POWER – PERFORMANCE OPTIMIZATION METHODS FOR DIGITAL CIRCUITS

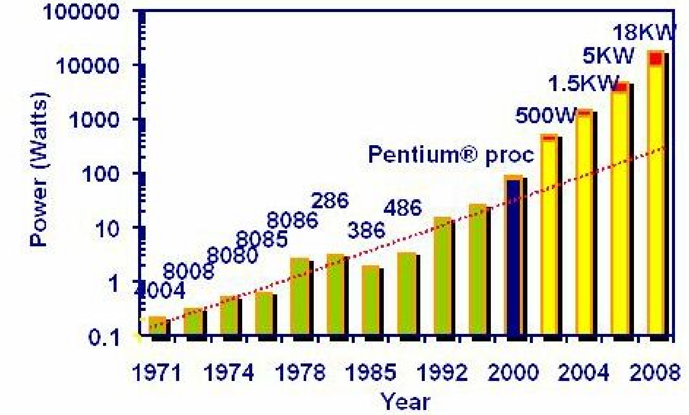
Radu Zlatanovici

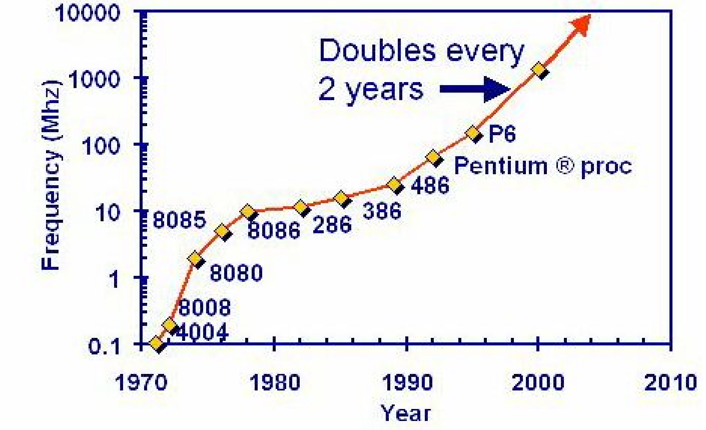
zradu@eecs.berkeley.edu
http://www.eecs.berkeley.edu/~zradu

Department of Electrical Engineering and Computer Sciences
University of California, Berkeley

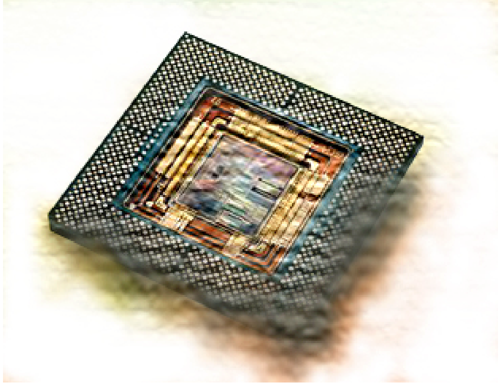
Power and Performance with Scaling

- **If we continue doing business as usual, both dynamic and leakage power will be a problem…**





…chips are getting hot…

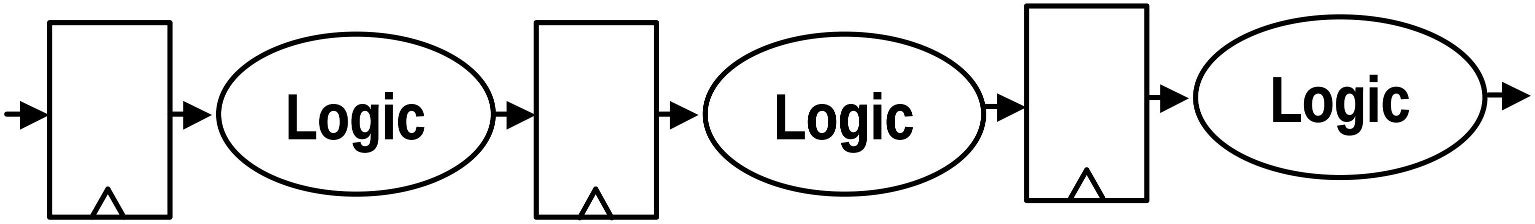


…and phones leaky!

- Need to deliver maximum performance under power constraints

From S. Borkar, Intel

A Common Problem



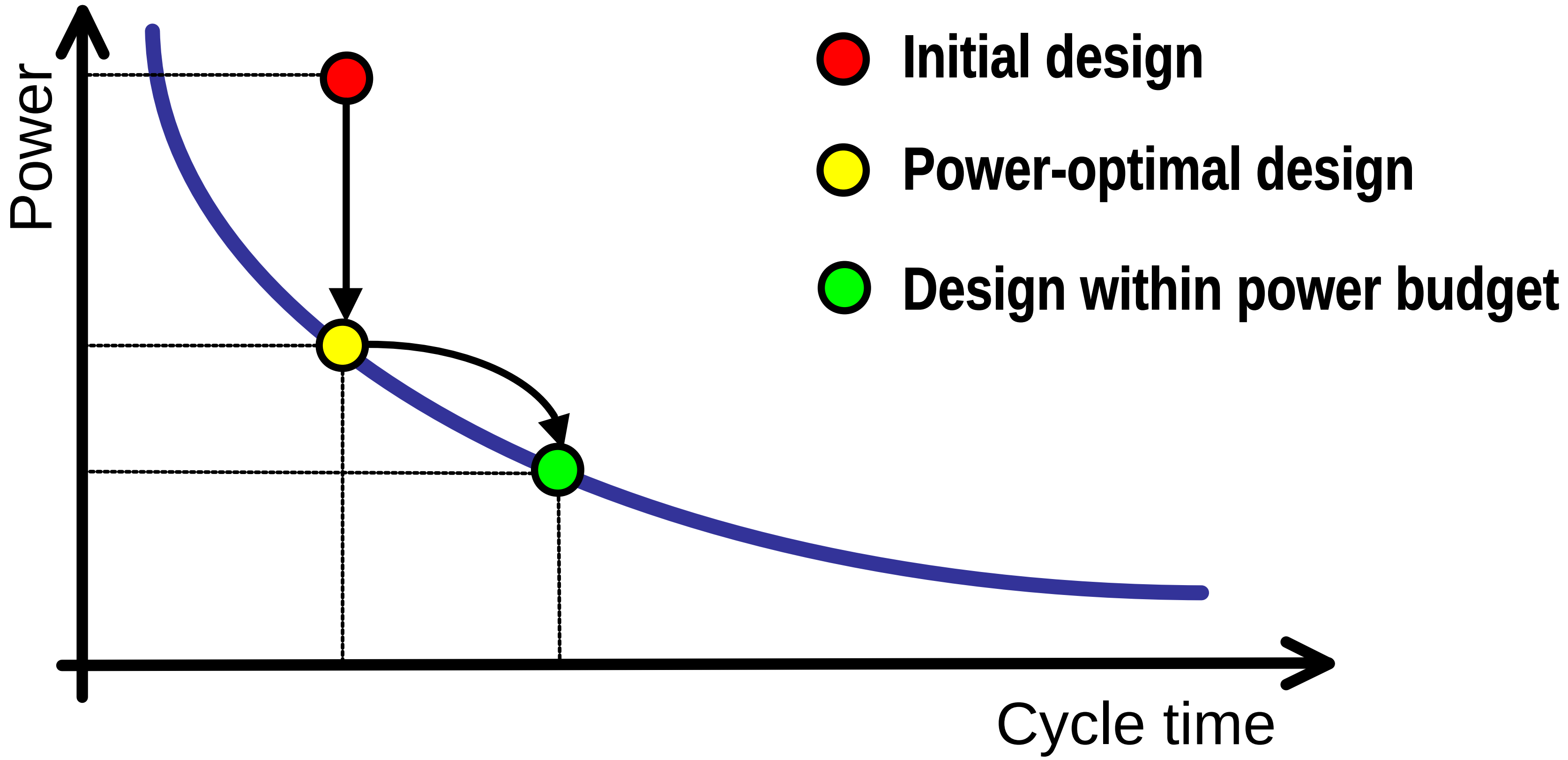
- Need to reduce power by 30%, while willing to give up 3% of performance
- What to do:
  - Decrease supply?
  - Increase thresholds?
  - Downsize?
  - Downsize latches or logic?
  - Use dual supplies?
  - Re-pipeline?
  - Parallelize?

Outline

- Design as a power – performance optimization problem
- Fundamentals of circuit optimization
- Design examples
- Dealing with variations
- Conclusions

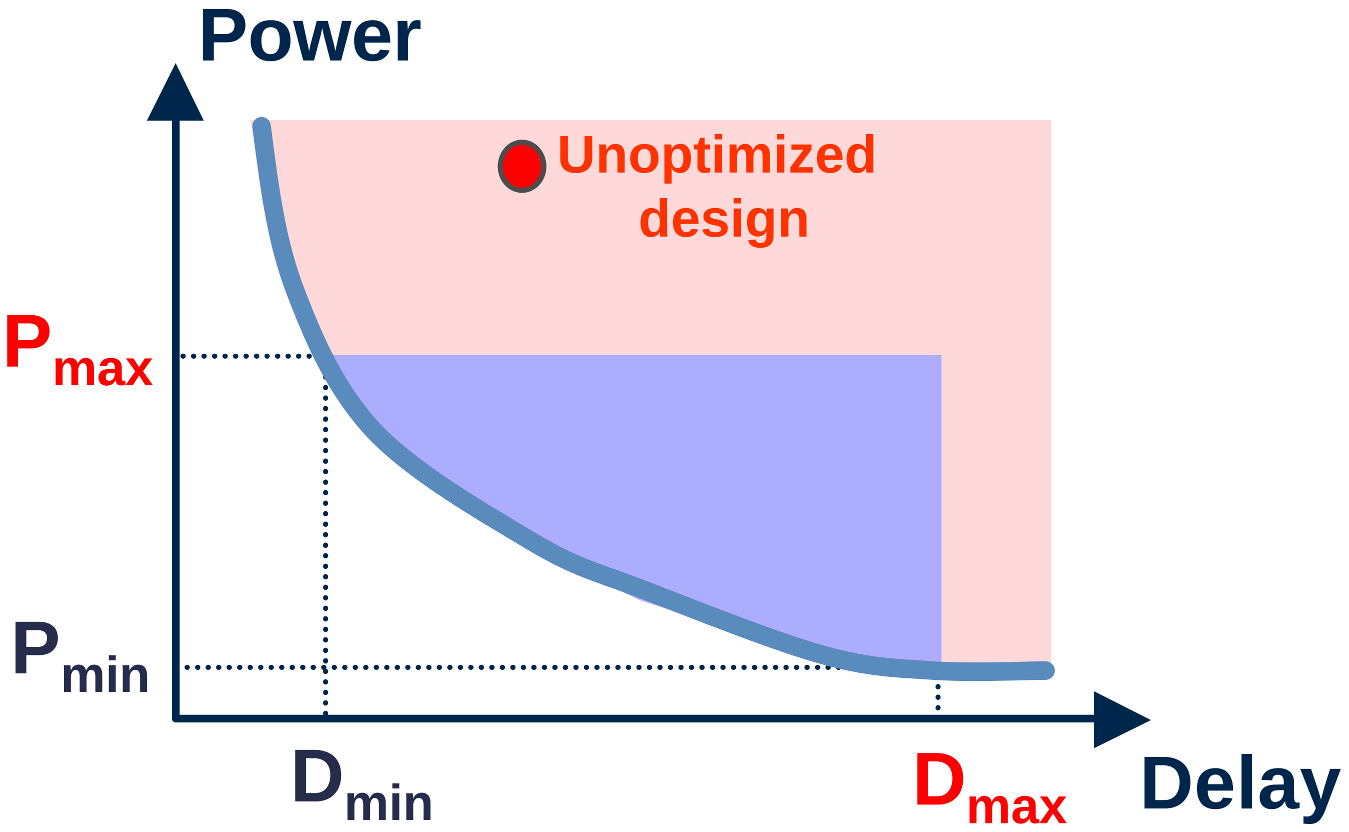
- **Collaborative effort:**
  - Students: R. Zlatanovici, D. Markovic, L.-T. Pang, J. Garrett, S. Kao
  - Faculty: B. Nikolic, R. Brodersen

Power – Performance Optimization



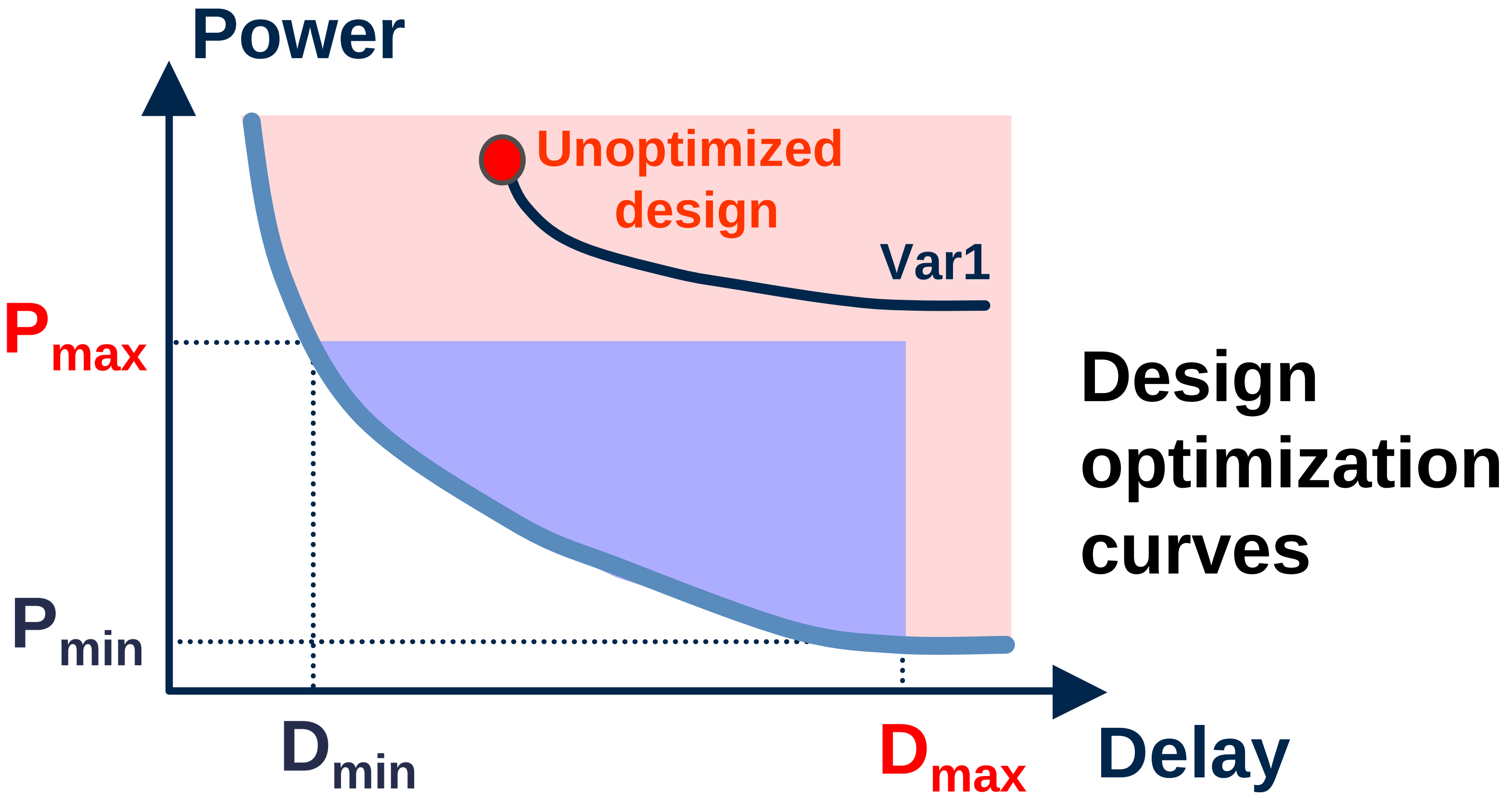
OPTIMAL POWER – PERFORMANCE TRADEOFF CURVE

Power Limited Operation



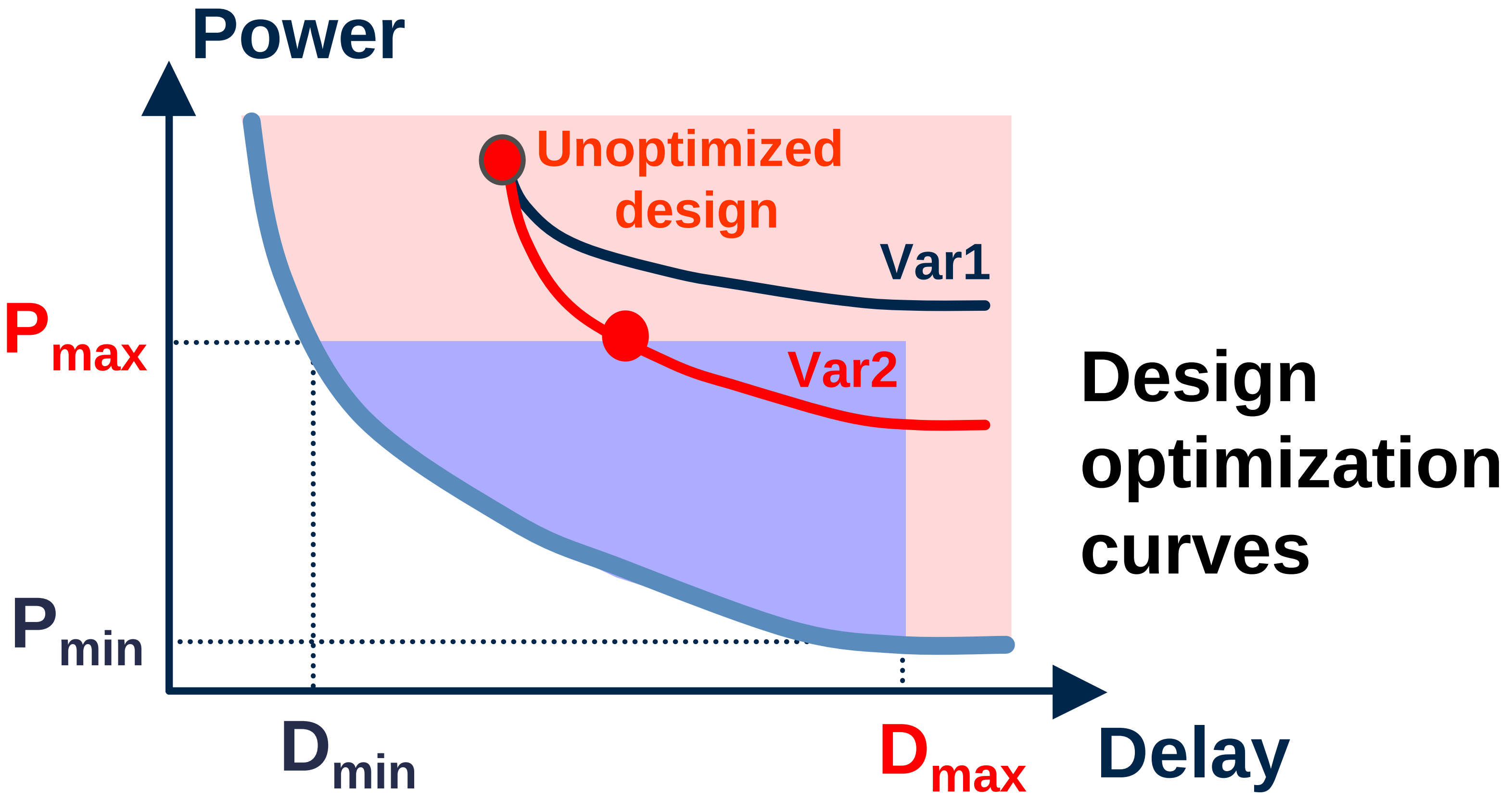
Achieve the highest performance under the power cap

Power Limited Operation



Achieve the highest performance under the power cap

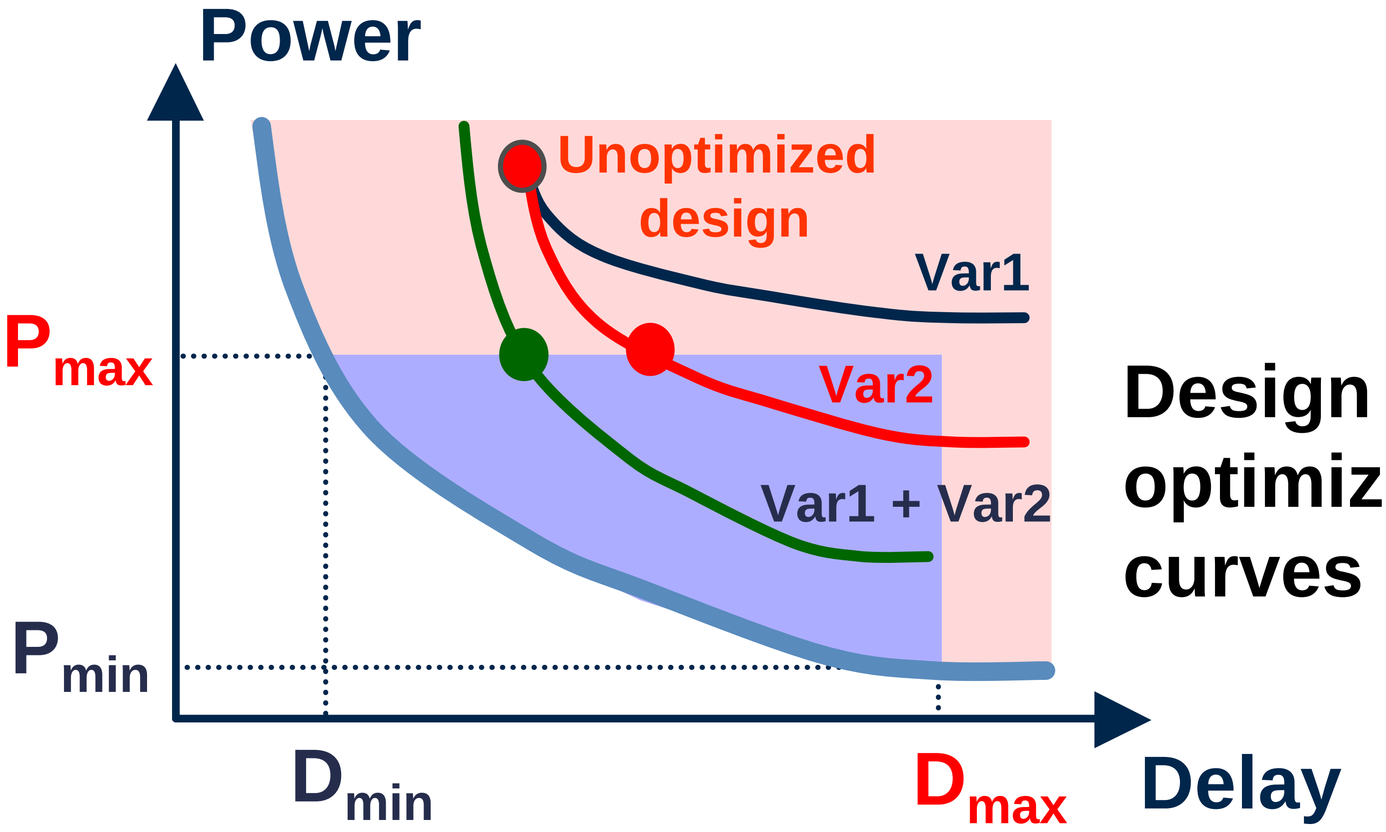
Power Limited Operation



Design optimization curves

Achieve the highest performance under the power cap

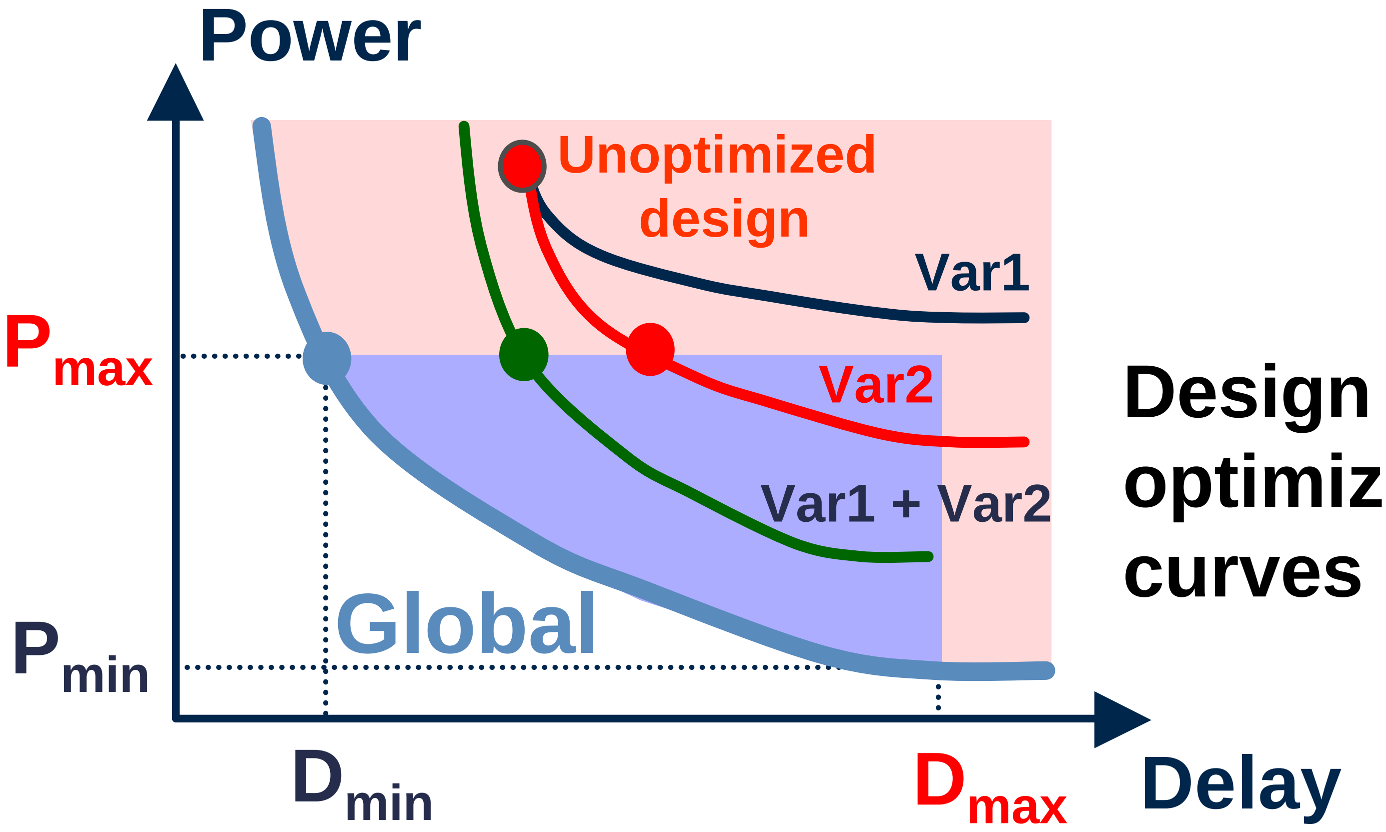
Power Limited Operation



Design optimization curves

How far away are we from the optimal solution?

Power Limited Operation

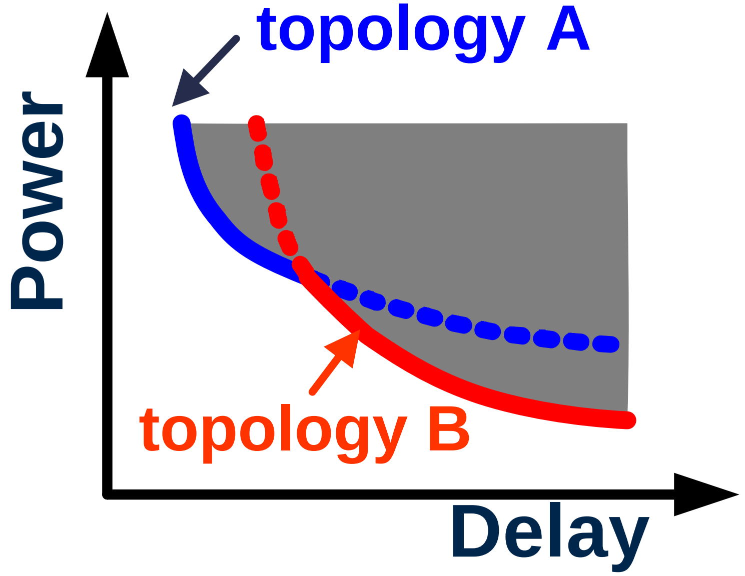


Design optimization curves

Global optimum – best performance

Design Optimization

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



Design Optimization

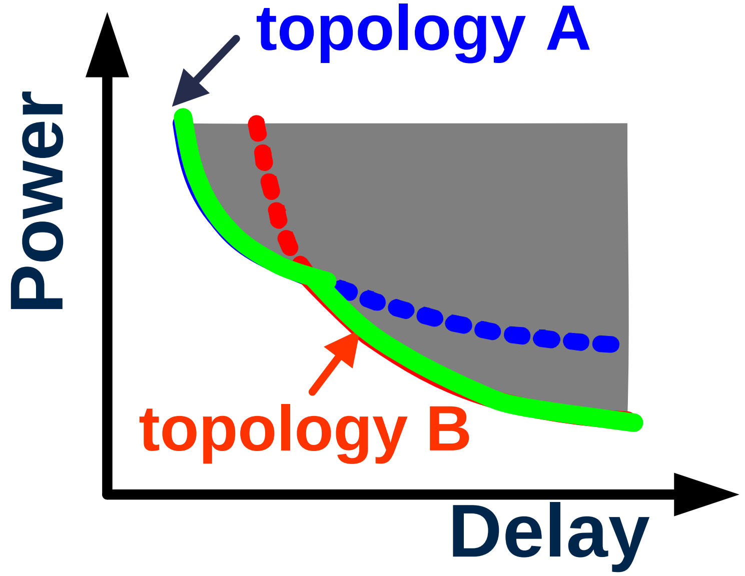
- **There are many sets of parameters to adjust:**
  - Circuit

    (sizing, supply, threshold)
  - Logic style

    (domino, static, pass-gate, …)
  - Block topology

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

    (parallel, pipelined)



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

Outline

- Design as a power – performance optimization problem
- **Fundamentals of circuit optimization**
- Design examples
- Dealing with variations
- Conclusions

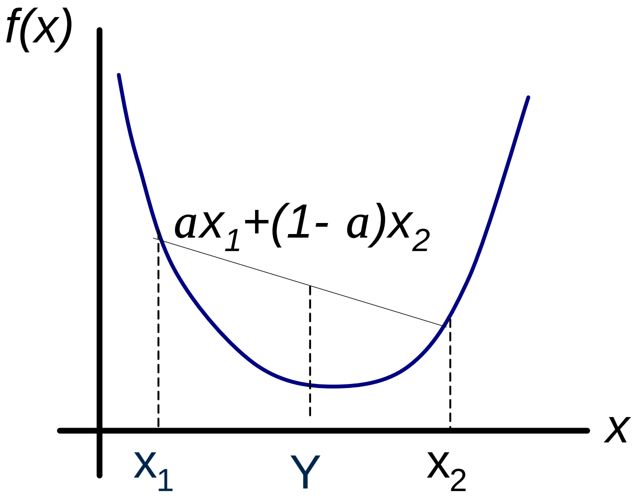
Optimization Problem

$$\min_{x \in R^n} f(x)$$

subject to

$$g_i(x) \leq 0 \quad i = 1..m$$

$$h_j(x) = 0 \quad j = 1..p$$

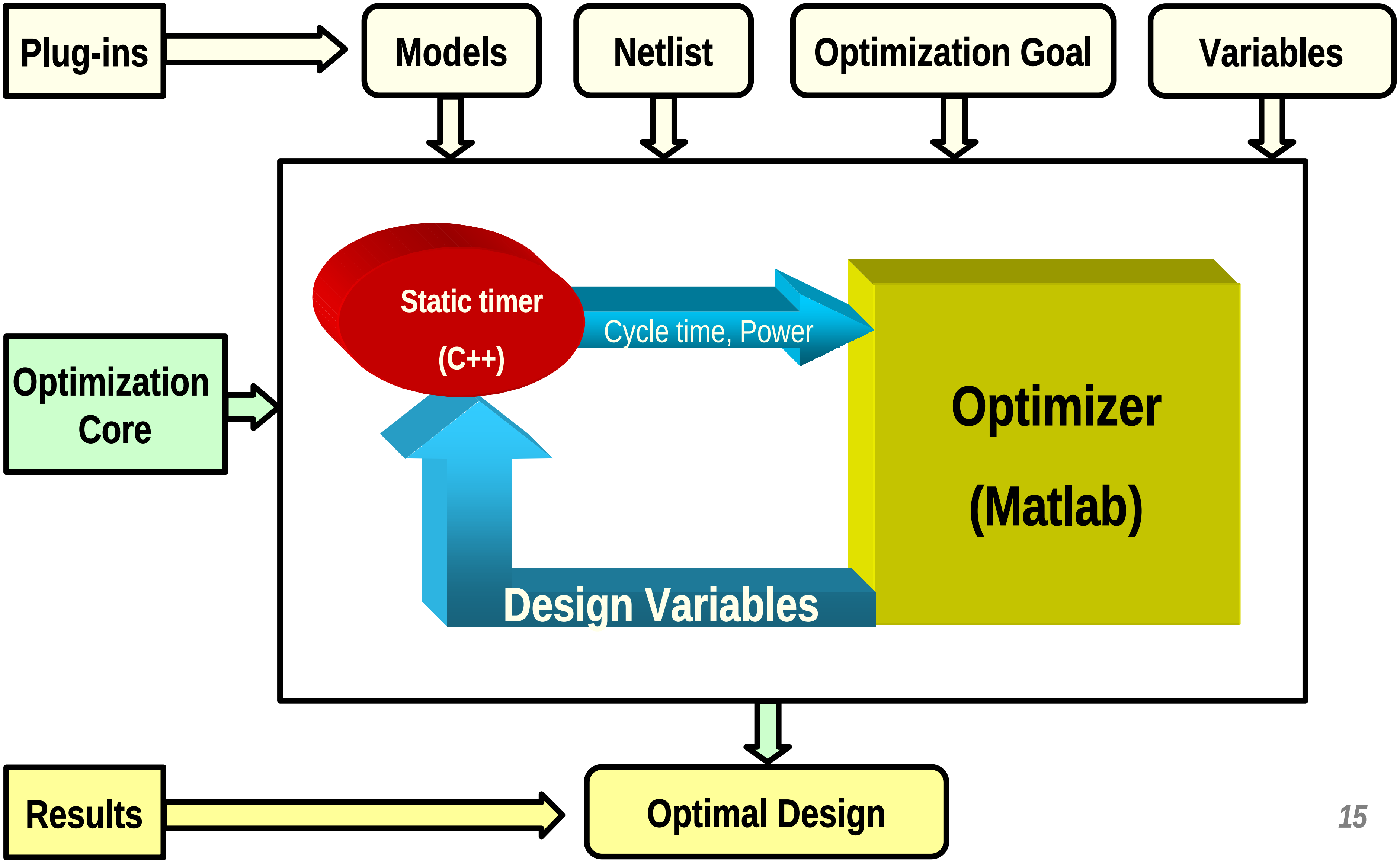


- Very difficult in the general case
- Optimality not guaranteed

Convex Optimization

- ***f***, ***g***$_i$ – convex, ***h***$_j$ – linear
- Key property: every local minimum is a global minimum
- Optimality guaranteed

Circuit Optimization



Plug-ins

Models

Netlist

Optimization Goal

Variables

Optimization Core

Static timer

(C++)

Cycle time, Power

Optimizer

(Matlab)

Design Variables

Results

Optimal Design

Power Constrained Optimization Problem

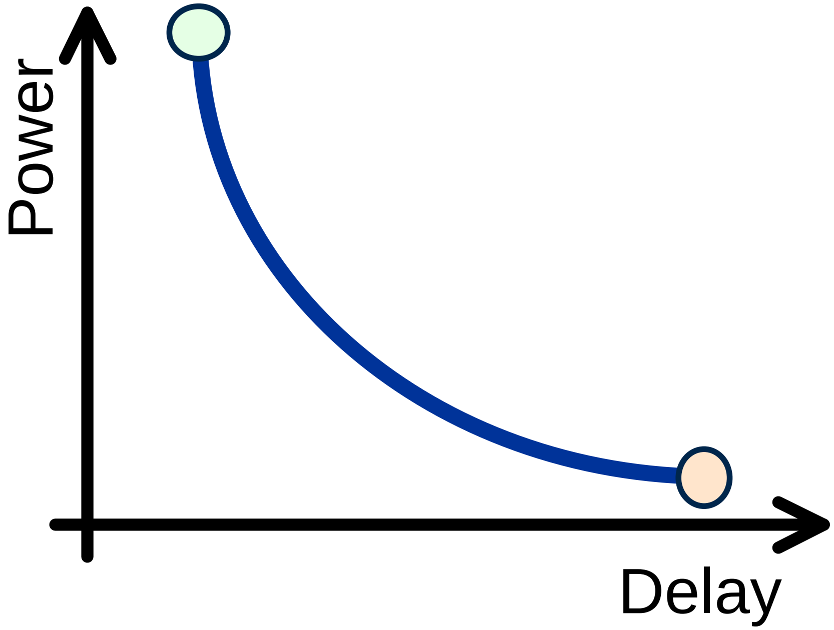
Minimize DELAY

subject to

Maximum POWER

Basic Result:

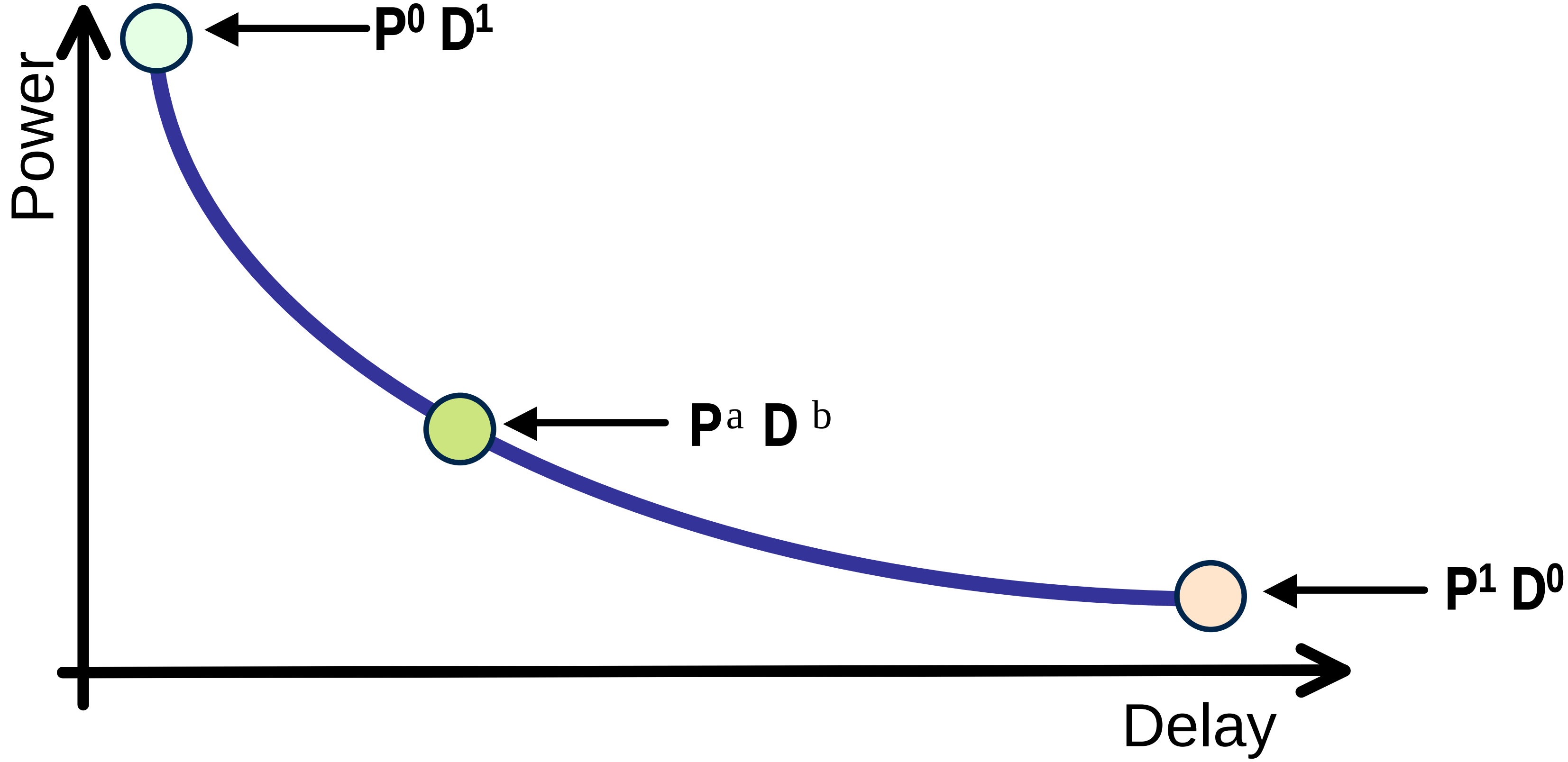
- Power - Performance tradeoff curve



Constraints:

- **Maximum output slew**
- **Maximum internal slew**
- **Maximum input capacitance**
- **Minimum sizes**

PD, PD^2, P^2D ... : Obsolete



ALL P-D METRICS ARE INCLUDED IN THE TRADEOFF CURVE

Choice of Models

ANALYTICAL

- - Limited accuracy
- + Fast parameter extraction
- + Provide insight in the operation of the circuit
- + Can exploit their mathematical properties to help optimization
  - Target: convex optimization

TABULATED

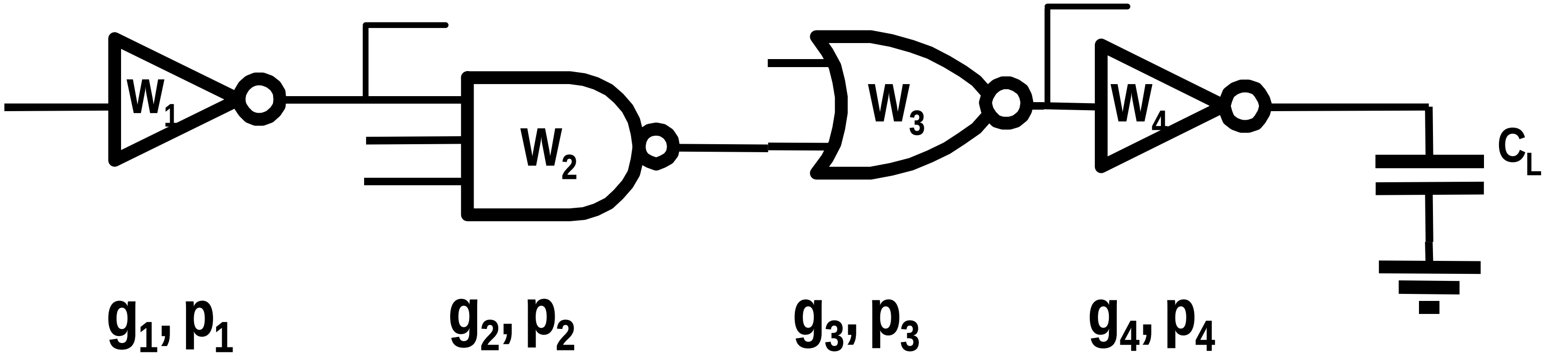
- + Very accurate
- - Slow to generate
- - No insight in the operation of the circuit
- - Can’t guarantee convexity
- - Optimization is “blind”
- ☞ If convex models are any good, the optimization problem is not very “non-convex”

Logical Effort: Delay of a Logic Path

$$t_D = p + g \cdot \frac{C_L}{C_{in}}$$

$$C_{in} = g \cdot C_{inv} \cdot W$$

$$b = \frac{C_{on-path} + C_{off-path}}{C_{on-path}}$$



$$D = \sum_{i=1}^{N} p_i + g_i \cdot \frac{C_{load,i}}{C_{in,i}} = \sum_{i=1}^{N} p_i + \sum_{i=1}^{N} g_i \cdot \frac{b_i g_{i+1} W_{i+1}}{g_i W_i}$$

Posynomial Functions

- **Definition of a posynomial:**

$$p(x) = \sum_{j} g_{j} \prod_{i=1}^{n} x_{i}^{a_{ij}} \quad a_{ij} \in R \;\; g_{j} \in R^{+}$$

- **A posynomial can be converted to a convex function using a simple change of variables: $x_i = e^{Z_i}$ for x_i ³ 0**
- **Logical effort delay is posynomial**
  - Fishburn, ICCAD ’85: Elmore delay formula can be written as a posynomial
- **Switching energy is linear in W_i ➔ posynomial**
- **Convex optimization with posynomials is called geometric programming**

V_{DD}- Dependent Analytical Delay Model

- **Gate equivalent resistance can be computed from analytical saturation current models (a reduced form of the BSIM3v3 equation)**

$$R_{EQ}=\frac{1}{V_{DD}/2}\int_{V_{DD}/2}^{V_{DD}}\frac{V_{DS}dV_{DS}}{I_{DSAT}}=\frac{3}{4}\frac{V_{DD}(b_1V_{DD}+b_0+V_{DD}-V_{TH})}{W\cdot K(V_{DD}-V_{TH})^2}(1-\frac{7V_{DD}}{9V_A})$$

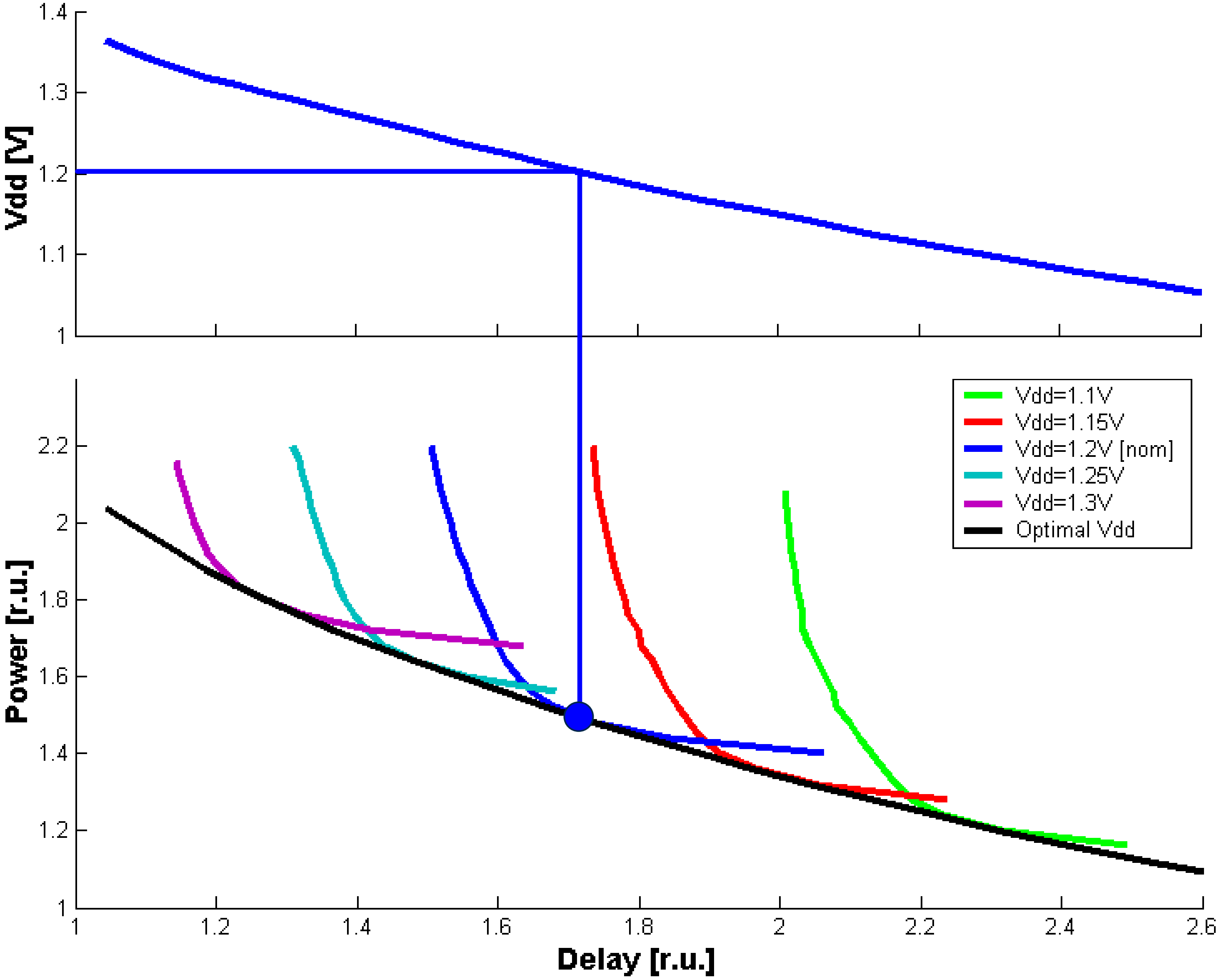
- **Include supply and threshold dependency in the delay model:**

$$d=c_2R_{EQ}(W,V_{DD},V_{TH})+c_1R_{EQ}(W,V_{DD},V_{TH})\cdot\frac{C_L}{C_{in}}+(h_0+h_1V_{DD})t_{s,in}$$

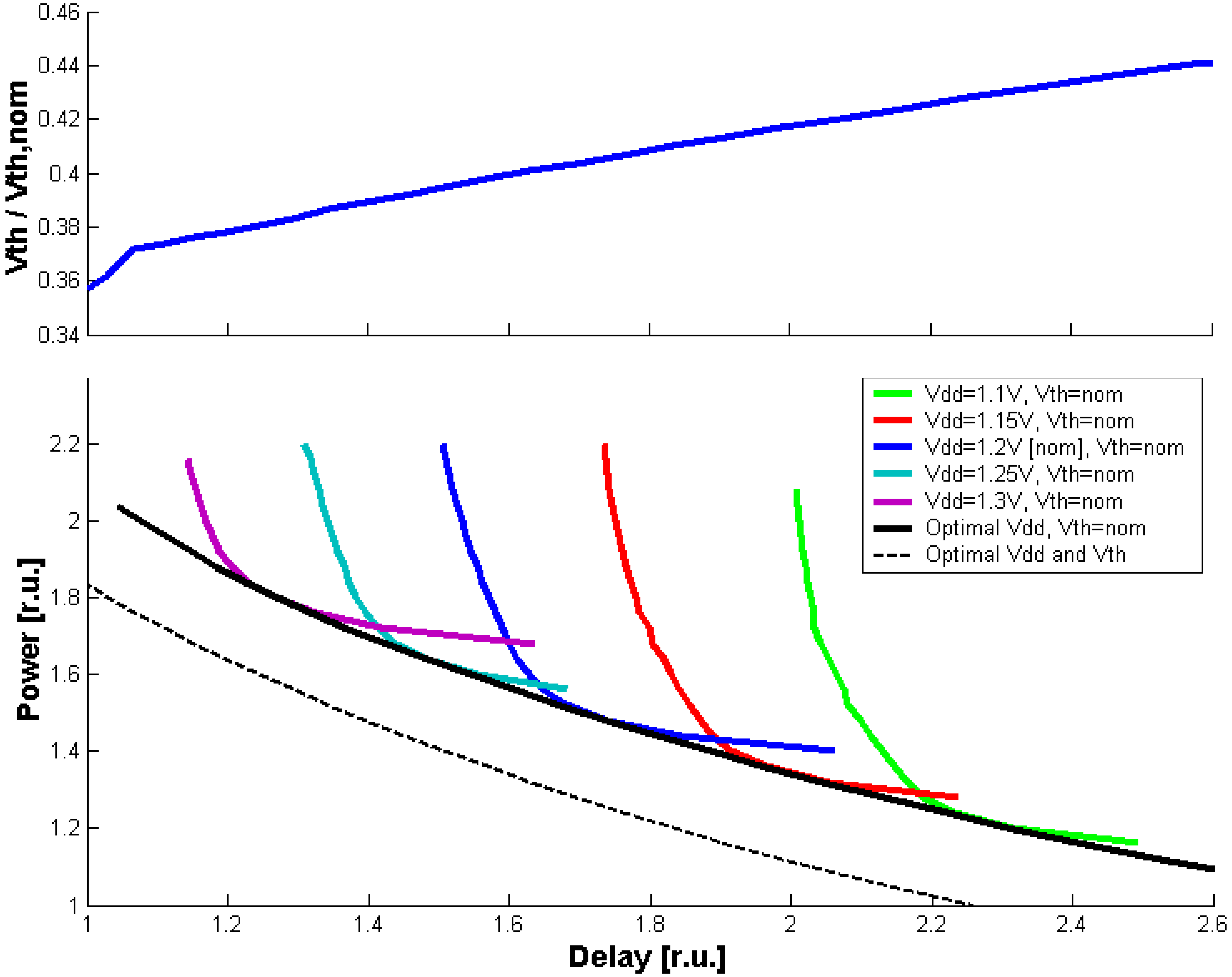
- **Accurate over a reasonable yet limited range of fanouts (2.5-6), supplies and threshold (+/- 30%)**
  - Most datapath blocks are within this range
- **Compatible with convex optimization**
  - Captures dependencies on V_{DD} and V_{TH} ➔ they can be optimization variables

Work by Joshua Garrett

Vdd and W Optimization



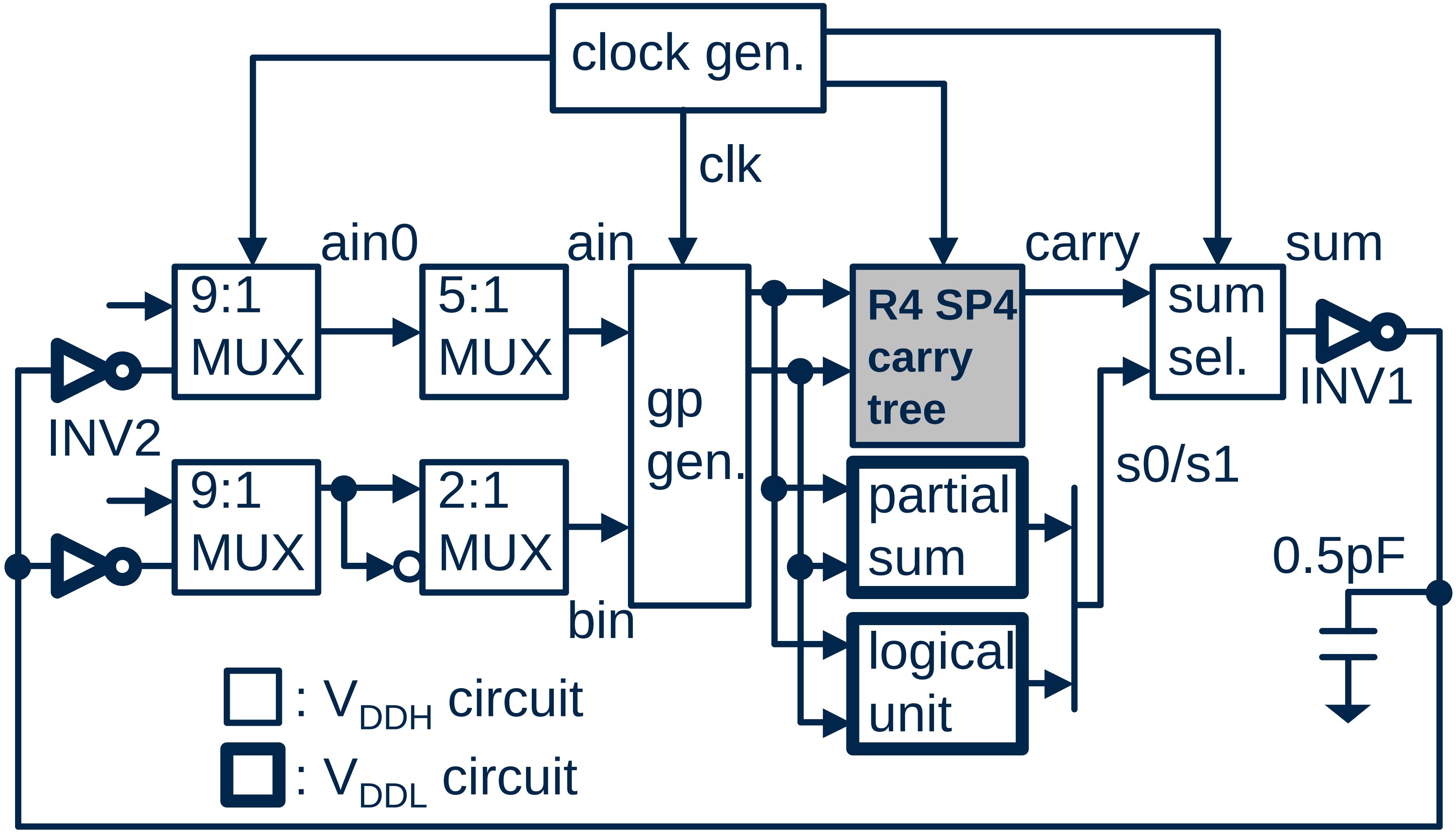
Vth,Vdd and W Optimization



Outline

- Design as a power – performance optimization problem
- Fundamentals of circuit optimization
- **Design examples**
- Dealing with variations
- Conclusions

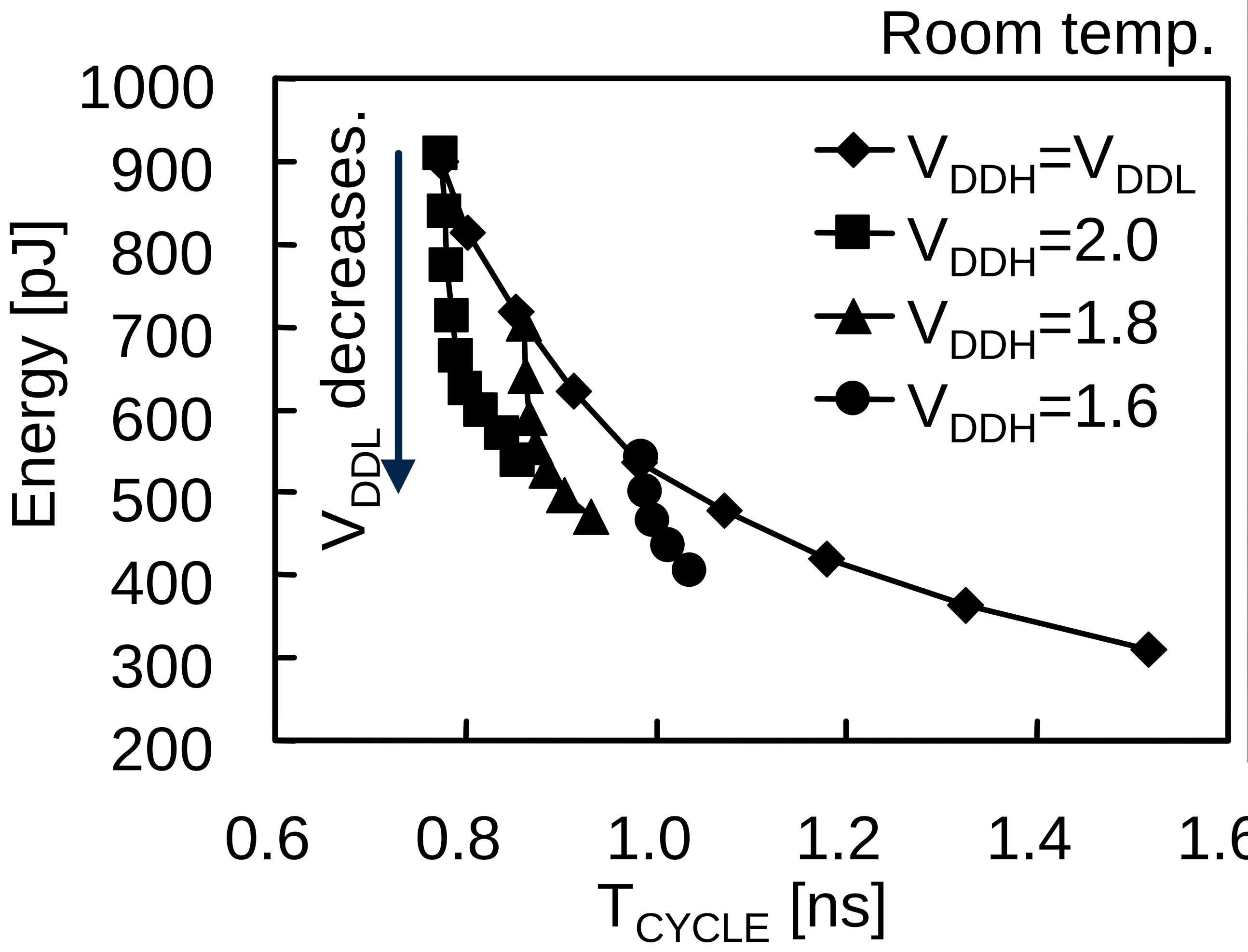
Dual V_{DD} ALU in Domino Logic

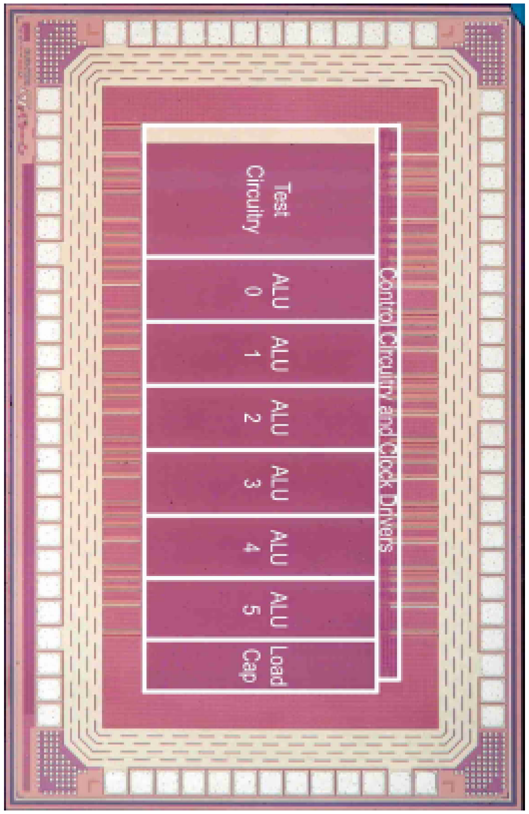


Y. Shimazaki, R. Zlatanovici, B. Nikolic: A Shared – Well Dual – Supply – Voltage 64-bit ALU

ISSCC’03, JSSC 03/2004

Extending the Space: Dual Supply





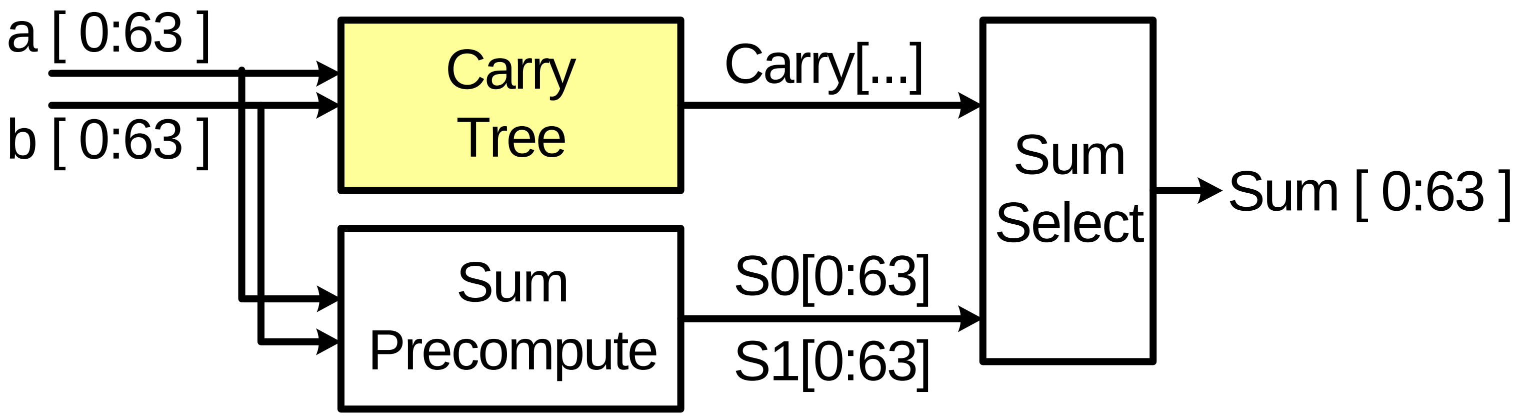
CLA Adders in E-D Space

- **Adders are common in critical paths**
- **CLA adders:**
  - Many designs, commonly used in practice
  - Recent interest in sparse adders
  - No fair comparison in energy – delay space
- **This work:**
  - Optimization of representative 64-bit adders in energy – delay space
  - Optimal 64-bit adder design

R. Zlatanovici, B. Nikolic: Power – Performance Optimal 64-Bit Carry-Lookahead Adders, ESSCIRC 2003

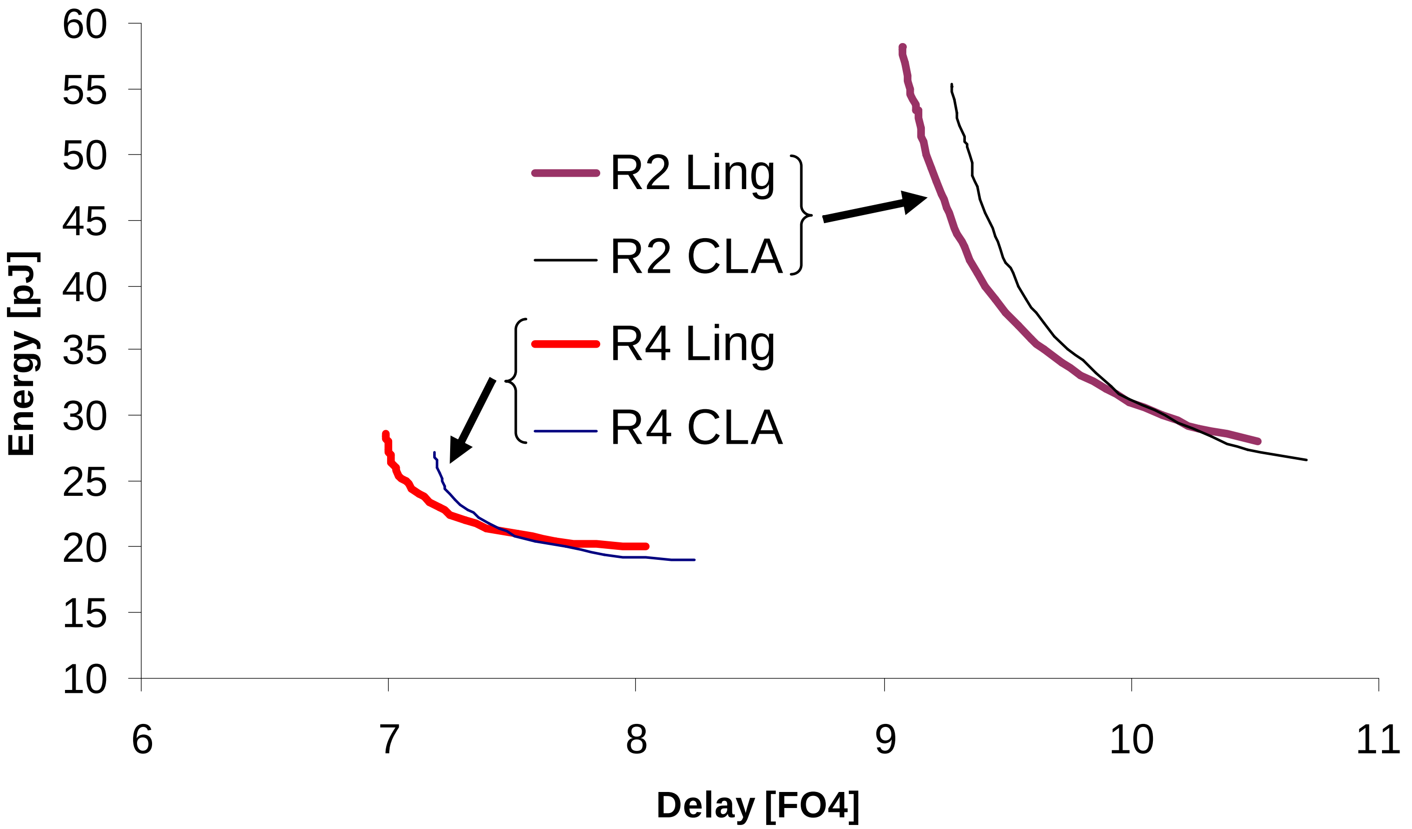
64-Bit CLA Adders

- Generic 64-bit adder block diagram



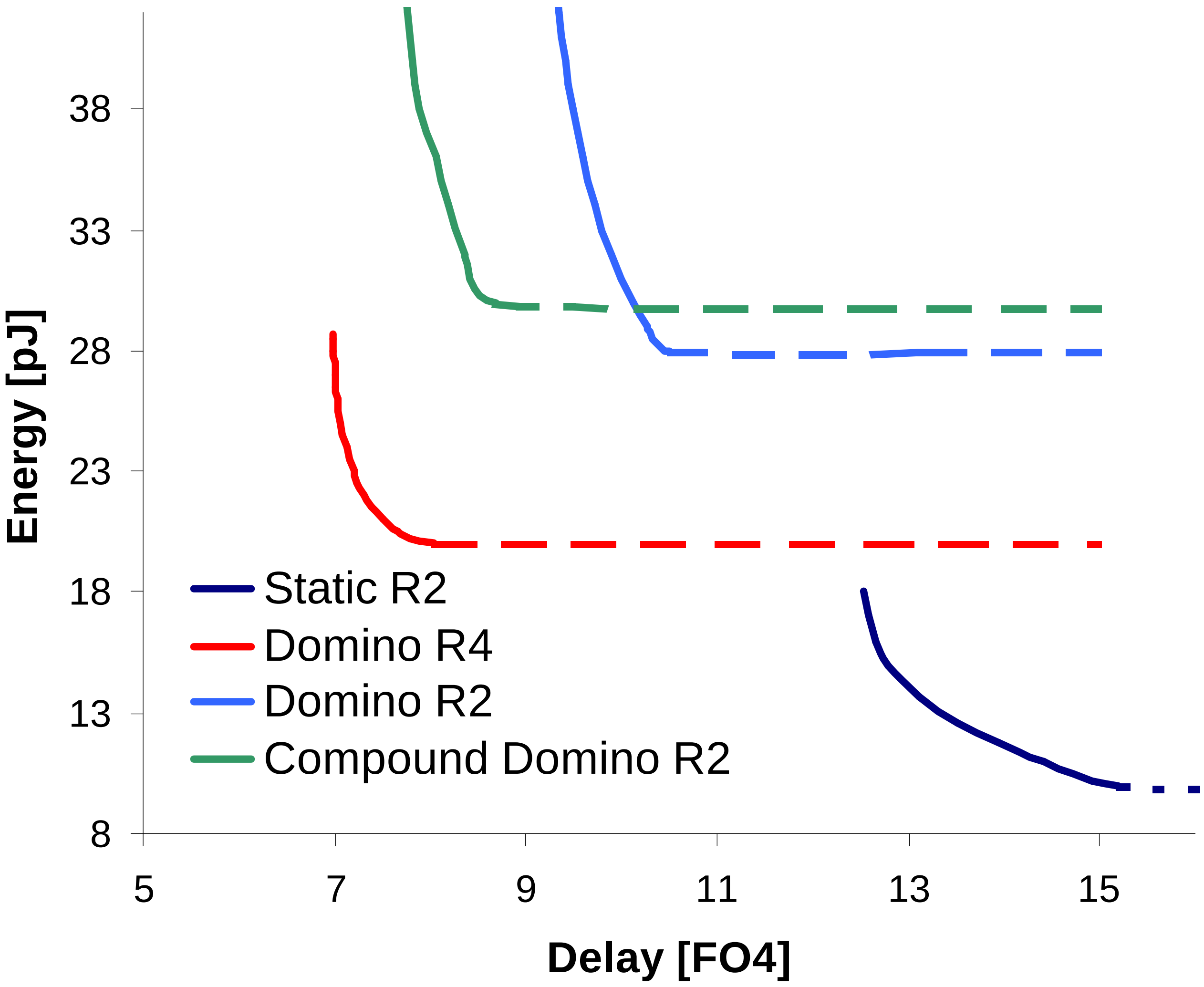
- Classical CLA, Ling equations
- Static, single-rail domino, compound domino logic
- Radix-2 and radix-4 carry trees
- Full and sparse trees (sparseness of 2 and 4)
- Use tabulated delay and analytical energy models (switching, leakage)

Equation Set Comparison



Ling adders achieve shorter delays
Radix-4 are faster than radix-2

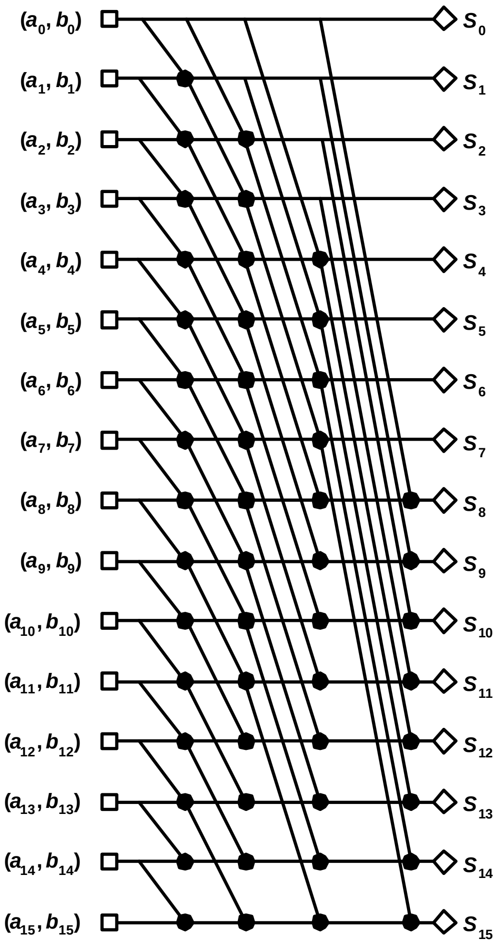
Chosing a Logic Style



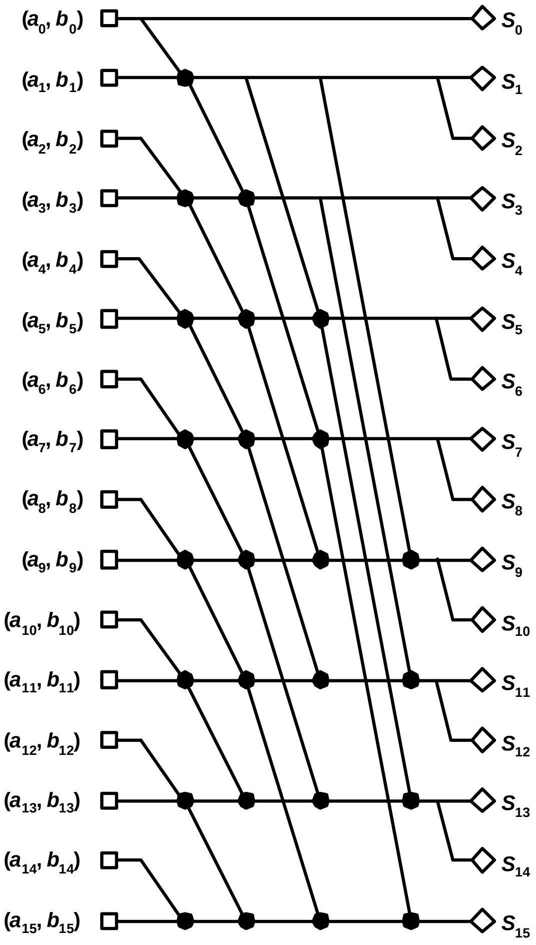
- **Static adders are low power but slow**
- **Domino logic is the choice for short cycle times**

Full vs. Sparse Trees

R2 Full



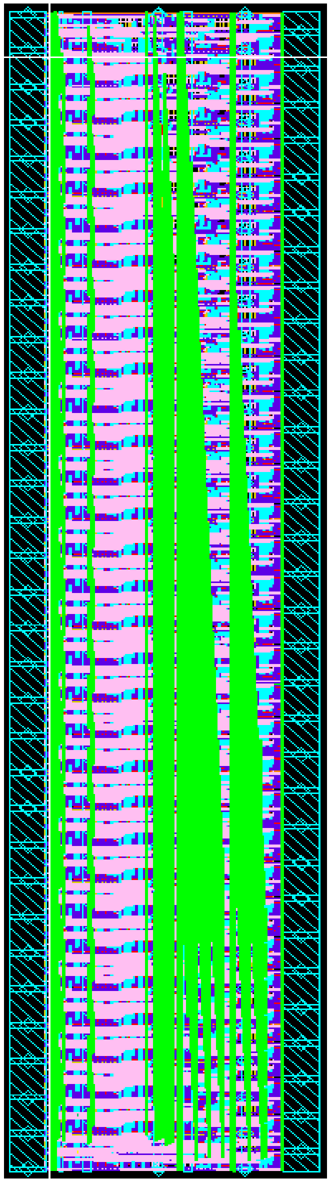
R2 SP2



<u>Sparse trees:</u>

- ➘ Heavier load on the carry tree
- ➚ Reduced input loading
- ➘ More complex sum precompute gates

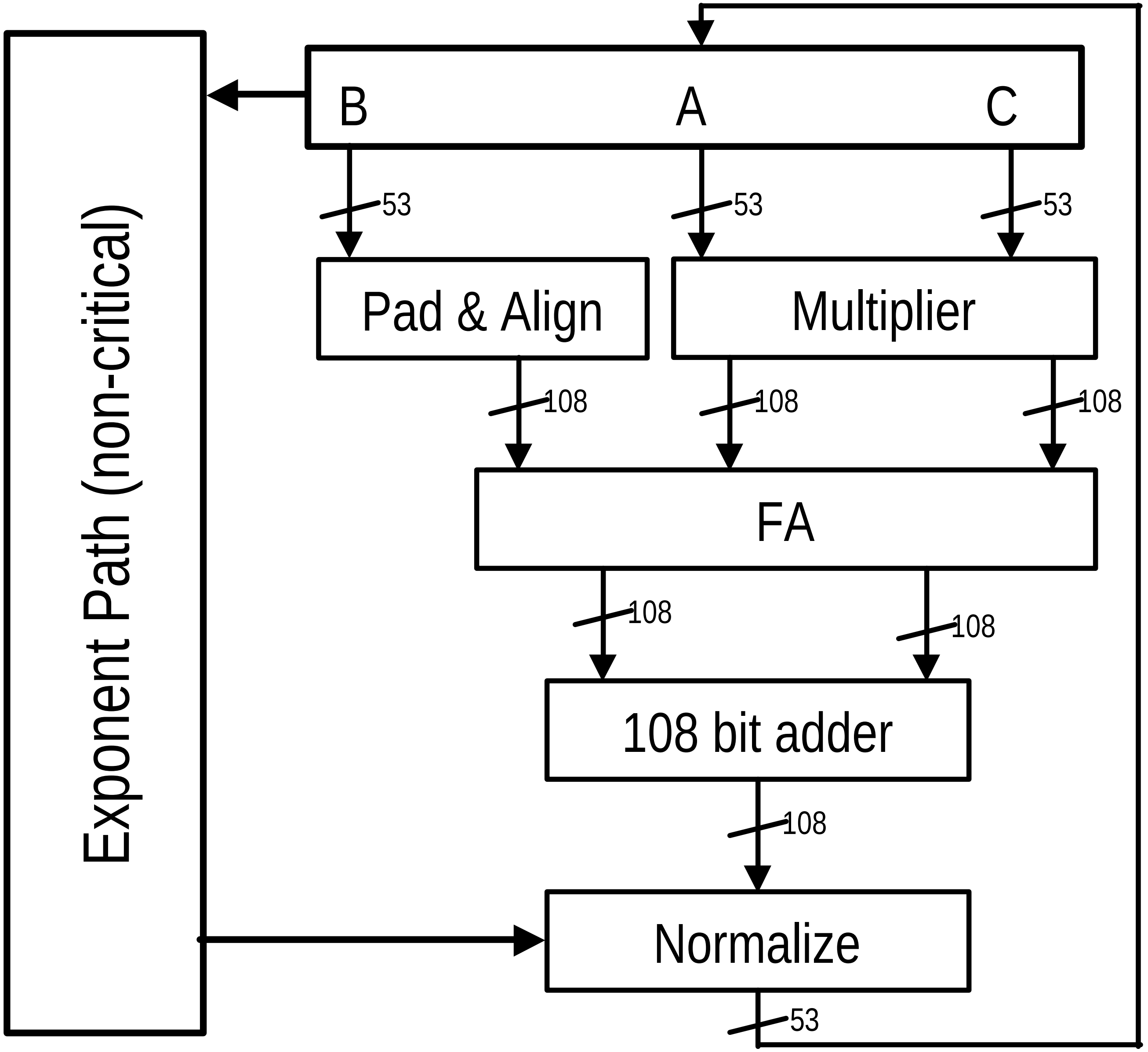
Proof of Existence: Fastest Adder



- **Radix-4 sparse-2 domino Ling adder**
- **Technology:**
  - 90 nm 1P 7M
  - V_{DD} = 1 V
- **Performance:**
  - Delay: 210 ps (post – layout simulation)
  - Energy: 9.1 pJ / cycle (optimization tool)
- **Core dimensions: 417.3 mm x 75.3 mm**
- **Chip to be taped out 11/1/04**

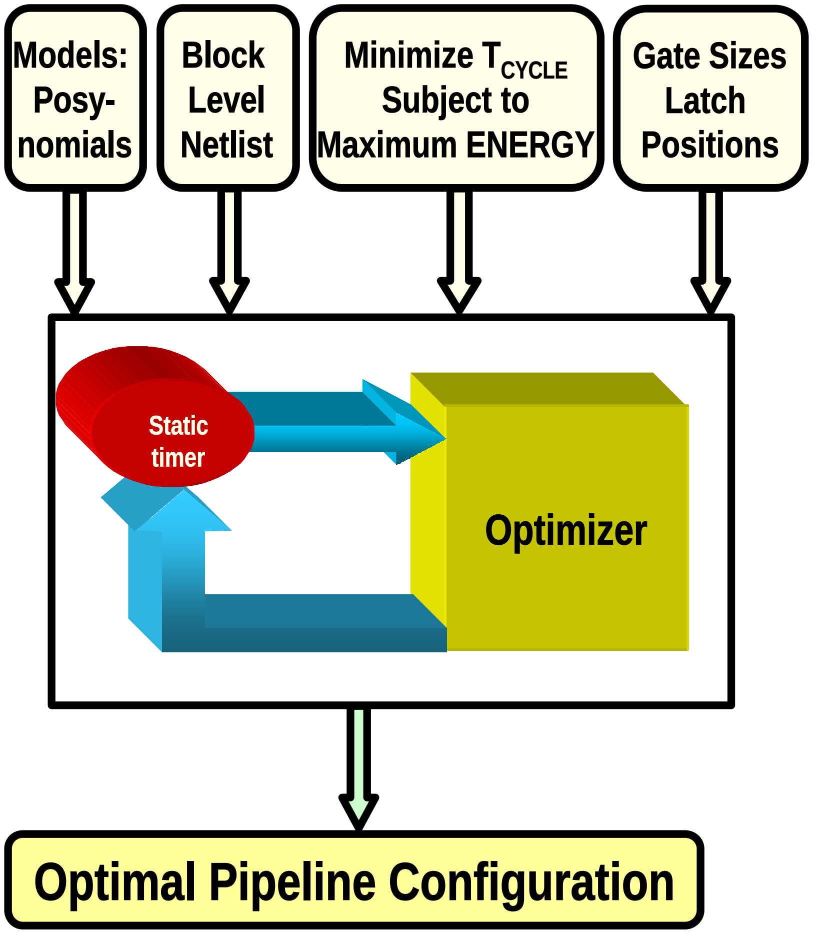
*with Sean Kao

Micro-Architecture Optimization: Power4 FPU



- **5 cycles**
- **2-phase clocking**
- **Static CMOS**

Optimizing Pipelined Circuits



Models: Posy-nomials

Block Level Netlist

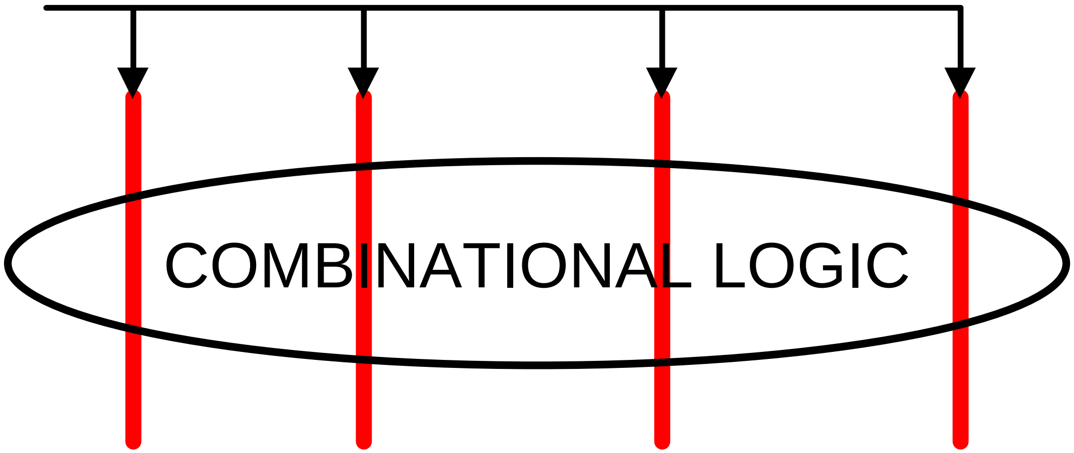
Minimize T_{CYCLE} Subject to Maximum ENERGY

Gate Sizes Latch Positions

Static timer

Optimizer

Optimal Pipeline Configuration



Cycle boundaries: transparent latches

COMBINATIONAL LOGIC

- Fix pipeline depth
- Find shortest cycle time for fixed cutset
- Search for optimum cutset

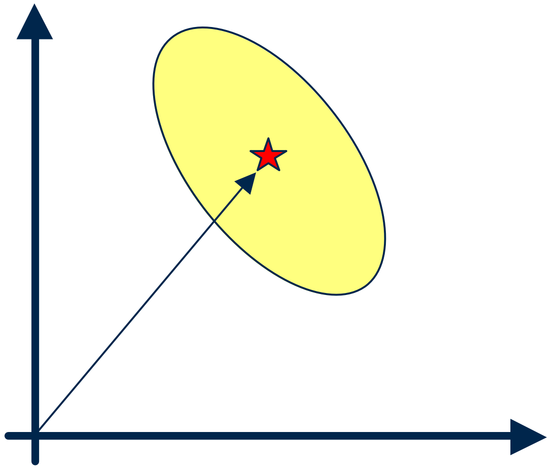
Work in progress…

Outline

- Design as a power – performance optimization problem
- Fundamentals of circuit optimization
- Design examples
- **Dealing with variations**
- Conclusions

Robust Optimization

- Parameters are within an ellipsoid centered on the nominal values
- Optimize the worst case



- Compatible with convex optimization for the presented analytical models
- Problem: computing the ellipsoid

Stochastic Optimization for Yield

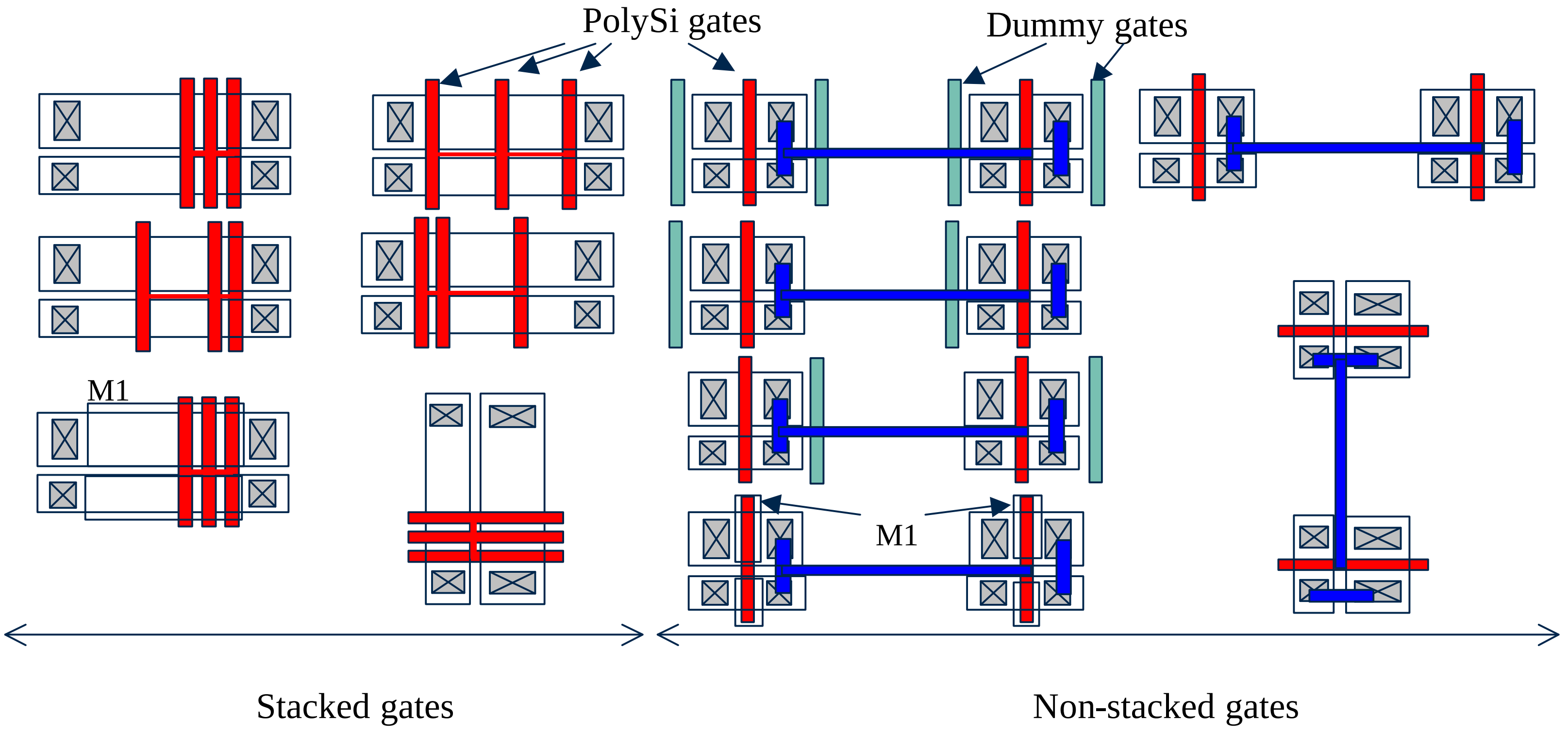
- **Parameters are random variables centered on the nominal values**
- **Optimize for a desired yield**

$$\boxed{\begin{array}{l} \min\limits_{x \in R^n} f(x) \\ \text{subject to} \\ g_i(x) \le 0 \quad i = 1..m \end{array}} \longrightarrow \boxed{\begin{array}{l} \min\limits_{x \in R^n} f(x) \\ \text{subject to} \\ P(g_i(x) \le 0, \quad i = 1..m) \ge h \end{array}}$$

- **Compatible with convex optimization under certain conditions**
  - Convex analytical models
  - Jointly Gaussian parameters
- **Problem: finding the distributions of the parameters, especially the correlations**

Impact of Layout on Variations

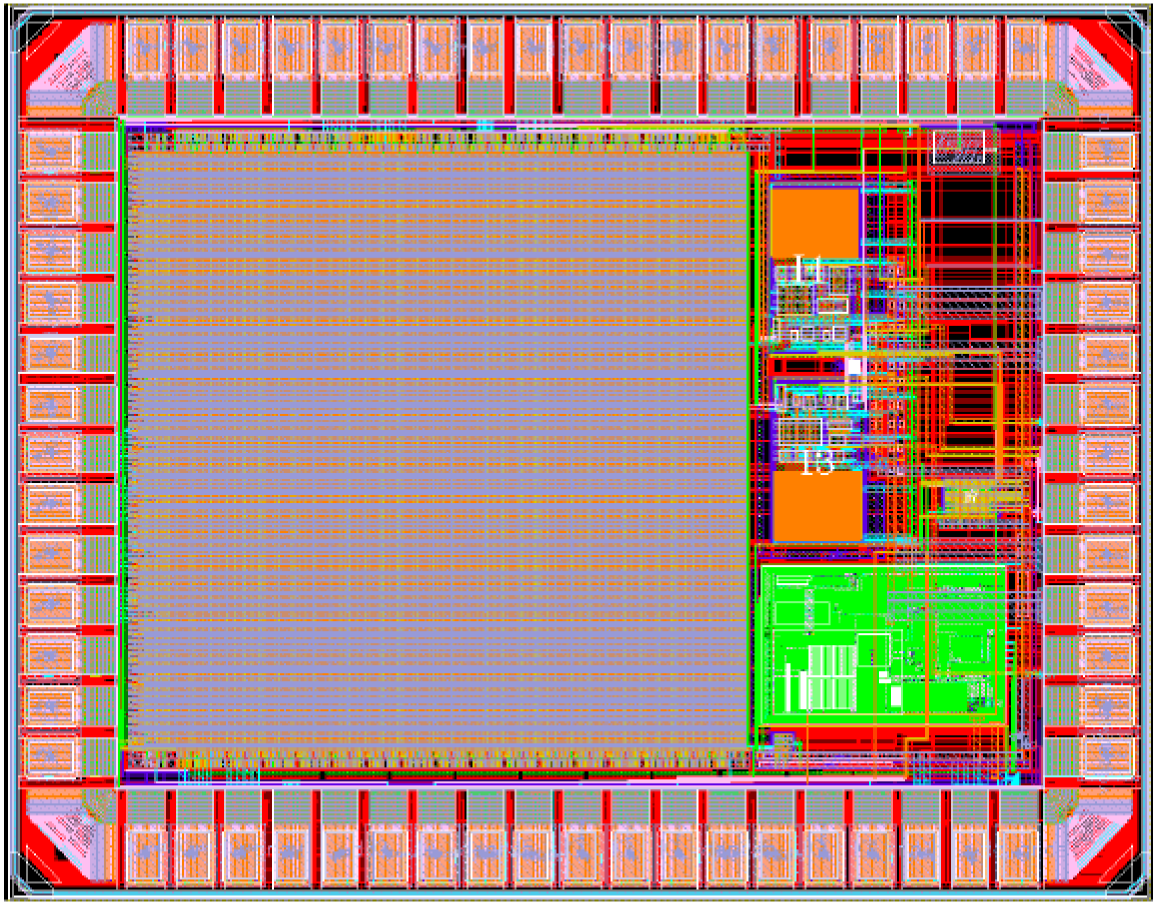
- **Stacked gates vs. non-stacked gates (e.g. gates vs. buffers)**
- **Proximity effects, orientation of gates, metal layer above gate (annealing)**



Work by Liang – Teck Pang

Test Chip

- 90nm 1P 7M
- V_{DD}= 1 V
- 1.55 x 1.17 mm^2
- Taped out 9/04
- To be packaged and tested
- Measurements will provide statistical data for the stochastic optimizer



Conclusions

Power and performance are the two sides of the same coin. The connection: the power – performance tradeoff curve.

- **Built a suite of stackable tools to design "power – performance optimal circuits"**
- **Design space explorations:**
  - Different optimization variables
  - Various levels of abstraction
  - Impact of variations
- **Experimented the tool suite on various designs**
  - Dual-supply ALU
  - CLA Adders in the E-D space