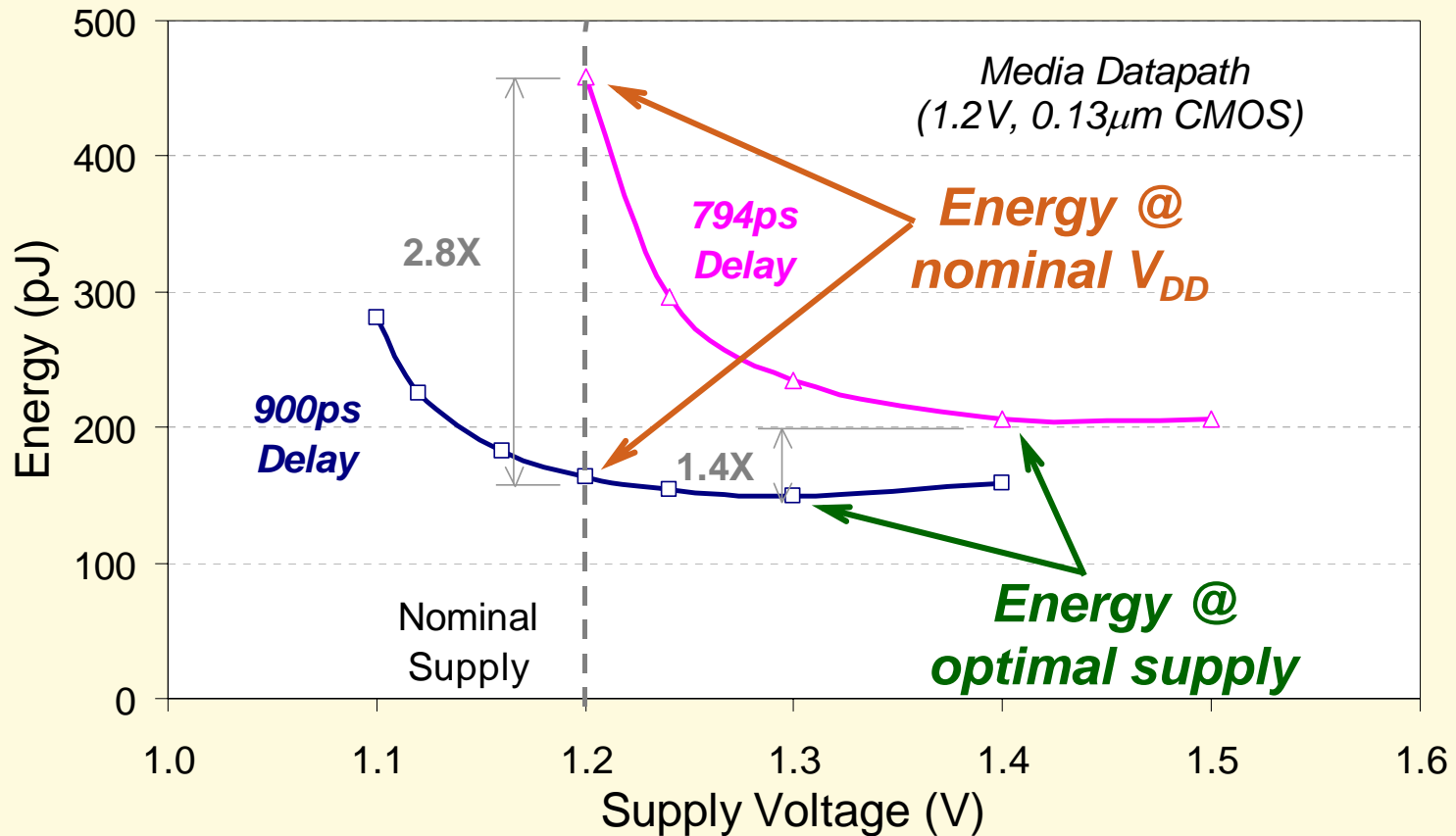


- Efficient $\{E, D, V_{DD}\}$ are dependent

- Optimal criterion:

$$\left. \frac{D}{E} \frac{\partial E}{\partial D} \right|_{sizing} = \left. \frac{D}{E} \frac{\partial E}{\partial D} \right|_{supply} \quad \text{or} \quad \eta_{system} = \theta$$

Potential Saving of System Energy

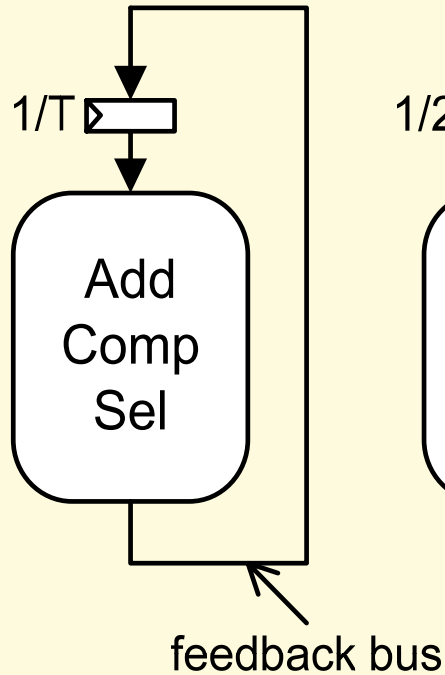


- Significant energy saving with correct supply or delay selection

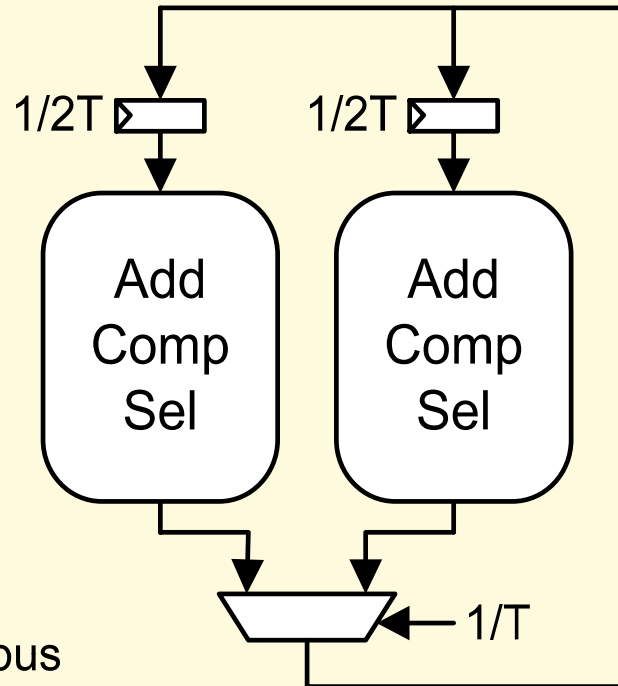
Energy-Delay Improvement of Pipelined Stages



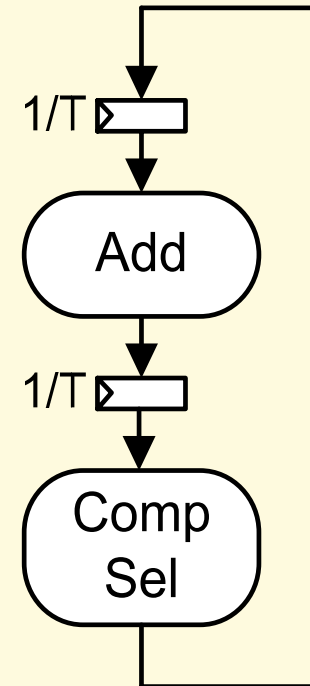
Architectural Advantages



(a) Reference



(b) Parallelism



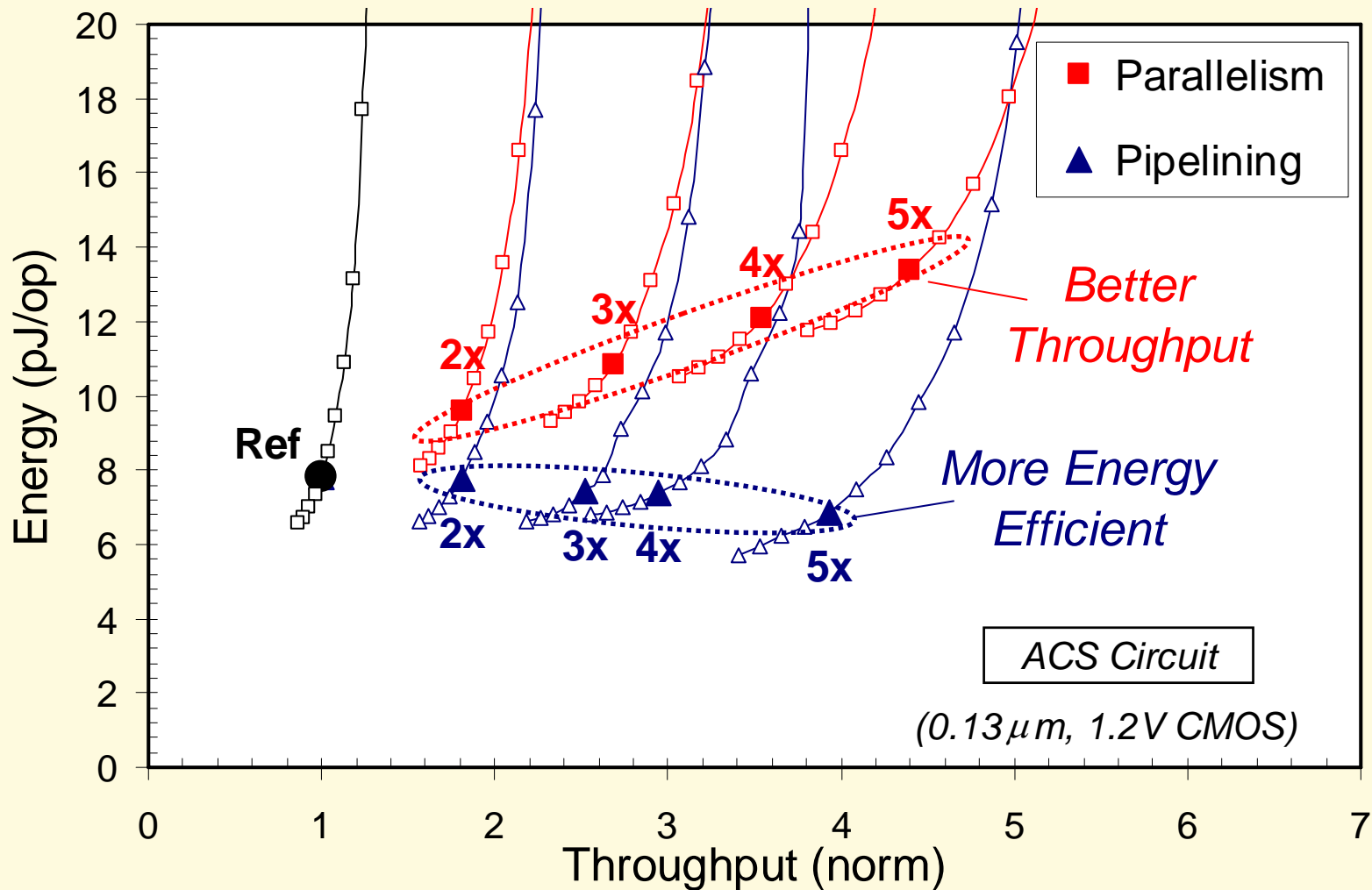
(c) Pipelining

- Feedback bus is typical
- Length = $128\mu\text{m}$

- $\cong N \cdot \text{area}$
- $\cong N \cdot \text{bus length}$
- $1/N$ clock rate

- $\cong \text{area}$
- $\cong \text{bus length}$
- Same clock rate

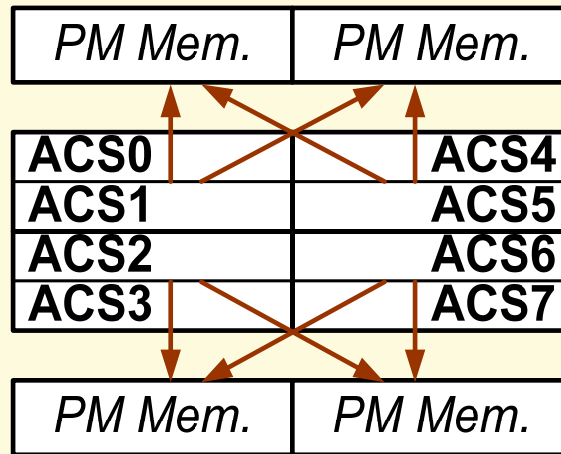
Energy-Throughput Comparison



- Pipelining is mostly more efficient in E-D domain!

ACS Unit Implementation

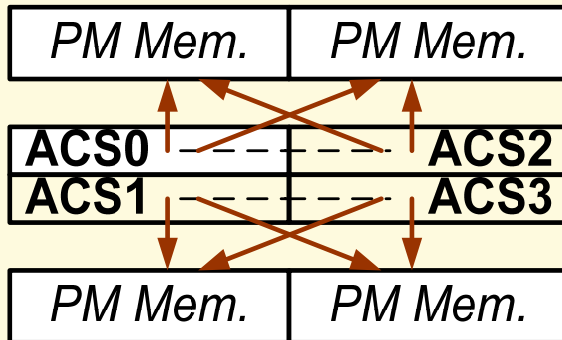
Application:
64-state rate- $\frac{1}{2}$
Viterbi decoder



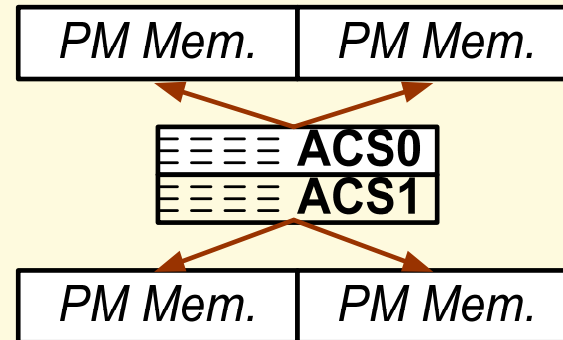
(a) Non-Pipelined

- ❖ Not pipelined
- ❖ 8 ACS circuits
- ❖ Largest area
- ❖ Least PM entries per ACS

- ❖ Pipelined-2
- ❖ 4 ACS circuits
- ❖ Midway area
- ❖ More PM entries / ACS



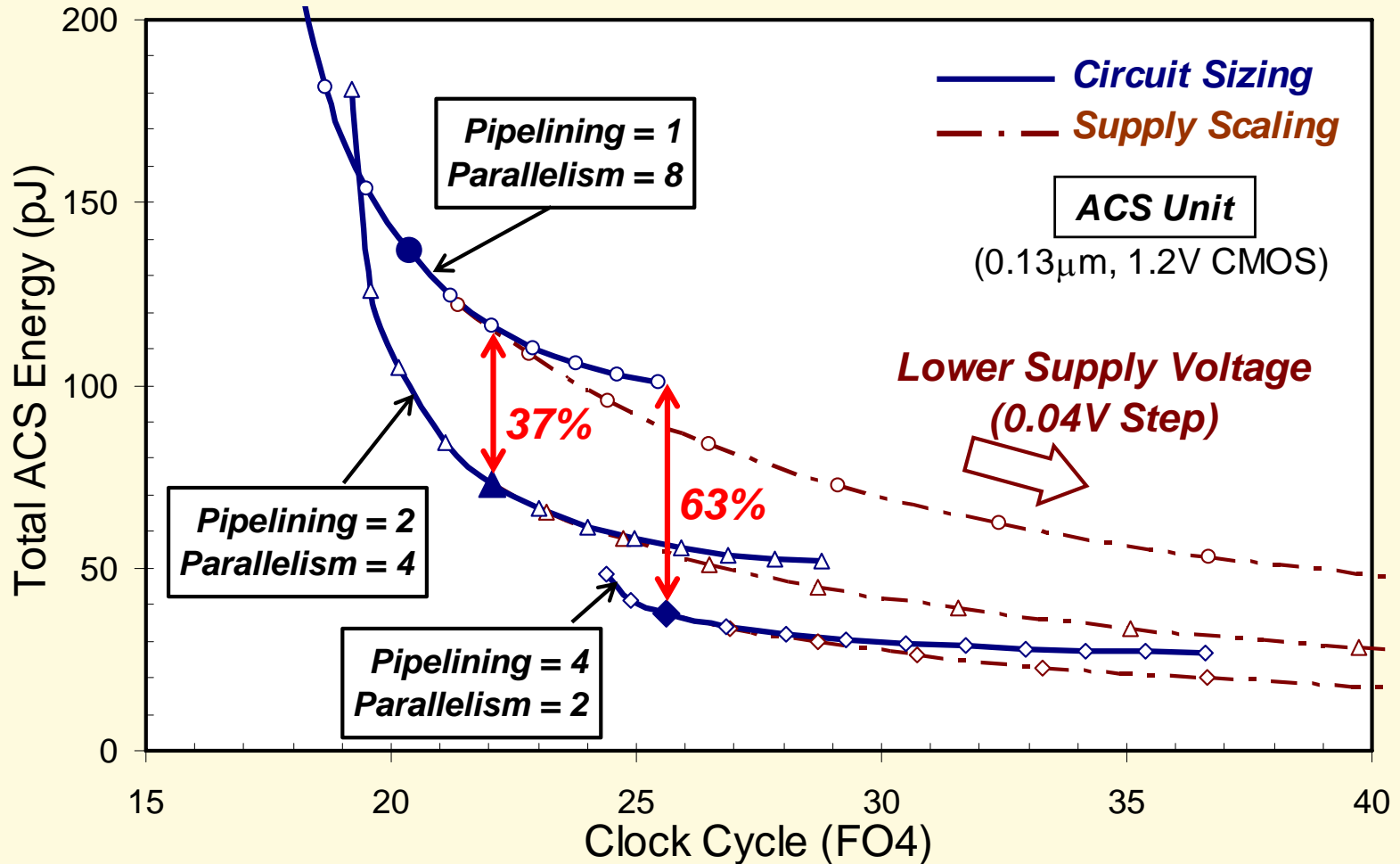
(b) Pipelined-2



(c) Pipelined-4

- ❖ Pipelined-4
- ❖ 2 ACS circuits
- ❖ Small area
- ❖ Most PM entries / ACS

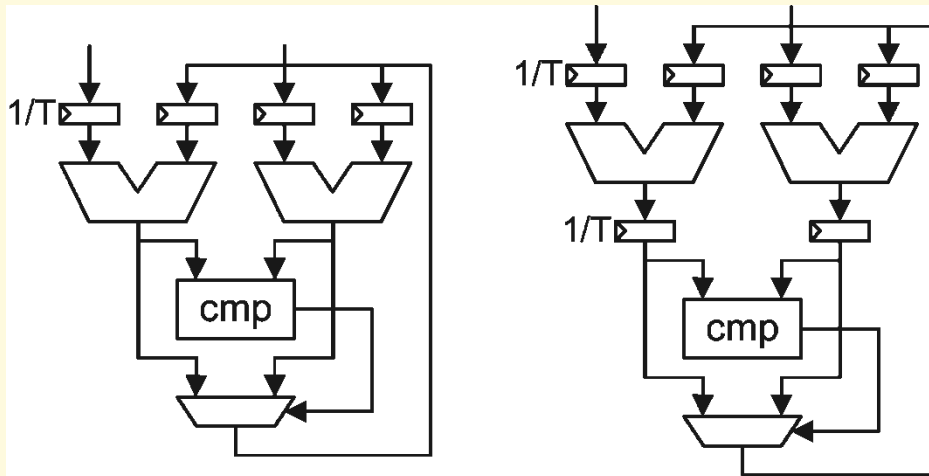
Energy-Throughput Comparison



- Less energy for deeper pipeline at given throughput

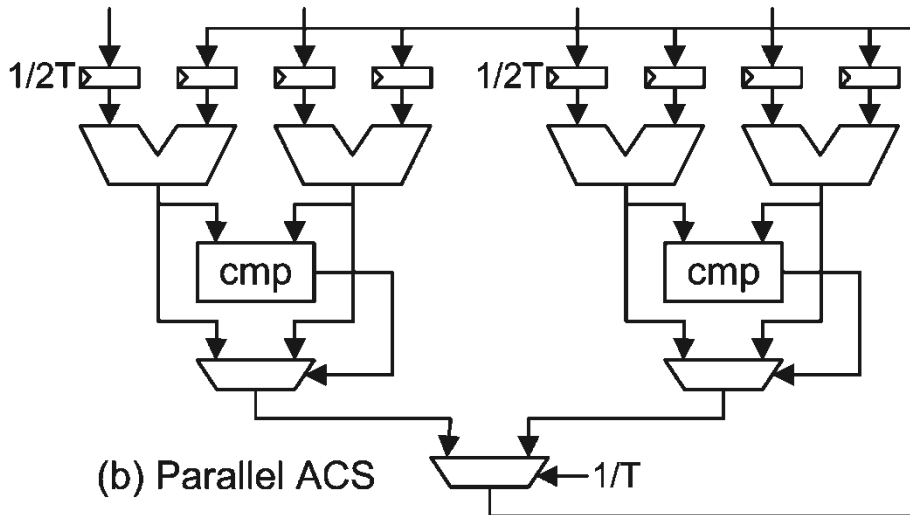
Designing a System for a Fixed Performance in the Energy-Delay Space

Pipelining and Parallelism for an ACS Circuit



(a) Single ACS

(c) Pipelined ACS



(b) Parallel ACS

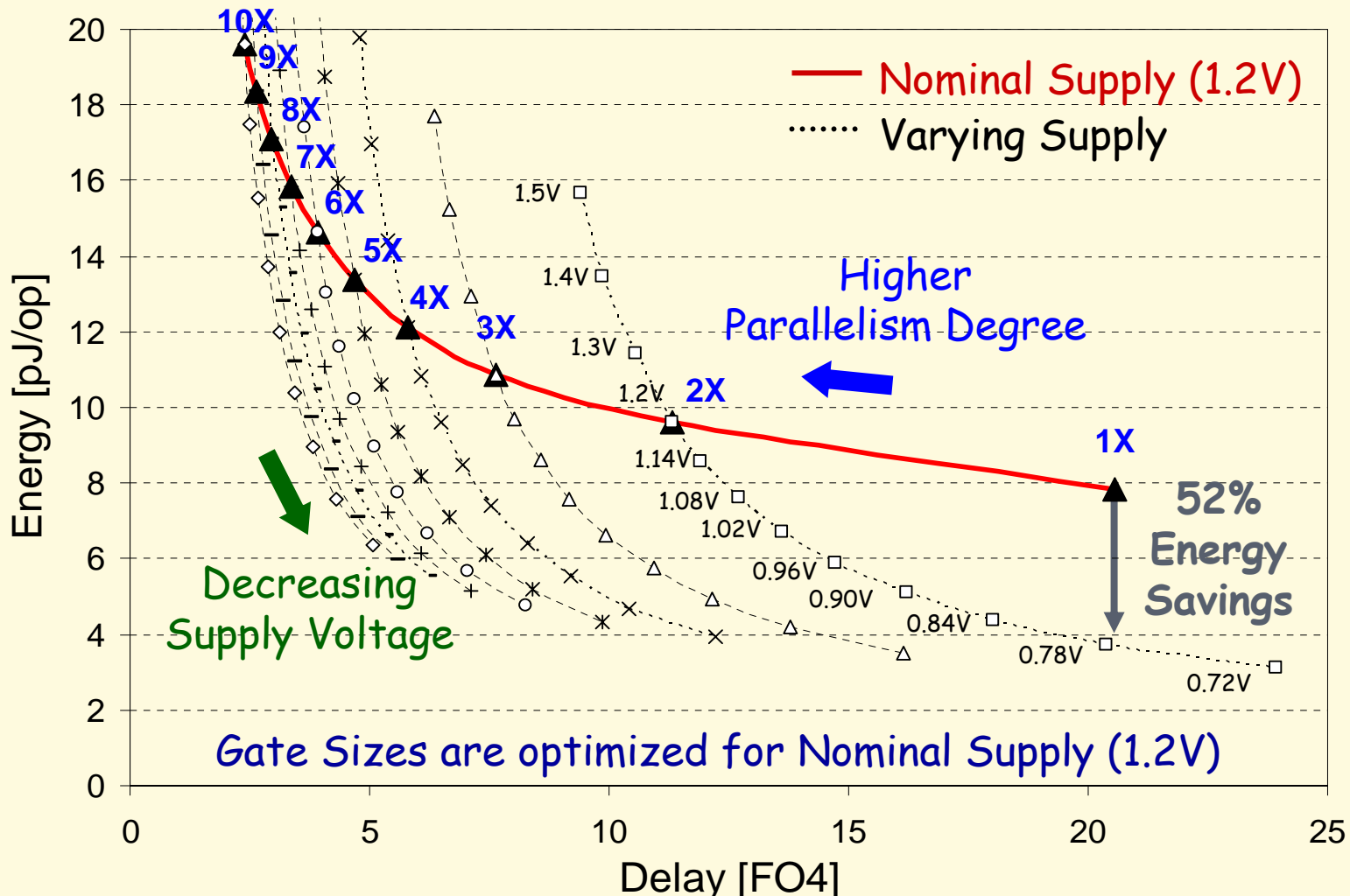
Ideal Degree-N Parallelism

- N-time throughput
- (Area, routing) overhead

Ideal Depth-N Pipelining

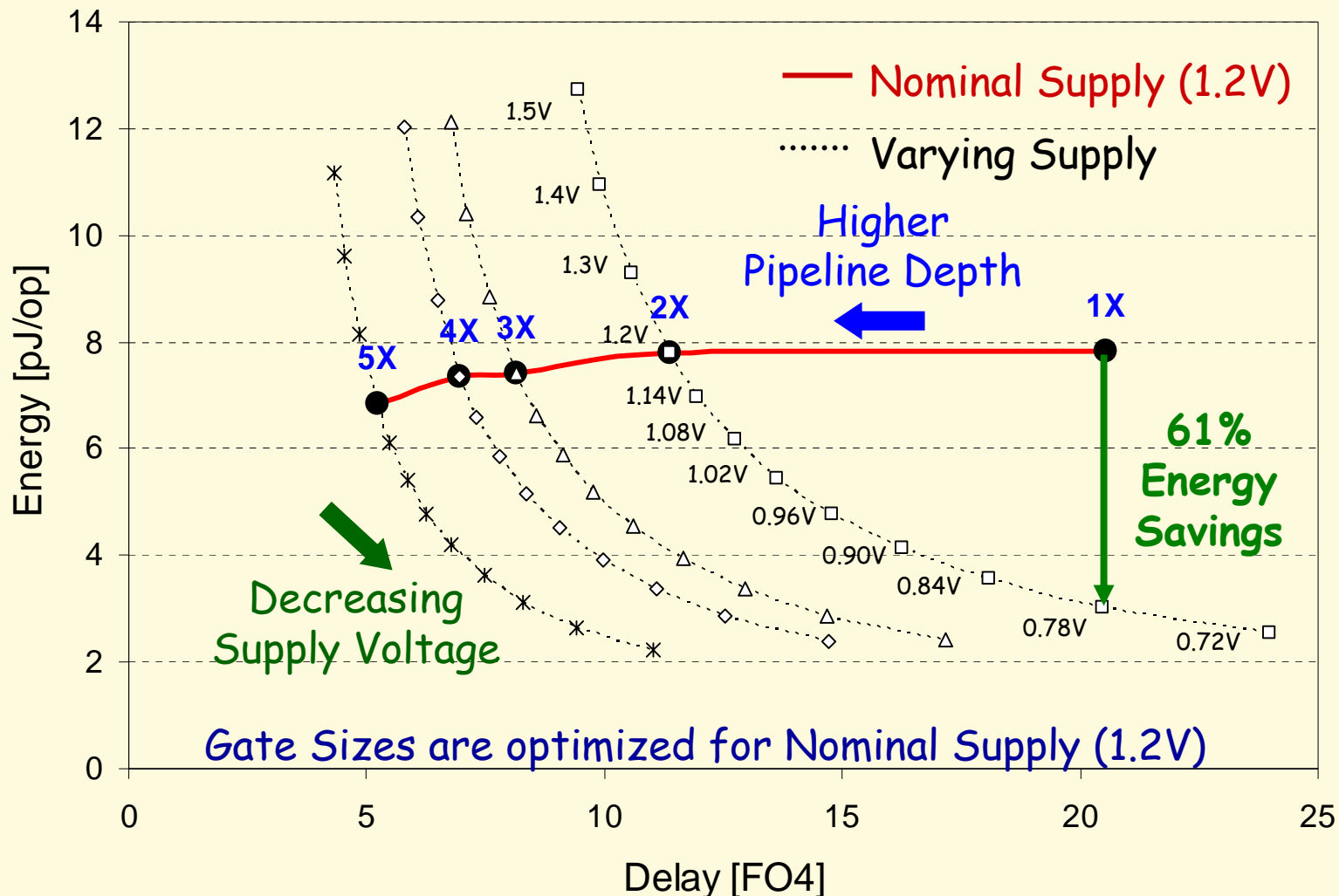
- N-time throughput
- CSE overhead

Energy-Delay Results for Parallelism



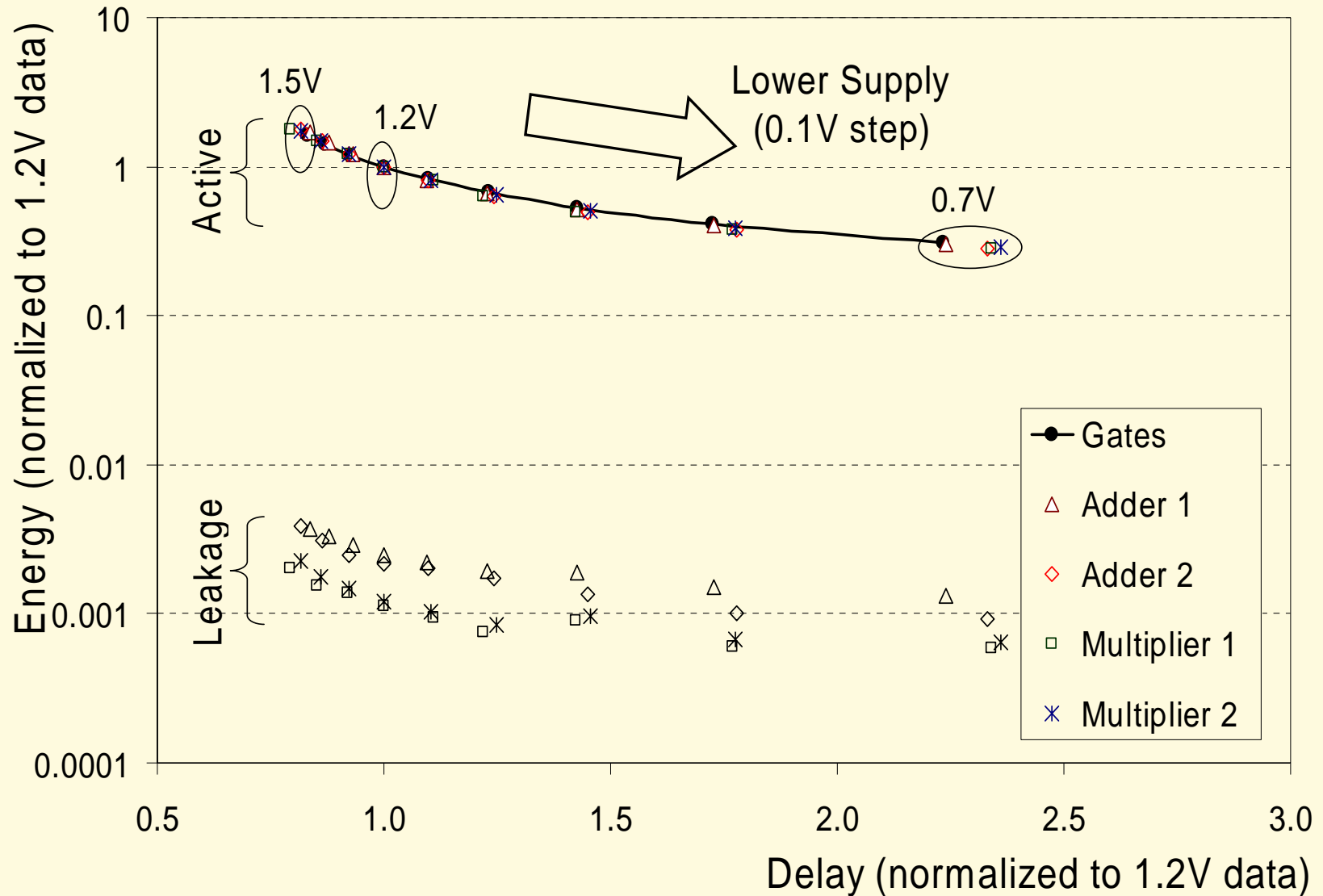
H. Q. Dao, B. R. Zeydel, V. G. Oklobdzija, "Energy-Efficient Optimization of the Viterbi ACS Unit Architecture", Proceedings of the Asian Solid-State Circuit Conference, A-SSCC 2005, Hsinchu, Taiwan, November 1-3, 2005.

Energy-Delay Results for Pipelining

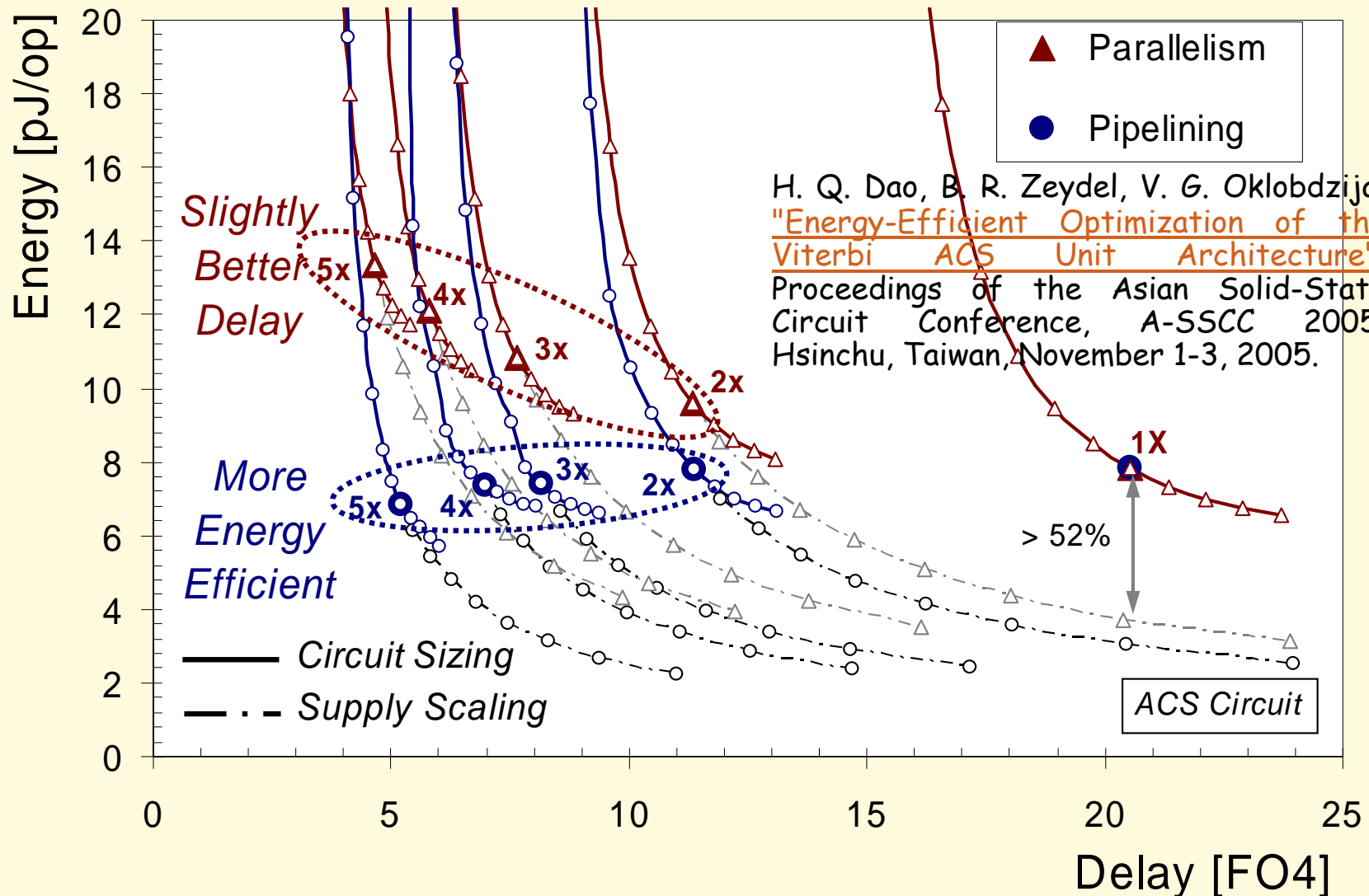


H. Q. Dao, B. R. Zeydel, V. G. Oklobdzija, "Energy-Efficient Optimization of the Viterbi ACS Unit Architecture", Proceedings of the Asian Solid-State Circuit Conference, A-SSCC 2005, Hsinchu, Taiwan, November 1-3, 2005.

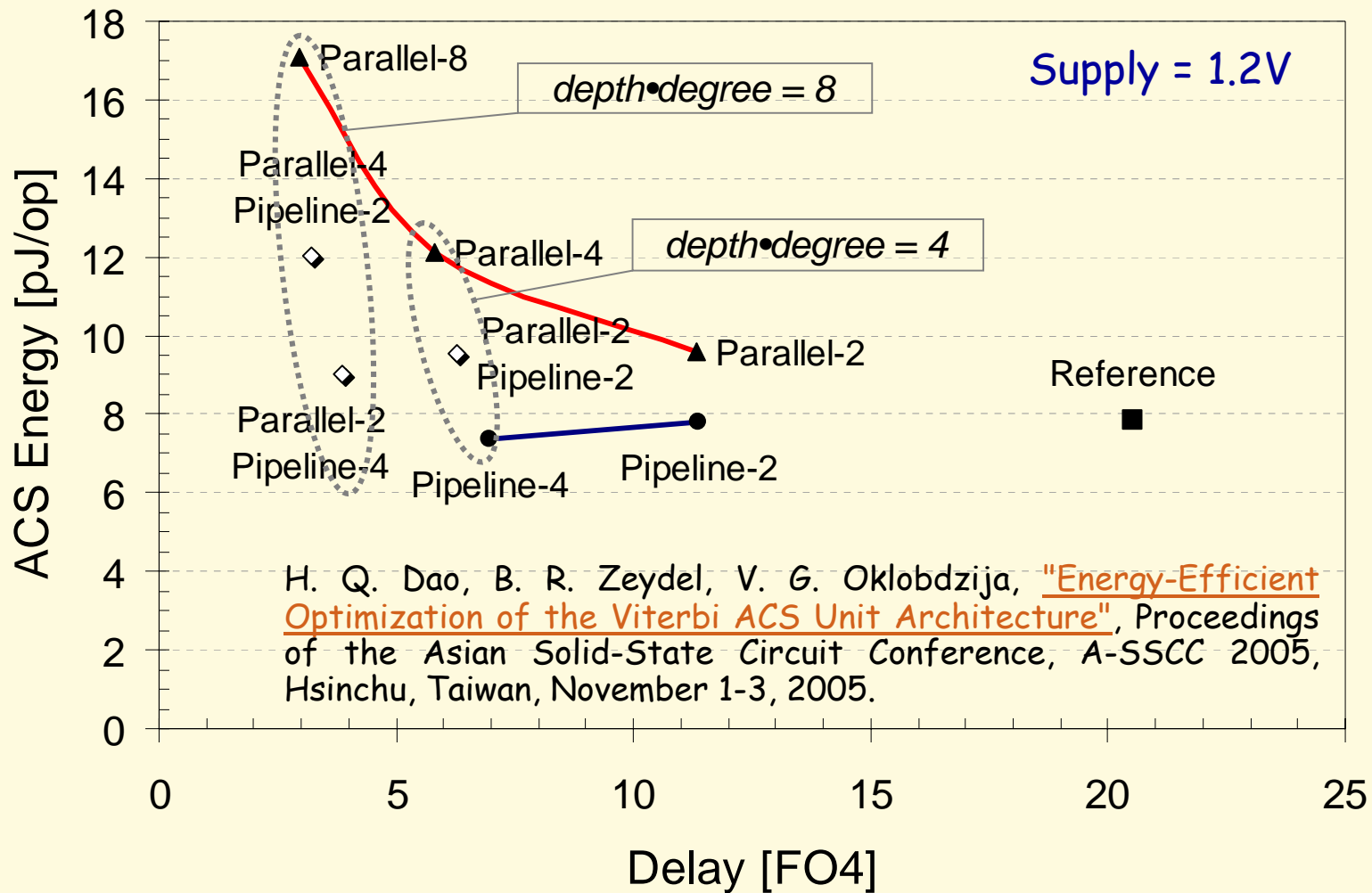
Energy-Delay Estimation Matches Complex Circuit Simulation



Energy Efficiency of Architectural Choices (including supply scaling and circuit sizing)

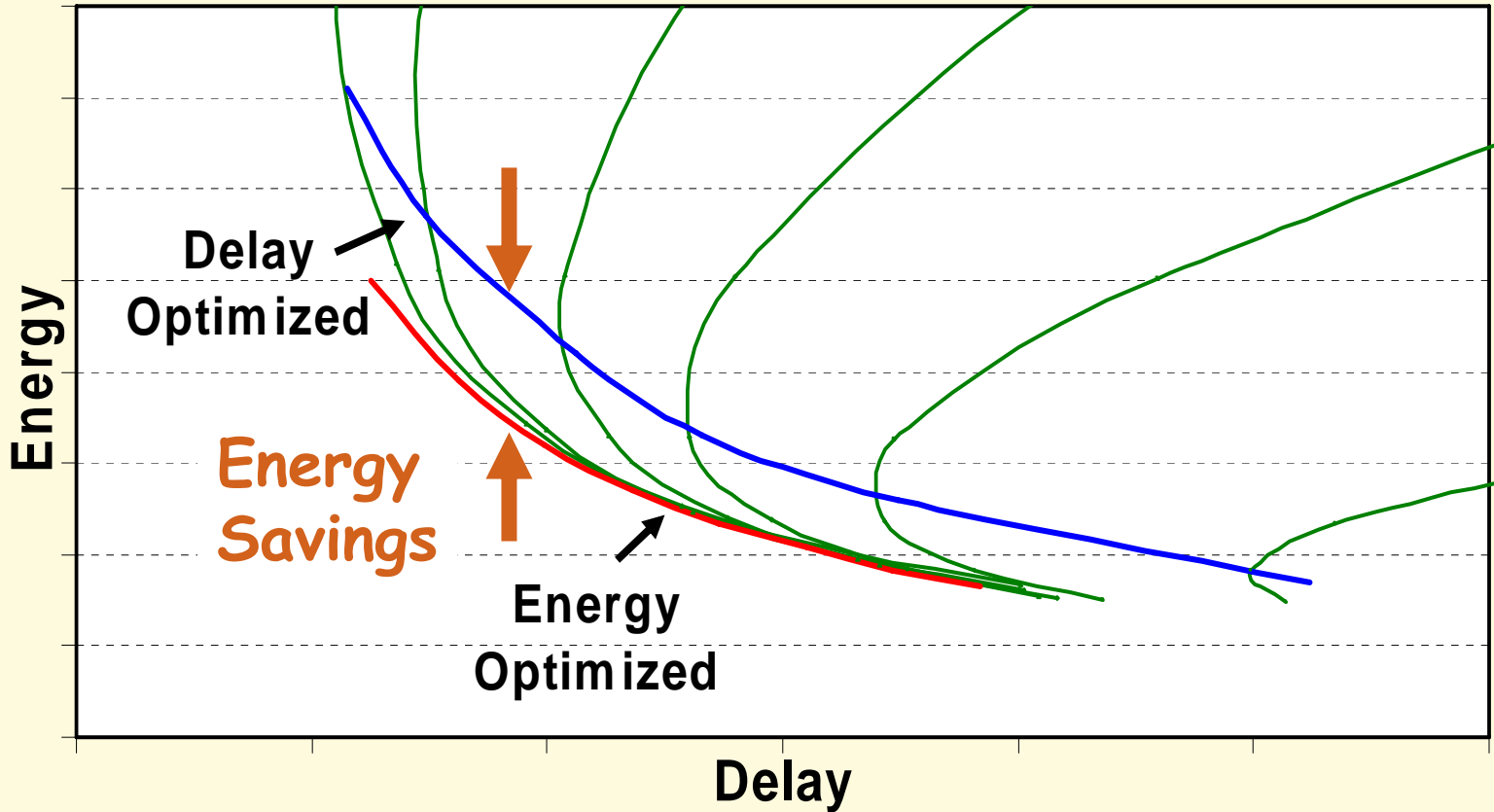


Energy-Delay Results for Parallelism and Pipelining



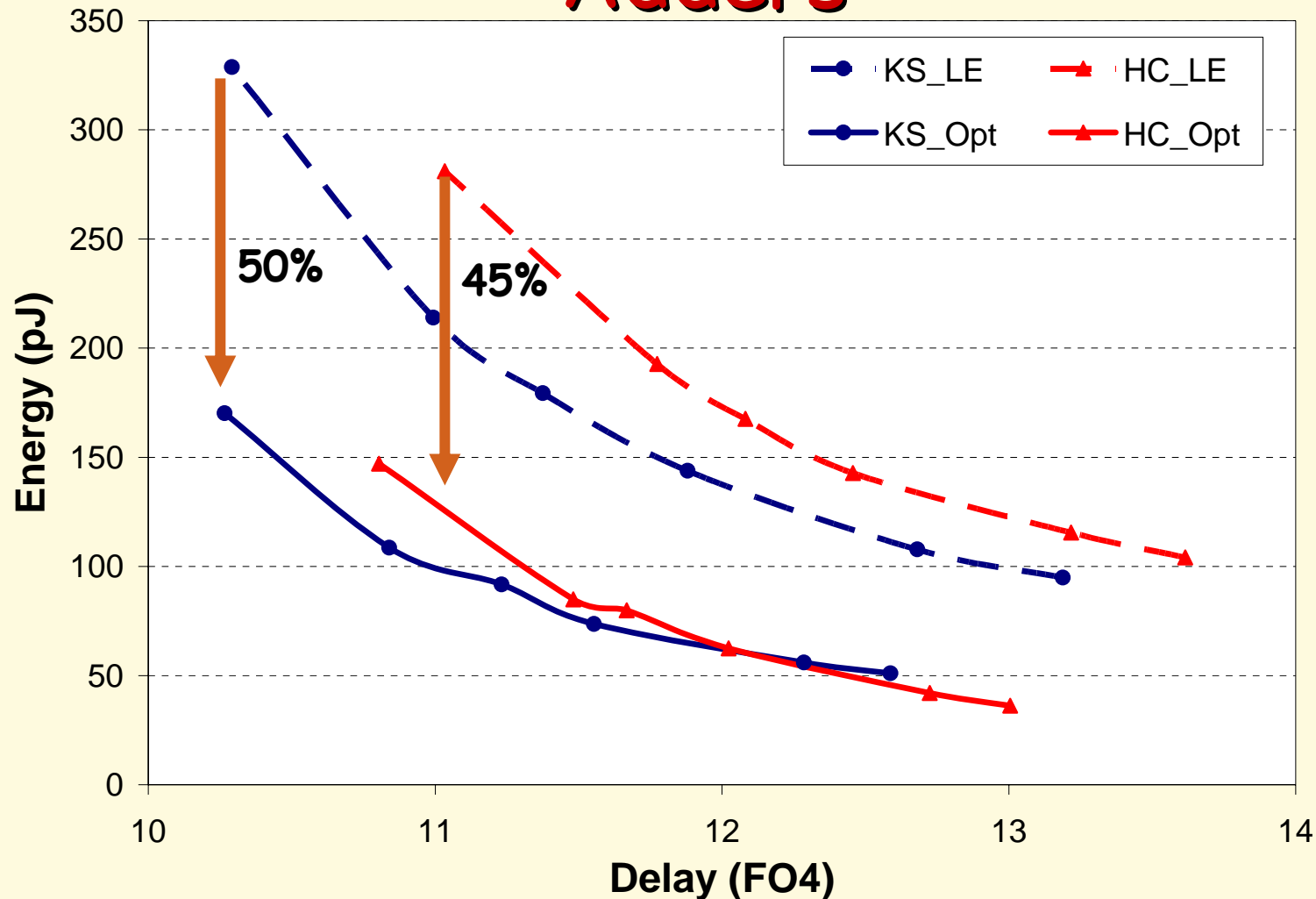
Pipelining has the Best Energy-Delay Tradeoff

Is it possible to lower the Energy ?



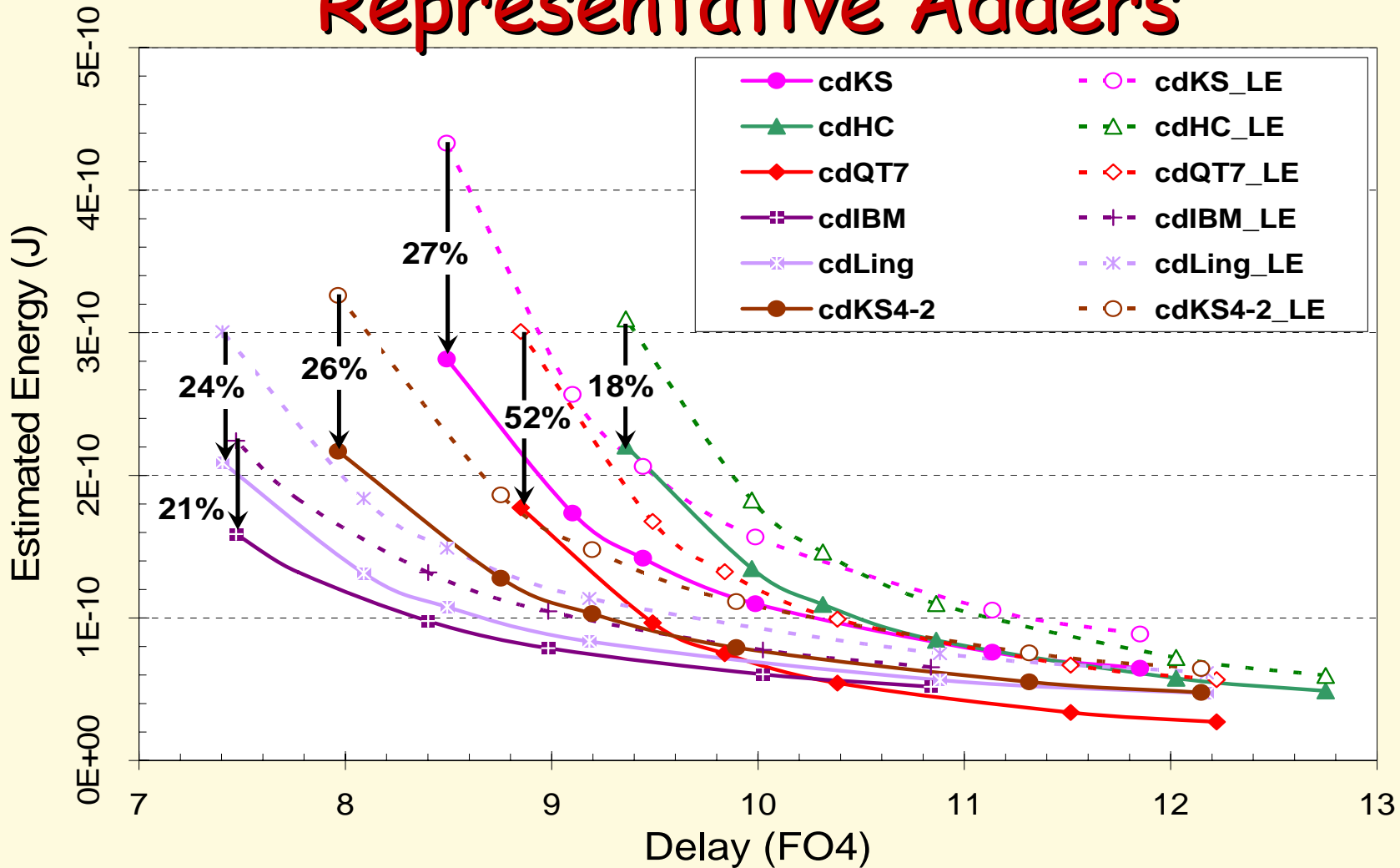
- Reduce Energy for same Delay!
- Improve Delay for same Energy!

Achieved Energy Savings in KS and HC Adders



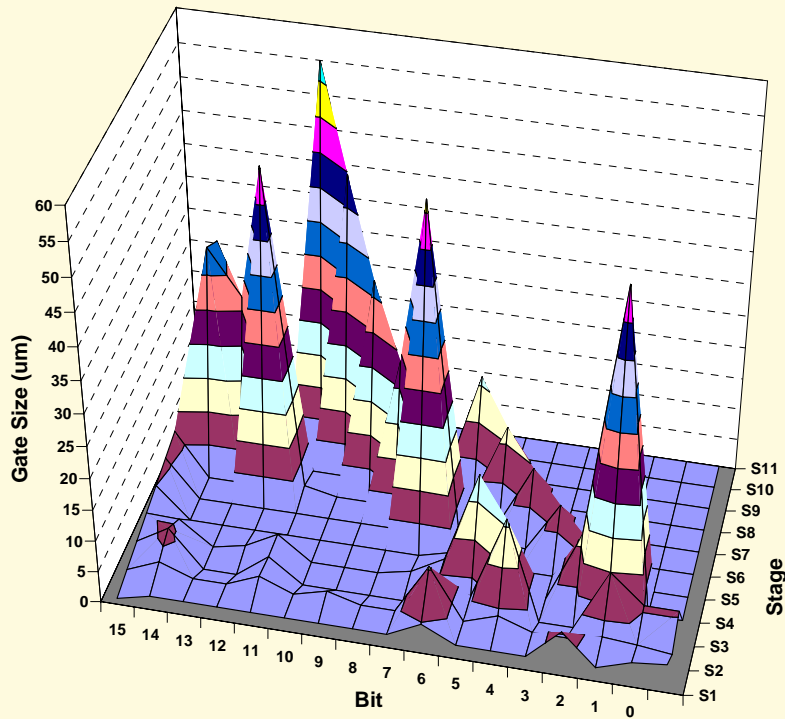
- Simulation of 64-bit static adders confirms saving!

Achieved Energy Savings in Representative Adders

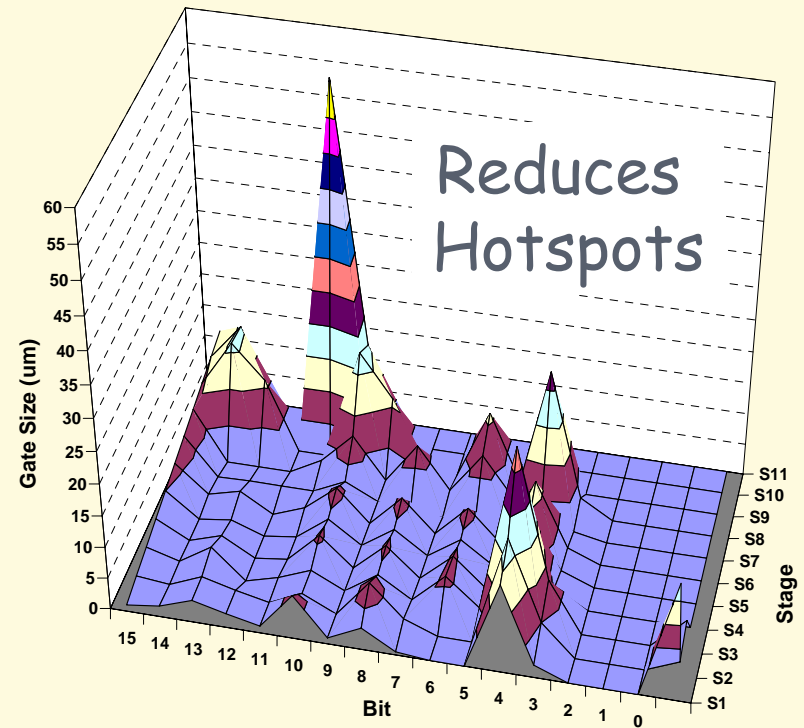


- Comparison of high-performance 64-bit adders
- Delay and Energy Optimized

Reduction of Hot-Spots



Delay Optimized

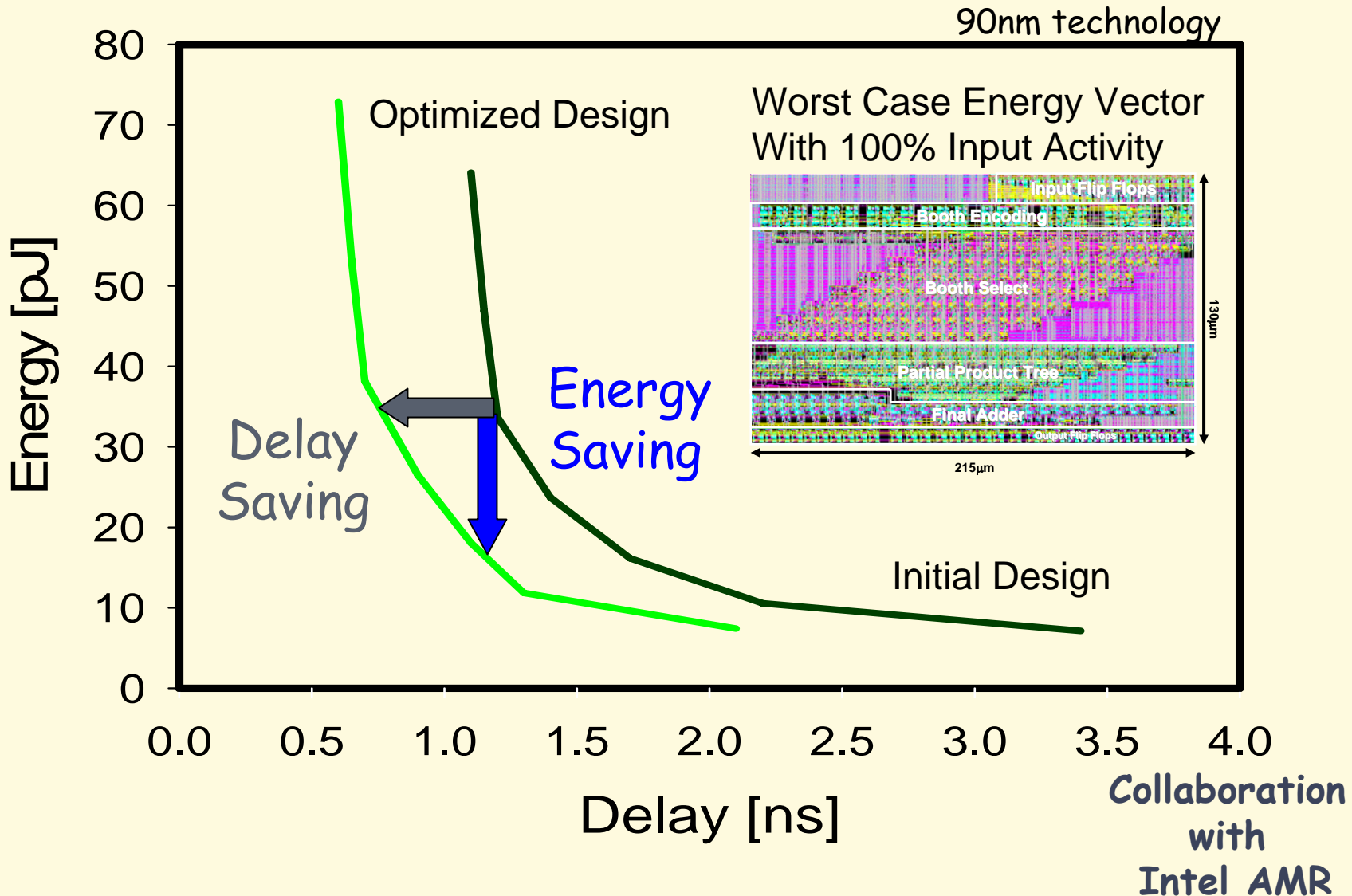


Energy Optimized

- Energy minimization improves hotspots!

Accomplishments

(June 30, 2003 to June 30, 2004)



Summary

- Energy-Efficient Design requires:
 - Early structure comparison in energy-delay space
 - Early Layout/Floorplanning
 - Optimization using energy minimization objective function
- LE does not guarantee a good design
- Our method of **energy-minimization** focuses on reducing power
- The same principles hold for other logic functions

Future Work

- Methodology improvement
 - General design rules
 - Algorithms with fast convergence
 - Guidelines for close-to-optimal solutions
- CAD tool
 - Custom vs. standard cell library
- Improve gate modeling & characterization
 - Worst-case vs. single-switching,... or between?
 - Process variation

Publications on Energy-Delay:

- Vojin G. Oklobdzija, Bart R. Zeydel, Hoang Dao, Sanu Mathew, Ram Krishnamurthy, "Energy-Delay Estimation Technique for High-Performance Microprocessor VLSI Adders", Proceedings of the International Symposium on Computer Arithmetic, ARITH-16, Santiago de Compostela, SPAIN, June 15-18, 2003.
- Hoang Q. Dao, Bart R. Zeydel, Vojin G. Oklobdzija, "Energy Minimization Method for Optimal Energy-Delay Extraction", Proceedings of the European Solid-State Circuits Conference, ESSCIRC 2003, Estoril, PORTUGAL, September 16-18, 2003.
- V. G. Oklobdzija, B. R. Zeydel, H. Q. Dao, S. Mathew, R. Krishnamurthy, "Comparison of High-Performance VLSI Adders in Energy-Delay Space", *IEEE Transaction on VLSI Systems*, Volume 13, Issue 6, pp. 754-758, June 2005.
- Hoang Q. Dao, Bart R. Zeydel, Victor Zyuban, and Vojin G. Oklobdzija, "On Energy Optimization of Digital Systems", The fourth annual IBM Austin Conference on Energy-Efficient Design, ACEED 2005, Austin, Texas, March 1-3, 2005.
- H. Q. Dao, B. R. Zeydel, V. G. Oklobdzija, "Energy-Efficient Optimization of the Viterbi ACS Unit Architecture", *Proceedings of the Asian Solid-State Circuit Conference, A-SSCC 2005, Hsinchu, Taiwan, November 1-3, 2005*.
- H. Q. Dao, B. R. Zeydel, V. G. Oklobdzija, "Energy Optimization of Pipelined Digital Systems Using Circuit Sizing and Supply Scaling", *IEEE Transaction on VLSI Systems*, Vol. 14, Issue 2, Feb. 2006 pp. 122-134.
- S. K. Hsu, S. K. Mathew, M. A. Anders, B. R. Zeydel, V. G. Oklobdzija, R. K. Krishnamurthy, S. Y. Borkar, "A 110 GOPS/W 16-bit Multiplier and Reconfigurable PLA Loop in 90-nm CMOS", *IEEE Journal of Solid-State Circuits*, Vol.41, No.1, January 2006.
- B. R. Zeydel, D. Baran, V. G. Oklobdzija, "Energy Efficient Design of High-Performance VLSI Adders", *IEEE Journal of Solid-State Circuits*, June 2010.