Modeling for Statistical Timing Applications

Joel Phillips, Cadence Berkeley Laboratories

Collaborators:

L. Miguel Silveira (INESC-ID/CBL)

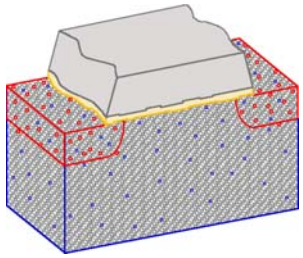
Luis Guerra e Silva (INESC-ID/CBL), Zhenhai Zhu (CBL)

Outline

- Background: Statistical Static Timing Context
- Incorporating Variability via Affine Delay Models
- Cell Modeling Under Parameter Variation
- Interconnect Modeling Under Parameter Variation

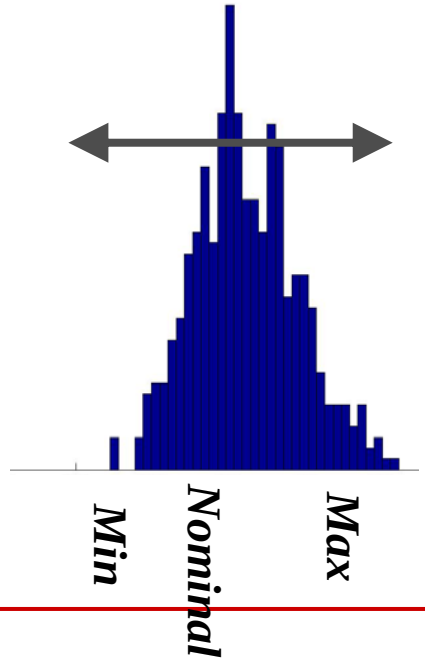
Worrying About Variability

- What it is
  - Undesired fluctuation of circuit figure of merit
    - e.g. gate delay, leakage, ….
- Why do we care ?
  - (It's widely believed that) relative magnitude of variability is increasing
    - approaching physical limits of MOS technology
  - Increased susceptibility: Devices in small geometries at low voltages are more sensitive to perturbation



A 22 nm MOSFET
In production 2008

[Asenov et al, "Simulation of Intrinsic Parameter Fluctuations in Decananometer and Nanometer-Scale MOSFETs",
ELECTRON DEVICES, SEPT 2003]

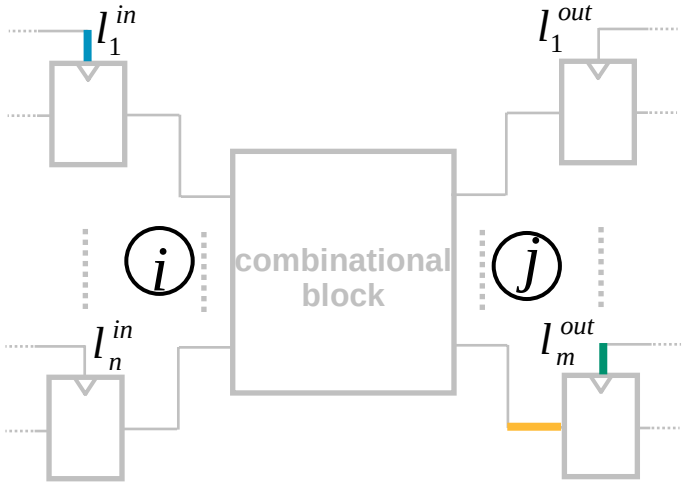


The Usual Suspects

- Processing condition shift (“dial spinning”)
  - Wafer to wafer, fab to fab, same fab over time
- Environmental – thermal gradient, IR drop
- Wafer level parameter nonuniformity
  - Oxide thickness, etch rate vary on wafer scale
- Across-die systematics
  - Optical lens aberration
- Local systematic – tied to (and predictable from) layout
  - Sub-wavelength lithography effects, CMP
- Stochastic – random device-to-device variation
  - Channel dopant fluctuation effects, line edge roughness

Analyzing Variability Effects

- **Analog folks….old news!**
  - SPICE, Monte Carlo, cocktail napkins ….
- **Digital folks**
  - Meet timing constraints?
  - Meet power budget?
  - Scalability, automation, ….

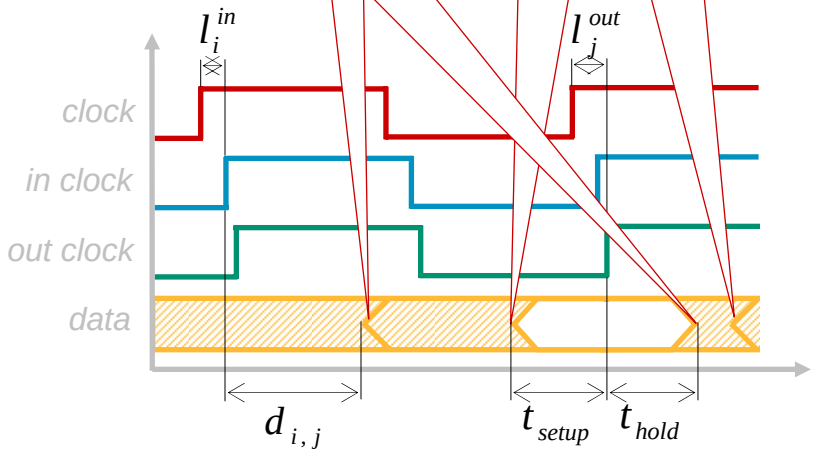


- **Setup**

$$l_i^{in} + d_{i,j} \leq T + l_j^{out} - t_{setup}$$

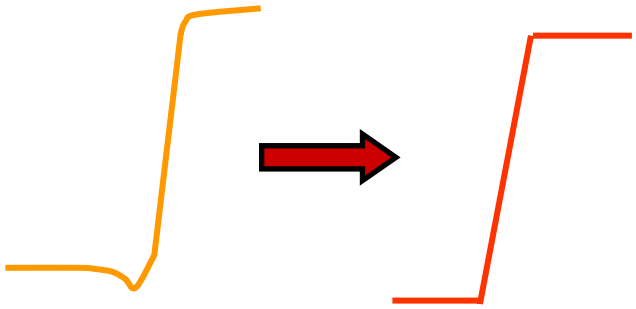
- **Hold**

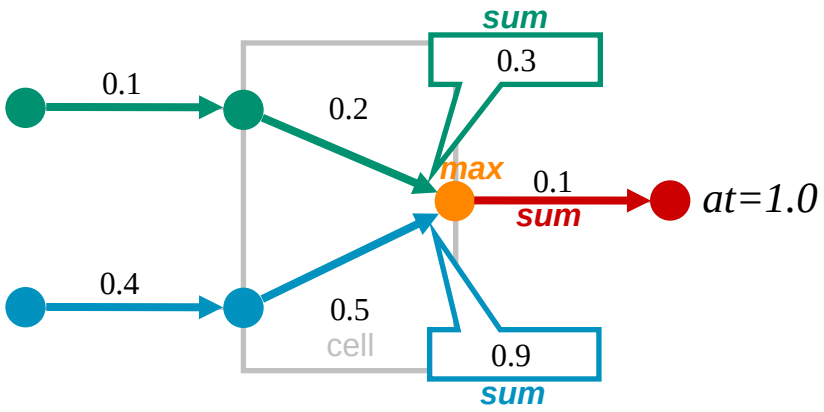
$$l_j^{out} + t_{hold} \leq l_i^{in} + d_{i,j}$$



Static Timing Analysis (STA)

Waveform →
slope + arrival time

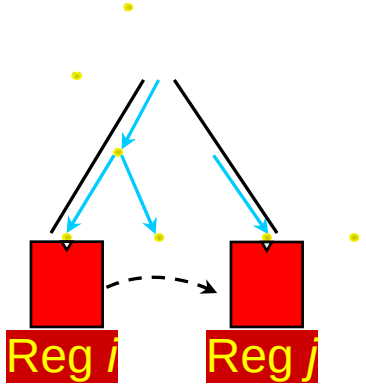




- Abstract timing analysis to propagation *(T,s)* through graph
- Logic operation reduced to *max* computation : simple but conservative analysis
- Delays from wires, cells represented notated on arcs of graph
- From arrival times, compute slacks, critical paths, …
- Note: delays depend on selected operating and process conditions

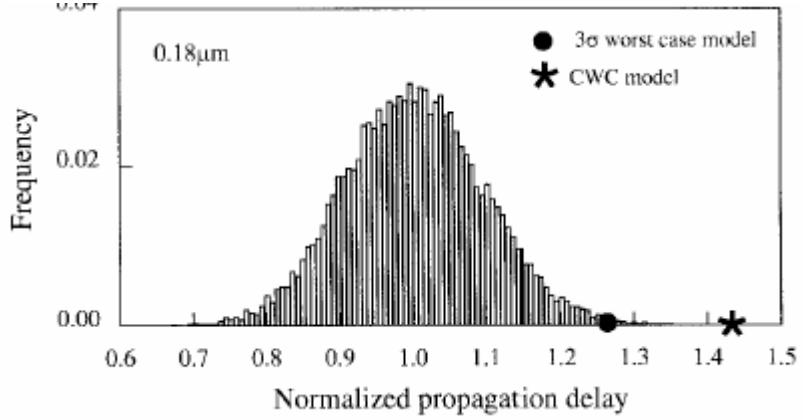
The Case For "Statistical Timing" (SSTA)

- "Traditional" worse-case analysis
  - Pick a set of operating conditions (corners)
  - Run STA for each "corner"
- Which corners to pick? Exhaustive analysis may not be possible!

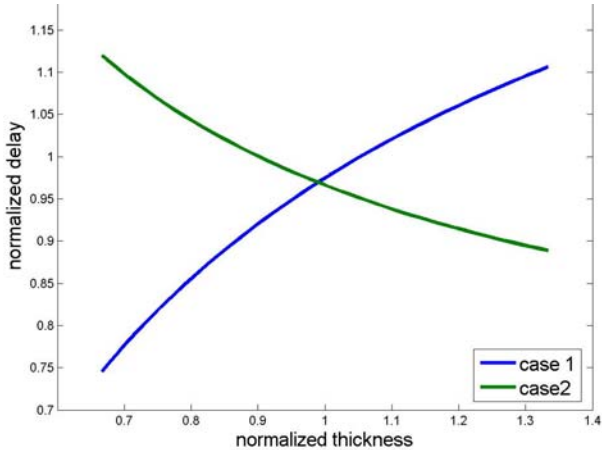


- may not catch all problems

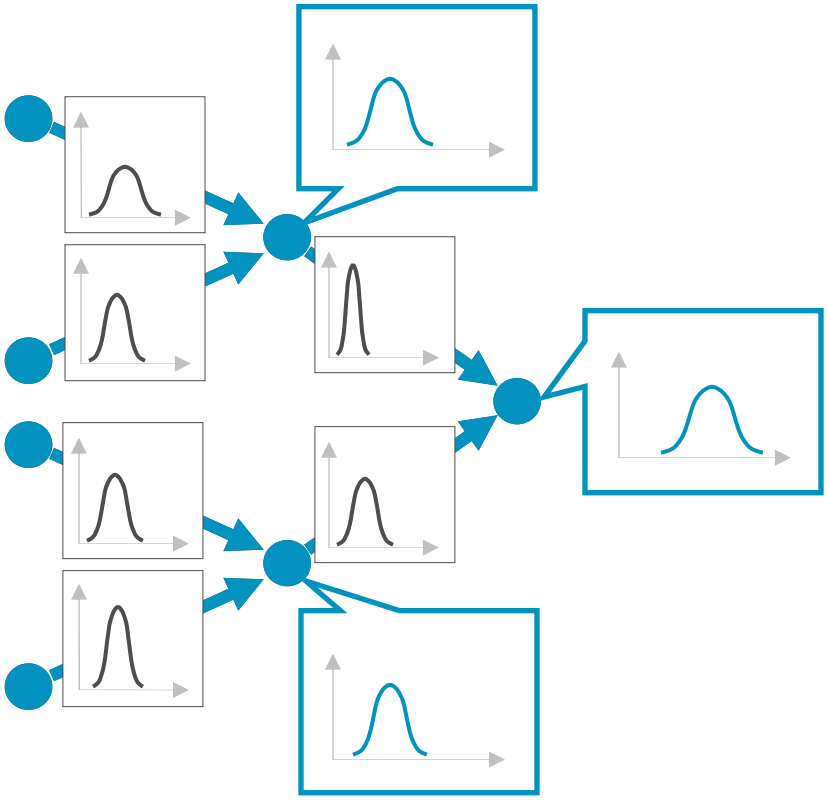
- may be overly pessimistic



Source: Nardi et al, Trans. Sem Manuf '99



Timing Analysis in Presence of Process Variation

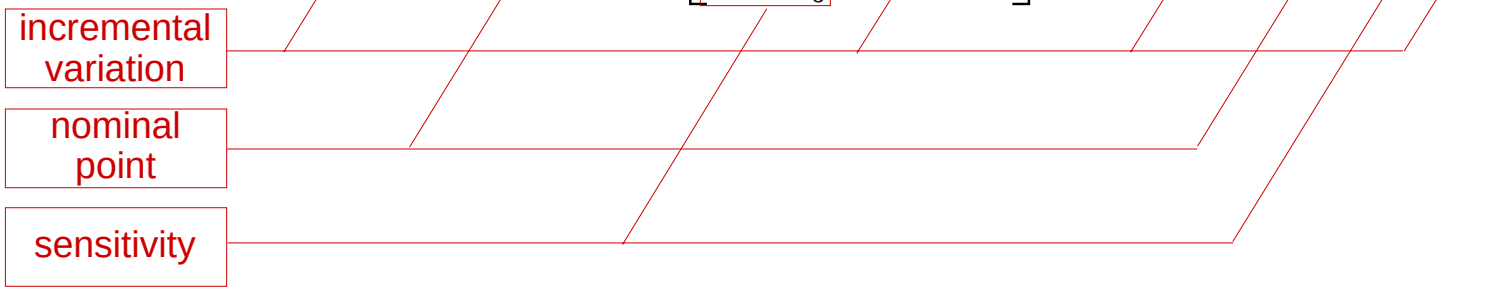


- "Statistical Timing" methodologies intensely researched in past few years
- Intellectual paradigm: propagate some parametric quantity (e.g., distributions) through timing graph
- Key issues
  - Max function computation
  - Tracking parameter correlations
  - Delay models

Affine Delay Models

- Why model delays as **linear functions of parameters** ?
  - Digital circuits are strongly nonlinear with respect to circuit inputs, but cell delays are often close to linear with respect to process parameters over relevant parameter ranges.
  - Introduces explicit dependence of all quantities on specific variation sources
  - Simplifies certain computations in analysis
- Affine parametric delay formulation:

$$x(\lambda - \lambda_0) = x(\lambda_0) + \sum_{i=1}^{p} \left[\left. \frac{\partial x}{\partial \lambda_i} \right|_{\lambda_0} (\lambda_i - \lambda_0) \right] \Leftrightarrow x(\Delta\lambda) = x_0 + s_x^T \Delta\lambda$$

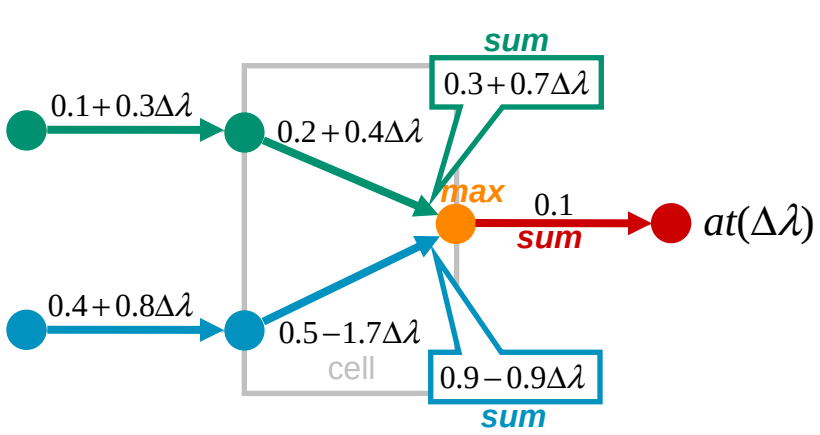


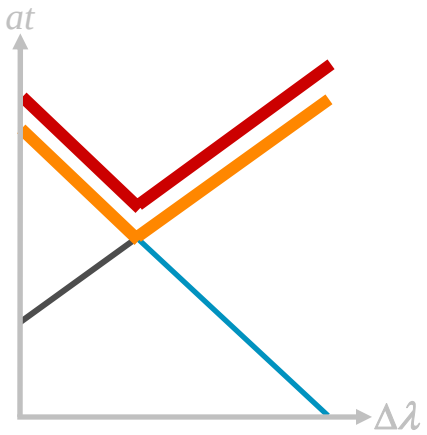
Timing Analysis with Parametric Delays

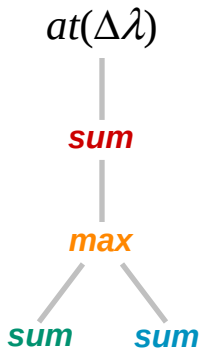
- Typical model: Gate and interconnect **delays** are **affine** functions of $\Delta\lambda$.

$$d = 5 + 0.24\Delta\lambda_1 - 0.32\Delta\lambda_2 + 0.07\Delta\lambda_3 - 0.11\Delta\lambda_4$$

- **Circuit delays** and **arrival times** are **piecewise-affine** functions of $\Delta\lambda$.
  - Result from ***sum*** and ***max***.

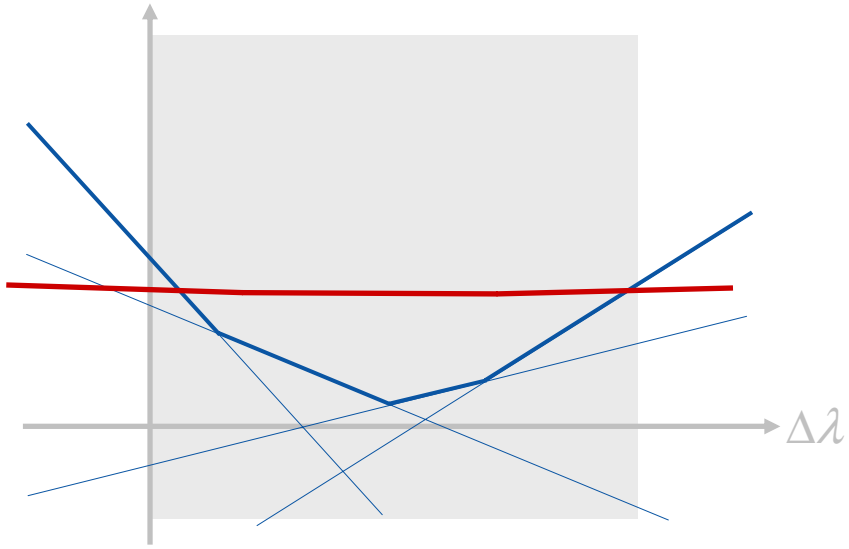






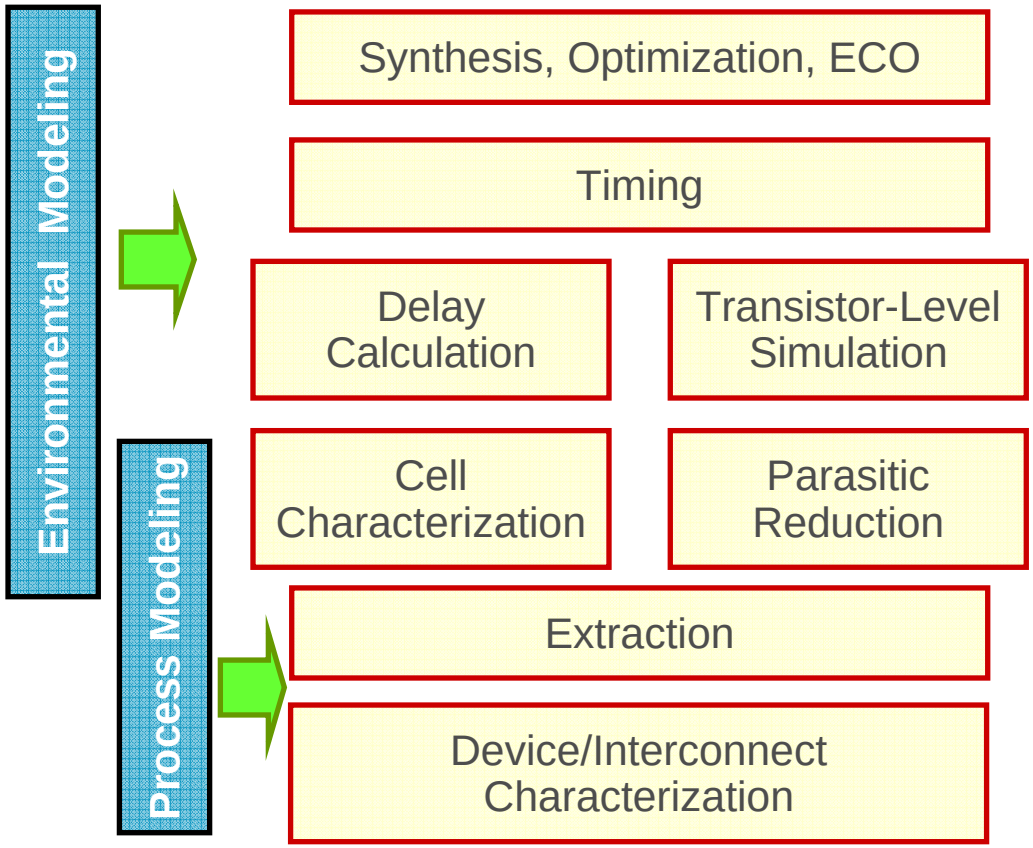
Tractable *max* computations

- **Given:** convex ***n*-piece affine function** $X(\Delta\lambda) = \max_{i=1,\ldots,n} (x_{0i} + x_i^T \Delta\lambda)$.
- **Compute:** single **affine function** $Y(\Delta\lambda) = y_{0j} + y_j^T \Delta\lambda$.
- **Such that:** $Y(\Delta\lambda) \geq X(\Delta\lambda)$ in some sense

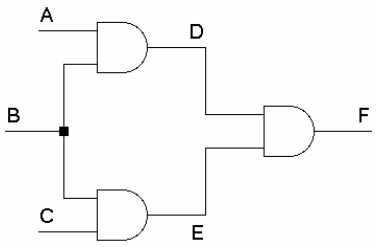


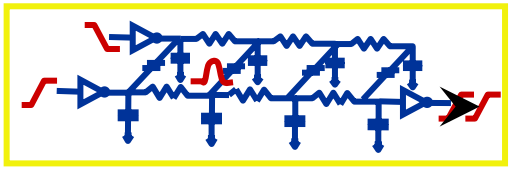
- Example: Assume Gaussianity for all parameters. Compute Y to match first two moments of X. Lots of papers.

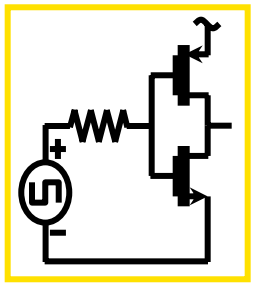
Electrical Analysis Stack

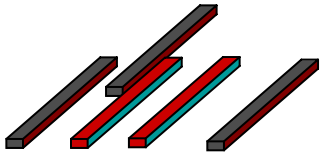


$$T_{rc} = 5 + 0.24\Delta\lambda_1$$







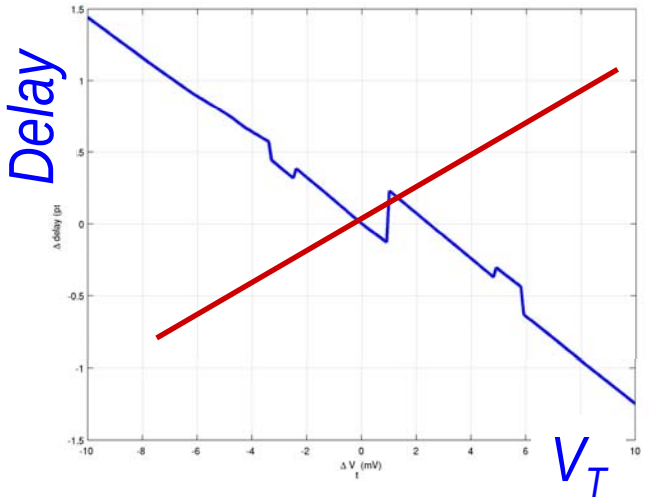


$$\leftarrow \Delta W \rightarrow$$

- Practical question: how to get the parametric delay models from all this *stuff?*

Model Construction Challenge

- Black-box approaches: post-process analysis data
  - Finite-differencing for sensitivities
  - Response-surface models (RSMs) and other curve-fitting approaches



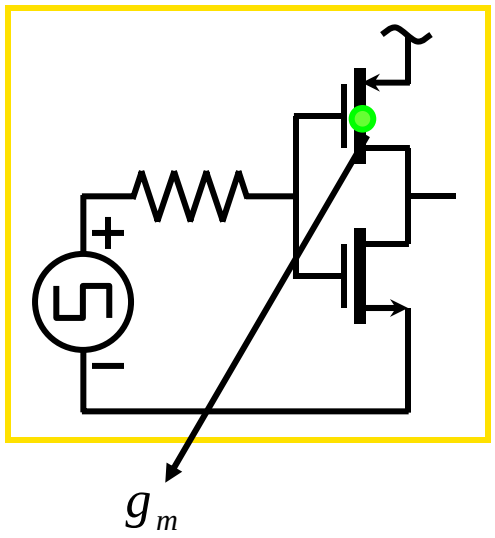
- Problems
  - Most ***programs*** produce non-smooth output [timing, SI, circuit simulation, field solvers, extraction, parasitic reduction]
  - Slow. Multiple runs per coefficient in best case. Numerical noise requires heavy over-sampling → excessive number of analysis runs.

Tools in the Bag

- Perturbation Theory
  - Taylor series, Volterra series, orthogonal polynomial representations
- Data Compression
  - Principal Components (PCA) / Singular Value Decomposition (SVD)
  - Adjoint methods

Perturbation Theory

Basic Perturbation Analysis, I



- Consider solving nonlinear circuit equations (KCL)

$$i(v) = u$$

- Suppose input is decomposed into nominal plus perturbation (assumed small)

$$u = u_0 + \Delta u$$

- Assume response (circuit node voltages) can be decomposed into operating point (bias) + perturbation (assumed small)

$$v = v_0 + \Delta v$$

- Assume current-voltage relation can be treated perturbatively as well [e.g., Taylor series]

$$i(v) = i(v_0) + \frac{\partial i}{\partial v}\Delta v$$

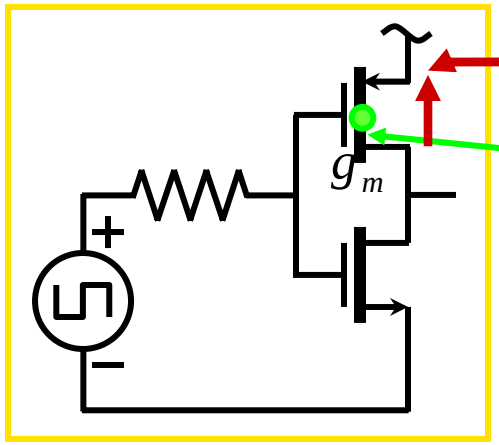
Basic Perturbation Analysis, II

- Substitute series expansions into original equation

$$\left[i(v_0) + \frac{\partial i}{\partial v} \Delta v \right] = u + \Delta u$$

- Collect zero-order terms → equation for bias point

$$i(v_0) = u_0$$



- Collect first-order terms → equation for perturbation

$$\frac{\partial i}{\partial v} \Delta v = \Delta u$$

- Physical interpretation:
  - LHS: small-signal circuit model
  - RHS: equivalent “perturbation” sources

Perturbation Analysis, Generalized

- Nonlinear Response (higher-order terms)

DIstortion Sources

$$i(v) = i(v_0) + \frac{\partial i}{\partial v}\Delta v + \frac{\partial i}{\partial \lambda}\Delta\lambda + \frac{\partial^2 i}{\partial v^2}(\Delta v \otimes \Delta v) + h.o.t.$$

- Parameter Variation

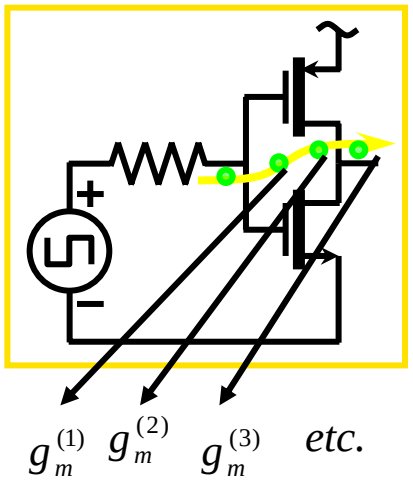
Mismatch Sources

$$i(v) = i(v_0) + \frac{\partial i}{\partial v}\Delta v + \frac{\partial i}{\partial \lambda}\Delta\lambda$$

- Time-Varying Operating Point

$$i(v,t) = i(v_0(t)) + \left.\frac{\partial i}{\partial v}\right|_{v0(t)} \Delta v(t)$$

$$q(v,t) = q(v_0(t)) + \left.\frac{\partial q}{\partial v}\right|_{v0(t)} \Delta v(t)$$



Notable Applications

- AC Noise Analysis
- RF Noise Analysis
- Volterra-based distortion
- Time-varying / Weakly nonlinear automated macromodeling
- Parametric interconnect reduction
- Mismatch analysis
- Mismatch parameter extraction
- Parametric cell delay characterization
- Parametric cell delay calculation

General Observations

- Decomposition of Nonlinearity
  - Assumes circuit responds to perturbations (noise, process variation) in weakly nonlinear way, but includes strongly nonlinear circuit biasing
- Matrix Equation Structure

Zero-Perturbation Admittance Matrix

$$[Y(v_0)]\Delta v^{(k)} = RHS(v_{0,} \ldots, \Delta v^{(k-1)})$$

Sources from response up to order k-1

Reponse to order k

  - Basically similar for all orders of expansion (recursive structure)
  - Basically similar for all applications
  - Basically similar regardless of derivation

Interactions of Models and Statistics

- Given parametrized models…

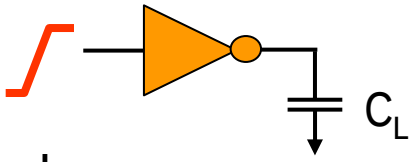
$$T_d = T_d^{(0)} + c_L \Delta L + c_{vt} \Delta V_t + \cdots$$

- Statistics can be obtained
  - Through final analysis / simulation
    - Often using standard linear algebra identities

- Please do not embed statistical assumptions in the model derivation process!!!!

Cell Models

Table-Based Delay Model (.lib)



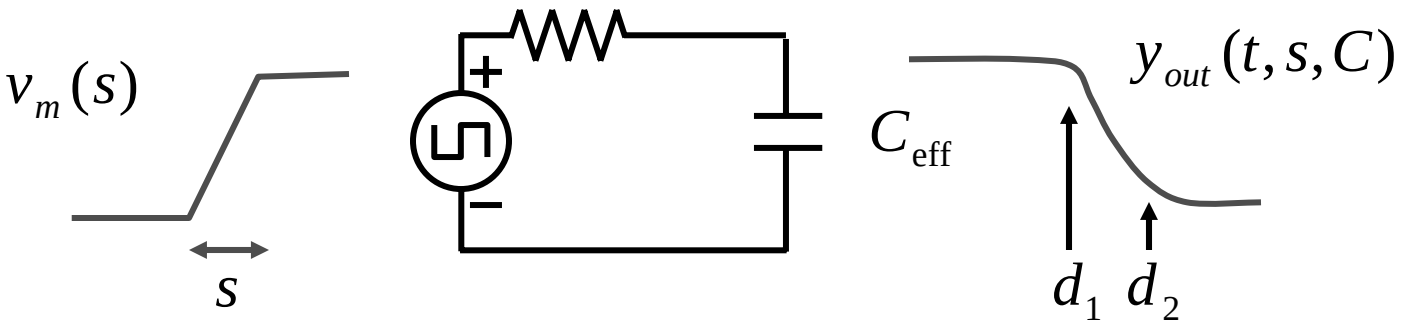
slew s

C_L \ s	0.03	0.1	0.3
0.03	0.1498	0.1492	0.1440
0.2	*0.3597*	0.3588	0.3528
0.6	*0.9594*	0.9576	0.9492

- Waveform model: linear ramp
- Load model: lumped capacitance
- Cell model
  - Pin-pin delay
  - Output slew
- Delays, output slews tabulated for each pin-pin arc as functions of input slew, output load

- Issues:
  - Lumped capacitance bad model of interconnect
  - Real waveforms not ramps (tails, non-monotonic)
- Hack: $C_{eff..}$. Try to transform interconnect into "effective" capacitance seen by gate depending on…..

Ceff Methodology (idealized)



- Variables: slew, effective cap, output waveform points
- Model output must match at table characterization points

$$y_{out}(d_1(C), s, C) = a_1 Vdd$$

- Require equality of average current drawn by "actual" load and effective capacitance model (as a function of interconnect parameter variation) !

$$\langle I_C(C,s) \rangle = \langle I_\pi(\lambda) \rangle$$

Perturbation Analysis: Non-Intuitive, but Possible!

- Base equations
  - Assume we solve these for some s, C

$$y_{out}(d_2(C), s, C) - y_{out}(d_1(C), s, C) = \alpha V dd$$

$$\langle I_C(C,s)\rangle = \langle I_\pi(\lambda)\rangle$$

- Assume perturbation solution

$$y_{out}(d(C), s, C) \rightarrow y(d(C_0), s_0, C_0) + \frac{\partial y}{\partial d}\frac{\partial d}{\partial C}\Delta C + \frac{\partial y}{\partial s}\Delta s + \frac{\partial y}{\partial C}\Delta C$$

$$\langle I_C(C,s)\rangle \rightarrow \frac{\partial \langle I_C(C,s)\rangle}{\partial C}\Delta C + \frac{\partial \langle I_C(C,s)\rangle}{\partial s}\Delta s$$

$$\langle I_\pi(\lambda)\rangle \rightarrow \sum_{k=1}^{p}\frac{\partial \langle I_\pi(\lambda)\rangle}{\partial \lambda_k}\Delta\lambda_k$$

Analytic Ceff equations

- Solve small matrix equation

$$\begin{bmatrix} \frac{\partial y}{\partial s} & \frac{\partial y}{\partial d}\frac{\partial d}{\partial C} + \frac{\partial y}{\partial C} \\ \frac{\partial I_C}{\partial s} & \frac{\partial I_C}{\partial C} \end{bmatrix} \begin{bmatrix} \Delta s_{drv} \\ \Delta C_{eff} \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

(model parameters)

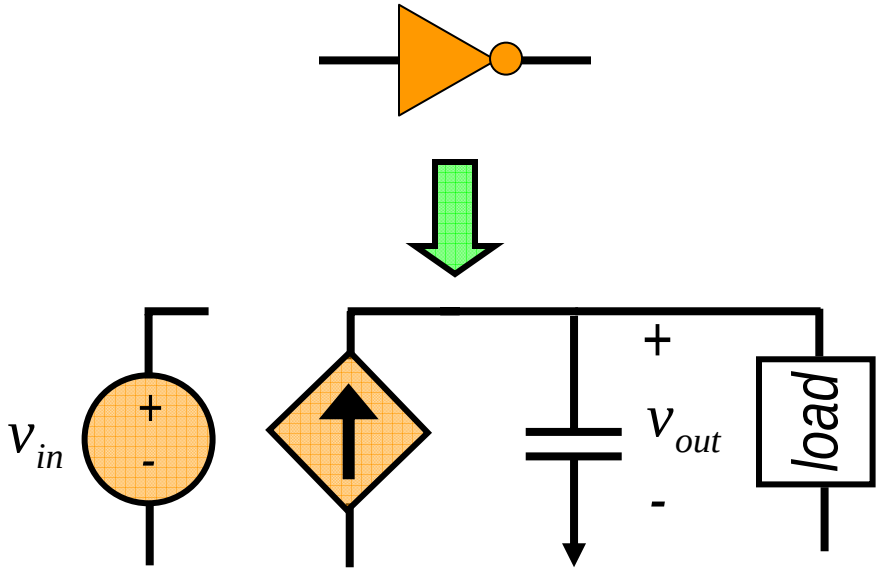
$$\frac{\partial \langle I_\pi(\lambda) \rangle}{\partial \lambda_k}$$

(from interconnect sensitivity computation)

- Final Computations
  - Ceff perturbation $\frac{\partial s}{\partial \lambda}, \frac{\partial d}{\partial \lambda}$
  - from tables $\frac{\partial d}{\partial C}, \frac{\partial d}{\partial s}$

$$delay(parameters)$$
$$slew(parameters)$$

Current-Source Driver Models



VCCS: $i(v_{out})$ or $i(v_{in}, v_{out})$

- Much better at dealing with real-world interconnect effects

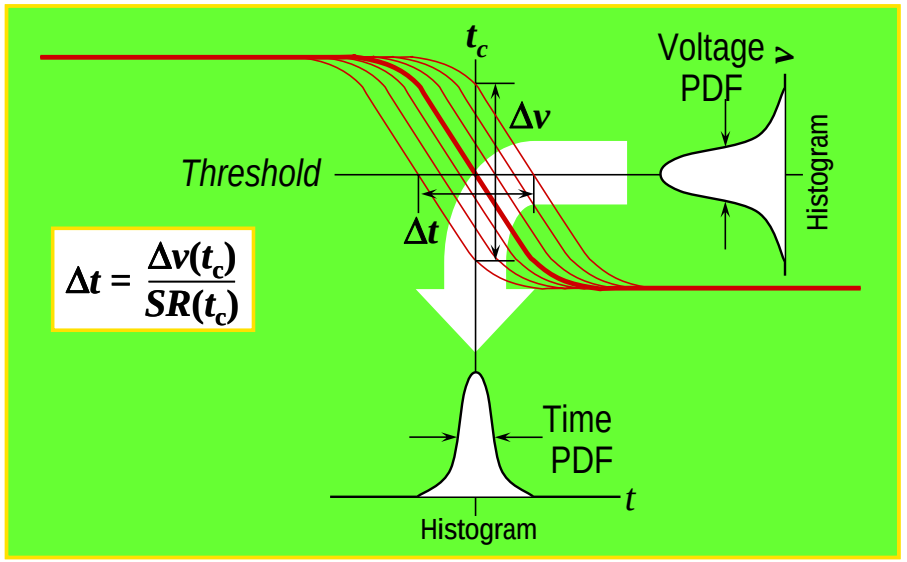


Fig. courtesy K. Kundert

- Parametric analysis can be performed by similar exercise in perturbation theory to obtain Δv_{out} as function of parameters

Interconnect Models

Variation-Aware Model Order Reduction

- Base Equations for RC Model

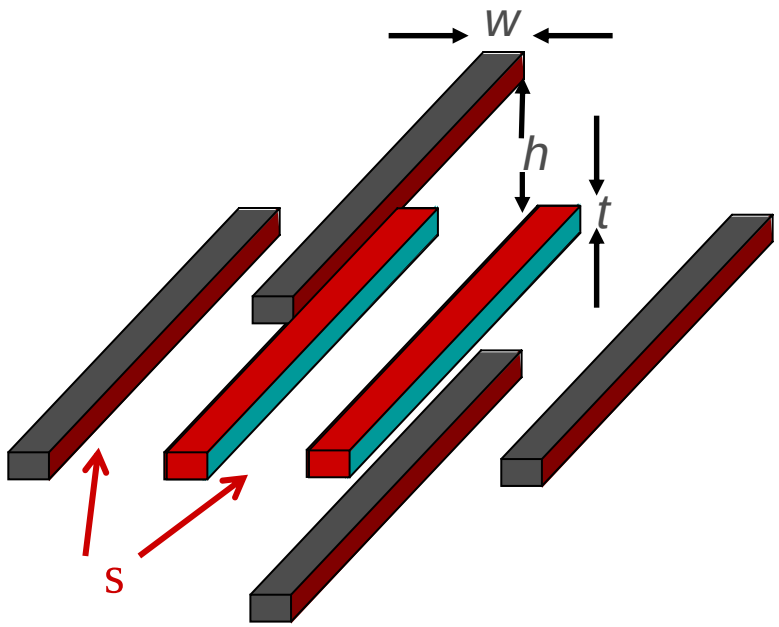
$$C(\lambda)\frac{dx}{dt} + G(\lambda)x = Bu$$

$$y = Lx + Du, \quad \lambda = \begin{bmatrix} w & s & t & h \end{bmatrix}$$

- Example parametric form:

$$C(\lambda) = C_0 + C_1\Delta\lambda_1 + C_2\Delta\lambda_2 + \cdots$$

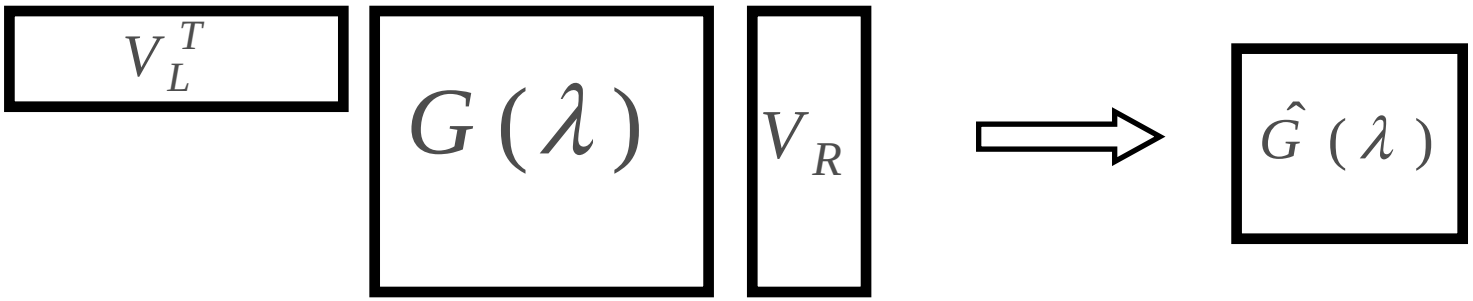
$$G(\lambda) = G_0 + G_1\Delta\lambda_1 + G_2\Delta\lambda_2 + \cdots$$



- Goal is to compute reduced model in similar form that is amenable to fast delay calculation (previously described)
  - Reduced interconnect models by themselves are not so useful.

Generating Parametric Reduced Models

- Projection-based reduction approach



- Consider constant matrices

$$\hat{G}(\lambda) = V^T\left[G_0 + G_1\Delta\lambda_1 + G_2\Delta\lambda_2 + \cdots\right]V$$
$$= V^T G_0 V + V^T G_1 V\Delta\lambda_1 + V^T G_2 V\Delta\lambda_2 + \cdots$$
$$= \hat{G}_0 + \hat{G}_1\Delta\lambda_1 + \hat{G}_2\Delta\lambda_2 + \cdots$$

- Affine models in → affine models out!
  - (yes, intra-die is still possible)

Picking Projection Matrices : Moment Matching

- Krylov family (interpolation-like)
  - Choose projection matrices to match moments of transfer function at selected frequency points

$$\mathrm{colsp}\{\mathrm{V}\} \supset \{(s_1 I - A)^{-1} p, (s_2 I - A)^{-1} p, \ldots\}$$

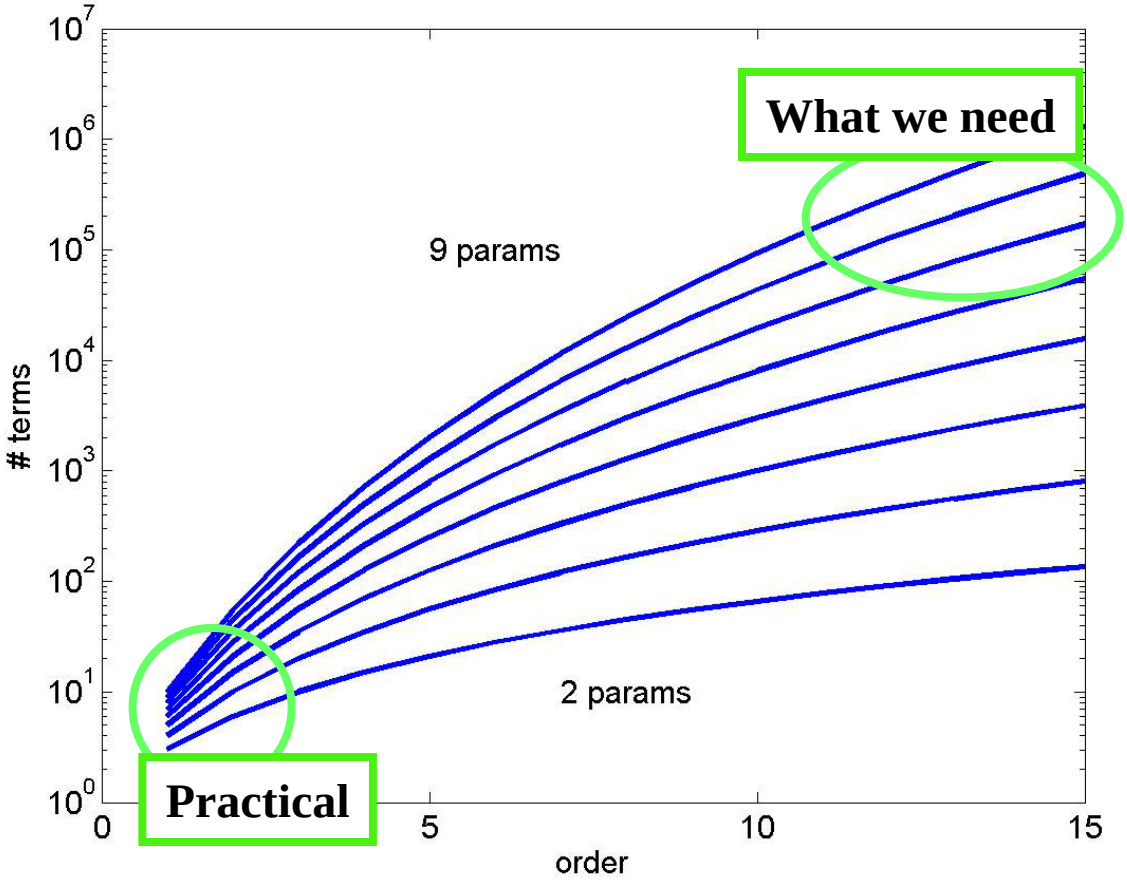
  - Extensions to parametric case conceptually easy, problematic in computational application

$$x(s, \lambda_1, \ldots \lambda_n)$$
$$= \left[sC(\lambda) + G(\lambda)\right]^{-1} B$$
$$= \sum_{k,l,m,\ldots} M_{k,l,m,\ldots} s^k \lambda_1^l \lambda_2^m \cdots$$

$$\mathrm{colsp}\{\mathrm{V}\} \supset M_{k,l,m,\ldots}$$

Projection Spaces Via n-D Moments

- Issues
  - Exponentially increasing cost with order if all moments kept
  - Not clear how to "prune" moments
  - Hard to achieve good error / effort tradeoff

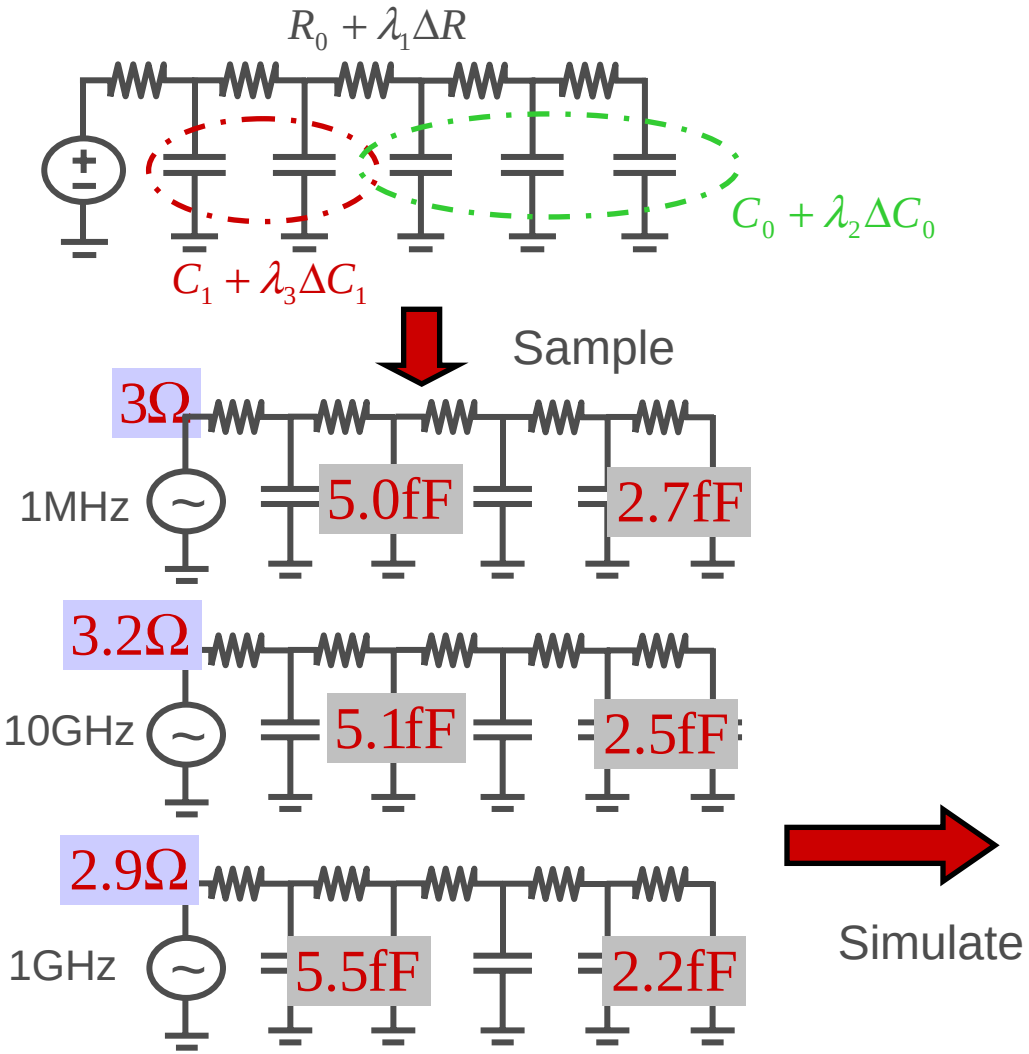


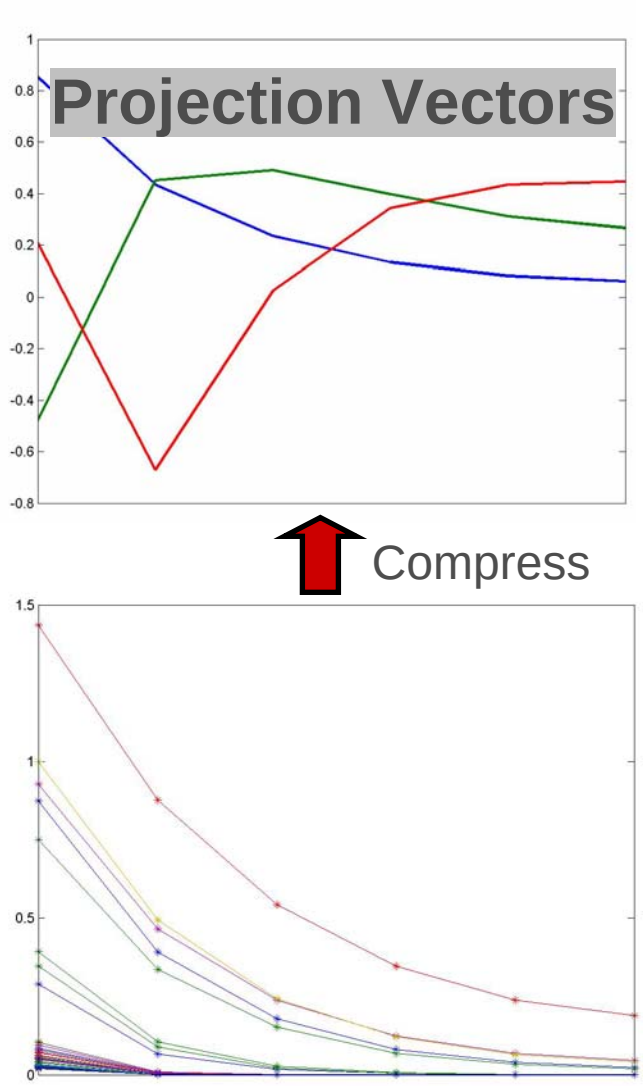
Avoiding Moment Explosions

- Option 1: Ignore the problem.
  - Use the nominal-case projection matrices and hope for the best. Works more often than you might think.
- Option 2: CORE (X. Li et al)
  - Combination of n-D moment matching and Volterra series (perturbation)
  - Exposes special structure in n-D moments
- Option 3: V-PMTBR (Phillips et al)
  - Compress data in operator range
  - Based on approximate computation of stochastic Grammian

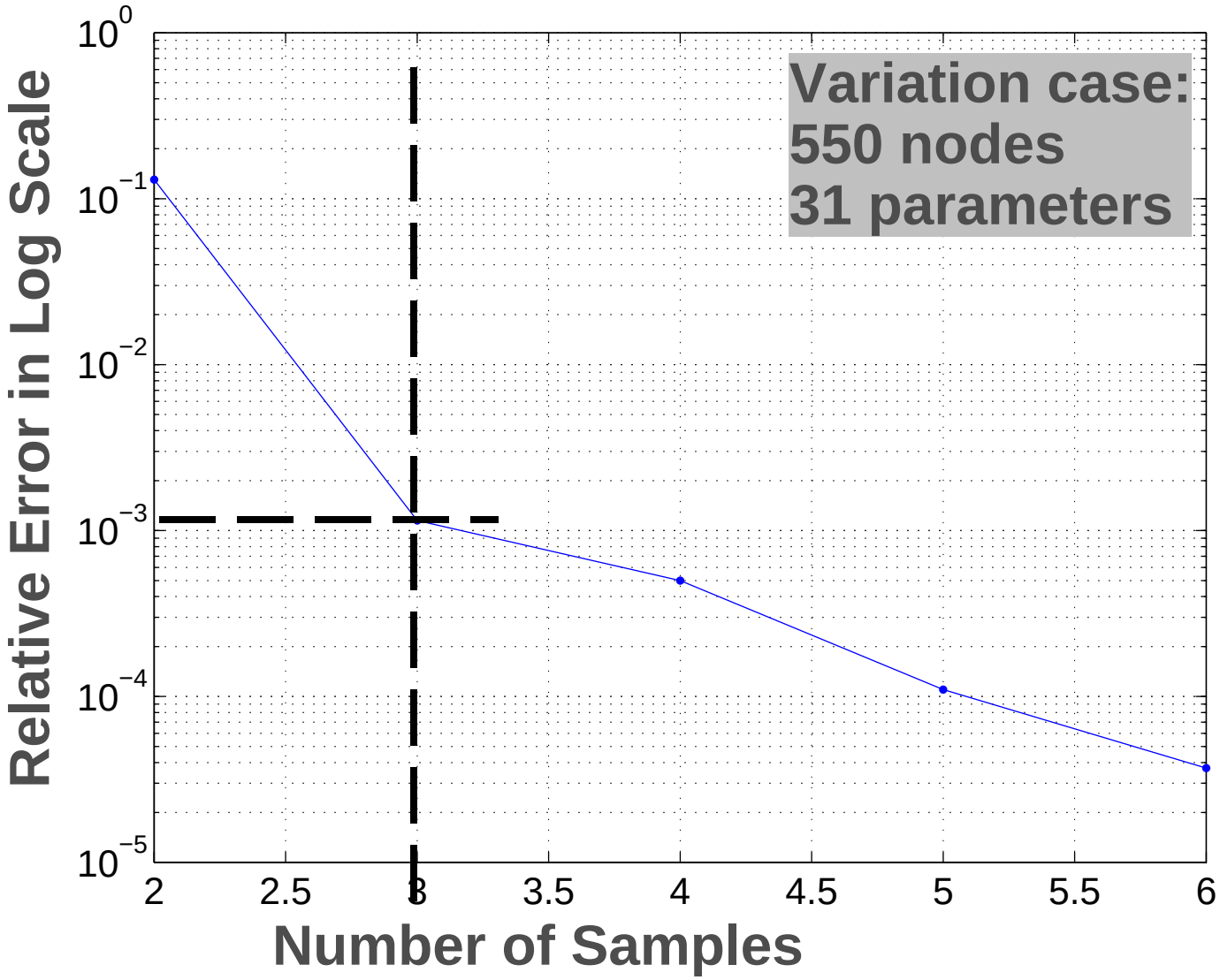
$$X = \int_{-\infty}^{\infty} \int \left[sC(\lambda) + G(\lambda)\right]^{-1} \left[sC(\lambda) + G(\lambda)\right]^{-H} p(\omega, \lambda) d\omega d\lambda$$

PMTBR Visually Speaking

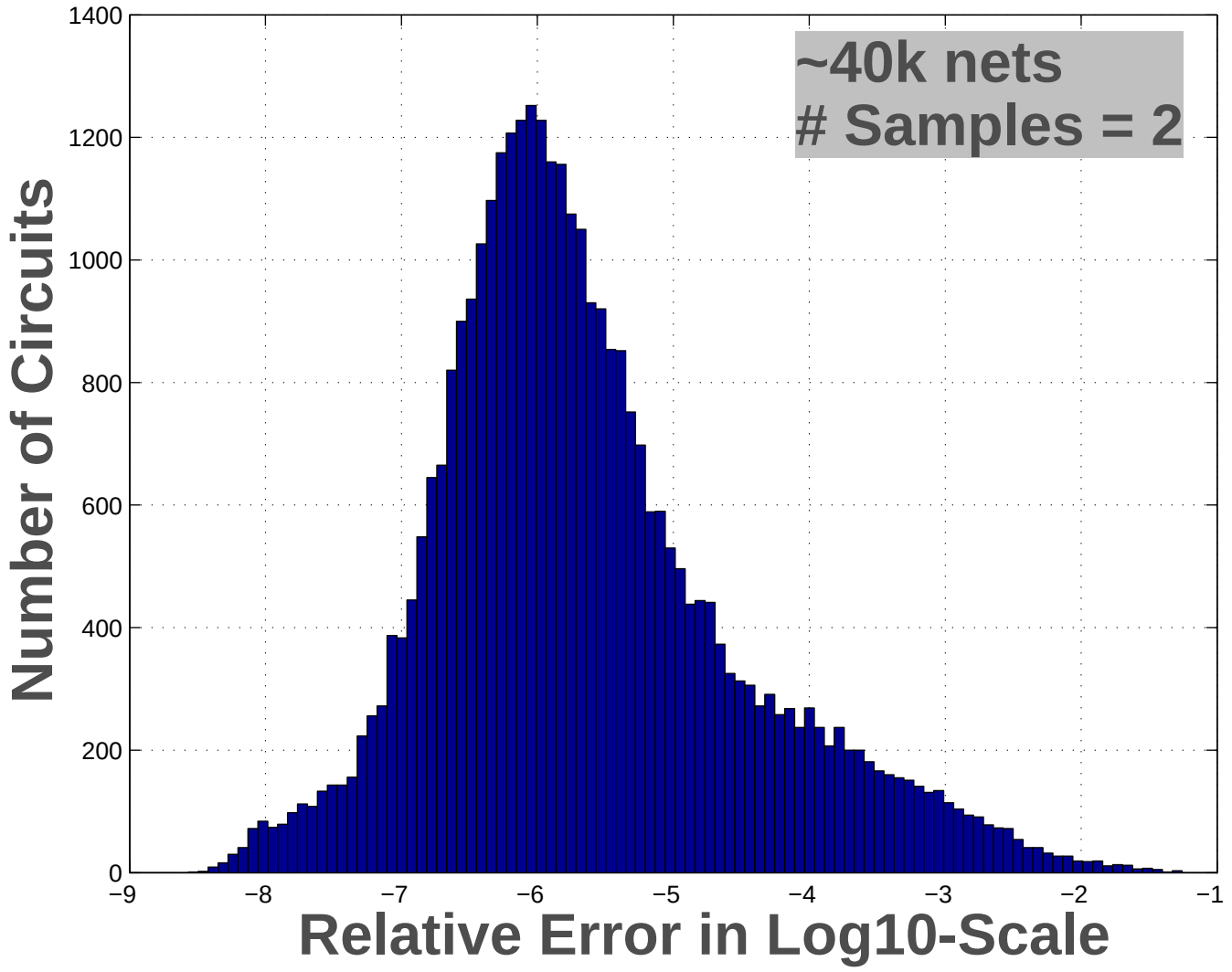




PMTBR Convergence Behavior

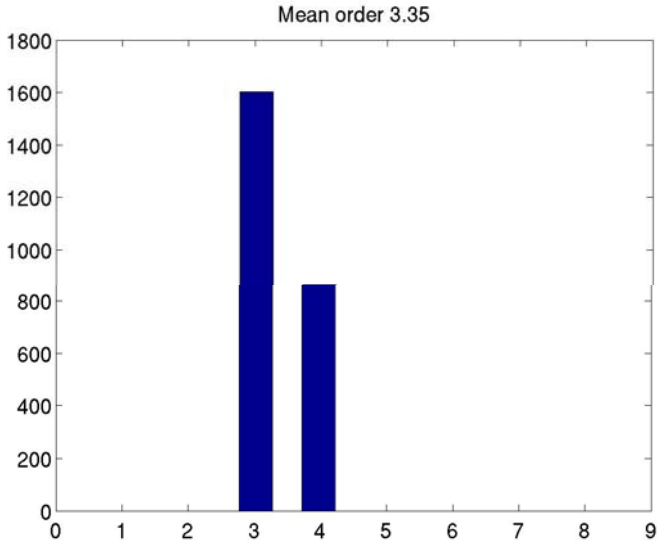


Error Histogram With Small Fixed Order

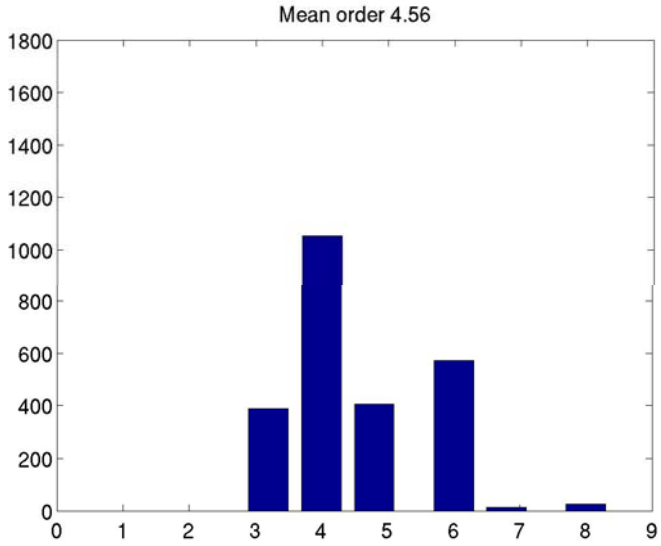


Model Complexity Comparison

Nominal



Parametric



- PMTBR algorithm with automatic order selection compared on several thousand large nets with >> 20 process parameters
- Model size increases 40-70% on average (depends on parameter ranges)
  - Roughly correlates to computational overhead of parametric models in parasitic reduction for SSTA

Summary

- Variability analysis is becoming a first-order concern
  - Power (esp. leakage), timing, physical/functional yield interaction
- Methodologies are still being developed
  - "Full statistical" is a big change –
    some value to it, but not the whole story
- Inevitable that *models* will become richer
  - While keeping analysis times under control
  - While fitting into *very complex* analysis/optimization stack